IMT School for Advanced Studies, Lucca
Lucca, Italy

Beyond the Ideal: Multimodal Learning with Imperfect Data Conditions

PhD Program in Systems Science

Track in Software Quality

XXXVIII Cycle

By

Muhammad Irzam Liaqat

2026

The dissertation of Muhammad Irzam Liaqat is approved.

PhD Program Coordinator: Prof. Mirco Tribastone, IMT School for Advanced Studies Lucca

Advisor: Gabriele Costa, IMT School for Advanced Studies Lucca

Co-Advisor: Fabio Pinelli, IMT School for Advanced Studies Lucca

The dissertation of Muhammad Irzam Liaqat has been reviewed by:

Prof. Sunil Vadera, University of Salford, Manchester, England

Dr. Usman Naseem, Macquarie University, Sydney, Australia

IMT School for Advanced Studies Lucca
2026

Contents

List of Figures x

List of Tables xiv

Acknowledgements xviii

Vita and Publications xix

Abstract xxi

1 Introduction 1

1.1 Background and motivation 3

1.2 Major Contributions 4

1.3 Thesis Organization 5

2 Related Work 7

2.0.1 Literature Search and Filtering Strategy 8

2.1 Multimodal Learning 9

2.1.1 Traditional Multimodal Learning 10

2.1.2 Multimodal Learning in the Era of Transformers . . 11

2.1.3 Benchmark Datasets 13

2.1.4 Key Findings 15

2.2 Dissecting Multimodal Learning Under Missing Modalities 15

2.2.1 Missing Modality Taxonomy 16

2.2.2 Reconstruction Approaches 16

2.2.3 Architectural Approaches 20

2.2.4 Hybrid Approaches . . . 26
2.2.5 Benchmark Datasets . . . 28
2.2.6 Key Findings . . . 28
2.3 Dissecting Multimodal Learning Under Corrupted Modalities 30
2.3.1 Data Processing Methods . . . 31
2.3.2 Architectural Methods . . . 32
2.3.3 Training Strategies . . . 37
2.3.4 Post-hoc Methods . . . 40
2.3.5 Benchmark Datasets . . . 42
2.3.6 Key Findings . . . 42
2.4 Chapter Summary . . . 44

3 Unraveling Multimodal Systems: Presenting Chameleon for Missing Modalities 45
3.1 Rendering Non-visual Modalities into Images . . . 48
3.2 Methodology . . . 49
3.2.1 Problem Formulation . . . 49
3.2.2 Encoding Scheme . . . 50
3.2.3 Modality Passing and Joint Training Strategy . . . 51
3.2.4 Robustness to Missing Modalities . . . 53
3.3 Experiments . . . 54
3.3.1 Evaluations Under Modality-complete Setting . . . 54
3.3.2 Evaluations Under Modality-missing Setting . . . 55
3.3.3 Analysis and Discussion . . . 58
3.4 Chapter Summary . . . 64

4 RobustA: Multimodal Robustness Under Corrupted Modalities 66
4.1 Weakly Supervised Video Anomaly Detection and Multimodality . . . 69
4.1.1 Addressing Compromised Modalities . . . 69
4.2 Methodology . . . 71
4.2.1 Background and Overview . . . 71
4.2.2 Preliminaries . . . 72
4.2.3 Learning a Shared Representation Space . . . 72
4.2.4 Dynamic Weighting During Inference . . . 73

4.2.5 Modality Corruptions 75
4.3 Experiments 77
4.3.1 Implementation Details 77
4.3.2 Experimental Results 78
4.4 Analysis and Discussion 81
4.4.1 Importance of Dynamic Weighting 81
4.4.2 Is Our Approach Plug-and-Play? 82
4.4.3 Does Multimodal Robustness Translate to Better Zero-shot Performance 82
4.4.4 Is Linear Projection Necessary? 83
4.4.5 Qualitative Results 84
4.5 Chapter Summary 84

5 Key Insights and Future Work **86**
5.1 Analysis and Key Insights 86
5.2 Limitations 89
5.3 Challenges & Future Research 90

Conclusion **93**

List of Figures

1 Overview of multimodal representation learning under imperfect data conditions. (a) Unimodal learning using a single modality. (b) Standard multimodal learning with complete modality inputs. (c) Multimodal learning with missing modalities, where one or more inputs are unavailable at inference time. (d) Multimodal learning under corrupted modalities, where inputs are present but degraded in quality. This thesis focuses on developing robust multimodal methods that can operate reliably across scenarios (b), (c) and (d), bridging the gap between idealised multimodal settings and real-world conditions. 2

2 Taxonomy of the existing works addressing MML with MMP 16

3 Taxonomy of the existing works addressing multimodal learning (MML) with corrupted modality problem (CMP) . 31

4 Comparison of Chameleon with ViLT [210] & Ma et al. [271] on UPMC Food-101 dataset. Chameleon demonstrates better multimodal and unimodal performances and retains significantly superior performance when textual modality is missing at test time. While transformer-based methods such as Ma et al. achieve competitive performance under complete modality settings, they exhibit greater sensitivity to missing modalities. In contrast, Chameleon maintains more stable performance across varying levels of modality availability, including strong unimodal performance when evaluated with image-only inputs 47

5 Architecture of Chameleon. (a) Multimodal input consists of images and texts. (b) Embeddings of texts are encoded as images. (c) Visual and encoded text inputs are arbitrarily input to a weight-sharing visual network in a mini batch. Note that the figure represents the case of textual-visual modality pair. In the case of audio-visual pair, the same encoding is applied to the audio embedding. 49

6 (a) Orginal image. (b) Unimodal (image only) training and testing. (c) Multimodal training, unimodal testing (image only), i.e., 100% text missing. As seen, in unimodal image-only training (b), the model focuses on distinct features of the object. When the text modality (c) is missing during testing, the model focuses on the available modality to make correct final predictions demonstrating the success of our approach in training the multimodal system robust to missing modalities. 52

7 Comparison of Chameleon with FOP [275], and SBNet [353] on audio-visual dataset (e.g., VoxCeleb) under different levels of missing modality at test time. 59

8 Comparison of Chameleon with SMIL [219], Autoencoder (AE) [133], and Generative Adversarial Network (GAN) on avMNIST; an audio-visual dataset. (Left) training with 100% Image + n% Audio and testing with Image Only. (Right) training with 100% Image + 20% Audio and testing with Image + Audio. 60

9 Comparisons of Chameleon with ViLT [210] and Ma et al. [271] on various levels of missing textual modality during testing on textual-visual datasets (e.g., UPMC Food-101 and Hateful Memes). A smaller drop in performance by our approach in most cases signifies its effectiveness towards training multimodal Transformer resilient to missing modalities without dataset-centric fusion strategies. 60

10 t-SNE visualizations of the embedding space of Chameleon (a - c) and ViLT (d - f) along with accuracy on test set of UPMC Food-101. Compared to ViLT, Chameleon not only enhances the classification boundaries when complete modalities are available at test time but also retains these boundaries when the textual or visual modality is completely missing during test time. Note that classes are selected randomly from the test set. 65

11 Existing multimodal anomaly detection approaches [180], [237], [426] are not robust when subjected to corruptions in either modality, including vision (e.g., fog and motion blur) or audio (e.g., babble). 67

12 Architecture of our approach. Modality-specific features are extracted using pre-trained audio and visual encoders. A linear projection is used to match the feature dimensions. Modality embeddings are independently mapped to learn representations in a shared space. During inference, the shared learning space helps mitigate the adverse effects of modality corruptions. Moreover, a dynamic weighting scheme is utilized to adjust the weights of the corrupted modality for better anomaly detection. 71
13 Examples of a few visual and audio corruptions demonstrating the challenging scenarios presented in our proposed benchmark. 75
14 Comparison of weighting schemes (average and dynamic) with baseline concatenation approach [180] on two visual and one audio corruption. 79
15 Qualitative results of `RobustA` and the baseline on three videos taken from XD-Violence dataset. Blue represents baseline anomaly scores whereas green represents `RobustA` results. Dotter represents clean test samples whereas solid highlights corruption cases. As seen, our approach generally outputs comparable anomaly scores for clean and corruption cases. The baseline, while generating competitive anomaly scores for clean samples, demonstrates deteriorated performance when the input is corrupted. Red shaded area represents anomaly ground truth. 84

List of Tables

1 Search and filtering strategy with number of screened studies at each stage. **C:** Number of Corruption Studies **M:** Missing Studies 9
2 Studies leveraging multimodal learning across the five technical challenges defined by [101]. 12
3 Existing benchmark datasets for MML. Modalities: Audio (A), Video (V), Image (I), and Text (T). 14
4 Summary of Generative methods for Missing Modality Problem 17
5 Summary of Alignment methods for Missing Modality Problem 19
6 Summary of Modular design methods for Missing Modality Problem 21
7 Summary of Selective fusion methods for Missing Modality Problem 22
8 Summary of Co-learning methods for Missing Modality Problem 22
9 Summary of Distillation methods for Missing Modality Problem 23
10 Summary of Attention-based methods for Missing Modality Problem 25
11 Summary of Prompt learning methods for Missing Modality Problem 26
12 Summary of Hybrid methods for Missing Modality Problem 27

13 Existing benchmark datasets for MML for MMP. Modalities: Audio (A), Video (V), Image (I), and Text (T). †One or more modalities are *synthetically* dropped or naturally absent in derived splits. 29

14 Summary of Denoising methods for Corrupted Modality Problem 32

15 Summary of Noise-Aware methods for Corrupted Modality Problem 33

16 Summary of Confidence estimation methods for Corrupted Modality Problem 35

17 Summary of Robust fusion methods for Corrupted Modality Problem 36

18 Summary of Augmentation-based methods for Corrupted Modality Problem 38

19 Summary of Adversarial training methods for Corrupted Modality Problem 39

20 Summary of Error detection methods for Corrupted Modality Problem 40

21 Summary of Recovery based methods for Corrupted Modality Problem 41

22 Existing benchmark datasets for MML for CMP. Modalities: Audio (A), Video (V), Text (T), and Image (I). Corruption (C): Noise (N), Perturbations (P), Rephrasing (R), and Variance (V). 43

23 Statistics of the datasets used in our experiments. 54

24 Comparison of `Chameleon` with SOTA multimodal methods on textual-visual (e.g., UPMC Food-101, Hateful Memes, MM-IMDb, Ferramenta) and audio-visual (e.g., avMNIST and VoxCeleb) datasets using modality-complete data. Best results are bold; second best are underlined. 55

25 Comparison of Chameleon with different levels of available modality at test time using textual and visual datasets (e.g., UPMC-Food-101, Hateful Memes, MM-IMDb, and Ferramenta). Comparison is provided with multimodal Transformers (ViLT [210]*, Ma et al. [271], and CLIPPO[357]). *ViLT values are taken from Ma et al. [271]. Boldface and underline denote, respectively, the best and second-best results. † indicates our implementation. 56

26 Comparison of Chameleon with ViLT [210]*, Lee et al. [334], and CLIPPO [357] on UPMC Food-101 [59], Hateful Memes [165], and MM-IMDb [79] under different missing modality settings during training and testing. *ViLT values are taken from Lee et al. [334]. Boldface and underline denote, respectively, the best and second best results. † indicates our implementation. 58

27 Performance analysis of Chameleon with various visual networks using UPMC-Food-101, and Hateful Memes datasets. 58

28 Cross-modal verification results on unseen-unheard configurations of Chameleon and existing state-of-the-art methods. Best results are highlighted in bold text whereas the second best are underlined. 61

29 Comparison of different embeddings to encode textual modality under different levels of missing modality at test time. 62

30 Comparison of word- and sentence-level embeddings using UPMC-Food-101, and Hateful Memes datasets. 63

31 Comparison of our approach and baseline under various corruption types and corruption levels (percentages). AP(%) is used as the evaluation metric (higher is better). Bold represents best performance in each comparison. 76

32 Comparison of baseline vs. proposed when one modality is completely missing and the other is corrupted. Higher accuracy is better. 80
33 Plug-and-play performance of our approach for mixed audio and visual distortions at 0% and 100% levels of corruption on multiple baselines including XDVioDet, HyperVD, and RTFM. Bold: best performance. 81
34 AUC comparison on the test set of UCF-Crime dataset under zero-shot setting where the models are trained on XD-voilence dataset. Our approach demonstrates better zero-shot generalization, highlighting the importance of our proposed shared space representation learning. As seen, our approach also outperforms the existing unsupervised approaches trained and tested on UCF-crime [122]. . 83
35 Importance of linear projection compared to padding and concatenation (baseline) under different levels (0% to 100%) of an arbitrarily selected corruption case (pitch shift) for conciseness. 83

Acknowledgements

For the family who believed in me

To Shumaiza, Shanfa, Shaiza, Hazam, and my beautiful parents — you never asked me to explain what I was working on, yet never stopped believing in me. You gave me the kind of support that does not need words. You managed to love someone who was perpetually "almost done," who missed weddings, birthdays, a handful of other holidays, and who communicated primarily through the medium of "sorry, I have a deadline." I love you all so much and owe you years. This dissertation is, in no small part, yours.

For the friends who kept me going

To God's strongest soldiers[1] — my people, my friends, and my sweetest companions in this long, strange journey. You showed up when it mattered most: you sat with me in the doubt, shared your food and drinks, and refused, stubbornly and repeatedly, to let me give up. I thank you for surviving the "I don't want to do this anymore" spirals, the "That's what she said" puns, and for keeping me fed when I forgot that was something I needed to do. You were also, it must be said, an *impeccable* source of gossip — the freshest tea, delivered at exactly the right moment.[2] You know who you are. I love you more everyday and every chapter of this thesis carries your warmth, your stubbornness, and your refusal to let me be dramatic alone.

A scholarly tribute

An acknowledgment is owed to whoever first invented coffee. That decision has done more for academic output than any funding body in history. I am grateful. Deeply, chronically, irreversibly GRATEFUL.

AI, Publications and co-authors

Minor post-processing of parts of this thesis (limited to language refinement and formatting) was assisted by AI-based tools. The material incorporated from my co-authored works is included with the full consent of all co-authors and in accordance with the publishing policies of the respective venues; detailed acknowledgments of those contributions appear in the respective chapters. Briefly: portions of this thesis draw from *Chameleon: A Multimodal Learning Framework Robust to Missing Modalities*; from *RobustA: Robust Anomaly Detection in Multimodal Data*; and from *Multimodal Learning Under Imperfect Data Conditions: A Survey*. To every collaborator named in those works: thank you — for the ideas, the patience, and the late-night review rounds.

[1] Alice, Andrea, Angy, Ci, Domi, Fede, Gabri, Gio, GP, Hamza, Ju, Lala, Laeeq, Ludo, Ludo, Mala, Marianos, Miyase, Moli, Nigar, Qaiser, Ravi, Rici, Saugy, Talha, and et al.

[2] The author wishes to clarify that all tea consumed was strictly off the record.

Vita

2024	PhD - Visiting Researcher - 03 months Mohamed Bin Zayed University, UAE
2024	PhD - Visiting Researcher - 06 months Johannes Kepler University, Austria
2022 - 2026	PhD in Systems Science IMT School of Advance Studies Lucca, Italy
2019 - 2021	M.Sc. in Computer Science University of Engineering and Technology, Pakistan
2015 - 2019	B.S. in Computer Science University of the Punjab, Jhelum Campus, Pakistan

Publications

1. Moscati, M., Abdullah, A., Saeed, M. S., Nawaz, S., Das, R. K., Zaheer, M. Z., **Liaqat, M. I.**, & Schedl, M. (2025). "Face-voice Association in Multilingual Environments (FAME) 2026 Challenge Evaluation Plan." *Proceedings of the 32nd ACM International Conference on Multimedia*.
2. Hannan, A., Manzoor, M. A., Nawaz, S., **Liaqat, M. I.**, Schedl, M., & Noman, M. (2025). "PAEFF: Precise Alignment and Enhanced Gated Feature Fusion for Face-Voice Association." *Interspeech*.
3. **Liaqat, M. I.**, Nawaz, S., Zaheer, M. Z., Saeed, M. S., Sajjad, H., De Schepper, T., ... & Schedl, M. (2025). "Chameleon: A Multimodal Learning Framework Robust to Missing Modalities." *International Journal of Multimedia Information Retrieval*, 14(2), 21.

Submitted Works:

1. AlMarri, S., **Liaqat, M. I.**, Zaheer, M. Z., Nawaz, S., Nandakumar, K., & Schedl, M. "RobustA: Robust Anomaly Detection in Multimodal Data." `https://arxiv.org/abs/2511.07276` Submitted to *IEEE Transactions in Image Processing*.
2. **Liaqat, M. I.**, Abbas, Q., Nawaz, S., Zaheer, Z., Moscati, M., Hou, Y., Khan, M. H., Khan, S., Andre, E., & Schedl, M. "Multimodal Learning Under Imperfect Data Conditions: A Survey. `10.36227/techrxiv.176410566.65375877/v1` Submitted to *ACM Computing Surveys*.
3. Abbas, Q., **Liaqat, M. I.**, Qayum, S., & Hina, S. "Rethinking Architectural Complexity in Deep Vision Models for Computational Histopathology." Submitted to *Visual Computing*.
4. Abbas, Q., Khan, F., **Liaqat, M. I.**, Khan, M. L., & Sajjad, H. "From Inductive Priors to Learned Attention: A Comprehensive Evaluation of Deep Learning Architectures for Remote Sensing Image Classification." Submitted to *Remote Sensing*.

Abstract

Multimodal learning combines heterogeneous data sources such as vision, audio, and text to improve perception and decision-making. Despite its empirical success, this thesis argues that multimodal learning is not inherently robust: most existing approaches assume the availability and reliability of all modalities, leading to significant performance degradation when modalities are missing or corrupted. Our work investigates how robustness in multimodal learning can be explicitly modeled and enforced under imperfect data conditions. It addresses three interconnected aspects: architectural design for modality integration, learning with missing modalities, and robustness to corrupted modalities. The contributions include unified taxonomies that reveal structural limitations of existing methods, a modality-agnostic learning framework that enables reliable inference under missing modalities, and principled benchmarks and models for evaluating and improving robustness under modality corruption. Overall, our work demonstrates that robustness in multimodal learning must be deliberately designed rather than assumed, and provides foundations for building reliable multimodal systems in real-world environments.

Chapter 1

Introduction

Multimodal Learning (MML) leverages multiple data streams (e.g. audio, visual, or textual) to perform various complex perception tasks. Drawing inspiration from human cognition, these methods mimic how individuals naturally integrate information from multiple sensory modalities to understand and interact with the world [466]. While traditional unimodal methods are designed for individual independent modalities, MML has emerged to model and learn from heterogeneous data streams, notably enhancing performance across tasks [398]. Existing MML models have demonstrated remarkable success in domains ranging from medical imaging [396] and emotion recognition [424] to autonomous vehicles [382].

Despite these advances, the transformative potential of MML remains constrained due to fundamental challenges arising from the heterogeneity and complex integration of diverse modalities. Each modality possesses distinct statistical characteristics, noise patterns, and temporal dynamics that cause inherent complexities towards MML [398]. More critically, real-world MML rarely operates under ideal conditions where all modalities are simultaneously available and noise-free. Instead, multimodal networks must withstand sensor failures, environmental interference, bandwidth limitations, and asynchronous data pipelines that introduce uncertainty and deteriorate model performance [434]. This motivates us to focus on three interconnected challenges that are essential for deploy-

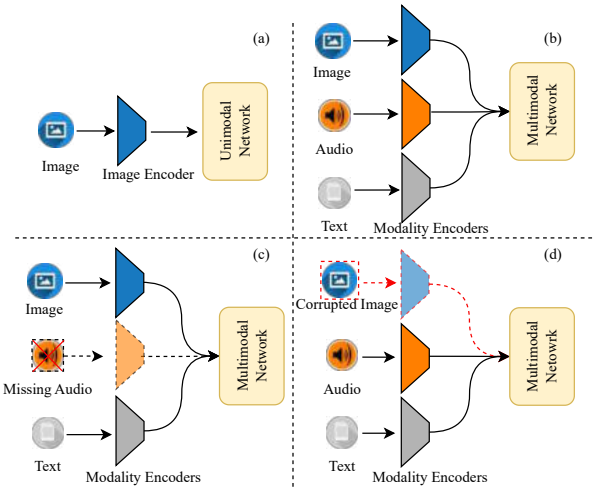


Figure 1: Overview of multimodal representation learning under imperfect data conditions. (a) Unimodal learning using a single modality. (b) Standard multimodal learning with complete modality inputs. (c) Multimodal learning with missing modalities, where one or more inputs are unavailable at inference time. (d) Multimodal learning under corrupted modalities, where inputs are present but degraded in quality. This thesis focuses on developing robust multimodal methods that can operate reliably across scenarios (b), (c) and (d), bridging the gap between idealised multimodal settings and real-world conditions.

ing multimodal methods in real-world settings. First, design architectures that can effectively integrate and adaptively prioritize multiple modalities while maintaining semantic coherence between diverse modalities. Second, develop strategies to handle missing modalities when one or more data streams become unavailable. Third, ensure resilience to corrupted modalities where data streams are present but compromised by noise, misalignment, or deteriorated.

1.1 Background and motivation

Architectural design for modality integration. A fundamental challenge lies in designing multimodal architectures that can effectively learn representations across multiple data types [466]. Optimal performance requires not only effective integration of diverse modalities but also adaptability to the varying importance of each modality. In real-world settings, the contribution of each modality varies dynamically based on context and task, noise levels, or data availability [265]. Multimodal networks must therefore handle heterogeneous data characteristics, including different temporal and spatial resolutions, while maintaining semantic coherence across modalities. This requires innovative attention mechanisms, cross-modal transformers, or graph-based strategies that enable fine-grained alignment and multimodal learning.

Robustness to missing modalities. Despite its advantages over unimodal methods, MML may fail when one or more modalities are missing during inference [415]. This missing data may arise from sensor failure or incomplete data pipelines. Consequently, real-world data frequently contain missing modalities, and this vulnerability can lead to deteriorated performance. Designing robust multimodal architectures to address the missing modality problem (MMP) represents a major research challenge. Such architectures must not only impute or reconstruct missing modalities but also dynamically reconfigure the inference process based on available data. Recent approaches explore generative modeling [391], modality dropout training [270], [404], and conditional learning to improve model flexibility [320].

Robustness to corrupted modalities. Beyond simply missing modalities, a more realistic yet often overlooked challenge is handling corrupted modality problem (CMP) [423]. Real-world MML methods frequently encounter scenarios where data streams are partially available but suffer from noise, misalignment, or degradation. Such data conditions may stem from sensor malfunction, environmental interference, bandwidth limitations, or communication errors, drastically reducing the reliability and effectiveness of MML methods [328]. Unlike completely missing

data, corrupted modalities appear present but contribute misleading or conflicting information, making detection and mitigation more complex. Ensuring model robustness under CMP is therefore critical for practical deployment and reliable performance.

In principle, these challenges are interconnected, for instance, architectural design choices directly influence how networks respond to missing data [338]. Meanwhile, strategies for handling missing data often rely on assumptions about the quality of the remaining modalities [259]. Conversely, corrupted modalities can mimic missing data, blurring the line between the two [440].

1.2 Major Contributions

To this end, this thesis addresses three core technical challenges in multimodal learning, exploring innovative methods for model design and evaluating proposed architectures under missing and corrupted data conditions. Our key contributions include:

1. **Unified Survey and Taxonomies for Multimodal Learning Under Imperfect Conditions:** We present a comprehensive survey that consolidates the traditionally separate viewpoints of multimodal architectural design, missing modality learning, and corrupted modality robustness into a single, unified framework. We introduce clear taxonomies that organize methods across these three interconnected dimensions, and synthesize benchmark datasets, application domains, and methodological gaps.

   Why it matters: Researchers and practitioners lack a single reference that spans all three dimensions; this survey fills that gap and accelerates informed method selection and benchmark comparison.

2. **Chameleon: A Unified Learning Framework for Robust Multimodal Representation Under Missing Modalities:** We introduce `Chameleon`, a multimodal learning framework comprising (a) a modality-agnostic shared representation space; (b) a modular encoder architecture that maintains independence across modalities;

and (c) an adaptive routing and fusion mechanism that dynamically restructures the forward pathway under missing inputs. Extensive evaluation across diverse benchmarks demonstrates consistent performance gains over existing baselines.

Why it matters: Real-world deployments rarely guarantee complete modality inputs; `Chameleon` provides a principled and deployable solution that remains stable under severely degraded conditions.

3. **RobustA Benchmark and Resilient Anomaly Detection:** We introduce `RobustA`, the first comprehensive benchmark, to the best of our knowledge, for evaluating multimodal anomaly detection under modality corruptions, covering 8 visual and 8 audio corruption types at varying severity levels. We further propose a novel detection approach that learns a shared representation space and dynamically reweighs modality contributions based on estimated corruption levels during inference.

   Why it matters: No prior benchmark systematically assessed multimodal anomaly detection under corruption; `RobustA` establishes the first rigorous foundation for both evaluation and future method development in this setting.

By addressing these challenges in a systematic manner, this thesis provides a unified perspective on robust multimodal learning, bridging model design, missing modality handling, and corruption robustness. It serves as a practical and methodological foundation for developing multimodal systems that remain reliable under the imperfect and dynamic conditions encountered in real-world deployments.

1.3 Thesis Organization

The remaining work is organized into the following chapters, each addressing a core aspect of multimodal learning under imperfect data conditions.

- Chapter 2 presents a comprehensive review of related literature. The chapter examines the evolution of multimodal learning from traditional architectures to transformer-based approaches, surveys benchmark datasets, and provides a synthesis of key insights. It further dissects multimodal learning under missing modalities and introduces the proposed taxonomy, reviewing reconstruction-based, architectural, and hybrid approaches. Finally, it reviews multimodal robustness under corrupted modalities, covering preprocessing, architectural, training, post-hoc methods, and associated datasets.

- Chapter 3 introduces `Chameleon`, the proposed multimodal learning framework designed to operate under missing modalities. The chapter presents the framework through a unified representation of non-visual modalities as images and outlines the problem formulation and architectural components. A thorough experimental evaluation is provided under both modality-complete and modality-missing settings, followed by a discussion of results and insights.

- Chapter 4 presents `RobustA`, a systematic study of multimodal robustness under corrupted modalities, with a particular focus on weakly supervised video anomaly detection. It introduces our proposed methodology for learning a shared representation space and dynamically reweighting modalities during inference. Furthermore, it describes the corruption types considered, outlines implementation details, and reports extensive experimental results.

- Chapter 5 consolidates insights from the previous chapters and discusses broader challenges in developing robust multimodal systems. It highlights limitations and open research questions while outlining promising directions for future work.

- Finally, We conclude the thesis by summarizing the major contributions and reflecting on their implications for the design of real-world multimodal systems capable of operating under missing or corrupted data conditions.

Chapter 2

Related Work

The goal of MML is to leverage information across multiple modalities to improve the performance of various tasks such as recognition, retrieval, and verification [145], [159]. While modalities often provide complementary information, they may also be contradictory (e.g., under corrupted or adversarial conditions) or redundant [271]. In this context, MML aims to define methods that can understand diverse and heterogeneous modalities and representations. [1] These modality representations, when combined, are called multimodal representations. However, this integration involves several challenges such as addressing the heterogeneity of complex modalities, and, most critically, dealing with scenarios involving imperfect modalities. Therefore, defining a strategy for the integration of modalities into a multimodal representation is a crucial step towards a robust multimodal system. Machine learning methods rely heavily on effective feature representations to achieve better performance. As noted by Bengio et al. [23], compact feature representations should possess qualities such as sparsity, smoothness, natural clustering, and spatio-temporal coherence. Additionally, Sarivastava et al. [20] emphasize that effective

The contents of this chapter are taken from our work "Multimodal Learning Under Imperfect Data Conditions: A Survey", available at `https://www.techrxiv.org/doi/full/10.36227/techrxiv.176410566.65375877`

[1] Following the guidelines of [23], in this work, we interchangeably use the words representation, feature and embedding, all defining a tensor or a vector in order to represent an entity from image, text, audio or other modalities.

feature representations should associate instances in the same latent space with identical concepts, enable easy extraction of features even when some modalities are missing, and allow for the reconstruction of missing modalities based on the available ones. Feature extraction for individual independent modalities have been studied extensively in the past, however, recent years have seen a surge from hand-crafted features to feature encoding with deep neural networks [50]. In the early 2000s, traditional methods such as Scale-Invariant Feature Transform (SIFT) were popular for image representation [7]. However, the focus has since shifted towards Convolutional Neural Networks (CNNs) [90]. Similarly, in acoustic feature extraction, mel-frequency cepstral coefficients (MFCC) have been largely replaced by deep neural networks and recurrent neural networks for capturing acoustic and paralinguistic representations [17], [71]. In natural language processing, static word embeddings [29] have been largely superseded by large pre-trained contextualized models (such as BERT and GPT [350]), which provide richer semantic and syntactic representations.

2.0.1 Literature Search and Filtering Strategy

To construct a comprehensive related work, we designed a multi-stage search and filtering pipeline across widely used academic databases, including IEEE Xplore, ACM Digital Library, CVF Open Access, AAAI, Elsevier, Springer, Google Scholar, Dblp, and arXiv. Our keyword strategy combined core terms related to MML with descriptors of incompleteness and robustness. Specifically, we used queries such as: *"multimodal AND missing"*, *"multimodal AND incomplete"*, *"multimodal AND corrupted"*, *"multimodal AND noisy"*, *"robust multimodal"*, *"resilient multimodal"*, and *"degraded modalities"* to ensure broader coverage. We applied a publication time window of 2020–2025, while also including a small number of earlier studies that were influential or foundational to the field. The retrieved works were initially screened by title and abstract, followed by full-text review when necessary. They included if they explicitly addressed missing or corrupted modalities in MML or inference pipelines. Secondary filters were applied to remove duplicates (common across IEEE, CVF,

and Google Scholar), exclude tangential uses of robustness (e.g., general noise in unimodal data), and ensure that both peer-reviewed venues and relevant arXiv preprints were represented. Table 1 outlines the details of filtering at each stage. The final screening left us with 84 papers on MML with corrupted modalities and 133 works on missing modalities, spanning journals, conferences, workshops, and preprint repositories [2]. This distribution highlights both the technical grounding of robustness research in engineering venues and the growing interdisciplinary interest in incomplete MML across computer vision, speech, and general AI communities.

Table 1: Search and filtering strategy with number of screened studies at each stage. **C:** Number of Corruption Studies **M:** Missing Studies

Filtering Stage	C	M	Description
Initial keyword retrieval	1,243	1,487	Raw query results
Duplicate removal	947	1,196	Removed overlapping entries
Title/abstract screening	352	426	studies mentioning defined keywords
Full-text relevance check	164	181	Excluded unrelated studies
Final inclusion	**84**	**135**	Final Selection

2.1 Multimodal Learning

This section offers a concise overview structured around two complementary perspectives. The first follows the taxonomy of Baltrušaitis et al. [101], which organizes corresponding research into five major challenges: Representation, Translation, Alignment, Fusion, and Co-learning. Although well-established, these categories remain central to understanding modality interaction, and we revisit them in light of recent developments. The second perspective [368] highlights the emerging trend of transformer-based MML. It includes single-stream, multi-stream, and hybrid-stream architectures that leverage learning paradigms such as contrastive, masked, and generative modeling. Together, these perspectives summarize the evolution of MML and provide the foundation for understanding crucial MML methodologies.

[2]Complete list of resources available at: `https://github.com/qaixerabbas/awesome-multimodal-learning-with-imperfect-data`

2.1.1 Traditional Multimodal Learning

Representation learning concerns how features from different modalities are modeled. For instance, *joint representation* methods aim to map multiple modalities into a unified embedding space. Such representations have been widely explored in multimodal classification and verification tasks. *Kernel-based methods* such as multiple kernel learning have been applied in object classification, disease detection, and affect recognition [30]. *Probabilistic models* including dynamic Bayesian networks facilitate joint modeling of temporal sequences [46], while deep learning-based fusion with CNNs and RNNs has reached state-of-the-art in Visual Question Answering (VQA), sentiment analysis, and multimodal emotion recognition [163], [189]. In contrast, *coordinated representations* maintain modality-specific spaces and enforce consistency through cross-modal objectives. These include contrastive learning approaches [164] and distance-based methods [26].

Translation focuses on transforming information from one modality into another. Translation approaches can be *retrieval-based* that uses one modality to retrieve corresponding samples in another modality. For example, image–text retrieval frameworks embed both modalities in a joint space to enable cross-modal search [157]. Meanwhile, *generative* methods explicitly produce one modality from another. For instance, encoder–decoder architectures map videos to textual descriptions [193] or generating image captions [205], [391]. *Hybrid approaches* combine retrieval and generation, balancing diversity with accuracy [203], [220]. Despite progress, translation faces challenges related to modality heterogeneity, subjectivity of generated outputs, and evaluation reliability.

Alignment addresses the problem of discovering correspondences across modalities, whether temporal, spatial, or semantic. These approaches can be categorized as implicit or explicit. *Implicit* alignment leverages probabilistic graphical models and neural architectures that learn latent correspondences. For instance, attention-based networks have been applied to VQA tasks to dynamically align visual and textual cues [190]. *Explicit* alignment relies either on supervision, where paired annotations

guide the alignment [41], [42] or on unsupervised learning, which exploits co-occurrence statistics or structural constraints [106], [182]. These methods are particularly critical for downstream tasks where modality synchronization is essential, such as video understanding, cross-modal retrieval, or audio–visual speech recognition.

Fusion combines information from multiple modalities to improve prediction and robustness. *Early fusion* directly concatenates raw or low-level features from different modalities before model training [5]. *Late fusion* aggregates unimodal predictions, often through voting or weighted averaging, which can mitigate the impact of noisy modalities [226], [383]. *Hybrid fusion* strategies integrate intermediate-level representations with final predictions, offering robustness in tasks such as multimedia event detection and emotion recognition [159], [222]. *Model-assisted fusion* employs more principled learning approaches: *kernel-based methods* capture modality-specific similarities [30], *probabilistic models* handle temporal dependencies [46], and neural architectures learn complex integration functions across heterogeneous inputs [189].

Co-learning leverages one modality to enhance learning in another and involves iteratively updating classifiers across modalities to improve generalization [352]. *Transfer learning* transfers knowledge from a well-labeled modality to improve performance in a resource-scarce one, with applications in audio-visual recognition [388], disease classification [202], and action recognition [365]. *Conceptual grounding* maps symbolic linguistic concepts to perceptual modalities that aids semantic representation learning [87]. *Zero-shot learning* has also emerged as an important approach that enables models to generalize to unseen modality combinations using shared semantic embeddings [272]. In cross-lingual transliteration, *hybrid bridging* approaches use intermediate representations to facilitate knowledge transfer [73].

2.1.2 Multimodal Learning in the Era of Transformers

Transformers have revolutionized MML by leveraging unified frameworks for processing and integrating multiple modalities, representing

Table 2: Studies leveraging multimodal learning across the five technical challenges defined by [101].

Method	Task	Modalities	References
Representation			
Joint Representations			
Early Fusion	Emotion Recognition	Audio, Image	[5], [10]
Late Fusion	Emotion Recognition	Audio, Image	[8], [16], [226], [383], [388]
Hybrid Fusion	Event Detection	Audio, Image	[34], [74]
	Emotion Recognition	Audio, Image	[159], [171], [220]
Kernel-Based	Classification	Image, Text	[12], [27], [85]
Probabilistic	Speech Recognition	Audio, Image	[11], [78]
Neural Network	VQA	Text, Image	[14], [324]
	Emotion Recognition	Audio, Image	[115], [189]
Coordinated Rep.			
Alignment	Similarity Learning	Image, Text	[26], [57]
Contrastive	Multimodal Learning	Audio, Image	[164]
Translation			
Retrieval-based	Media Retrieval	Image, Text	[38], [60]
	Image Captioning	Image, Text	[15], [47]
Generative	Image Captioning	Image, Text	[40], [205], [391]
	Visual Speech	Audio, Image	[21], [22]
	Text Description	Image, Text	[24], [39]
Hybrid	Image Captioning	Image, Text	[18], [49]
Alignment			
Implicit Alignment	VQA	Image, Text	[47], [75], [190]
	Speech Segmentation	Audio, Text	[1], [4]
Explicit Alignment	Classification	Image, Text	[54], [67], [154], [192], [372]
	Classification	Image, Text	[56], [106], [182], [292]
	Action Recognition	Audio, Image	[70]
Fusion			
Model-Agnostic	Emotion Recognition	Audio, Image	[10], [16], [383]
	Event Detection	Audio, Image	[34]
Model-Assisted	Classification	Image, Text	[27], [85]
	Emotion Recognition	Image, Text	[30], [45], [238]
	Speech Recognition	Audio, Image	[11], [46]
Co-learning			
Co-training	Classification	Text, Image	[2], [6], [168], [349]
Transfer Learning	Lip Reading	Audio, Image	[36]
	Classification	Text, Image	[25], [51], [66], [185]
Conceptual Grounding	Word Similarity	Audio, Text	[48], [87]
Zero-shot Learning	Classification	Text, Image	[25], [31], [272], [285]
Bridging	Transliteration	Text, Image	[19], [73]

a significant shift from traditional approaches that relied on modality-specific encoders and fusion mechanisms [368]. The success originates from transformers' ability to treat different modalities as sequences of tokens that enables uniform processing and bridging semantic gaps between modalities, as demonstrated by models like CLIP [228] and DALL-E [229]. Modern architectures can be categorized into three **Paradigms** [368]: *Single-stream Architectures* (e.g., UNITER [157], FLAVA [282], CLIPPO [357])

that process all modalities through shared transformer backbones; *Multi-stream Architectures* exemplified by CLIP's dual-encoder design and extended by ALIGN [207] and Florence [243]; and *Hybrid Architectures* like ALBEF [213], BLIP [266], and InstructBLIP [314] that combine modality-specific encoders with cross-modal fusion layers.

The **Learning Strategies** have evolved around three dominant directions to address cross-modal learning challenges. *Contrastive Learning*, popularized by CLIP [228] and extended by ALIGN [207], Chinese-CLIP [288], and domain-specific applications in medical imaging [302] and scientific literature [211], maximizes agreement between corresponding multimodal pairs. *Masked Modeling leverages* Masked Language Modeling (MLM) with Masked Region Modeling (MRM) as demonstrated by VinVL [244], OSCAR [167], BEiT [194], and MAE [260]. *Generative Modeling* encompasses cross-modal content generation, e.g., DALL-E [229], [273], Flamingo [254], and GPT-4V [350]. Applications span Vision-Language Models (e.g., BLIP-2 [335], VideoBERT [150]), Audio-Visual Models (e.g., AVBert [233], SpeechT5 [255], LAVISH [339], AV-HuBERT [279]), and Multimodal Large Language Models (e.g., LLaVA [341], Video-ChatGPT [347]), which represent the current frontier.

Moreover, modern **Training Strategies** include *small-scale specialized models* like LayoutLM [183] and CLIP-Score [203], that focus on domain-specific tasks with computational efficiency. Similarly, *Large-Scale Foundation Models* including CLIP [228], DALL-E2 [273], and Flamingo [254] leverage massive datasets for general-purpose representations with emergent capabilities. *Instruction-Tuned Models* like InstructBLIP [314] and LLaVA [341] leverage the foundational models by incorporating instruction-following capabilities and demonstrate how instruction tuning techniques can be effectively adapted from natural language processing to multimodal scenarios.

2.1.3 Benchmark Datasets

The advancement of MML has been accompanied by the emergence of several benchmark datasets that serve as standard testbeds for evaluating

model performance across diverse tasks and modalities. General-purpose datasets, such as MS-COCO [42], Visual Genome [89], and LAION-5B [276], have played a central role in image captioning, visual question answering, and large-scale vision–language pretraining. These datasets collectively support the development and benchmarking of models for representation learning, fusion, and alignment, providing the foundation for studying scalability and generalization in multimodal systems. A detailed overview of these datasets and their applications is provided in Table 3.

Table 3: Existing benchmark datasets for MML. Modalities: Audio (A), Video (V), Image (I), and Text (T).

Dataset	A	V	I	T	Application Area
Image/Video Captioning and Retrieval					
ActivityNet Captions [88]	✗	✓	✗	✓	Video Understanding
COCO Captions [42]	✗	✗	✓	✓	Computer Vision, Language
Flickr30K [41]	✗	✗	✓	✓	Vision–Language
MS COCO [35]	✗	✗	✓	✓	Object Recognition
MSR-VTT [76]	✗	✓	✗	✓	Video–Language Alignment
UPMC Food-101 [59]	✗	✗	✓	✓	Food Recognition
VATEX [151]	✗	✓	✗	✓	Multilingual Video Captioning
WebVid-10M [193]	✗	✓	✗	✓	Video–Text Pretraining
WIT [234]	✗	✗	✓	✓	Multilingual Multimodal
Medical Imaging and Specialized Domains					
MMIST-ccRCC [406]	✗	✗	✓	✓	Medical Imaging
OBELICS [332]	✗	✗	✓	✓	Document Understanding
Multimodal and Vision–Language Pretraining					
CC-12M [195]	✗	✗	✓	✓	Internet-scale Vision–Language
CLIP Benchmark [228]	✗	✗	✓	✓	Zero-Shot Learning
Conceptual Captions [119]	✗	✗	✓	✓	Vision–Language Pretraining
LAION-5B [277]	✗	✗	✓	✓	Large-scale Pretraining
LAION-5B [276]	✗	✗	✓	✓	Vision–Language Pretraining
UNITER eval. splits [157]	✗	✗	✓	✓	Vision–Language Pretraining
Sentiment and Emotion Analysis					
CMU-MOSEI [99]	✓	✓	✗	✓	Affective Computing
CMU-MOSI [77]	✓	✓	✗	✓	Affective Computing
Video Analysis and Audio–Visual Tasks					
AVA-ActiveSpeaker [175]	✓	✓	✗	✗	Video Analysis, AV Sync
HowTo100M [138]	✓	✓	✗	✓	Instructional Learning
XDViolence [181]	✓	✓	✗	✗	Violence Recognition
Visual Question Answering and Grounding					
CLEVR [86]	✗	✗	✓	✓	Reasoning
GQA [130]	✗	✗	✓	✓	Scene Understanding
MultiModalQA [236]	✗	✗	✓	✓	QA, Multimodal Reasoning
NLVR2 [121]	✗	✗	✓	✓	Visual Reasoning
VQA 2.0 [84]	✗	✗	✓	✓	Question Answering
Visual Genome [89]	✗	✗	✓	✓	Scene Understanding

2.1.4 Key Findings

MML has evolved significantly from relying on hand-crafted features like SIFT and MFCC [7] to employing deep learning-based methods using CNNs or RNNs [17], [29], [90]. This shift has enabled more sophisticated multimodal representations and evolved beyond simple feature concatenation [382]. Recently, the transformer architecture has revolutionized the field and enabled unified processing of modalities as token sequences and giving rise to powerful models such as CLIP and ViLBERT [136], [228], [368]. These models form the backbone of modern Vision-Language and Multimodal Large Language Models and excel at tasks including contrastive learning, masked modeling, and generative modeling [335], [341]. However, the integration of modalities involves challenges such as managing inherent heterogeneity, and most critically, addressing scenarios with imperfect modalities [20]. The progression towards large-scale foundation [228], [254] and instruction-tuned models [314] promises more general-purpose, robust, and interpretable systems capable of reasoning across an ever-expanding set of tasks and domains.

2.2 Dissecting Multimodal Learning Under Missing Modalities

Multimodal methods have demonstrated superior performance compared to their unimodal counterparts. However, such methods are predominantly designed to operate under modality-complete conditions. Meanwhile, real-world data is often incomplete due to factors such as occlusion, sensor failure, storage issues, and missing streams, therefore, introducing significant challenges. Due to their reliance on complementary information from multiple modalities, MML exhibit substantial performance deterioration when evaluated under missing modality scenarios. To assess the robustness of MML under MMP, a number of studies have been conducted to evaluate their performance with incomplete-modality settings. In recent works, [216], [271], [334] defined benchmarks for MMP with systematical drop of varying fractions of missing labels or modalities

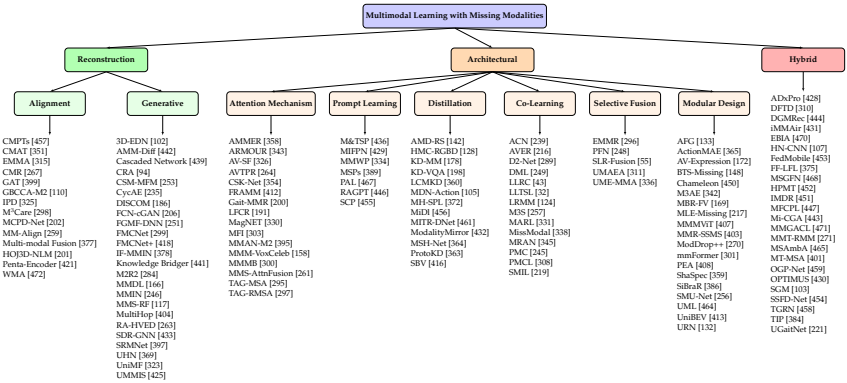


Figure 2: Taxonomy of the existing works addressing MML with MMP

either during train or test stage. These studies demonstrate that existing MML architectures experience severe performance deterioration, at times under-performing their unimodal counterparts.

2.2.1 Missing Modality Taxonomy

A wide range of approaches have been explored in the existing literature to address this issue and enhance the robustness of MML against missing modalities. Leveraging the filtering strategy outlined in Section 2.0.1, we classified the selected methods into eight distinct categories. To provide better readability and understanding for the readers, we have defined a taxonomy, as illustrated in Fig. 2. Our taxonomy classifies these approaches into three major categories: reconstruction, architectural and hybrid approaches, discussed in the following.

2.2.2 Reconstruction Approaches

Reconstruction approaches aim to regenerate or estimate the missing modality representation from the existing modalities. While these methods are inherently complex and require challenging implementation, they rely on cross-modality interactions and relationships during training while preserving the integrity of the original feature representations. Re-

construction approaches can be broadly classified into generative and alignment methods.

Generative Approaches

These methods leverage deep learning techniques, particularly Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), to generate the absent modality by learning the underlying joint distribution of multimodal data [102], [186], [246], [299], [369], [378], [418]. By capturing the statistical dependencies among modalities, generative methods are capable of generating semantically consistent and plausible representations of missing data [94], [117], [206], [235], [251], [253], [323], [425]. For instance, M2R2 [284] leveraged a data imputation strat-

Table 4: Summary of Generative methods for Missing Modality Problem

Method	Modalities	Application Area
3D-EDN [102]	Images, Genomics	Alzheimer Detection
AMM-Diff [442]	Multimodal MRI	MRI Synthesis
Cascaded Network [439]	Audio, Video, Thermal	Person Classification
CRA [94]	Images, Text	Object Recognition
CSM-MFM [253]	Multimodal MRI	Tumor Segmentation
CycAE [235]	Multimodal Images	Alzheimer Detection
DISCOM [186]	Images, Genomics	Alzheimer Prediction
FCN-cGAN [206]	MRI, Gene Expression	Medical Imaging
FGMF-DNN [251]	Multimodal MRI	Tumor Segmentation
FMCNet [299]	Images, Infrared	Person Re-Identification
FMCNet+ [418]	Images, Infrared	Person Re-Identification
IF-MMIN [378]	Audio, Video, Text	Emotion Recognition
Knowledge Bridger [441]	Video, Audio	Object Detection
M2R2 [284]	Audio, Video, Text	Emotion Recognition
MMDL [166]	Multisensor Data	Traffic Data Imputation
MMIN [246]	Audio, Video, Text	Emotion Recognition
MMS-RF [117]	Multisensor Data	Context Recognition
MultiHop [404]	Images, Text	Recommendation System
RA-HVED [263]	Multimodal MRI	Tumor Segmentation
SDR-GNN [433]	Video, Audio	Emotion Recognition
SRMNet [397]	Multimodal MRI	Tumor Segmentation
UHN [369]	Multimodal MRI	Medical Imaging
UniMF [323]	Audio, Video, Text	Sentiment Analysis
UMMIS [425]	Multimodal MRI	Medical Imaging

egy using common representation learning (CRL) for reconstructing the absent modality representations during inference. Similarly, Jeong et al. [263] utilized a variational encoder-decoder architecture coupled with a joint discriminator to enhance segmentation performance under modality-

incomplete settings. Recent advancements further push the boundaries of generative imputation. AMM-Diff [442] integrated a diffusion-based generative model with an Image-Frequency Fusion Network for high-fidelity modality reconstruction. Ke et al. [441] explored the use of large multimodal models to generate, rank, and select the most appropriate reconstructions for missing modalities. John et al. [439] proposed a metric-learning framework to ensure the consistency of generated multimodal features.

Graph-based and autoencoder-driven approaches also contribute to this growing field. SDR-GNN [433] reconstructed missing modalities by aggregating spectral features through a spectral domain reasoning graph neural network. Li et al. [166] applied stacked autoencoders for effective modality imputation in traffic data scenarios, while Li et al. [397] introduced a deformation-aware encoder that infers missing modality information to recover the original image representation. Collectively, these generative frameworks demonstrate the efficacy and versatility of learning-based synthesis techniques in overcoming modality incompleteness across diverse domains. Table 4 provides an overview of recent studies that leverage different generative methods.

Alignment

Unlike generative methods, alignment-based approaches do not aim to explicitly reconstruct missing modalities. Instead, they seek to align available modalities within a shared latent space by leveraging cross-modal interactions and semantic associations. Correlation analysis and contrastive learning are commonly employed techniques to ensure that different modalities are semantically consistent while retaining their inherent characteristics [201], [202]. For instance, MM-Align [259] explored optimal transport theory with deep learning modules to effectively align modalities in a common latent space. M3Care [298] utilized auxiliary information derived from present modalities to assist the alignment process through the construction of similarity matrices. Leveraging Bayesian methods, Matsuura et al. [110] proposed a Bayesian Canonical Correlation Analysis (CCA) framework that projects multiple modalities into a shared

Table 5: Summary of Alignment methods for Missing Modality Problem

Method	Modalities	Application Area
CMPTs [457]	Images, Text	Robust Learning
CMAT [351]	Images, Skeleton	Human Action Recognition
CMR [267]	Multimodal Data	Cross-modal Retrieval
EMMA [315]	Text, Images	Object Retrieval
GAT [399]	MM Knowledge Graph	Representation Learning
GBCCA-M2 [110]	Images, Tabular	Image Retrieval
HOJ3D-NLM [201]	Video, Audio	Human Action Recognition
IPD [325]	Audio, Video, Text	Sentiment Intensity Analysis
M^3Care [298]	Multimodal EHRs	Ocular Disease Diagnosis
MCPD-Net [202]	Images, Accelerometer	Parkinson Classification
MM-Align [259]	Multimodal Sequences	Model Robustness
Multi-modal Fusion [377]	Images	Facial Recognition
Penta-Encoder [421]	Multimodal MRI	Brain Tumor Segmentation
WMA [472]	Video, Text, Audio	Hate Speech, Movie Genre

space and enhances robustness against modality incompleteness. In addition, several alignment strategies based on sparse regression, covariance matrix modeling, and modality compensation have shown promising results in maintaining cross-modal consistency [267], [315], [325], [351], [377].

More recent innovations further expand the scope of alignment-based strategies. Liang et al. [399] introduced a graph attention network (GAT) using transfer learning to align heterogeneous data sources. Yu et al. [421] employed a modality-dynamic encoder to capture tumor semantics across incomplete modalities. Reza et al. [457] applied cross-modal proxy tokens and alignment loss to guide the fusion process in the presence of missing data. Contrarily, Zhi et al. [472] incorporated Wasserstein distance into self-attention mechanisms to enhance modality fusion and alignment. Table 5 presents an overview of the reviewed works that employ alignment methods for improving model resilience towards missing modalities.

In summary, generating and aligning multimodal representations before passing them to the model layers enables reconstruction approaches to exhibit resilience against missing modalities. However, the major challenges include handling misalignment caused by outliers in the data and overcoming noise in the synthesized data.

2.2.3 Architectural Approaches

Architectural approaches have gained significant attention within the research community and have been extensively explored in the literature. Unlike complex reconstruction-based methods, these approaches focus on modifying existing multimodal architectures to enhance their robustness against missing modalities. They leverage various techniques that enable multimodal systems to maintain strong performance, even in the presence of missing modalities. These techniques include model design manipulation, selective fusion, co-learning, distillation methods, attention-based architectures, and prompt learning.

Modular Design

Model adaptation through architectural design is a key strategy for enhancing the robustness of multimodal systems in the presence of missing modalities. This approach focuses on modifying the model architecture to ensure that reliable predictions can still be produced using only the available modalities. Common techniques include modality masking, modality dropout, and modular architectural components that allow each modality to contribute independently during both training and inference [133], [172], [217]. Table 6 highlights the key studies that introduced resilient architectural designs. For example, M3L [403] employed a semi-supervised framework that integrates modality masking to improve model generalization. Similarly, MMMVIT [407] leveraged Vision Transformers to learn a shared latent representation and extracted multi-scale features via intra-model fusion. Reza et al. [408] improved the robustness of pre-trained multimodal networks by adapting the modulation of intermediate feature representations which enabled the architecture to maintain performance despite missing inputs. UniBEV [413] introduced Channel Normalized Weights to learn flexible combinations of input modalities resulting in increased modular adaptability towards missing data scenarios.

Furthermore, transformer-based models have recently gained attention for their architectural flexibility in handling missing modalities. mmFormer [301] introduced a hybrid design with modality-specific encoders

Table 6: Summary of Modular design methods for Missing Modality Problem

Method	Modalities	Application Area
AFG [133]	Audio, Video	Action Recognition
ActionMAE [365]	RGB, Depth, Infrared	Action Recognition
AV-Expression [172]	Audio, Video	Expression Recognition
BTS-Missing [148]	Multimodal MRI	Brain Tumor Segmentation
Chameleon [450]	Image, Text, Audio	Representation Learning
M3AE [342]	Multimodal MRI	Brain Tumor Segmentation
MBR-FV [169]	Facial Video	Biometrics Recognition
MLE-Missing [217]	Audio, Video	Emotion Recognition
MMMViT [407]	Multimodal MRI	Brain Tumor Segmentation
MMR-SSMS [403]	Images, Depth Imaging	Semantic Segmentation
ModDrop++ [270]	Multimodal MRI	Sclerosis Lesion Segmentation
mmFormer [301]	Multimodal MRI	Brain Tumor Segmentation
PEA [408]	Depth, Thermal, Images	Semantic Segmentation
ShaSpec [359]	Multimodal MRI	Brain Tumor Segmentation
SiBraR [386]	Image, Text, Audio	Recommendation System
SMU-Net [256]	Multimodal MRI	Brain Tumor Segmentation
UML [464]	Multimodal OCT	Visual Acuity Prediction
UniBEV [413]	LiDAR, Images	Object Detection
URN [132]	Multimodal MRI	Brain Tumor Segmentation

and inter-modal transformers, complemented by auxiliary regularizers that guide the learning of modality-invariant features. UML [464] applied attentional masking for handling missing inputs and incorporated auxiliary contrastive learning between image and text features, effectively combining masked modality modeling with cross-modal alignment tasks. Other architectural innovations include the use of modality dropout and regularization techniques during training [132], [148], [365], masked autoencoders for unsupervised feature reconstruction [169], [342], modular network designs with independently functioning sub-networks [256], [359], and dynamic co-learning mechanisms [270] that adaptively balance modality contributions.

Selective Fusion

Fusion methods aim to mitigate the impact of missing modalities by utilizing different techniques such as ensemble learning and weighted voting mechanisms [296], [311], [336]. Table 7 summarizes the studies that enable selective fusion dynamically to improve multimodal performance. For instance, Zheng et al. [248] proposed progressive fusion that learns representations from single to multiple modalities. It involves taking the

Table 7: Summary of Selective fusion methods for Missing Modality Problem

Method	Modalities	Application Area
EMMR [296]	Text, Audio, Video	Sentiment Analysis
PFN [248]	Video, Sensor	Person Re-identification
SLR-Fusion [55]	Images, Depth	Face Recognition
UMAEA [311]	Images, Text	Multimodal Entity Alignment
UME-MMA [336]	Audio, Video, Text	Audio–Video Event Localization

predictions from different modalities which are fused for decision making. Shao et al. [55] investigated the decision time fusion through weighted voting mechanism for representations. These studies indicates that fusion empowers the multimodal systems by considering the most contributing modalities and leverage that information for better performance.

Co-Learning

Co-learning involves learning inter- and intra-modal relationships from the participating modalities. These learned relationships compensate for missing modalities during inference and contribute heavily towards improving model performance. Common techniques under this category include transfer learning [43], [219], zero-shot learning [124], and conceptual grounding [331]. Moreover, as shown by Table 8, these methods are leveraged to improve modular performance across a variety of application areas. In recent literature, Bao et al. [308] proposed a fed-

Table 8: Summary of Co-learning methods for Missing Modality Problem

Method	Modalities	Application Area
ACN [239]	Multimodal MRI	Brain Tumor Segmentation
AVER [216]	Audio, Video	Emotion Recognition
D2-Net [289]	Multimodal MRI	Brain Tumor Segmentation
DML [249]	Multisensor Data	Weather Mapping
LLRC [43]	Images, Text	Face Recognition
LLTSL [32]	Images, Text	Face Recognition
LRMM [124]	Text, Images	Recommendation Systems
M3S [257]	Text, Audio, Video	Sentiment Analysis
MARL [331]	Multimodal MRI	Brain Tumor Segmentation
MissModal [338]	Text, Audio, Video	Sentiment Analysis
MRAN [345]	Text, Audio, Video	Sentiment Analysis
PMC [245]	Images, Text	Object Recognition
PMCL [308]	Multiple Sensors	Prototype Probing
SMIL [219]	Video, Audio, Text	Movie Genre Classification

erated learning-based architecture that leverages a masking strategy to

find, learn and compensate the missing information. Konwer et al. [331] proposed a meta-learning methodology in combination with conceptual grounding that enables the approximation of missing features from the learned shared latent space. MissModal [338] leveraged shared learning by implementing geometric contrastive loss with distribution distance loss that allows the model to learn a shared representation which assists under missing modality settings. MRAN [345] introduced resilience in multimodal systems by reconstructing and aligning the participating modalities in a textual feature domain for effective sentiment analysis. Similar approaches include meta sampling [257], duel disentanglement [289], knowledge sharing [216], [249], progressive modality cooperation [245], adversarial co-training [239], and low-rank transfer [32].

Distillation

Distillation networks leverage student-teacher frameworks, where a student model trained with incomplete modalities learns to emulate the predictions or representations of a teacher model trained on complete modality data. By transferring semantic knowledge from a modality-complete teacher to a modality-incomplete student, the resulting models achieve enhanced robustness and generalization [198].

Table 9: Summary of Distillation methods for Missing Modality Problem

Method	Modalities	Application Area
AMD-RS [142]	Multispectral/HSI	Remote Sensing Classification
HMC-RGBD [128]	Images, Depth	Video Action Recognition
KD-MM [178]	Multimodal MRI	Alzheimer's Diagnosis
KD-VQA [198]	Images, Text	Visual Question Answering
LCMKD [360]	Multimodal MRI	Brain Tumor Segmentation
MDN-Action [105]	Images, Depth, Skeleton	Human Action Recognition
MH-SPL [372]	Images, Text	Image Recognition
MiDl [456]	Video, Audio	Action Recognition
MITR-DNet [461]	Text, Audio, Video	Sentiment Intensity Analysis
ModalityMirror [432]	Audio, Video	Audio Classification
MSH-Net [364]	Multispectral/HSI	Remote Sensing Classification
ProtoKD [363]	Multimodal MRI	Brain Tumor Segmentation
SBV [416]	Audio, Video	Audiovisual Segmentation

Table 9 outlines recent studies that have explored various distillation

strategies. For instance, Zhang et al. [372] proposed cross-branch supervision through a bi-directional distillation architecture, where one branch learns unimodal representations while the other performs MML. Wu et al. [416] extended this concept to audio-visual semantic segmentation and leveraged a dual student-teacher design in which audio and visual student models independently learn from their respective teacher networks to handle missing modalities. Similarly, ProtoKD [363] transferred semantic knowledge from a complete-modality teacher to a student model that performs segmentation tasks under partial modality input. Cross-Model distillation techniques [178], [360] further demonstrate how knowledge from complete-modality models can be effectively leveraged to guide lightweight or specialized student networks. Similar methods for modality-aware distillation include joint adaptation distillation [364], adversarial distillation [142], and hallucination-based distillation [105], [128].

Attention Mechanisms

Attention involves focusing on the relevant information by dynamically adjusting the weight of each modality based on its relative importance. Learned attention mechanisms and weighted fusion strategies are commonly employed to achieve this dynamic adaptation, making multimodal systems more resilient to missing data [200], [354]. Table 10 outlines existing studies leveraging attention-based methods for modular robustness. In a recent study, MMAN-M2 [395] used a transformer-based encoder-decoder model with a multi-head mechanism for dynamic extraction and fusion of feature representations. Similarly, CTNet [326] proposed a transformer-based multi-head attention network for enhancing resilience in the task of audio-visual classification. The architecture enables the modality alignment and allows the network to perform better with and without missing modalities.

Meanwhile, multiple attention-based architectures have been proposed to handle missing modality in different application areas. For effective emotion recognition and classification, Vazquez et al. [358] enhanced model performance by introducing cross-model and self-attention

Table 10: Summary of Attention-based methods for Missing Modality Problem

Method	Modalities	Application Area
AMMER [358]	Audio, Video, Text	Emotion Recognition
ARMOUR [343]	Text, Table	Mortality Prediction
AV-SF [326]	Audio, Video	Person Recognition
AVTPR [264]	Audio, Video	Person Recognition
CSK-Net [354]	Optical, Infrared	Semantic Segmentation
FRAMM [412]	Clinical Trials Data	Clinical Trial Site Selection
Gait-MMR [200]	Images, Silhouettes, Depth	Gait Recognition
LFCR [191]	Multimodal MRI	Brain Tumor Segmentation
MagNET [330]	Multimodal MRI	Brain Tumor Segmentation
MFI [303]	Multimodal MRI	Brain Tumor Segmentation
MMAN-M2 [395]	Audio, Video	Emotion Recognition
MMM-VoxCeleb [158]	Audio, Video	Speaker Identification
MMMB [300]	Images, Text	Image Aesthetic Prediction
MMS-AttnFusion [261]	Multimodal MRI	Brain Tumor Segmentation
TAG-MSA [295]	Video, Audio, Text	Sentiment Analysis
TAG-RMSA [297]	Video, Audio, Text	Sentiment Analysis

in the underlying transformer architecture. Similar recent works include inter-model contrastive learning [343], masked cross model attention [412], and tag-assisted attention [295]. In addition, attention-based architectures are employed for sensor fusion [264], brain tumor segmentation [191], [252], [261], [303], sentiment analysis [158], [297], and image aesthetic prediction [300].

Prompt Learning

Prompt learning leverages modality-specific prompts that guide the model in adapting to varying input conditions. These learnable prompts enhance the generalizability by acting as auxiliary inputs that allow the underlying network to recognize and respond to the presence or absence of particular modalities [389]. Table 11 outlines the existing literature that leverages prompt learning. For modality-aware prompts, Lee et al. [334] leveraged dynamically learnable training where the proposed model learns adaptive prompt embeddings to indicate the availability of each modality. Pipoli et al. [455] proposed a semantically-driven prompt learning module that generates sample-specific prompts by querying a memory bank using semantic features from the available modalities. Similarly, Yue et al. [467] introduced modality-specific and task-aware prompts

Table 11: Summary of Prompt learning methods for Missing Modality Problem

Method	Modalities	Application Area
M&TSP [436]	Text, Audio, Video	Emotion Recognition
MIFPN [429]	Multimodal MRI	Brain Tumor Segmentation
MMWP [334]	Images, Text	Food Recognition
MSPs [389]	Images, Audio	Movie Genre Classification
PAL [467]	Audio, Video, Text	Food Recognition
RAGPT [446]	Text, Images	Hate Speech Detection
SCP [455]	Images, Text	Hate Speech Detection

that dynamically adjust the model's behavior by leveraging both intra- and inter-modal features, significantly enhancing robustness to missing data. Recent advancements have pushed the boundaries of prompt-based adaptation further. Lang et al. [446] employed context-aware retrieval mechanisms to generate instance-specific prompts that dynamically tailor the model's response to missing modality conditions. Similarly, Guo et al. [436] designed a system with modality-specific, task-specific, and task-aware prompts, capturing fine-grained modality-task interactions.
In summary, architectural techniques ranging from modular designs and distillation to prompt-based learning and attention, provide computationally efficient means for introducing robustness in multimodal systems and while these methods show promising success, key challenges remain in managing the trade-off between computational efficiency and performance, especially across diverse tasks and modality combinations. Furthermore, task-independent designs, refinements to prompt learning, and novel attention or fusion mechanisms to create more adaptable and resilient multimodal systems can be explored for enhanced model resilience.

2.2.4 Hybrid Approaches

Hybrid methods aim to overcome the limitations of individual techniques by combining their complementary strengths, often yielding significantly improved performance under diverse missing modality scenarios. For instance, MTMSA [401] combined modality translation with fusion mechanisms in an encoder-decoder setup to enhance sentiment analysis per-

formance under missing modality conditions. Zhou et al. [375] proposed a hybrid of cross-model fusion leveraging multi-scale fusion and multi-task learning, to improve robustness in multimodal tumor segmentation. Likewise, Chen et al. [310] proposed imputation of missing modalities through combined distillation networks with a disentanglement module to capture both intra- and inter-modal characteristics. Table 12 highlights the work leveraging hybrid strategies such as model design integrated with dynamic fusion [271], self-supervised learning using attention and pre-trained encoders [221], and semi-supervised reconstruction-based co-learning frameworks [103], [107]. Furthermore, recent hybrid mod-

Table 12: Summary of Hybrid methods for Missing Modality Problem

Method	Modalities	Application Area
ADxPro [428]	MRI, PET, CSF, Biomarkers	Alzheimer Prediction
DFTD [310]	MRI, PET	Alzheimer Diagnosis
DGMRec [444]	Text, Images, Audio	Recommendation System
EBIA [470]	Text, Audio, Video	Sentiment Analysis
FedMobile [453]	Multimodal Sensor Data	Human Activity Detection
FF-LFL [375]	MRI, CT, PET	Brain Tumor Segmentation
HN-CNN [107]	Optical, DSM	Remote Sensing, Classification
HPMT [452]	Text, Images	Long Document Classification
IMDR [451]	Fundus, OCT	Ophthalmic Diagnosis
iMMAir [431]	Images, Sensor Data	Air Quality Prediction
MSGFN [468]	Text, Audio, Video	Sentiment Analysis
MFCPL [447]	Audio, Video, Text	Hate Speech Detection
Mi-CGA [443]	Text, Audio, Video	Emotion Recognition
MMGACL [471]	Text, Images	Recommendation System
MMT-RMM [271]	Images, Text	Hate Speech Detection
MSAmbA [465]	Audio, Video	Human Action Recognition
MT-MSA [401]	Text, Audio, Video	Sentiment Analysis
OGP-Net [459]	Images, Infrared	Scene Segmentation
OPTIMUS [430]	MRI, PET, CSF	Alzheimer Prediction
SGM [103]	Image, Audio, Signals	Emotion Recognition
SSFD-Net [454]	Palmprint, Palmvein	Biometric Recognition
TGRN [458]	Text, Audio, Video	Sentiment Analysis
TIP [384]	Tabular, Images	Myocardial Prediction
UGaitNet [221]	Optical Flow, Silhouette	Gait Recognition

els continue this trend by combining diverse learning paradigms such as contrastive learning, disentanglement, proxy learning, and federated training. Dhivyaa et al. [428] and Zhang et al. [470] introduced hybrid models using VAEs, RNNs, and fuzzy-aware modules for healthcare and sentiment analysis. Kieu et al. [443] and Liu et al. [452] employed graph neural networks and hierarchical modeling to address missing modalities in emotion and document classification. More recent works such as Fan

et al. [431], SSFD-Net [454], IMDR [451], and FedMobile [453] integrated diffusion-based disentanglement, proxy learning, and federated MML. Others such as Zhao et al. [471], Le et al. [447], and Du et al. [384] used attention-enhanced fusion and contrastive alignment strategies.
In summary, hybrid methods present a promising way towards robustness with the issue of missing modalities by strategical combination of different methods that show great resilience. However, this introduces several challenges including complex architectural design, computationally expensiveness, and modality alignment when integrating different approaches. Future works can investigate these problems and can propose efficient hybrid designs that overcome these issues.

2.2.5 Benchmark Datasets

To facilitate research on robustness towards missing modalities in MML, multiple benchmark datasets have been explored in the existing literature. We summarize them in Table 13. For instance, in affective computing, SMIL-MOSI [219], MISA-MOSEI [163], and VQA-Missing [438] are used to evaluate architectural robustness under missing modalities. Table 13 summarizes the benchmark datasets that are explored in the existing literature for MML under missing modalities.

2.2.6 Key Findings

Addressing missing modalities remains a critical challenge in MML, driving the development of a diverse range of strategies. This section has surveyed the major ones employed through a comprehensive taxonomy presented in Fig. 2. The proposed taxonomy include Reconstruction, Architectural, and Hybrid approaches with each offering distinct mechanisms to handle missingness in MML. *Reconstruction* approaches focus on estimating or synthesizing the missing modality from existing ones by leveraging cross-modal relationships. Generative methods such as GANs, VAEs, and diffusion models have demonstrated success in producing semantically coherent and high-fidelity reconstructions by learning the joint distribution of multimodal data [263], [284], [433], [441], [442]. While pow-

Table 13: Existing benchmark datasets for MML for MMP. Modalities: Audio (A), Video (V), Image (I), and Text (T). †One or more modalities are *synthetically* dropped or naturally absent in derived splits.

Dataset†	A	V	I	T	Application Area
Action and Activity Recognition					
Kinetics-Sounds [80]	✓	✓	✗	✗	Action Recognition
NTU RGB-D [68]	✗	✗	✓	✗	Activity Recognition
Event Detection					
AudioSet (inc.) [83]	✓	✓	✗	✗	Event Classification
General Classification					
MM-IMDB [79]	✗	✓	✓	✓	Multi-label Classification
Hate Speech Detection					
Hateful Memes [165]	✗	✗	✓	✓	Multimodal Classification
Medical AI (Diagnosis, Prognosis, Imaging)					
BraTS 2018/2020 [100]	✗	✗	✓	✗	Image segmentation
M3Care [298]	✗	✗	✗	✓	Prognosis Prediction
MIMIC-III [65]	✗	✗	✗	✓	Clinical Diagnosis
Multimodal Translation and ASR					
How2 [118]	✓	✓	✗	✓	Machine Translation
Retrieval					
COCO-Missing [286]	✗	✗	✓	✓	Image–text Retrieval
Flickr30K-Missing [286]	✗	✗	✓	✓	Image–text Retrieval
MIT-States [44]	✗	✗	✓	✓	Zero-shot Learning
Recipe1M+/Food-101 [59]	✗	✗	✓	✓	Cross-modal Retrieval
Sentiment and Emotion Classification					
CMU-MOSEI [99]	✓	✓	✗	✓	Emotion Recognition
IEMOCAP [9]	✓	✓	✗	✓	Emotion Recognition
MELD [116]	✓	✓	✗	✓	Emotion Recognition
MISA-MOSEI [163]	✓	✓	✗	✓	Sentiment Analysis
SMIL-MOSI [219]	✓	✓	✗	✓	Sentiment Analysis
Visual Question Answering					
VQA-Missing [281]	✗	✗	✓	✓	Robust Learning

erful, these methods often entail high computational complexity and can be sensitive to noise or outliers in the training data [235], [246]. Alignment-based methods, in contrast, do not generate the missing modality directly but instead embed all available modalities into a shared latent space. This enables the model to infer robust multimodal representations without explicit imputation [110], [259], [298]. Though computationally lighter, these methods can struggle with modality misalignment and semantic drift under high missingness. *Architectural* approaches avoid reconstruct-

ing missing modalities altogether and focus instead on adapting model architectures or learning dynamics to be inherently resilient. Techniques such as modality-specific architectural designs [403], [407], late fusion [55], [248], and co-learning frameworks [331], [338] allow models to function robustly with only partial inputs. Knowledge distillation methods train student models with incomplete inputs under the guidance of complete-modality teacher models, which enhances generalization across varying input configurations [363], [372], [456]. Attention-based mechanisms further enhance adaptability by dynamically weighting modalities based on relevance [343], [395]. On the other hand, prompt learning offers a lightweight and modular approach particularly suited to transformer-based architectures [334], [455]. Despite their efficiency, these approaches may require extensive tuning or specialized designs to generalize across tasks and domains. *Hybrid* methods integrate multiple architectural methods, often combining the reconstruction, alignment, attention, distillation, and other methods to capitalize on their complementary strengths. For example, combining cross-modal imputation with fusion and disentanglement has yielded promising results in sentiment analysis and medical imaging [310], [375], [401]. Recent advances further incorporate contrastive learning, federated learning, and graph-based reasoning to address complex real-world scenarios [431], [453], [463], [465]. While hybrid strategies offer improved robustness and generalization, they often involve increased architectural complexity and higher computational demands, which may hinder scalability [428], [470].

2.3 Dissecting Multimodal Learning Under Corrupted Modalities

In real-world scenarios, collecting and processing high-quality multimodal data is inherently challenging due to the possibility of various forms of data corruption within the data. These corruptions can occur during data collection, transmission, or during storage stages, and their effects are often severely compounded in multimodal settings where multiple heterogeneous data streams must be synchronized and learned

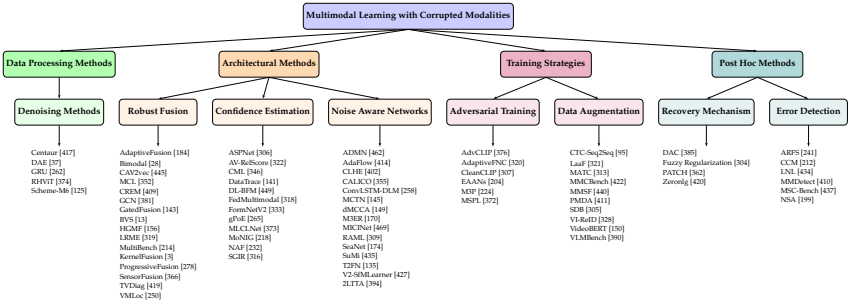


Figure 3: Taxonomy of the existing works addressing multimodal learning (MML) with corrupted modality problem (CMP)

effectively. Therefore, the presence of noise in any single modality can degrade the overall system performance, and when multiple modalities are simultaneously affected, the impact can be severe and unpredictable [423]. In such cases, the corruptions can propagate through the MML pipeline, leading to compounding errors or biased predictions. These corruptions disrupt the integration of diverse data modalities, ultimately decreasing both accuracy and robustness [381], [400]. Empirical studies have further demonstrated that multimodal models are particularly sensitive to corrupted inputs, underscoring the necessity for robust MML methodologies [129], [177], [196], [280], [348], [380], [460]. To address the CMP, a wide array of approaches have been proposed, which we classify into four categories according to our taxonomy of corruption handling methodologies as shown in Fig. 3. We also survey their application areas, outline corresponding benchmark datasets, and highlight open challenges in designing resilient multimodal systems.

2.3.1 Data Processing Methods

Data Processing methods seek to ensure input quality at the initial stage of the data pipeline, by preventing the propagation of corrupted data through the multimodal architecture. They often operate at the signal or feature level and attempt to restore a clean version of the input using statistical or learned techniques. Such methods are especially useful when

the corruption is detectable and relatively localized.

Denoising Methods

Denoising methods such as autoencoders are neural network architectures that are designed to remove the noise from the input modalities. In multimodal settings, they are either trained for each individual modality or jointly across modalities to exploit correlations. Table 14 presents different denoising methods have been explored across diverse application domains. In medical imaging, RHViT [374] proposed a Hierarchical Vision Transformer for brain tumor segmentation that leverages masked image modeling (MIM) with 3D convolutions during pre-training. For sensor noise handling, Centaur [417] leveraged a denoising autoencoder for reconstruction of noisy sensor signals prior to multimodal representation pipeline. In cloud computing, Ikhlas et al. [262] targeted the noise

Table 14: Summary of Denoising methods for Corrupted Modality Problem

Method	Modalities	Application Area
Centaur [417]	Sensors	Activity Recognition
DAE [37]	Images, acoustic bathymetry	Autonomous Vehicles
GRU [262]	vCPU, memory, storage	Resource Forecasting
RHViT [374]	3D multimodal MRI	Medical Imaging
Scheme-M6 [125]	Images	Video Classification

in time series data by leveraging the proposed stacked denoising autoencoder with bidirectional gated recurrent units (BiGRUs). Similarly, Yin et al. [125] denoised the impulse noise in the multimodal streams for multimodal image reconstruction. Such studies contribute towards robust, smart surveillance and video communication systems and underscore the importance of incorporating noise-resilient denoising methods within multimodal architectures for improved resilience against data corruption.

2.3.2 Architectural Methods

Architectural methods are designed to withstand noise during learning and inference. These methods aim to introduce the robustness into the model architecture such that it is dynamically able to actively tolerate the

Table 15: Summary of Noise-Aware methods for Corrupted Modality Problem

Method	Modalities	Application Area
ADMN [462]	Images, Audio, Sensor	Object Localization
AdaFlow [414]	Camera, LiDAR, GPS	Mobile Sensing
CLHE [402]	Text, tabular data	E-commerce
CALICO [355]	Camera, LiDAR	3D Object Detection
ConvLSTM-DLM [258]	Images	Medical Imaging
MCTN [145]	Text, Video, Audio	Sentiment Analysis
dMCCA [149]	Images, Text	Multimodal Learning
M3ER [170]	Images, Text, Audio	Emotion Recognition
MICINet [469]	Images, Text, Audio	Security
RAML [309]	Video, Text, Audio	Emotion Recognition
SeaNet [174]	Audio, Accelerometer	Speech Enhancement
SuMi [435]	Images, Text, Audio	Test-Time Adaptation
T2FN [135]	Text, Video, Audio	Speech Recognition
V2-SfMLearner [427]	Images, Vibration	Robotics
2LTTA [394]	Images, Text, Audio	Test-Time Adaptation

present corruption. Noise-aware Networks, Confidence Estimation, and Robust Fusion strategies are among the key strategies. Corresponding approaches are especially valuable when corruption is subtle, modality-specific, or occurs during deployment, making it impractical to rely solely on preprocessing.

Noise-aware Networks

They explicitly anticipate and introduce uncertainty of corrupted input into their model design. Instead of treating all inputs equally, these networks are structured to model and respond to noise during both training and inference. These architectures simulate the likelihood of noise by introducing modality masking, gating, and dropout during training to introduce resilience in the architecture. Table 15 highlights the existing studies that leverage novel design choices to withstand data corruption in multimodal networks. For instance, SuMi [435] addressed the stable test-time adaptation under complex multimodal noise by utilizing statistical smoothing and mutual information sharing towards noise-awareness in present modalities.

Zhang et al. [469] proposed MICINet, which targets inter-class con-

fusing information (ICI) by identifying and removing structured noise patterns across vision, audio, and text inputs. In network design, ADMN [462] proposed dynamic architectural adjustments in response to varying levels of Gaussian noise across RGB, depth, radar, audio, and WiFi inputs. Lei et al. [394] explored model parameter manipulation for test-time robustness through a two-level adaptation strategy. Leveraging contrastive learning, Ma et el. [402] proposed learning with noisy and sparse multimodal data in recommendation systems. Additional strategies including tensor rank regularization [135], cyclic translation between modalities [145], and cross-modal correlation learning [149] have been proposed to recover clean signals from noisy or missing data. These approaches show that incorporating noise handling directly into network design improves generalization and resilience in multimodal settings.

Confidence Estimation

Confidence estimation evaluates the trustworthiness of the input modalities dynamically by computing confidence score either learned or through heuristic methods. It estimates the perceived quality of the input and guides the architecture towards the contributions of the participating modalities towards the final multimodal prediction. Table 16 summarizes the recent studies leveraging diverse strategies for confidence estimation for CMP. For instance, Li et al. [449] proposed an adaptive attention-based architecture for mental health monitoring within Internet of Smart Things (IoST) systems that learns to adjust weight of each modality given the contextual relevance. In medical imaging, Zheng et al. [373] proposed MLCLNet for multi-level confidence estimation to assess both feature and label reliability. Similarly, Ma et al. [218] incorporated a Mixture of Normal-Inverse Gamma (MoNIG) distribution to model uncertainty for trustworthy multimodal regression and estimating per-modality confidence in noisy inputs. Calibrating Multimodal Learning (CML) [346] introduced a regularization method that mitigates compromised predictions in the presence of corrupted or missing modalities. FedMultimodal [318] explored federated learning with structured and unstructured noise and emphasized on the need for calibrated confidence signals across

Table 16: Summary of Confidence estimation methods for Corrupted Modality Problem

Method	Modalities	Application Area
ASPNet [306]	Video, Sensor	Action Segmentation
AV-RelScore [322]	Images, Audio	Speech Recognition
CML [346]	Video, Text, Audio	Classification
DataTrace [141]	Text, Sensor	Anomaly Detection
DL-BFM [449]	Text, Audio, Sensor	Health Monitoring
FedMultimodal [318]	Images, Text, Audio	Activity Recognition
FormNetV2 [333]	Text, Layout, Image	Document Analysis
gPoE [265]	Images, Sensor	Video Segmentation
MLCLNet [373]	Images, Sensor	Medical AI
MoNIG [218]	Images, Text	Sentiment Analysis
NAF [232]	Video, Audio	Speech Extraction
SGIR [316]	Images, Text, Audio	Classification

distributed and unreliable input sources. Hong et al. [322] proposed AV-RelScore, a reliability scoring module for audio-visual speech recognition (AVSR). Similarly, Ding et al. [316] proposed Star Graph Interaction for confidence-based weighting via centralized hubs and private pathways to maintain robustness under noisy conditions. In document understanding, FormNetV2 [333] leveraged graph contrastive learning for learning robust representations from noisy text and image data. Across these works, confidence estimation proves critical not only for accurate predictions but also for graceful degradation in the presence of real-world noise and modality-specific corruption.

Robust Fusion

Robust fusion strategies are designed to suppress the impact of corrupted inputs during the fusion process. Instead of assuming that all modalities are equally informative, these methods selectively combine modalities based on context, reliability, or mutual agreement. Some approaches fuse only those modalities that exceed a quality threshold, or use late fusion to maintain independent predictions that can be reconciled based on confidence. Table 17 highlights the recent works that leverage fusion for CMP. For instance, Kim et al. [445] proposed leveraging fusion with self-supervised learning for Audio-Visual Speech Recognition to

jointly address noise in audio and visual modalities. Similarly, Shi et al. [409] introduced context-aware fusion network for remote sensing that addresses the structured noise in Digital Surface Models (DSMs) by using a Context Representation Enhancement Module (CREM). Patel et al. [143] developed a gated architecture for Unmanned Ground Vehicle (UGV) navigation that fuses camera and LiDAR data towards ensuring fault tolerance to sensor dropout and noise. In the industrial domain,

Table 17: Summary of Robust fusion methods for Corrupted Modality Problem

Method	Modalities	Application Area
AFusion [184]	Images, Sensor	Emotion Detection
Bimodal [28]	Speech, Facial Images	Smart Surveillance
CAV2vec [445]	Audio, Video	Speech Recognition
MCL [352]	Gas Sensor, Images	Industry 5.0
CREM [409]	Optical Imagery, DSMs	Remote Sensing
GCN [381]	Images, Text	Scene Recognition
GatedFusion [143]	Images, LiDAR	Object Detection
BVS [13]	Images, Sensor	Biometric Security
HGMF [156]	Images, Audio, Video	Data Mining
LRME [319]	Sensor, Time-Series	Industry 5.0
MultiBench [214]	Text, Images, Video, Audio	HCI, Healthcare
KernelFusion [3]	Image, Text, Audio, Video	Content Analysis
PFusion [278]	Text, Images, Audio, Sensor	Sentiment Analysis
SensorFusion [366]	LiDAR, Images	Autonomous Driving
TVDiag [419]	Logs, Metrics, Traces	Failure Diagnosis
VMLoc [250]	RGB Images, Depth Maps	Robotics

Fu et al. [319] applied a low-rank multi-manifold embedding approach for monitoring incomplete sensor readings Rahate et al. [352] leveraged co-learning-based fusion for handling noisy sensor and thermal images and demonstrated that robust multimodal training can mitigate corruption in safety-critical applications. Graph-based approaches like HGMF by Chen et al. [156] integrate compromised multimodal data by using a heterogeneous graph representation and ensuring that only reliable modalities influence the final prediction. For facial action recognition, Yang et al. [184] proposed Adaptive Multimodal Fusion (AMF) that dynamically adjusts the architectural weights to handle modality-specific

occlusions and improve expression recognition. Similarly, Raghavendra et al. [13] enhanced person verification by combining palm print and hand vein information while using a noise-resistant edge mask for fusion under visual corruption. *In summary*, architectural approaches offer high flexibility and runtime adaptability that makes them suitable for dynamic and safety-critical systems where data quality may fluctuate. However, they often introduce new challenges such as architectural complexity and increased training demands. Moreover, estimating the confidence of noise accurately in real-time remains an open research challenge, especially when corruption is adversarial or multimodal in nature.

2.3.3 Training Strategies

Unlike preprocessing or architectural strategies that primarily operate at pre-training or model design stages, training strategies aim to induce models with generalization capabilities under noise and corruptions by modifying the input representations or learning objectives.

Data Augmentation

Corruption-based data augmentation involves augmenting the training data with artificially generated corrupted samples to simulate real-world noisy data. These methods infuse robustness by diversifying the input distribution and enable the model to learn features invariant to noise. In audiovisual domain, frame dropout and motion blur have been applied to simulate sensor-induced artifacts during video pretraining [150]. For speech recognition, time masking and additive noise improve model robustness to environmental corruptions [95]. In multimodal sentiment analysis, partial masking of audio or textual inputs has been employed to force the model to rely on complementary visual signals [313].

Recent studies have expanded corruption-based data augmentation to increasingly complex and realistic multimodal settings as shown in Table 18. Kaplan et al. [390] investigated the effect of structured text corruptions in long-form vision-language tasks, revealing their influence on hallucination behaviors and offering insights into robustness under

Table 18: Summary of Augmentation-based methods for Corrupted Modality Problem

Method	Modalities	Application Area
CTC-Seq2Seq [95]	Audio, Video	Speech Recognition
LaaF [321]	MRI, Clinical Data	Medical Imaging
MATC [313]	Text, Audio	Emotion Recognition
MMCBench [422]	Images, Text, Audio	Multimodal Classification
MMSF [440]	Images, Infrared	Smart Surveillance
PMDA [411]	Images, Infrared	Smart Surveillance
SDB [305]	Images, Sensor	Industry 5.0
VI-ReID [328]	Images, Infrared	Surveillance Systems
VideoBERT [150]	Audio, Video	Action Recognition
VLMBench [390]	Text, Images	Long Form Generation

systematic textual noise. Zhang et al. [422] provided a broad evaluation of large multimodal models prone to common data corruptions across vision, text, and speech modalities. In industrial and robotics contexts, Altinses et al. [305] leveraged augmentation techniques specifically tailored to realistic sensor failures. In medical imaging, Hager et al. [321] applied corruption-aware contrastive learning to handle noisy or incomplete cardiac imaging and clinical tabular data. These studies collectively demonstrate how corruption-based data augmentation not only enhances model robustness but also promotes generalization across a wide array of multimodal applications, from surveillance and healthcare to industrial automation.

Adversarial Training

Adversarial training introduces perturbations that are artificially curated to degrade model performance, with the aim of improving robustness against worst-case inputs. In the multimodal setting, perturbations may target individual modalities, cross-modal alignments, or shared latent spaces. By incorporating adversarial examples during training, the model learns to recognize and resist subtle but harmful perturbations that may not be captured through simple data augmentation. Table 19 summarizes studies that have explored a variety of adversarial training techniques

to enhance the resilience of multimodal models against data corruption. For instance, Huang et al. [204] leveraged evolutionary adversarial attention networks (EAANs) to withstand the random noise in visual data. Their method incorporates adversarial perturbations during training to refine multimodal feature representations. *CleanCLIP* [307] specifically targets data poisoning attacks in multimodal contrastive learning, where poisoned images are introduced to implant backdoors in vision-language models. Similarly, *AdvCLIP* [376] leveraged artificially curated adversarial examples that target both image and text modalities. Similarly, Gupta

Table 19: Summary of Adversarial training methods for Corrupted Modality Problem

Method	Modalities	Application Area
AdvCLIP [376]	Image, Text	Contrastive Learning
AdaptiveFNC [320]	Image, Video	Video Classification
CleanCLIP [307]	Image, Text	Adversarial Learning
EAANs [204]	Image, Video, Text	Multimodal Classification
M3P [224]	Text, Image	Captioning, Retrieval
MSPL [372]	Images	Image Classification

et al. [320] explored adversarial manipulations in self-supervised contrastive learning by mitigating false negative pairs, showing that careful negative sampling improves adversarial robustness in both image and video classification tasks.

In summary, Training-based strategies are generally model-agnostic and can be integrated with a wide range of multimodal architectures. However, their effectiveness is highly dependent upon the type and severity of the corruptions employed during training. Additionally, adversarial training methods, while powerful, introduce significant computational overhead and may suffer from reduced generalization if the perturbation space is not carefully controlled. Nevertheless, when appropriately designed, these strategies significantly enhance a model's robustness to modality-specific or cross-modal corruptions.

2.3.4 Post-hoc Methods

Post-hoc correction methods are designed to operate after the intermediate architectural stages to detect and mitigate the impact of noise. They are especially valuable in safety-critical systems where runtime correction can enhance model reliability without requiring complete reprocessing.

Error Detection

Error detection models seek to distinguish between clean and noisy modality signals, often leveraging consistency checks across modalities or statistical deviations from expected representations. For instance, confidence-based approaches assess internal model uncertainty to flag potentially corrupted data segments, while cross-modal agreement techniques detect inconsistencies in predicted semantic content between modalities. Table 20 presents the recent studies that have explored various post-hoc error detection techniques for CMP. For instance, Gou et al. [434] leveraged internal model signals to distinguish between clean and corrupted samples. They utilized a small clean dataset for recovery and training noise-aware model that actively tolerates corruption without external supervision. Lee et al. [212] proposed a "Detect, Reject, Correct" framework

Table 20: Summary of Error detection methods for Corrupted Modality Problem

Method	Modalities	Application Area
ARFS [241]	Images, Audio, Text	Multimodal Learning
CCM [212]	Images, Depth, Sensor	Autonomous Vehicles
LNL [434]	Images, Text	Language Understanding
MMDetect [410]	Images, Text	LLM Evaluation
MSC-Bench [437]	Camera, LiDAR	3D Object Detection
NSA [199]	Images	Activity Recognition

for detecting corrupted sensor modalities by evaluating reconstruction inconsistencies. Cross et al. [199] applied the Negative Selection Algorithm (NSA) to detect anomalous inputs in multimodal activity recognition pipelines. The proposed network identifies the noisy sensor data without prior labeling, making it particularly useful in health monitoring applications.

Table 21: Summary of Recovery based methods for Corrupted Modality Problem

Method	Modalities	Application Area
DAC [385]	Images, 3D Point Clouds	Cross-Modal Retrieval
MMFF [304]	Image, Time Series, Sensor	Industrial Automation
PATCH [362]	Sensor, Video, Infrared, Thermal	Autonomous Driving
Zeronlg [420]	Images, Videos, Text	Video Captioning

Recovery Mechanisms

Recovery mechanisms seek to restore the reliability of the multimodal representation after corrupted inputs have been detected. These mechanisms can range from simple heuristics, such as excluding low-confidence modalities from the final fusion stage, to more advanced techniques like learned reconstruction mappings that regenerate corrupted modality features from clean ones. As shown by Table 21, several cross-modal translation networks and redundancy-aware imputation methods have been explored to recover corrupted modality streams using information from intact modalities. For instance, Gan et al. [385] leveraged a divide-and-conquer method that mitigates the effects of noisy labels in 2D-3D retrieval via credibility modeling. Similarly, PATCH [362] presented a plug-in framework that incorporates masked autoencoders for data recovery, a feature pair ranking module to optimize imputation strategies, and a time-insensitive alignment mechanism to synchronize heterogeneous data streams. In the industrial automation domain, Altinses et al. [304] leveraged fuzzy regularization to handle CMP. Yang et al. [420] addressed the absence of supervision in zero-shot multimodal and multilingual natural language generation. By aligning unpaired data in a shared latent space, they reconstructed target outputs without direct supervision. Together, these methods illustrate diverse strategies for post-hoc correction and enhance model resilience through the recovery of degraded or absent modality signals.

In summary, post-hoc correction methods provide a flexible and computationally efficient layer of robustness, particularly beneficial when upstream components fail to fully account for corruption effects. How-

ever, their effectiveness depends heavily on the accuracy of the error detection mechanisms and the redundancy across modalities. Moreover, while post-hoc methods may successfully mitigate mild to moderate corruption, they may be insufficient under conditions of severe multimodal degradation, emphasizing the need for their integration with upstream robust modeling techniques.

2.3.5 Benchmark Datasets

To this end, several benchmark datasets have been developed and are leveraged to evaluate multimodal robustness under corruption and noise. Datasets such as ImageNet-C [129], COCO-C [137], and PASCAL-C [137] introduce systematic perturbations to assess model stability against visual degradation. These resources, summarized in Table 22, enable consistent assessment of corruption-aware multimodal networks.

2.3.6 Key Findings

This section has surveyed the diverse strategies employed to enhance the robustness of multimodal systems under data corruption. Collectively, these approaches provide a multi-tiered defense against both localized and systemic corruptions commonly encountered in real-world multimodal applications. Data preprocessing methods such as denoising autoencoders [262], [417] improve data quality during the initial stage of processing pipeline. These methods are effective for structured, modality-localized noise and are especially beneficial in low-resource or high-noise environments. However, while not commonly used these days, they often struggle to generalize to complex or semantic corruptions that are harder to localize. Architectural strategies incorporate noise-awareness directly into the architecture through dynamic gating, uncertainty modeling, and confidence-guided fusion [53], [346], [435]. These methods demonstrate strong performance in critical domains such as medical imaging [374] and autonomous driving [462], where runtime resilience is essential. Nonetheless, they introduce higher architectural complexity and require accurate noise modeling which remains a key challenge particularly in multimodal

Table 22: Existing benchmark datasets for MML for CMP. Modalities: Audio (A), Video (V), Text (T), and Image (I). Corruption (C): Noise (N), Perturbations (P), Rephrasing (R), and Variance (V).

Dataset	A	V	T	I	C	Application Area
Audio and Event Detection						
Common Voice-N [155]	✓	✗	✓	✗	N	Speech Recognition
CrisisMMD [97]	✗	✗	✓	✓	N	Crisis Recognition
DESED [222]	✓	✗	✗	✗	N	Event Detection
MIMII [146]	✓	✗	✗	✗	N	Anomaly Detection
Data Retrieval						
Flickr30K-N [213]	✗	✗	✓	✓	N	Cross-modal retrieval
MSR-VTT-N [215]	✗	✓	✓	✗	N	Video–text retrieval
Winoground [283]	✗	✗	✓	✓	N	Cross-modal reasoning
Image Classification and Captioning						
CIFAR-10-C [129]	✗	✗	✗	✓	N	Image classification
Cityscapes-C [137]	✗	✗	✗	✓	N	Semantic segmentation
COCO-C [137]	✗	✗	✗	✓	N	Object detection
COCO-Noisy [197]	✗	✗	✓	✓	N	Image–text alignment
ImageNet-C [129]	✗	✗	✗	✓	N	Image classification
ImageNet-P [129]	✗	✗	✗	✓	P	Image classification
PASCAL-C [137]	✗	✗	✗	✓	N	Object detection
Sentiment and Emotion Classification						
CMU-MOSI-R [173]	✓	✓	✓	✗	R	Sentiment Classification
M3ED [269]	✗	✓	✓	✗	V	Emotion Recognition
VoxCeleb-N [92]	✓	✓	✗	✗	N	Speaker Identification
Visual Question Answering						
R-Bench [448]	✗	✗	✓	✓	N	Multimodal Reasoning
VQA-CP [96]	✗	✗	✓	✓	R	Bias Evaluation
VQA-Rephrasings [147]	✗	✗	✓	✓	R	Linguistic Variation
VQA-Robust [438]	✗	✗	✓	✓	N	Noise-tolerant VQA

scenarios with asynchronous or unbalanced signals. Training-time approaches build robustness proactively by simulating noisy environments through augmentation or adversarial perturbations. Corruption-based augmentation has been shown to improve generalization across modalities [95], [440], while adversarial training strategies [307], [376] enhance robustness under worst-case scenarios. However, such methods demand careful corruption design to avoid overfitting and may incur significant computational cost during training. Lastly, Post-hoc correction techniques, such as error detection and recovery mechanisms [212], [362], [434], operate downstream to detect and mitigate the influence of corrupted modalities. These are lightweight and adaptable to runtime systems, but their success is contingent on redundancy among modalities and the precision of error detection modules. In combination, these strategies form a layered robustness framework that addresses different stages of the multimodal

pipeline. Meanwhile, effective systems often integrate these methods across different architectural layers to balance robustness, efficiency, and scalability. Despite these advances, several open challenges remain. For instance, the lack of standardized benchmarks for multimodal corruption [380], [422] hinders consistent evaluation across tasks and domains.

2.4 Chapter Summary

In this chapter, we provided a comprehensive overview of multimodal learning with a focus on three fundamental and critical aspects: general architectural design, performance deterioration under missing modalities, and robustness to corrupted modalities. We review state-of-the-art methods, highlight the key application areas, and outline the recent advances in addressing these challenges. Furthermore, we presented benchmark datasets and comprehensive taxonomized literature review to strengthen the fundamentals of understanding MML. In essence, this chapter serves as a valuable resource for researchers aiming to deepen their understanding and strengthen the robustness of multimodal learning.

Chapter 3

Unraveling Multimodal Systems: Presenting Chameleon for Missing Modalities

In recent years, there has been a surge in the use of multimodal data for various applications. For instance, users combine text, image, audio, or video modalities to sell a product over an e-commerce platform or express views on social media platforms. Two or more of these modalities are often combined to solve different tasks such as multimodal classification [108], [165], cross-modal retrieval [72], cross-modal verification [113], multimodal named entity recognition [111], [126], visual question answering [63], [98], image captioning [58], semantic relatedness [33], segmentation [356], and multimodal machine translation [61], [69]. In all these tasks, multimodal methods can experience scenarios where some modalities are missing, e.g., due to failures in data acquisition pipelines. Several researchers have recently concluded that multimodal learning

The contents of this chapter are taken from our work "Chameleon: A Multimodal Learning Framework Robust to Missing Modalities" available `https://link.springer.com/article/10.1007/s13735-025-00370-y`

is not *inherently* robust to missing modalities and can result in a significantly deteriorated performance when modalities are missing [219], [271], [334]. For example, as seen in Fig. 4, ViLT [210], a baseline multimodal Transformer demonstrates significant performance drops when the textual modality is missing at test time. The performance deterioration is attributed to the fundamental design that assumes the presence of all modalities concurrently for training, thus inadvertently developing a dependency on having a complete set of modalities to make a prediction. Therefore, if a modality is missed during inference, the performance deteriorates significantly [271]. Recently, Ma et al. [271] proposed a strategy to improve robustness of ViLT against missing modality by incorporating modal-incomplete data. As seen in Fig. 4, the performance on missing modality scenarios improves notably using this approach. Though it reduces the dependency of ViLT on having a complete set of modalities to make a prediction, it requires a data-centric fusion strategy to handle modal-incomplete data which is cumbersome to optimize [271].

To resolve this, it is pertinent to have an approach that may be trainable using the *complete set* of modalities but resilient to missing modalities. To this end, we propose `Chameleon`, a framework that adapts a common-space visual learning network to align all input modalities. This way, the network is able to process either of the two inputs, visual or non-visual, as well as both together, all as images. Therefore, when a modality is missing, `Chameleon` leverages the available modality to output the correct prediction scores. Moreover, mapping different modalities into a single shared embedding space not only simplifies the input interface but also allows harnessing architectures or training procedures designed for one domain into another domain. As seen in Fig. 4, `Chameleon` not only outperforms the baseline ViLT by a notable margin but also demonstrates better robustness than Ma et al. [271] without any data-centric optimization. To explore the general applicability of `Chameleon`, we incorporate various kinds of modalities (including, audio and text) for training the multimodal network by proposing an encoding scheme that transforms non-visual modalities into a visual format. We evaluate `Chameleon` on six benchmark datasets including four textual-visual datasets (i.e., UPMC

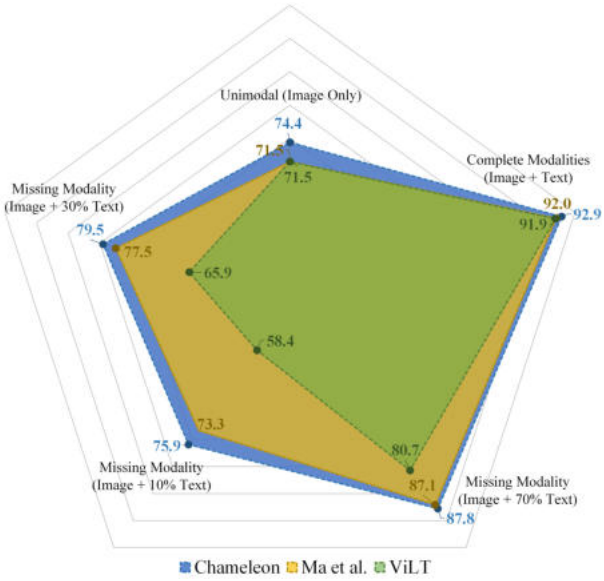


Figure 4: Comparison of `Chameleon` with ViLT [210] & Ma et al. [271] on UPMC Food-101 dataset. `Chameleon` demonstrates better multimodal and unimodal performances and retains significantly superior performance when textual modality is missing at test time. While transformer-based methods such as Ma et al. achieve competitive performance under complete modality settings, they exhibit greater sensitivity to missing modalities. In contrast, `Chameleon` maintains more stable performance across varying levels of modality availability, including strong unimodal performance when evaluated with image-only inputs

Food-101 [59], Hateful Memes [165], MM-IMDb [79], and Ferramenta [81]) and two audio-visual datasets (i.e., avMNIST [123], VoxCeleb [92]). Our approach outperforms state-of-the-art (SOTA) multimodal methods on complete modalities and demonstrates superior robustness against missing modalities during training and testing. The key highlights of our work are as follows:

1. We propose `Chameleon`, a multimodal learning framework robust to missing modalities.

2. To eliminate the dependency of the learning network on complete

set of modalities, we propose a novel encoding scheme that transforms non-visual modality into visual representations to carry out multimodal training using shared weights.

3. We evaluate our framework extensively on Convolutional Neural Networks (CNNs), Vision Transformers (ViT), and Adapters, demonstrating its general applicability across various networks. Notably, unlike previous study [271], we demonstrate that the robustness of Transformers against missing modalities can be improved significantly by using an appropriate learning framework such as `Chameleon`.

4. A wide range of experiments performed on four textual-visual and two audio-visual datasets under different multimodal, and missing modality settings during training and testing demonstrate the significance of `Chameleon` in application scenarios with missing modalities.

3.1 Rendering Non-visual Modalities into Images

Recently, some researchers have explored the idea of transforming text information into a visual format to perform visual recognition task. For example, Gallo et al. [82] leveraged Word2Vec word embeddings to reconstruct the semantics associated with text as an image to train a visual network for text classification. Salesky et al. [230] rendered the raw text directly into visual format, divide the rendered image into overlapping slices, and produce representations with optical character recognition to train a machine translation model. Similarly, Rust et al. [274] proposed a Masked Autoencoding Visual Transformer to reconstruct the pixels in masked image patches to train a language model. More recently, Tschannen et al. [357] proposed CLIPPO that renders text information as images to train a pure pixel-based model performing multimodal learning. All approaches rendering text into visual format are, in essence, related to `Chameleon` as we also encode text into visual format. However, instead

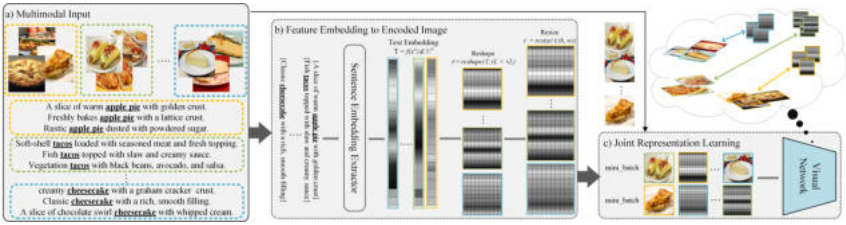


Figure 5: Architecture of `Chameleon`. (a) Multimodal input consists of images and texts. (b) Embeddings of texts are encoded as images. (c) Visual and encoded text inputs are arbitrarily input to a weight-sharing visual network in a mini batch. Note that the figure represents the case of textual-visual modality pair. In the case of audio-visual pair, the same encoding is applied to the audio embedding.

of the common approach of transforming raw text directly into images, we use embeddings to encode audio or textual information which is found to be effective in our experiments for the missing modality problem (Section 3.3.2). In addition, we acknowledge the existing approaches [92], [140], [153] that encode audio modality into visual format (e.g., spectrogram [92]) to train on a given task. However, we propose a generic encoding scheme instead of using a particular modality (like text in [357]) directly as input making our approach generalizable to any modality.

3.2 Methodology

`Chameleon` relies on the intuition that a common feature space across modalities enable learning of multimodal representations robust to missing modality. Each component of `Chameleon` is visualized in Fig. 5 and discussed subsequently in this section:

3.2.1 Problem Formulation

Formally, given $\mathcal{D} = \{(x_i^a, x_i^b)\}_{i=1}^N$ is the training set where N is the number of pairs of modality a and modality b and x_i^a and x_i^b are the individual modality samples of the i^{th} instance respectively. Moreover, each pair (x_i^a, x_i^b) has a class label y_i. Typical existing multimodal methods

take multiple modalities as input by using a multi-branch network $\mathcal{C}_m$ to perform the classification task:

$$\tilde{y}_i = \mathcal{C}_m(\{x^a, x^b\}, y_i) \tag{3.1}$$

Training such a multi-branch configuration inadvertently depends on modality-complete data to perform a given task. This results in a significant performance deterioration during inference in case of missing modality (as seen in Fig. 4). To alleviate this issue, we propose to encode non-visual modality as image and perform the multimodal task entirely in the visual domain, both during training and testing. In other words, a visual network is trained on multimodal data consisting of images and non-visual modality through a common interface of visual information. Building on the motivation from section 3.1, we hypothesize that unifying the modalities to an identical input format makes the model independent of modality-specific branches, thus eliminating the need of modality complete data and results in a framework robust to missing modalities. Therefore, in `Chameleon`, Eq. (3.1) takes the following form:

$$\tilde{y}_i = \mathcal{C}_v(\{\hat{x}^a, x^b\}, y_i), \tag{3.2}$$

where $\mathcal{C}_v$ is a visual classifier, $\hat{x}^a = \mathcal{E}(x^a)$, and $\mathcal{E}$ is the encoding scheme to transforms x^a into visual format $\hat{x}^a$ [1].

3.2.2 Encoding Scheme

To transform non-visual modality (e.g., audio or text) into a common visual format ($\hat{x}^a$), we extract the modality-specific embeddings (T). Formally:

$$T = f(x^a) \in \mathbb{R}^d \tag{3.3}$$

This results in an embedding of length d, which is reshaped into a square matrix:

[1] While the description in Eq. 3.2 holds for arbitrary modality types, in our approach, one of the modalities is always visual.

$$\tau = \text{reshape}(T, \lceil\sqrt{d}\rceil \times \lceil\sqrt{d}\rceil) \tag{3.4}$$

The square matrix is then resized to match the corresponding accompanying image dimensions.

$$\tilde{\tau} = \text{resize}\left(\tau, (h, w)\right) \tag{3.5}$$

where h and w are the height and width of the image respectively. This results in an additional image containing only the encoded text as visual information. For training, the two images are input arbitrarily to a weight-shared network.

3.2.3 Modality Passing and Joint Training Strategy

`Chameleon` supports two distinct training and evaluation strategies, each differing in how the visual and encoded non-visual modalities are combined.

Fused Training. In the fused setting, the encoded non-visual representation $\hat{x}^a$ (e.g., the text embedding rendered as an image) is overlaid directly onto the original image x^b, producing a single composite image that encapsulates both modalities. This fused image is passed through the visual classifier $\mathcal{C}_v$ as a single forward pass:

$$\tilde{y}_i = \mathcal{C}_v\left(\hat{x}^a \oplus x^b\right) \tag{3.6}$$

where $\oplus$ denotes the overlay operation. During evaluation, the same fused representation is constructed and passed through the network. In the missing modality setting, only the available modality image is presented to the network — since the model has been trained on a single-image interface, it handles the unimodal input naturally, resulting in strong missing modality robustness without any architectural modification.

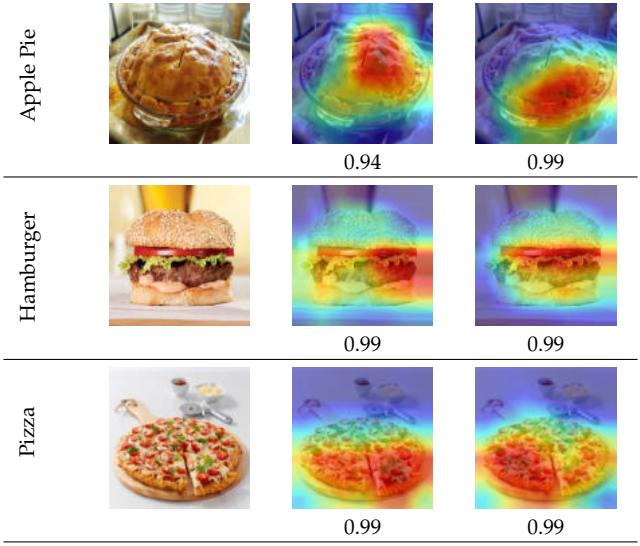


Figure 6: (a) Orginal image. (b) Unimodal (image only) training and testing. (c) Multimodal training, unimodal testing (image only), i.e., 100% text missing. As seen, in unimodal image-only training (b), the model focuses on distinct features of the object. When the text modality (c) is missing during testing, the model focuses on the available modality to make correct final predictions demonstrating the success of our approach in training the multimodal system robust to missing modalities.

Joint Training. In the joint setting, the two modality representations — the original image x^b and the encoded image $\hat{x}^a$ — remain separate and are passed *sequentially* and *independently* through the shared-weight visual classifier within each mini-batch, each carrying the same class label y_i. The training loss is computed independently for each forward pass and accumulated:

$$\mathcal{L} = \mathcal{L}\left(\mathcal{C}_v(\hat{x}^a),\, y_i\right) + \mathcal{L}\left(\mathcal{C}_v(x^b),\, y_i\right) \tag{3.7}$$

This effectively doubles the number of training samples per instance

and encourages the shared network to learn representations that are independently discriminative for each modality. During evaluation under the modality-complete setting, the logits from both forward passes are combined via a simple average:

$$\tilde{y}_i = \frac{1}{2}\left(\mathcal{C}_v(\hat{x}^a) + \mathcal{C}_v(x^b)\right) \tag{3.8}$$

In the missing modality setting, only the logits from the available modality are used directly for prediction — no imputation or substitution is required, as the shared-weight training has already equipped the network to perform well on either input independently.

3.2.4 Robustness to Missing Modalities

`Chameleon` inherently learns shared representations across modalities, aligning them into a common latent space. This results in a richer representation that is independent of the presence of both modalities, thus yielding a robust model in case one of the modality is missing. When a modality is missing, the available modality enables the use of the multimodal knowledge from the shared latent space to make a prediction. We observe this by visually comparing the activation maps [93] of cases taken from UPMC Food-101 for image-only unimodal training/testing and multimodal training with missing text modality testing in Fig. 6. In image-only unimodal training and testing (Fig. 6b), as expected, the model focuses on the distinct features of the object to predict the classification score. In the case of missing modality at test time, a multimodal architecture robust to missing modalities should ideally be able to retain the focus on the available modality and behave like a unimodal network. We observe this case in Fig. 6c for `Chameleon`, when the text modality is missing at test time. As seen, the model shifts its focus on the object itself and behaves comparably to the unimodal network. These visualizations highlight the internal working of `Chameleon` in successfully learning multimodal representations while demonstrating resilience against missing modalities. More examples are provided in Supplementary Section 2.

Table 23: Statistics of the datasets used in our experiments.

Dataset	Modalities	# of Classes	Train	Test
UPMC Food-101		101	67,988	22,716
Hateful Memes		2	8,500	2,000
MM-IMDb		23	15,552	7,799
Ferramenta		52	66,141	21,869
VoxCeleb		1,251	145,265	8,251
avMNIST		10	1,125	375

3.3 Experiments

Datasets. We evaluate `Chameleon` on the multimodal classification task using four textual-visual datasets (UPMC Food-101 [59], Hateful Memes [208], MM-IMDb [123], and Ferramenta [81]) and two audio-visual (VoxCeleb [92] and avMNIST [123]) datasets. Table 23 summarizes statistics of the datasets.

Evaluation Metrics. Following existing SOTA methods on complete and missing modalities [59], [123], [165], [271], [334], we report classification accuracy for UPMC Food-101, VoxCeleb, avMNIST and Ferramenta, area under the receiver operating characteristic (AUROC) for Hateful Memes and F1 macro-avaraged for MM-IMDb.

Implementation Details. Unless specified otherwise, ViT (vit-base-patch16-224) is used as the default choice of visual learning network in `Chameleon`. We utilize AnglE [337] for encoding textual modality into embeddings of size 1024 and Ecapa-tdnn [160] for encoding audio modality into embeddings of size 196. We provide further analysis on these choices in Section 3.3.3.

3.3.1 Evaluations Under Modality-complete Setting

To evaluate whether `Chameleon` is capable of learning from multiple data sources, we first experiment when complete modalities are present during training and testing. As seen in Table 24, `Chameleon` achieves SOTA performance on five of the six datasets. Specifically, `Chameleon` achieves classification performances of 92.9%, 73.1%, 51.5%, 96.9%, 99.5%, and 96.9%, respectively, on UPMC Food-101, Hateful memes, MM-IMDb,

Table 24: Comparison of Chameleon with SOTA multimodal methods on textual-visual (e.g., UPMC Food-101, Hateful Memes, MM-IMDb, Ferramenta) and audio-visual (e.g., avMNIST and VoxCeleb) datasets using modality-complete data. Best results are bold; second best are underlined.

UPMC Food-101 [59]

Method	Acc.
Wang et al. [59]	85.1
Fused Rep. [114]	85.7
CLIPPO [357]	91.2
CentralNet [123]	91.5
ViLT [210]	91.9
Ma et al. [271]	92.0
MMBT [131]	92.1
BL [162]	<u>92.5</u>
Chameleon	**92.9 ± 0.3**

Ferramenta [81]

Method	Acc.
Ferramenta [81]	92.9
Fused Rep. [114]	94.8
IeTF [104]	95.2
Two-Branch [275]	96.2
MHFNet [371]	<u>96.5</u>
Chameleon	**96.9 ± 0.4**

Hateful Memes [165]

Method	AUROC
MMBT-G [131]	67.3
ViLT [210]	70.2
Clippo [357]	70.7
Ma et al.[271]	71.8
MMBT-R [131]	72.2
Visual BERT [134]	73.2
ViLBERT [136]	**73.4**
Chameleon	<u>73.1</u> ± 0.6

avMNIST [123]

Method	Acc.
Moddrop [53]	94.8
Gated Multimodal Unit [79]	94.1
CentralNet [123]	95.0
SMIL [219]	98.2
Chameleon	**99.5 ± 0.2**

MM-IMDb [123]

Method	F1 Macro
CBGP [108]	52.9
CBP [62]	53.2
GMU [79]	53.9
Lee et al. [334]	54.0
ViLT [210]	55.3
MFAS [144]	<u>55.7</u>
CentralNet [123]	**56.1**
Chameleon	51.5 ± 0.5

VoxCeleb [92]

Method	Acc.
FOP [275]	92.9
SBNet. [353]	94.8
Chameleon	**96.9 ± 0.4**

Ferramenta, avMNIST, and VoxCeleb. Overall, the results demonstrate that Chameleon is capable of realizing a performance comparable with that of existing SOTA multimodal methods when learning from multiple modalities.

3.3.2 Evaluations Under Modality-missing Setting

In this section, we provide detailed comparisons with existing SOTA methods by extensively evaluating Chameleon on missing modality scenarios.

Missing Modalities During Testing

Table 25 compares Chameleon with other methods: multimodal Transformers (ViLT [210], Ma et al. [271], CLIPPO [357]), and large vision-language model (LLaVA [340]) for varying amounts of missing modality on textual-visual datasets including UPMC Food-101, Hateful Memes, MM-IMDb, and Ferramenta during testing. As seen, Chameleon outperforms compared methods on UPMC Food-101, Hateful Memes, and Ferramenta demonstrating resilience against missing modalities. For example, in the case when only 10% of text modality is available on the

Table 25: Comparison of Chameleon with different levels of available modality at test time using textual and visual datasets (e.g., UPMC-Food-101, Hateful Memes, MM-IMDb, and Ferramenta). Comparison is provided with multimodal Transformers (ViLT [210]*, Ma et al. [271], and CLIPPO[357]). *ViLT values are taken from Ma et al. [271]. Boldface and underline denote, respectively, the best and second-best results. † indicates our implementation.

Data	Training		Testing		ViLT*	Ma et al.	CLIPPO†	LLaVA†	Chameleon
	Image	Text	Image	Text					
Food-101	100%	100%	100%	100%	91.9	92.0	91.2	92.2	**92.9**
	100%	100%	100%	90%	88.2	90.5	89.6	90.3	**90.5**
	100%	100%	100%	70%	80.7	87.1	86.3	87.2	**87.8**
	100%	100%	100%	50%	73.3	82.6	81.9	83.0	**83.7**
	100%	100%	100%	30%	65.9	77.5	78.1	78.9	**79.5**
	100%	100%	100%	10%	58.4	73.3	74.1	75.4	**75.9**
Hateful Memes	100%	100%	100%	100%	70.2	71.8	70.7	70.7	**73.1**
	100%	100%	100%	90%	68.8	69.7	70.3	70.5	**71.9**
	100%	100%	100%	70%	65.9	66.6	70.0	70.3	**71.5**
	100%	100%	100%	50%	63.6	63.9	69.3	69.9	**70.0**
	100%	100%	100%	30%	60.2	61.2	69.0	69.8	**70.5**
	100%	100%	100%	10%	58.0	59.6	68.1	69.5	**69.7**
MM-IMDb	100%	100%	100%	100%	55.3	55.0	28.4	**59.3**	51.5
	100%	100%	100%	90%	51.8	53.8	27.5	**56.7**	49.5
	100%	100%	100%	70%	45.1	**52.0**	25.6	51.3	45.3
	100%	100%	100%	50%	38.9	**46.6**	23.6	44.0	41.7
	100%	100%	100%	30%	31.2	**41.8**	21.2	36.7	38.7
	100%	100%	100%	10%	23.1	**37.3**	18.9	27.4	36.7
Ferramenta	100%	100%	100%	100%	95.9	-	93.4	94.8	**96.9**
	100%	100%	100%	90%	68.1	-	91.6	94.0	**95.3**
	100%	100%	100%	70%	60.8	-	89.9	93.4	**95.0**
	100%	100%	100%	50%	60.4	-	88.7	93.0	**93.7**
	100%	100%	100%	30%	54.2	-	86.1	92.7	**92.9**
	100%	100%	100%	10%	51.3	-	84.6	92.0	**92.3**

UPMC Food-101 dataset, Chameleon demonstrates an accuracy of 75.9%. In comparison, ViLT [210], Ma et al. [271], CLIPPO [357] and LLaVA demonstrate performances of 58.4%, 73.3%, 74.1%, and 75.4%, respectively. Similar trends are noticeable on Hateful Memes and Ferramenta datasets. In the case of MM-IMDb dataset, though the performance of Chameleon is lower when all modalities are available, it outperforms ViLT, CLIPPO and LLaVa when a modality is severely missing. Overall, considering all four textual-visual datasets presenting 24 scenarios of complete and missing modality, Chameleon outperforms all compared methods in 18 while second best in 2 scenarios. Another interesting observation is seen when results of Chameleon are compared to CLIPPO which can be considered a baseline that renders text directly as an im-

age. The superior performance of Chameleon demonstrates that directly rendering text as input may not be an optimal choice compared to our proposed feature encoding approach. Furthermore, we compare results with a large-scale vision-language model (LLaVA), which converts image features to textual tokens. Though the methodology is different compared to Chameleon, the essence of converting one modality into the other is similar. Comparing results directly with CLIPPO and LLaVA demonstrates the effectiveness of the design choices in Chameleon. Furthermore, Fig. 7 compares Chameleon with other methods: FOP [275] and SBNet [353] for varying amounts of missing modality on audio-visual dataset (VoxCeleb). As seen, Chameleon significantly outperforms compared methods demonstrating resilience against missing modalities.

The overall persistent robustness to missing modality of Chameleon compared to existing methods across various challenging datasets and scenarios is attributed to proposed framework that enables the training of all input modalities in a common learning space. This way, if one modality is missing during testing, Chameleon shifts its focus to the available modality, thus yielding higher performance leveraging multimodal knowledge. More insights on this are provided in Fig. 6 where Grad-CAM visualizations clearly show the shifting of focus from multimodal data to the available modality for predictions. More visual examples are provided in the supplementary material.

Missing Modalities During Training

Derived from the motivation that modality can be missing in any data sample, Lee et al. [334] recently extended the evaluation protocol on textual-visual datasets by introducing scenarios of missing modality during training and testing, i.e., 30% of one modality is available against 100% of the other modality at train and test time. Table 26 compares Chameleon with existing methods ViLT [210], Lee et al. [334], and CLIPPO [357] using this protocol on UPMC Food-101, Hateful Memes, and MM-IMDb datasets. In most scenarios, Chameleon demonstrates comparable or better performance than existing methods. Notably, compared to CLIPPO, Chameleon demonstrates robustness to missing modalities during train-

Table 26: Comparison of Chameleon with ViLT [210]*, Lee et al. [334], and CLIPPO [357] on UPMC Food-101 [59], Hateful Memes [165], and MM-IMDb [79] under different missing modality settings during training and testing. *ViLT values are taken from Lee et al. [334]. Boldface and underline denote, respectively, the best and second best results. † indicates our implementation.

Data	Training		Testing		ViLT	Lee et al.	CLIPPO†	Chameleon
	Image	Text	Image	Text				
Food-101	100%	100%	100%	100%	91.9	<u>92.0</u>	91.2	**92.5**
	100%	30%	100%	30%	66.3	<u>74.5</u>	77.5	**78.0**
	30%	100%	30%	100%	76.7	**86.2**	83.6	<u>83.8</u>
Hateful Memes	100%	100%	100%	100%	70.2	<u>71.0</u>	70.7	**73.9**
	100%	30%	100%	30%	60.8	59.1	<u>60.9</u>	**69.4**
	30%	100%	30%	100%	61.6	**63.1**	58.4	<u>62.1</u>
MM-IMDb	100%	100%	100%	100%	46.0	**54.0**	28.4	<u>51.5</u>
	100%	30%	100%	30%	35.1	**39.2**	23.6	<u>36.6</u>
	30%	100%	30%	100%	37.7	<u>46.3</u>	23.3	**50.6**

Table 27: Performance analysis of Chameleon with various visual networks using UPMC-Food-101, and Hateful Memes datasets.

Dataset (Performance Metric)	ResNet-101	Adapter	ViT
UPMC Food-101 (Accuracy)	89.5	90.1	**92.9**
Hateful Memes (AUROC)	70.9	71.3	**73.9**

ing and testing thus reiterating the importance of our encoding scheme. Furthermore, Fig. 8 compares Chameleon with other methods: SMIL [219], Autoencoder (AE) [133], and Generative Adversarial Network (GAN) for varying amounts of missing modality on audio-visual (avMNIST). As seen, Chameleon significantly outperforms compared methods on avMNIST demonstrating resilience against missing modalities.

3.3.3 Analysis and Discussion

We provide further analysis on Chameleon including its application to different visual networks, embedding extractors, dataset specific optimizations, and visualizations for further insights.

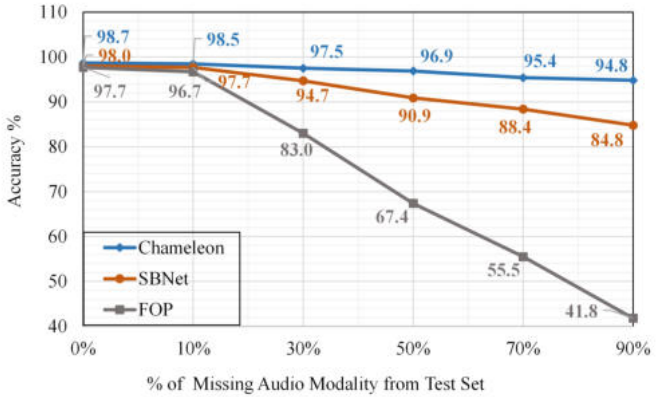


Figure 7: Comparison of `Chameleon` with FOP [275], and SBNet [353] on audio-visual dataset (e.g., VoxCeleb) under different levels of missing modality at test time.

Is Chameleon Agnostic to the Visual Network?

`Chameleon` can generally adapt any visual network for training. Investigating the extent to which the visual classification network influences performance, we conduct a series of experiments, using CNN (ResNet-101)[64], ViT [161], and Adapter [405]. The results of these experiments are summarized in Table 27 on a complete set of modalities. While adapting ResNet-101 yields 89.5% accuracy and 70.9% AUROC on UPMC Food-101 and Hateful Memes respectively, using Adapter [405] boosts the performance to 90.1% and 71.3%. Finally, using ViT yields the best performance of 92.9% and 73.9% on corresponding datasets. While the trend conforms with the existing literature comparing ResNets, ViTs, and Adapters, the overall competitive performance demonstrates that `Chameleon` is agnostic to the vision network.

Is Chameleon Suitable for Other Multimodal Task?

We conduct experiments to perform cross-modal verification task with `Chameleon` to establish face-voice association using VoxCeleb [92] dataset.

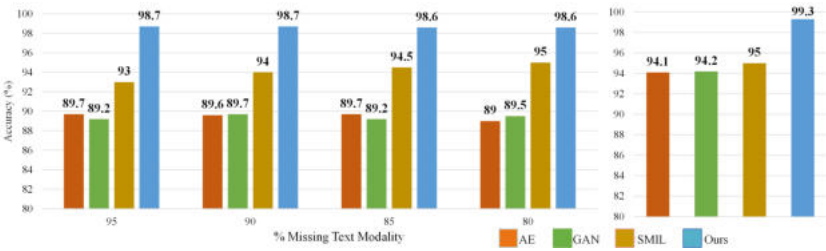
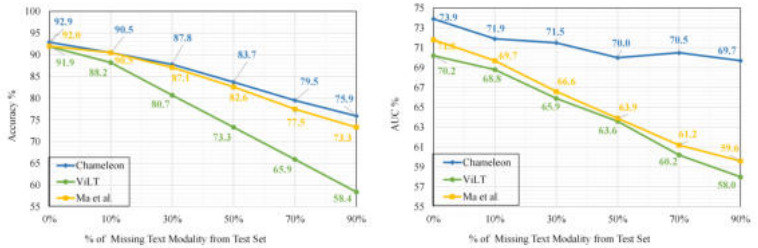


Figure 8: Comparison of `Chameleon` with SMIL [219], Autoencoder (AE) [133], and Generative Adversarial Network (GAN) on avMNIST; an audio-visual dataset. (Left) training with 100% Image + n% Audio and testing with Image Only. (Right) training with 100% Image + 20% Audio and testing with Image + Audio.



(a) UPMC Food-101 (b) Hateful Memes

Figure 9: Comparisons of `Chameleon` with ViLT [210] and Ma et al. [271] on various levels of missing textual modality during testing on textual-visual datasets (e.g., UPMC Food-101 and Hateful Memes). A smaller drop in performance by our approach in most cases signifies its effectiveness towards training multimodal Transformer resilient to missing modalities without dataset-centric fusion strategies.

Table 28: Cross-modal verification results on unseen-unheard configurations of Chameleon and existing state-of-the-art methods. Best results are highlighted in bold text whereas the second best are underlined.

Method	EER ↓	AUC ↑
	Unseen-Unheard	
DIMNet [152]	<u>24.9</u>	-
Learnable Pins [112]	29.6	78.5
MAV-Celeb [223]	29.0	78.9
Single Stream Net. [140]	29.5	78.8
Multi-view [231]	28.0	-
AML [247]	-	80.6
DRL [225]	25.0	**84.6**
FOP [275]	<u>24.9</u>	83.5
SBNet. [353]	25.7	82.4
Chameleon	**24.0**	<u>83.6</u>

We compare the performance of Chameleon against several SOTA face-voice association methods including DIMNet [152], Learnable Pins [112], MAV-Celeb [223], Single Stream Network [140], Multi-view [231], Adversarial Metric learning [247] (AML), Disentangled Representation Learning [225] (DRL), FOP [275], and Single Branch Network [353] (SBNet) Table 28 shows the results of cross-modal verification task on unseen-unheard configuration. We observe that Chameleon outperforms the other SOTA methods in terms of EER metric by achieving the EER scores of 24.0% whereas the existing best FOP [275] obtains the EER score of 24.9%. In terms of AUC metric, Chameleon obtains the favorably better score of 83.6% as compared to the DRL [225] that achieves the score of 84.6%.

Embedding Extractors

Chameleon can generally encode any audio or textual embedding for training. In Table 29, we compare the performance of various textual embeddings (e.g., AngIE [337], CLIP [227], MPNet [176], Bert [127] and Doc2Vec [29]) on UPMC Food-101 and Hateful Memes datasets under different levels of missing modality. Though, best performance is obtained with AngIE embeddings, other representations such as CLIP and MPNet also produces competitive results on complete as well as various levels of

Table 29: Comparison of different embeddings to encode textual modality under different levels of missing modality at test time.

Data	Testing Image	Testing Text	Doc2Vec	Bert	MPNet	CLIP	AnglE
Food	100%	100%	85.5	89.4	92.2	92.3	**92.9**
	100%	90%	87.5	84.4	90.3	**90.5**	**90.5**
	100%	50%	79.4	79.9	82.9	83.4	**83.7**
	100%	10%	73.4	74.4	75.0	75.2	**75.9**
Memes	100%	100%	69.8	69.9	70.8	73.4	**73.9**
	100%	90%	69.1	69.4	70.8	71.5	**71.9**
	100%	50%	68.9	69.0	69.3	69.9	**70.0**
	100%	10%	68.3	68.9	68.9	69.6	**69.7**

missing modalities.

Are Dataset-specific Transformers Necessary?

Ma et al. [271] have observed that multimodal Transformers are not only sensitive to missing modalities but different design choices, such as fusion strategy, should be tailored to the dataset. They further conclude that it may not be possible to design a general multimodal Transformer architecture to be used across datasets. In contrast, with our proposed design of encoding audio or text modality into visual format and then training a visual network such as ViT, `Chameleon` becomes independent of dataset-centric components. To evaluate this, we train two separate ViTs on the UPMC Food-101 and Hateful Memes datasets. Each model is then evaluated on different levels of missing textual modality at test time. Fig. 9 shows the corresponding results and compares the performances with ViLT [210] and Ma et al. [271]. As Ma et al. have proposed dataset-optimal fusion strategies, their approach demonstrates robustness to missing modality compared to ViLT. `Chameleon`, on the other hand demonstrates superior resilience against missing modality while training on similar ViTs without any changes across datasets. As seen, `Chameleon` enables the training of dataset-agnostic transformers capable of handling severe missing modalities.

Table 30: Comparison of word- and sentence-level embeddings using UPMC-Food-101, and Hateful Memes datasets.

Dataset (Performance Metric)	Word	Sentence
UPMC Food-101 (Accuracy)	92.5	**92.9**
Hateful Memes (AUROC)	**73.9**	73.1

Comparison Between Word-level and Sentence-level Embeddings

`Chameleon` utilizes non-visual-modality embeddings that are subsequently encoded into a common input modality. This enables any non-visual modality (e.g., audio or text) to be encoded into a visual format to train the network. However, in the case of text, another possibility of extracting embeddings would be to utilize word-level encoders (see Supplementary Section 3 for more details on word-level embedding). Table 30 compares word and sentence-level embedding performance on UPMC Food-101 and Hateful Memes datasets. As seen, the performance on both word and sentence embeddings are comparable, either one of them can be used depending on the task and application.

Embedding Space Analysis

We plot results of t-SNE projections to take a peek into the embedding space of `Chameleon` trained on UPMC Food-101 dataset and provide comparisons with ViLT [210] in Fig. 10. Comparisons are provided under the following three settings: Fig. 10a & d) complete modalities during training and testing, b & e) complete modalities during training but 100% missing textual modality during testing, and c & f) complete modalities during training but 100% missing visual modality during testing. In the case of multimodal training/testing, compared to ViLT, `Chameleon` is able to project modalities belonging to the same classes closer to each other while demonstrating inter-class separability (Fig. 10a & d). This highlights the success of `Chameleon` in training a robust multimodal learning method. Finally, in the case of multimodal training but 100% missing modality during testing (Fig. 10b, c, e & f), although some distortions are noticeable, compared to ViLT, `Chameleon` successfully retains

the overall separability of the classes under such severe missing modality case. This demonstrates the resilience of `Chameleon` towards missing modalities.

Limitations

While advantageous in most missing modality cases, utilizing text as an image has its limitations as well. For example, in the case of longer text descriptions, the network struggles to learn joint representation of multiple modalities as well as achieving robustness to missing modalities. This phenomenon can be observed in Tables 25 & 26 where CLIPPO, utilizing rendered text as an image, performs notably lower than its counterpart networks. In the case of `Chameleon`, this problem is partially handled as we convert the textual embedding into encoded visual representations before training a network. This way, longer texts are summarized into compact encoded visual representations thus yielding superior performance.

3.4 Chapter Summary

In this chapter, we presented a multimodal learning framework that encodes embeddings of non-visual modality into visual format to learn modality-independent representations of the input modalities. The common input format facilitates joint representation learning by sharing weights across multiple modalities resulting in multimodal learning robust to missing modalities. Extensive experiments are performed on UPMC Food-101, Hateful Memes, MM-IMDb, Ferramenta, avMNIST, and VoxCeleb. The proposed method is thoroughly evaluated on complete modalities as well as missing modalities during training and testing. The experimental results indicate that the proposed framework obtains superior performance when a complete set of modalities is available. In the case of missing modalities, the performance deterioration is noticeably smaller than that of the existing multimodal learning methods, indicating significant robustness of `Chameleon`.

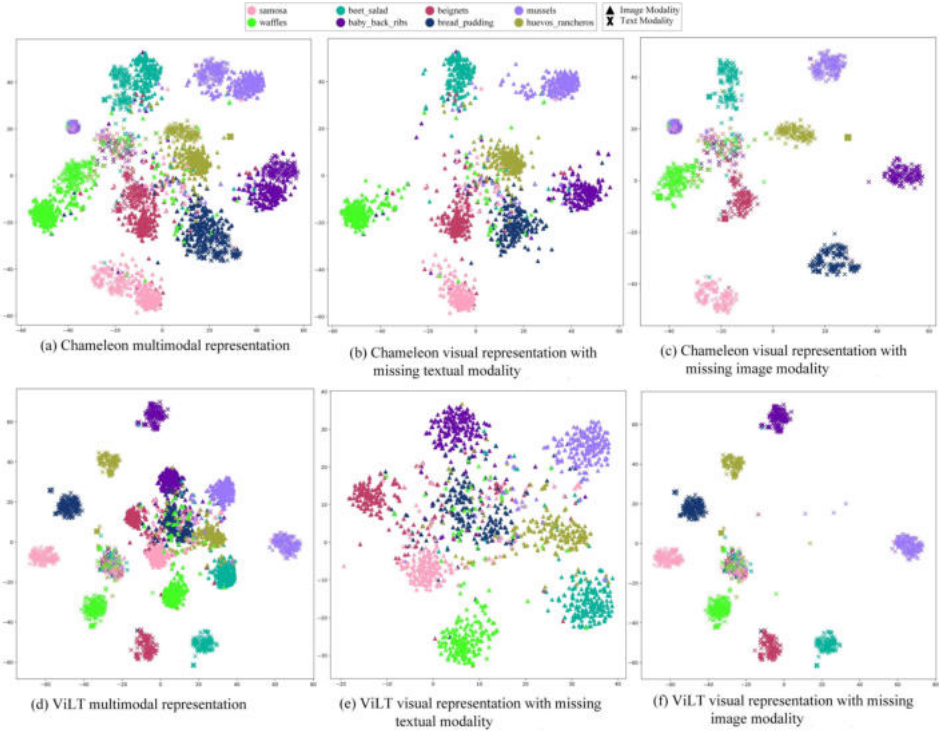


Figure 10: t-SNE visualizations of the embedding space of `Chameleon` (a - c) and ViLT (d - f) along with accuracy on test set of UPMC Food-101. Compared to ViLT, `Chameleon` not only enhances the classification boundaries when complete modalities are available at test time but also retains these boundaries when the textual or visual modality is completely missing during test time. Note that classes are selected randomly from the test set.

Chapter 4

RobustA: Multimodal Robustness Under Corrupted Modalities

Video anomaly detection (VAD) is a challenging computer vision task with various real-world applications including surveillance, autonomous navigation, packaging, and biomedical imaging [344]. Generally, VAD methods aim to predict high anomaly scores for the frames in a video that deviate significantly from the norm, where the application context determines the norm. For example, events such as shoplifting or violence can be considered anomalies in the CCTV surveillance context [180]. Since it is laborious to obtain fine-grained (pixel-level or frame-level) labels of anomalies in videos, a weakly-supervised VAD (WS-VAD) setting that learns to detect anomalous frames using only video-level binary labels is typically used. In WS-VAD, a training video is labeled as normal if no anomalous event is present, whereas it is labeled as an anomaly if any anomalous event is present [122], [294], [312]. In recent years, WS-VAD has gained considerable popularity and became a mainstream VAD

The contents of this chapter are taken from our work "RobustA: Robust Anomaly Detection in Multimodal Data", available at `http://arxiv.org/abs/2511.07276`. The corresponding code and data resources are accessible via GitHub at: `https://github.com/salemalmarri/RobustA`.

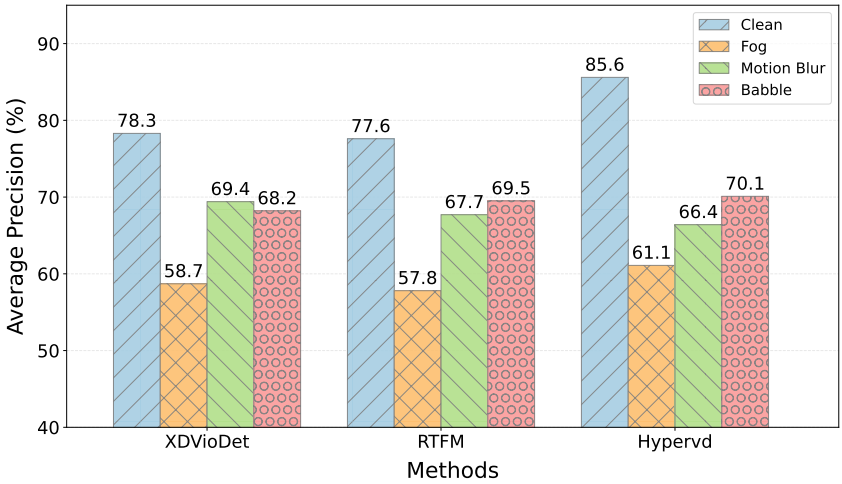


Figure 11: Existing multimodal anomaly detection approaches [180], [237], [426] are not robust when subjected to corruptions in either modality, including vision (e.g., fog and motion blur) or audio (e.g., babble).

paradigm [120], [122], [139]. More recently, WS-VAD has been transformed into a multimodal learning task by including other informative modalities such as audio to discriminate and locate events better [180], [291]. For instance, it is challenging to rely only on visual signals in shaky videos covering an explosion event. However, if accompanied by the audio modality, the event can be easily classified. Prior works have successfully shown the effectiveness of using video and audio pairs to improve the overall performance of WS-VAD [180], [287], [387].

Despite their many advantages, deep neural network models are vulnerable to typical corruptions in the input data (e.g., degradations, distortions, and disturbances caused by weather changes, system errors, etc.) [327]. Though multimodal learning systems achieve better performance on multimodal data [415] compared to their unimodal counterparts, they are also susceptible to such vulnerabilities. Multimodal VAD systems, in particular, are prone to such issues (see Fig. 11), as they are often subjected to extreme environmental conditions [242]. For this reason, it

is essential to design methods that can withstand corrupted data during deployment. While it may be possible to achieve this goal by including corrupted data during the training process, *this work aims to evaluate the robustness of multimodal VAD systems trained only on clean data, when encountering corrupted data during deployment*. Existing literature lacks a detailed study and appropriate dataset that investigates the impact of corrupted modalities on WS-VAD and multimodal VAD systems.

In this work, we empirically investigate this problem with a proposed dataset, `RobustA`, specifically to study this issue. We observe that existing multimodal anomaly detection models deteriorate dramatically under corrupted data. For example, in Fig. 11, the performance of XD-VioDet [180] drops from 78.36% to 58.70% when exposed to fog in visual modality during testing. To alleviate this issue, we propose a robust anomaly detection method that is resilient against corrupted modalities. The proposed method independently maps multiple modalities into a shared representation space. This shared space can be particularly beneficial in scenarios where one modality is compromised. For example, if one modality becomes distorted, inaccurate, noisy, or is entirely missing, the shared space can leverage information from the remaining modalities to compensate for the loss. Results reported in Section 4.3.2 suggest that our approach not only yields better multimodal performance but also demonstrates resilience against corruptions.
The key contributions of our work are as follows:

- We present `RobustA`, a first-of-its-kind, carefully curated dataset comprising 8 visual corruptions and 8 audio corruptions, enabling evaluation of the multimodal anomaly detection methods against various types and levels of modality corruptions.

- We propose an anomaly detection approach that independently maps multiple modalities to learn representations in a shared space while dynamically adjusting the weights of individual modalities to reduce the impact of the compromised modality towards multimodal anomaly detection.

- We study the plug-and-play nature of our proposed approach,

demonstrating its resilience against compromised modalities when used in conjunction with existing approaches.

4.1 Weakly Supervised Video Anomaly Detection and Multimodality

Weakly Supervised Video Anomaly Detection (WS-VAD) refers to the process of identifying abnormal events in a given video while the training is carried out using only video-level binary labels. The problem was first introduced by Sultani et al. [122] where the authors utilized multi-instance learning based ranking to carry out the overall training. The research was advanced by several researchers [187], [188], [293]. However, most of these studies are focused on unimodal (video-only) training and testing.

More recently, anomaly detection task is transformed into multimodal learning where more than one modalities (usually vision and audio) are present during training and testing. For example, XDVioDet [180] utilized a multi-branch neural network comprising holistic, localized, and score branches to effectively leverage multimodal feature representations. Similarly, HyperVD [426] proposed a method using two graph-based hyperbolic convolutional networks for spatio-temporal feature extraction from multimodal input. While the multimodal anomaly detection task has gained popularity due to superior performance compared to the unimodal counterpart approaches, existing literature lacks rigorous studies understanding the adverse effects of compromised data on multimodal learning systems. As, CCTV surveillance cameras equipped with anomaly detection systems are highly prone to environmental corruptions, it is pertinent to explore this novel research direction.

4.1.1 Addressing Compromised Modalities

Corrupted Modalities. Data collected from diverse multiple sources may contain corrupted samples due to various factors, such as sensor failure, obstructions in the video stream, noise in audio signals, data storage errors, and many more. Recent years have seen an increased

interest in investigating the vulnerability of deep models against modality corruptions [91], [268], [329], [367], [380]. For example, Hendrycks et al. [129] built benchmarks to evaluate the performance of various Convolutional Network Networks on the image classification task. Similarly, Yi et al. [242] curated robust video classification benchmark to evaluate state-of-the-art Convolutional Networks and Transformers against video corruptions. The growing interest has also broadened to multimodal learning where handling compromised modalities is critical to improve performance and robustness. For example, Beemelmanns et al. [380] introduces MultiCorrupt framework to evaluate the robustness of multimodal 3D object detectors against various corruption categories. Similarly, Hong et al. [322] proposed a multimodal input corruption modeling to develop robust audio-visual speech recognition models. These studies demonstrate that the performance of unimodal or multimodal methods deteriorate dramatically when a modality is corrupted.

Missing Modalities Multimodal learning has shown remarkable performance improvements over unimodal methods. However, such methods exhibit deteriorated performances if one or more modalities are missing [219], [271], [386], [450]. Considering the significance of multimodal learning, recent years have seen an increased interest in studies addressing missing modalities for various tasks including classification, such as [334], [338], [439].

While these methods have notably enhanced model robustness in addressing compromised modalities across various application areas, to the best of our knowledge, there remains a gap in the literature for the multimodal anomaly detection task. In this work, we investigated the robustness of multimodal anomaly detection under a carefully crafted dataset of compromised modalities. Moreover, we propose a robust approach that maintains superior performance when faced with extreme missing and corrupted modalities.

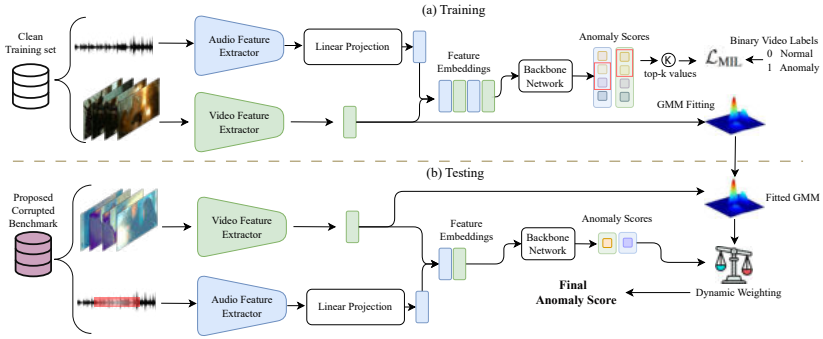


Figure 12: Architecture of our approach. Modality-specific features are extracted using pre-trained audio and visual encoders. A linear projection is used to match the feature dimensions. Modality embeddings are independently mapped to learn representations in a shared space. During inference, the shared learning space helps mitigate the adverse effects of modality corruptions. Moreover, a dynamic weighting scheme is utilized to adjust the weights of the corrupted modality for better anomaly detection.

4.2 Methodology

4.2.1 Background and Overview

Existing methods generally concatenate audio and visual embeddings to learn fused representations for the multimodal anomaly detection task [180], [287], [393], [426]. While effective under modality-complete settings, these methods rely heavily on the complete availability of modalities and suffer dramatic performance deterioration when a modality is corrupted or missing. Since surveillance data may involve corruptions due to weather conditions, connection latency, broken/tempered equipment, etc., a lack of robustness against such scenarios may deter the deployment of these methods in real-world applications. In this work, we develop an approach to *learn multimodal representations by mapping audio and visual modalities independently into a common feature space*. This enables the model to maintain *robustness when a modality is compromised* by leveraging a shared representation of cross-modal interactions, effectively

compensating for missing or corrupted modality data. The proposed approach is illustrated in Fig. 12 and the details are discussed next.

4.2.2 Preliminaries

Given a dataset of n videos, each video V is preprocessed into m non-overlapping audio and visual segments and passed to a pre-trained feature extractor. The audio and video feature extractors generate the feature embeddings $E^A \in \mathbb{R}^{m \times d_A}$ and $E^V \in \mathbb{R}^{m \times d_V}$ for each video, respectively with the feature dimensions are denoted as d_A and d_V. The class label $y \in \{1, 0\}$ indicates the presence (1) or absence (0) of anomalous events in the video. The goal is to learn an anomaly detector $\mathcal{A}_\theta(E^*)$ that takes a feature vector (audio or visual) as input and predicts an anomaly score in the range of 0 and 1.

4.2.3 Learning a Shared Representation Space

Prior multimodal anomaly detection methods rely on concatenation of audio and visual representations and learning the anomaly detector based on the concatenated representation. Given the audio and visual feature embeddings E^A and E^V for n training videos in a dataset, Eq. 4.1 outlines the traditional fusion-based learning approach utilizing Multi-Instance Learning (MIL) loss [180]:

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{n} \mathcal{L}_{\text{MIL}}(\mathcal{A}_\theta(E^A||E^V), y_i), \tag{4.1}$$

where $E^A||E^V$ is the fused multimodal representation obtained by concatenation of E^V and E^A, allowing the architecture to capture the inter-modality relationship. Consequently, the success of these methods is highly dependent on the availability of both modalities during inference. To alleviate this dependence on modality completeness, we propose a method that learns multimodal representations by mapping each modality independently into a shared space.

Specifically, we map individual modality independently into a common representation space, enabling the model to capture inter- and intra-

modality relationships through cross-modal data interactions. As a result, when a modality is compromised, the model compensates by leveraging the shared representation space. Thus, Eq. 4.1 needs to be modified as follows:

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{2n} \mathcal{L}_{\text{MIL}}(\mathcal{A}_\theta(E_i^*), y_i) \tag{4.2}$$

where E^* is randomly sampled from either E^V or E^A. However, in existing methods [180], depending on the choice of feature extractor, the output dimensions of E^V and E^A are usually inconsistent. To address this issue, we introduce a linear projection module $\mathcal{P}_\phi$ that transforms E^A to a higher dimensional space, matching the dimension of E^V. Thus, E^* in Eq. (4.2) will be randomly sampled from either E^V or $\mathcal{P}_\phi(E^A)$. The parameters θ of the anomaly detector and ϕ of the projection module are learned by minimizing the total loss $\mathcal{L}_{\text{total}}$. Intuitively, the proposed approach attempts to learn a single modality-agnostic anomaly detector, which implicitly forces the two modalities to share a common representation space.

4.2.4 Dynamic Weighting During Inference

As the anomaly detector is agnostic to the input modality, it is possible to obtain an anomaly score using both audio and visual features during inference. Hence, a strategy to optimally combine individual anomaly scores is needed to achieve good multimodal performance. A straightforward approach would be to compute the weighted average of the anomaly scores produced by the anomaly detector $\mathcal{A}$ for each modality in a late fusion manner. For instance, the final anomaly score S for a given video segment based on audio-visual features can be obtained as:

$$S = \lambda_A \mathcal{A}(\overline{E}^A) + \lambda_V \mathcal{A}(\overline{E}^V), \tag{4.3}$$

where $\overline{E}^A$ and $\overline{E}^V$ are the audio and visual feature embeddings of the test sample, respectively, which may also contain a compromised modality.

Here, λ_A and λ_V are the weights assigned to the audio and visual modalities, respectively, and these weights are linearly normalized such that they add up to 1. In a naive approach, λ_A and λ_V can be set to 0.5 giving equal weightage to both modalities. While, as shown in our experiments, this approach works reasonably well due to the shared representation space mitigating the impacts of an individual compromised modality, a more sophisticated approach may be required to handle extreme corruption in one of the modalities.

Since the training data does not have prior information about the compromised modality (since training is performed only on clean data), we devise a dynamic weighting scheme based on the clean data distribution. Specifically, we fit a Gaussian Mixture Model (GMM) to the clean (non-corrupted) training visual features and evaluate if the test data matches with this clean data distribution. Let $\mathcal{G}^{\mathcal{V}}$ denote a K-component GMM with parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ representing the clean distribution of visual features, where π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ denote the mixture probability, mean, and covariance matrix, respectively, of the k-th Gaussian component.

The negative log-likelihood ℓ of the test features $(\overline{E}^V)$ is then computed using the fitted GMM as follows:

$$\ell\big(\overline{E}^V\big) \;=\; -\log\left[\sum_{k=1}^{k} \pi_k \,\mathcal{N}\Big(\overline{E}^V \,\big|\, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\Big)\right] \tag{4.4}$$

where $\mathcal{N}$ denotes the likelihood of observing the feature vector $\overline{E}^V$ under the k^{th} Gaussian component of the mixture model. A higher ℓ value indicates that the sample is more likely to be compromised, while a lower value of ℓ suggests the sample is clean. Finally, λ_V is computed using a sigmoid function as:

$$\lambda_V \;=\; \frac{0.5}{1 \;+\; \exp\Big(\, c\big[\,\ell\big(\overline{E}^V\big) + x_0\big]\Big)}, \tag{4.5}$$

where c and x_0 are the scale and shift hyperparameters that shape the sigmoid function, respectively. The above equation maps the negative log-likelihood $\ell\left(\bar{E}^V\right)$ to a value of λ_V between 0 and 0.5. Similarly, the

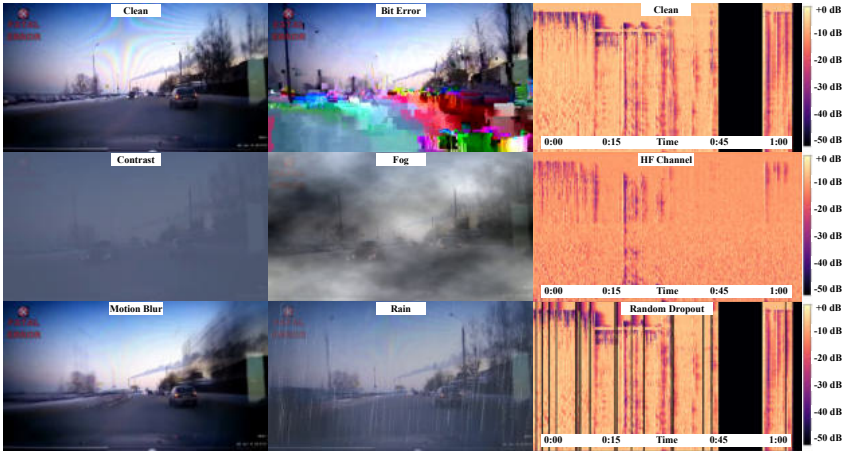


Figure 13: Examples of a few visual and audio corruptions demonstrating the challenging scenarios presented in our proposed benchmark.

value of λ_A can also be computed and the two weights can be linearly normalized so that they sum to 1. Extensive ablation studies on the proposed dynamic weighting scheme are provided in Section 4.4.1, which demonstrate the effectiveness of our approach. When one of the modalities is completely missing, its corresponding weight is set to zero because the anomaly score for that modality cannot be computed.

4.2.5 Modality Corruptions

In this work, we propose a carefully curated dataset, `RobustA`, to evaluate the critical impacts of audio/visual corruptions on the overall performance of multimodal anomaly detection systems. To the best of our knowledge, `RobustA` is the first rigorous attempt to study anomaly detection in such scenarios. We leverage XD-Violence multimodal anomaly detection dataset as the base to create several corruption scenarios, by varying the type and severity level, for both audio and visual modalities. We choose these modalities because of their globally standard availability in the existing real-world CCTV systems, while other modalities, such as text, lack practical application. Examples of some visual corruptions and

audio corruption mel spectograms [52], [109] are provided in Figure 13.

Types of Corruptions in Visual Modality. We consider the following corruptions in the visual modality: Bit error, brightness, contrast, fog, rain, motion blur, saturation, and shot noise. Details about each corruption and its implementation are provided in the Supplementary.

Types of Corruptions in Audio Modality. We consider the following corruptions in the audio modality: Overlay, bit error, pitch shift, random dropout, and reverb. Details about each corruption and the implementation details are provided in the Supplementary.

Number of Corrupted Samples. We consider varying numbers of corrupted modality samples. In particular, we consider 0%, 10%, 30%, 50%, 70%, 90%, and 100% of samples in each modality as corrupted.

Table 31: Comparison of our approach and baseline under various corruption types and corruption levels (percentages). AP(%) is used as the evaluation metric (higher is better). Bold represents best performance in each comparison.

Corruption Levels		0%		10%		30%		50%		70%		90%		100%	
Corruption	**Modality**	[180]	Ours	[180]	Ours	[180]	Ours	[180]	Ours	[180]	Ours	[180]	Ours	[180]	Ours
Bit Error	**Visual**	78.36	**80.00**	77.41	**79.19**	72.57	**76.67**	72.48	**76.35**	68.21	**74.70**	66.61	**74.78**	65.55	**74.45**
Brightness	**Visual**	78.36	**80.00**	77.39	**79.30**	76.35	**77.74**	75.54	**77.02**	75.06	**76.23**	74.34	**75.42**	73.76	**74.72**
Contrast	**Visual**	78.36	**80.00**	77.08	**78.67**	69.84	**75.75**	66.82	**74.44**	59.85	**71.86**	54.54	**70.35**	52.47	**69.21**
Fog	**Visual**	78.36	**80.00**	76.45	**78.05**	69.21	**73.47**	68.87	**72.87**	62.32	**69.68**	60.48	**69.26**	58.70	**67.97**
Rain	**Visual**	78.36	**80.00**	77.73	**79.52**	74.27	**77.66**	72.88	**77.16**	70.94	**75.54**	69.43	**74.56**	68.69	**73.89**
Motion Blur	**Visual**	78.36	**80.00**	77.22	**78.73**	74.45	**77.02**	74.03	**76.46**	71.43	**74.55**	70.28	**73.80**	69.46	**73.08**
Saturate	**Visual**	78.36	**80.00**	77.02	**78.84**	73.64	**77.14**	69.94	**76.10**	68.20	**75.19**	65.57	**74.26**	65.23	**73.86**
Shot Noise	**Visual**	78.36	**80.00**	77.69	**79.10**	74.65	**76.08**	74.60	**75.62**	72.35	**73.62**	71.40	**72.72**	70.99	**71.68**
Average	**Visual**	78.36	**80.00**	77.25	**78.93**	73.12	**76.44**	71.90	**75.75**	68.54	**73.92**	66.58	**73.14**	65.61	**72.36**
Babble	**Audio**	78.36	**80.00**	77.16	**79.06**	75.18	**77.64**	72.65	**74.36**	70.94	**73.29**	68.57	**70.96**	68.21	**70.45**
Bitrate	**Audio**	78.36	**80.00**	77.03	**78.45**	75.60	**76.66**	**72.76**	72.37	**71.12**	70.21	**68.67**	65.48	**68.31**	65.11
HF Channel	**Audio**	78.36	**80.00**	76.37	**79.01**	73.73	**77.77**	69.36	**73.24**	67.59	**72.09**	65.35	**69.78**	65.56	**69.47**
Pink	**Audio**	78.36	**80.00**	76.32	**79.02**	73.45	**77.94**	68.97	**73.63**	67.30	**72.67**	64.99	**70.85**	65.36	**70.60**
Pitch Shift	**Audio**	78.36	**80.00**	76.24	**78.87**	73.49	**77.62**	69.53	**72.83**	67.01	**71.23**	63.66	**66.91**	63.07	**66.39**
Random Dropout	**Audio**	78.36	**80.00**	76.45	**78.93**	74.22	**77.70**	69.80	**73.05**	67.79	**71.81**	64.01	**66.92**	63.53	**66.48**
Reverb	**Audio**	78.36	**80.00**	76.65	**79.01**	75.02	**77.78**	71.99	**73.23**	70.01	**71.90**	67.01	**70.02**	66.84	**69.73**
White	**Audio**	78.36	**80.00**	76.48	**78.95**	73.69	**77.81**	69.24	**73.48**	67.88	**72.36**	66.02	**70.43**	66.60	**70.21**
Average	**Audio**	78.36	**80.00**	76.59	**78.91**	74.30	**77.74**	70.41	**73.27**	68.71	**71.94**	66.04	**68.92**	65.94	**68.56**
Missing Visual	**Visual**	78.30	**80.00**	75.10	**76.94**	69.50	**71.87**	63.10	**71.18**	58.20	**68.61**	52.10	**68.00**	49.50	**67.15**
Missing Audio	**Audio**	78.36	**80.00**	77.30	**78.86**	75.20	**77.96**	73.90	**74.75**	72.50	**73.91**	70.00	**72.74**	69.10	**72.55**
Mixed Corruptions	**Visual**	78.36	**80.00**	77.70	**79.54**	73.93	**78.15**	73.75	**78.65**	72.75	**77.35**	72.09	**77.15**	71.45	**76.27**
Mixed Corruptions	**Audio**	78.36	**80.00**	76.94	**79.09**	74.77	**77.62**	72.40	**74.13**	70.48	**72.99**	67.71	**70.50**	67.14	**69.97**
Motion Blur+Babble	**Audio+Visual**	78.36	**80.00**	75.71	**77.59**	70.91	**73.54**	67.19	**69.03**	61.94	**64.22**	56.82	**60.13**	54.66	**58.04**
Brightness+Rand. Dropout	**Audio+Visual**	78.36	**80.00**	75.42	**78.08**	71.66	**74.49**	65.20	**67.83**	61.62	**65.11**	56.05	**56.94**	54.19	**55.12**
Bit Error+HF-Channel Noise	**Audio+Visual**	78.36	**80.00**	74.56	**78.07**	69.43	**74.29**	62.89	**70.27**	55.53	**65.87**	47.73	**60.60**	45.50	**59.47**

4.3 Experiments

4.3.1 Implementation Details

Feature Extraction: For feature extraction of visual modality, we employ ResNet-I3D model pretrained on Kinetics-400 dataset. Following prior work [237], we perform 10-crop augmentation by using a 16 frame sliding window with a sample rate of 24 FPS. The audio embeddings are extracted by using the VGGish network [83] pre-trained on large-scale YouTube dataset. Moreover, the extraction is performed by pre-processing the audio signal into 960 ms of overlapping segments and setting 96×64 bin size for the mel-spectrogram.

Training Details: We trained the proposed architecture on Nvidia RTX $A100$ for 50 epochs using Adam optimizer with initial learning rate of 10^{-5} and dynamic adjustment using multi-step cosine scheduler. We use a batch size of 640 and weight decay of 0.00001.

Backbone Network: We adopt XDVioDet [180] as the primary backbone network in our experiments (also referred to as baseline hereafter) and provide extensive comparisons with it. In addition, to explore the general applicability of our proposed approach, we provide additional experiments on two other models as backbone networks, including RTFM [237] and HyperVD [426]. Nevertheless, unless specified otherwise, the results are reported with XDVioDet as the primary backbone network.

Dataset: We use XD-Violence multimodal anomaly detection dataset [180] as a stepping stone for our dataset. More specifically, we apply corruptions on the test set of the XD-Violence dataset based on the details provided in Section 4.2.5 to create `RobustA`. Similar to the original dataset, the train and test splits include $3,954$ and 800 anomalous and normal videos, with varying levels of corruptions based on the experiment. Following prior works [180], [426], we use Average Precision (AP) as the evaluation metric.

4.3.2 Experimental Results

Robustness Against Compromised Modalities

Table 31 provides extensive performance evaluation of our method and the baseline [180] under various types of audio and visual corruptions. It may be noted that the training set remains clean, as our goal is to achieve robustness against corruptions when trained with clean data.

Visual Corruptions: During testing, various types of corruptions, such as bit error, fog, rain etc., are present in the visual modality (as outlined in Section 4.2.5). We evaluate the performance of our method against the baseline on varying levels of corruption, which reflect the proportion of corrupted video in the test set. A corruption level of 0% means no corrupted video, while a level of 100% indicates that all test videos are corrupted. The results (Table 31) demonstrate that our method outperformed the baseline against various types and levels of corruptions. Notably, for extreme cases of corruptions such as contrast, our method achieves performance scores of 80.00% and 69.21%, whereas the baseline attains scores of 78.36% and 52.47% under 0% and 100% of corrupted videos. Although the performance of our method and the baseline method is comparable under 0% corruption, our method notably outperforms it under 100% corruption, highlighting its robustness against visual modality corruption.

Audio Corruptions. During testing, various types of corruptions, such as babble, bitrate, and hfchannel, are present in the audio modality (as described in Section 4.2.5). Consistent with the trends observed in the visual corruption section, we find similar patterns for the audio modality. The results highlight that our method is resilient against various levels of audio corruptions.

Missing Modalities. While corruptions interrupt the overall quality of a modality, some information may still be retained, which might be useful for the network to perform the predictions. To take the challenge up a notch, in this section, we consider the missing modality setting in which a modality is entirely removed. This may be considered as an extreme case where no information regarding the corrupted modality is available to

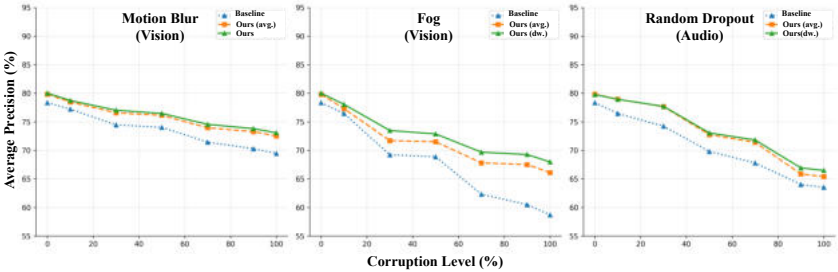


Figure 14: Comparison of weighting schemes (average and dynamic) with baseline concatenation approach [180] on two visual and one audio corruption.

the network. Although the performances of our method (80.00%) and the baseline (78.30%) are comparable when modalities are completely available, our method notably outperforms the baseline method under missing modalities. Notably, our method achieves performance of 67.15% and 72.55%, whereas the baseline method attains scores of 49.50% and 69.10% when removing visual and audio modality, respectively, highlighting the robustness of our approach.

Mixed Corruptions. As real-world environments may have several distortions present, we also explore the effects of the modalities compromised by mixed corruptions. To carry out this experiment, for each level of corruption, it is ensured that the test set contains all corruption types accounting for the total percentage of the corruption level. Performance trends similar to the previously discussed experiments are observed, where our approach outperforms the baseline in all compared cases, signifying the importance of our method against modality corruptions in real-world settings.

Audio and Visual Corruptions. Since real-world environments often contain multiple distortions at the same time, we investigate their impact on both modalities. Consistent with prior experiments, the corruptions cause significant drops in the performance of the baseline, whereas our approach demonstrates better resilience.

Table 32: Comparison of baseline vs. proposed when one modality is completely missing and the other is corrupted. Higher accuracy is better.

Corruption Levels		0%		30%		70%		100%	
Corruption	**Mod**	[180]	**Ours**	[180]	**Ours**	[180]	**Ours**	[180]	**Ours**
bit_error	V	69.17	**73.82**	63.76	**67.31**	61.57	**64.70**	59.23	**63.42**
brightness	V	69.17	**73.82**	65.64	**70.27**	63.82	**68.76**	60.84	**66.45**
contrast	V	69.17	**73.82**	59.51	**65.79**	48.12	**55.83**	38.65	**49.07**
fog	V	69.17	**73.82**	56.97	**60.76**	52.48	**54.49**	50.25	**50.41**
rain	V	69.17	**73.82**	63.89	**69.15**	59.60	**64.49**	58.11	**60.33**
motion_blur	V	69.17	**73.82**	63.21	**68.83**	58.21	**64.79**	54.61	**62.29**
saturate	V	69.17	**73.82**	63.27	**68.94**	57.61	**64.61**	54.16	**62.22**
shot_noise	V	69.17	**73.82**	65.11	**67.97**	**64.65**	63.23	**65.44**	59.28
average	V	69.17	**73.82**	62.67	**67.38**	58.26	**62.61**	55.16	**59.18**
babble	A	49.50	**67.65**	41.42	**59.27**	32.94	**42.12**	28.47	**31.47**
bitrate	A	49.50	**67.65**	43.10	**56.95**	34.36	**40.57**	28.74	**32.11**
hfchannel	A	49.50	**67.65**	40.92	**58.91**	29.05	**36.21**	**24.55**	24.01
pink	A	49.50	**67.65**	41.92	**59.12**	28.86	**36.80**	**25.93**	24.57
pitch_shift	A	49.50	**67.65**	36.68	**57.96**	26.11	**34.90**	22.23	**23.27**
dropout	A	49.50	**67.65**	38.74	**58.46**	27.53	**37.90**	23.62	**27.20**
reverb	A	49.50	**67.65**	40.07	**58.47**	30.83	**37.22**	26.09	**26.39**
white	A	49.50	**67.65**	42.38	**58.90**	29.77	**35.85**	**24.40**	23.40
average	A	49.50	**67.65**	40.65	**58.51**	29.93	**37.70**	25.50	**26.55**

Unimodal Testing with Corruptions

In a series of experiments, we explore a use case on our method in which one modality is completely missing while the other modality is corrupted. The intuition behind this experiment is to observe the importance of each modality for training and testing a robust multimodal anomaly detection system. Table 32 summarizes the results obtained during these experiments. It may be seen that the video modality is notably dominant than the audio modality. In the case of 0% corruption cases, the presence of video modality outperforms the presence of audio modality (AP of 73.82% vs. 67.65%). When the noise is added to the visual modality, a notable performance drop is observed in each corruption case (an average drop in AP from 73.80% to 59.18%). However, the performance drop increases dramatically when the system is exposed to 100% missing visual modality while adding noise to the audio modality (an average drop in AP from 67.65% to 26.55%). This series of experiments highlights that while audio

Table 33: Plug-and-play performance of our approach for mixed audio and visual distortions at 0% and 100% levels of corruption on multiple baselines including XDVioDet, HyperVD, and RTFM. Bold: best performance.

Corrupted Modality	Visual		Audio	
	0%	100%	0%	100%
XDVioDet (Baseline)	78.36	71.45	78.36	67.14
XDVioDet (Ours)	**80.00**	**76.27**	**80.00**	**69.97**
HyperVD (Baseline)	85.60	73.28	85.60	76.90
HyperVD (Ours)	**85.90**	**77.29**	**85.90**	**78.18**
RTFM (Baseline)	**77.66**	70.43	**77.66**	**77.44**
RTFM (Ours)	77.30	**72.67**	77.30	74.21

modality plays an important role in the overall performance, the visual modality is generally more informative, and the performance deteriorates dramatically if the visual modality is compromised.

4.4 Analysis and Discussion

4.4.1 Importance of Dynamic Weighting

During training, our method processes each modality independently, while at test time, we employ a dynamic weighting strategy (Section 4.2.4) to effectively combine the predictions of individual modalities for multimodal performance. In this section, we compare the empirical performance of three approaches: the naive averaging approach (as chalked in Eq. 4.3), dynamic weighting (as chalked in Eq. 4.5), and the concatenation-based inference (as proposed by the baseline XDVioDet [180]). Results on two visual corruptions (motion blur and Fog) and one audio Corruption (Random Dropout) provided in Fig. 14 show that the naive averaging approach with our method outperforms the baseline in all cases, demonstrated the importance of shared space representation learning. Furthermore, the best results are achieved with the proposed dynamic weighting approach.

4.4.2 Is Our Approach Plug-and-Play?

The components of our approach, such as projection layer, shared representation learning, and dynamic weighting, can be plugged into existing multimodal anomaly detection methods, including XDVioDet [180], HyperVD [426], and RTFM [237]. These methods usually incorporate multimodal information by concatenating audio-visual modalities at feature-level input, making them prone to corrupted modalities during testing. As seen in Table 33, when the components of our approach are added to XDVioDet and HyperVD, it outperforms the baseline in both cases of mixed corrupted modalities by a notable margin. In the case of RTFM as baseline, our method outperforms in the case of visual modality. However, in the case of corrupted audio modality, RTFM baseline performs better. This may be attributed to the projection layer used in our approach that upscale the audio features to match the dimensions of visual features. In contrast, RTFM by default uses the visual feature dimension of 2048, whereas the dimension of added audio features is 128.

4.4.3 Does Multimodal Robustness Translate to Better Zero-shot Performance

While the proposed shared space learning for multimodal data improves robustness against compromised modality, we explore its effectiveness in zero-shot setting. To this end, we utilize the models trained on XD-Violence dataset and evaluate on the test set of a benchmark anomaly detection dataset, UCF-crime [122]. As UCF-Crime is a visual-only dataset with no audio, this experiment poses two challenges: 1) Generalizability across datasets. 2) Missing audio data during testing. The results are summarized in Table 34. As seen, our method outperformed RTFM and XDVioDet baselines as well as existing unsupervised video anomaly detection methods [293], [392], [393].

Table 34: AUC comparison on the test set of UCF-Crime dataset under zero-shot setting where the models are trained on XD-voilence dataset. Our approach demonstrates better zero-shot generalization, highlighting the importance of our proposed shared space representation learning. As seen, our approach also outperforms the existing unsupervised approaches trained and tested on UCF-crime [122].

	Method	UCF-Crime
Unsupervised	Kim et al. [209]	52.00
	GCL [293]	65.32
	C2FPL [392]	65.85
	CLAP [393]	67.74
Zero-shot	RTFM [237] (Baseline)	59.6
	RTFM (Ours)	**68.9**
	XDVioDet [180] (Baseline)	63.3
	XDVioDet (Ours)	**68.4**

Table 35: Importance of linear projection compared to padding and concatenation (baseline) under different levels (0% to 100%) of an arbitrarily selected corruption case (pitch shift) for conciseness.

Method	0%	10%	30%	50%	70%	90%	100%
Baseline	78.36	76.24	73.49	69.53	67.01	63.66	63.07
Padding	76.21	73.46	70.52	64.30	60.15	56.83	55.86
Linear Proj.	80.00	78.87	77.62	72.83	71.23	66.91	66.39

4.4.4 Is Linear Projection Necessary?

In this section, we explore the importance of the linear projection layer used in our approach. To this end, we replace it with zero-padding to match the embedding dimension of visual and audio modalities. The results are reported in Table 35. As seen, the padding demonstrates lower performance than even the baseline, whereas our design choice of using projection layer notably helps achieve better performance. This demonstrates that projection layer complements the proposed shared space representation learning for robust anomaly detection under multimodal setting, as proposed in our approach.

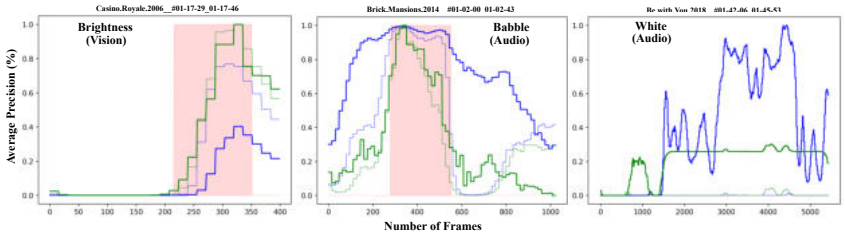


Figure 15: Qualitative results of `RobustA` and the baseline on three videos taken from XD-Violence dataset. Blue represents baseline anomaly scores whereas green represents `RobustA` results. Dotter represents clean test samples whereas solid highlights corruption cases. As seen, our approach generally outputs comparable anomaly scores for clean and corruption cases. The baseline, while generating competitive anomaly scores for clean samples, demonstrates deteriorated performance when the input is corrupted. Red shaded area represents anomaly ground truth.

4.4.5 Qualitative Results

Fig. 15 shows anomaly score plots of `RobustA` and the baseline on different videos highlighting cases with and without corruption. As seen, both `RobustA` and the baseline demonstrate reasonable anomaly scores when neither of the modalities is compromised. The performance of the baseline drops notably when modalities are corrupted, while `RobustA` retains the performance.

4.5 Chapter Summary

In this chapter, we presented a comprehensive study to investigate the adverse effects of corrupted modalities on multimodal anomaly detection task. To address this, we introduced a novel dataset `RobustA`, to systematically evaluate the impact of audio and visual corruptions on anomaly detection performance. Moreover, we proposed a robust multimodal anomaly method that demonstrate resilience against corrupted and missing data. Extensive experiments on the XD-Violence dataset, across various corruption types and levels, demonstrated the effectiveness and robustness of our proposed method. `RobustA` would be instrumental in

evaluating anomaly detection methods under the settings more closer to the real-world scenarios.

Chapter 5

Key Insights and Future Work

5.1 Analysis and Key Insights

This section synthesizes the findings presented throughout this thesis and provides a unified analysis of multimodal learning under imperfect data conditions. Although multimodal learning has demonstrated consistent performance gains over unimodal approaches under modality-complete settings, the results analyzed in this thesis reveal that such improvements do not translate to robustness under realistic conditions. In particular, multimodal learning systems exhibit substantial performance degradation when modalities are missing or corrupted, indicating that robustness is not an inherent property of multimodal architectures but rather a characteristic that must be explicitly addressed.

Architectural assumptions: A central observation emerging from our work is that most contemporary multimodal architectures implicitly assume the availability and reliability of all modalities during both training and inference. Existing MML models tend to couple semantic reasoning with modality presence. As a consequence, these systems learn inference pathways that are optimized for modality-complete inputs and do not generalize when one or more modalities are absent. Empirical evalua-

tions across datasets and tasks showcases that the removal of a modality frequently disrupts learned representations to the extent that multimodal models underperform even unimodal baselines. These findings suggest that performance degradation under missing modalities cannot be attributed solely to reduced input information, but rather to architectural dependencies formed during training. In addition, at times MML models tend to over-rely on dominant modalities when trained under ideal conditions, reinforcing modality-specific biases instead of learning modality-invariant representations.

Learning under missing modalities: The analysis of methods addressing missing modalities reveals a broad spectrum of strategies, ranging from reconstruction-based approaches to architectural and hybrid solutions. Reconstruction-based methods attempt to recover missing modalities by exploiting cross-modal correlations, often through generative or alignment-based techniques. While effective under controlled assumptions, these approaches rely on strong predictability between modalities and are sensitive to distribution shifts, noise, and partial missingness. Architectural approaches, in contrast, aim to mitigate the impact of missing modalities by enabling inference using only the available inputs. Our work demonstrates that enforcing a shared representation space and modality-agnostic inference pathways substantially improves robustness under missing-modality settings. By decoupling semantic reasoning from modality presence, such architectures exhibit reduced performance degradation and more stable behavior across varying modality configurations. These results indicate that robustness to missing modalities is most effectively addressed through architectural design rather than post-hoc reconstruction.

Reliability-Aware inference: Beyond missing modalities, corrupted modalities introduce a more challenging and less explored failure mode. Unlike missing data, corrupted modalities are present but unreliable, often containing misleading or noisy information that is propagated through the learning pipeline. Our results indicate that standard multimodal fusion mechanisms frequently fail to distinguish between reliable and corrupted inputs, resulting in overconfident and unstable predictions even when

clean modalities are available. This behavior highlights a fundamental limitation of existing multimodal architectures: the absence of explicit mechanisms for modeling modality reliability. The results obtained using corruption-aware benchmarks demonstrate that robustness under such conditions requires explicit confidence estimation, uncertainty modeling, and adaptive modality weighting. Architectural robustness alone is therefore insufficient to address corruption-related failures.

Robustness in prior work: An important insight arising from our unified analysis is that prior research has largely treated missing and corrupted modalities as separate problems. Reconstruction and architectural methods predominantly focus on modality absence, while denoising, augmentation, and noise-aware techniques target corruption. However, real-world multimodal systems often encounter both challenges simultaneously, with some modalities missing and others degraded. This fragmentation limits the applicability of existing solutions, as methods designed for one type of imperfection frequently fail under the other. Our work underscores the need for integrated robustness strategies that account for both absence and unreliability, and that are evaluated under realistic real-world conditions.

Implications for robust MML: Synthesizing evidence across the survey, architectural contributions, and benchmarking efforts, we identify several implications for the design of robust multimodal systems. First, robustness should be treated as a first-class design objective rather than an emergent property of multimodal fusion or model scale. Second, architectures should support modality-agnostic inference to enable graceful degradation under missing modalities. Third, robustness to corrupted modalities requires explicit modeling of modality reliability and adaptive fusion mechanisms. Finally, standardized benchmarks that reflect realistic data imperfections are essential for meaningful robustness evaluation.

In a nutshell, MML is inherently fragile under imperfect data conditions. Robustness does not arise naturally from combining multiple modalities, but must be explicitly infused through architectural constraints, learning objectives, and evaluation strategies. By integrating taxonomic analysis, robust architectural design, and principled bench-

marking, this thesis provides a coherent foundation for reliable MML.

5.2 Limitations

While this thesis provides a unified analysis and concrete advances toward robust MML, several limitations and scope constraints should be acknowledged. First, the architectural and algorithmic contributions primarily focus on robustness under structured forms of modality absence and corruption. Although the considered scenarios reflect common real-world conditions such as missingness, noise, and degradation, they do not entail the full space of possible multimodal imperfections. In particular, adversarially crafted corruptions, extreme distribution shifts, and long-term temporal failures remain outside the scope of our work. Second, the proposed robustness mechanisms assume the availability of at least one informative modality at inference time. In cases where all modalities are simultaneously missing or severely corrupted, reliable inference is inherently ill-posed. Addressing such scenarios would require incorporating external priors, memory-based reasoning, or active sensing strategies, that remain out of scope for our study. Third, while the benchmarks and evaluations span diverse datasets and modality combinations, the analysis primarily focuses on classification and anomaly detection tasks. Extending the presented insights to generative, sequential decision-making, or large-scale multimodal foundation models introduces additional challenges related to scalability, training dynamics, and evaluation that requires separate investigation. Another important observation relates to how multimodal representations are formed within existing architectures. While many approaches rely on implicit fusion mechanisms in latent spaces, recent works have explored more explicit strategies for modality unification, such as projecting heterogeneous inputs into a shared representation space. Our design is motivated by this broader line of research, as discussed in Section 3.1, where non-visual modalities are transformed to enable unified processing. In contrast to approaches that rely on modality-specific tokens or implicit alignment within hidden layers, our formulation minimizes modality-specific as-

sumptions and enforces a common representation space that is directly optimized for robustness under missing inputs. While alternative formulations, including token-based or hybrid fusion strategies, remain viable, their systematic evaluation under missing and corrupted modality settings presents an interesting direction for future work. Finally, robustness is evaluated under controlled and reproducible corruption protocols to enable systematic analysis. Although such protocols are necessary for principled benchmarking, they may not fully capture the complexity and unpredictability of real-world multimodal degradation. It requires tighter integration between benchmark design, real-world data collection, and evaluation, which remains an open direction for future work.

5.3 Challenges & Future Research

Finally, we synthesize the research gaps and outline concrete, evidence-based future directions in the following.

Efficiency, Scalability, and Generalization: Despite a proliferation of task-specific models, current methods lack generalization due to diverse modality configurations in different application areas. In addition, while effective, hybrid models such as DCFMNet [463], often suffer from computational bloat. For instance, AMM-Diff [442] and OPTIMUS [430] report performance improvements but entail a fourfold increase in training cost compared to standard encoders. Future work should focus on exploring parameter-efficient architectures, such as modular transformers with dynamic routing, that can activate only relevant sub-modules depending on modality availability. Similarly, as proposed in LMIM [465], cross-domain pretraining on heterogeneous multimodal datasets could improve transferability. Finally, benchmarking models under varying computational cost would provide clearer cost-benefit trade-offs.

Modality Drift and Misalignment: Han et al. [259] and Zhang et al. [469] are examples of alignment-based models that show performance deterioration when modalities are asynchronous or semantically divergent. Similarly, many architectures such as M3Care [298] struggle under temporal drift, where the statistical relationships or synchronization among

modalities change over time. Future studies can investigate these challenges and propose novel methods that integrate causal modeling or temporal attention mechanisms by dynamically realigning modalities as explored by SuMi [435]. Similarly, interested researchers can investigate meta-learning strategies that can adapt alignment functions dynamically under the contextual shifts in semantic meaning.

Adversarial Robustness: While adversarial robustness has been addressed in contrastive learning frameworks including CleanCLIP [307] and AdvCLIP [376], it is often confined to vision-language pairs and fail under open-set corruptions or cross-modal perturbations. Zhang et al. [422] show that large multimodal models lose up to 25–40% accuracy in case of corruptions in textual and visual input. Future research towards adversarial robustness should investigate cross-modal perturbation augmentation strategies, where noise in one modality affects alignment in another, to simulate real-world sensor malfunctions. Meanwhile, development of open-set corruption detectors and robust pretraining objectives also remains as a key challenge [320].

Standardized Benchmarks and Taxonomies: Many studies use custom or synthetic corruption setups, making cross-comparison difficult. While recent work such as Beemelmanns et al. [380], Zhang et al. [422], Dong et al. [317], and Yu et al. [370] take steps toward benchmark unification, the field still lacks holistic standards for missing and corrupted modalities across different domain areas. In addition, a critical challenge towards robust multimodal architecture remains due to unavailability of standardized taxonomy of corruption types, model-agnostic evaluation metrics, and challenge datasets.

Imperfect modalities in Multimodal Large Language models: The rise of multimodal large language models (MLLMs) such as GPT-4V, LLaVA, and Flamingo introduces new and largely unexplored vulnerabilities to missing and corrupted modalities. Unlike smaller task-specific architectures, MLLMs operate at scale and are increasingly deployed in open-world settings where input completeness cannot be guaranteed. Future work should systematically investigate how modality degradation propagates through the attention and cross-modal alignment mechanisms central to

these models, and whether robustness strategies developed in this thesis — such as shared representation spaces and adaptive fusion — can be adapted or scaled to the MLLM regime.

Interpretability and Explainability: Despite exceptional success across multiple application areas such as Cross-modal Information Retrieval [179], Multimodal Emotion Recognition [159], Sentiment Analysis [290], and Genere Classification [271], interpretability of existing robustness strategies is still rudimentary. This remains as a key challenge as most models offer no transparency into why a modality is ignored [361], compensated for [240], or hallucinated [379].

Dynamic Modality Estimation: While some modalities become more informative under specific environmental or task conditions (e.g., thermal sensing at night vs. RGB during daylight), most multimodal architectures treat all modalities uniformly. This often leads to degraded performance when some modalities are unreliable. Although recent works such as AV-RelScore [322] have introduced confidence scoring to weigh modalities, such strategies remain largely application-specific and lack generalizability across diverse domains. A promising direction is to develop model- and task-agnostic reliability estimators that can predict both modality-level contributions and resource costs.

Conclusion

This thesis presents a comprehensive exploration of multimodal robustness by integrating a detailed literature taxonomy, a novel missing-modality framework, and a corruption-robust evaluation methodology. `Chameleon` demonstrated that unified representations offer stability under missing-modality conditions, while `RobustA` showed that corruption-aware training significantly enhances real-world reliability. Together, these contributions advance the state of multimodal learning by bridging conceptual foundations with practical robustness requirements. The work presented here moves the field closer to designing multimodal systems that are accurate, resilient, flexible, and ready for deployment in environments where uncertainty is not the exception but the norm.

Bibliography

[1] S. Vogel, H. Ney, and C. Tillmann, "Hmm-based word alignment in statistical translation," in *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*, 1996.

[2] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. eleventh annual conference on Computational learning theory*, 1998, pp. 92–100.

[3] H. Aradhye and C. Dorai, "New kernels for analyzing multimodal data in multimedia using kernel machines," in *Proceedings. IEEE International Conference on Multimedia and Expo*, IEEE, vol. 2, 2002, pp. 37–40.

[4] K. Sjölander, "An hmm-based system for automatic segmentation and alignment of speech," in *Proceedings of fonetik*, Citeseer, vol. 2003, 2003, pp. 93–96.

[5] C. Busso, Z. Deng, S. Yildirim, *et al.*, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. ICMI*, 2004, pp. 205–211.

[6] M.-A. Krogel and T. Scheffer, "Multi-relational learning, text mining, and semi-supervised learning for functional genomics," *Machine Learning*, vol. 57, pp. 61–81, 2004.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[8] H. Gunes and M. Piccardi, "Affect recognition from face and body: Early fusion vs. late fusion," in *2005 IEEE international conference on systems, man and cybernetics*, IEEE, vol. 4, 2005, pp. 3437–3443.

[9] C. Busso, M. Bulut, C.-C. Lee, and *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources & Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[10] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: Face, body gesture, speech," *Affect and Emotion in Human-Computer Interaction: From Theory to Applications*, pp. 92–103, 2008.

[11] M. Gurban, J.-P. Thiran, T. Drugman, and T. Dutoit, "Dynamic modality weighting for multi-stream hmms inaudio-visual speech recognition," in *Proc. 10th international conference on Multimodal interfaces*, 2008, pp. 237–240.

[12] P. Gehler and S. Nowozin, "On feature combination for multiclass object classification," in *2009 IEEE 12th International Conference on Computer Vision*, IEEE, 2009, pp. 221–228.

[13] R. Raghavendra, M. Imran, A. Rao, and G. H. Kumar, "Multimodal biometrics: Analysis of handvein & palmprint combination used for person verification," in *ICETET*, IEEE, 2010, pp. 526–530.

[14] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng, *et al.*, "Multimodal deep learning.," in *ICML*, vol. 11, 2011, pp. 689–696.

[15] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," *NeurIPS*, vol. 24, 2011.

[16] G. A. Ramirez, T. Baltrušaitis, and L.-P. Morency, "Modeling latent discriminative dynamic of multi-dimensional affective signals," in *ACII*, Springer, 2011, pp. 396–406.

[17] G. Hinton, L. Deng, D. Yu, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[18] P. Kuznetsova, V. Ordonez, A. Berg, T. Berg, and Y. Choi, "Collective generation of natural image descriptions," in *Proc. 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2012, pp. 359–368.

[19] P. Nakov and H. T. Ng, "Improving statistical machine translation for a resource-poor language using related resource-rich languages," *Journal of Artificial Intelligence Research*, vol. 44, pp. 179–222, 2012.

[20] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," *NeurIPS*, vol. 25, 2012.

[21] S. L. Taylor, M. Mahler, B.-J. Theobald, and I. Matthews, "Dynamic units of visual speech," in *Proc. 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, 2012, pp. 275–284.

[22] R. Anderson, B. Stenger, V. Wan, and R. Cipolla, "Expressive visual text-to-speech using active appearance models," in *CVPR*, 2013, pp. 3382–3389.

[23] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE TPAMI*, vol. 35, no. 8, pp. 1798–1828, 2013.

[24] P. Das, C. Xu, R. F. Doell, and J. J. Corso, "A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching," in *CVPR*, 2013, pp. 2634–2641.

[25] A. Frome, G. S. Corrado, J. Shlens, *et al.*, "Devise: A deep visual-semantic embedding model," *NeurIPS*, vol. 26, 2013.

[26] B. Li and L. Han, "Distance weighted cosine similarity measure for text classification," in *IDEAL*, Springer, 2013, pp. 611–618.

[27] F. Liu, L. Zhou, C. Shen, and J. Yin, "Multiple kernel learning in the primal for multimodal alzheimer's disease classification," *IEEE journal of biomedical and health informatics*, vol. 18, no. 3, pp. 984–990, 2013.

[28] N. McLaughlin, J. Ming, and D. Crookes, "Robust multimodal person identification with limited training data," *IEEE Transactions on Human-Machine Systems*, vol. 43, no. 2, pp. 214–224, 2013.

[29] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *NeurIPS*, vol. 26, 2013.

[30] K. Sikka, K. Dykstra, S. Sathyanarayana, G. Littlewort, and M. Bartlett, "Multiple kernel learning for emotion recognition in the wild," in *Proc. ACM ICMI*, 2013, pp. 517–524.

[31] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, "Zero-shot learning through cross-modal transfer," *NeurIPS*, vol. 26, 2013.

[32] Z. Ding, S. Ming, and Y. Fu, "Latent low-rank transfer subspace learning for missing modality recognition," in *AAAI*, vol. 28, 2014.

[33] D. Kiela and L. Bottou, "Learning image embeddings using convolutional neural networks for improved multi-modal semantics," in *Proceedings of the 2014 Conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 36–45.

[34] Z.-z. Lan, L. Bao, S.-I. Yu, W. Liu, and A. G. Hauptmann, "Multimedia classification and event detection using double fusion," *Multimedia tools and applications*, vol. 71, pp. 333–347, 2014.

[35] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," *ECCV*, pp. 740–755, 2014.

[36] S. Moon, S. Kim, and H. Wang, "Multimodal transfer deep learning for au-dio visual recognition," *arXiv:1412.3121*, 2014.

[37] D. Rao, M. De Deuge, N. Nourani-Vatani, B. Douillard, S. B. Williams, and O. Pizarro, "Multimodal learning for autonomous underwater vehicles from visual and bathymetric data," in *ICRA*, IEEE, 2014, pp. 3819–3825.

[38] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 207–218, 2014.

[39] J. Thomason, S. Venugopalan, S. Guadarrama, K. Saenko, and R. Mooney, "Integrating language and vision to generate natural language descriptions of videos in the wild," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 1218–1227.

[40] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," *arXiv:1412.4729*, 2014.

[41] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," in *TACL*, vol. 2, 2014, pp. 67–78.

[42] X. Chen, H. Fang, T.-Y. Lin, *et al.*, "Microsoft coco captions: Data collection and evaluation server," *arXiv:1504.00325*, 2015.

[43] Z. Ding, M. Shao, and Y. Fu, "Missing modality transfer learning via latent low-rank constraint," *IEEE transactions on Image Processing*, vol. 24, no. 11, pp. 4322–4334, 2015.

[44] P. Isola, J.-Y. Lim, and E. Adelson, "Discovering states and transformations in image collections," in *CVPR*, 2015, pp. 1383–1391.

[45] N. Jaques, S. Taylor, A. Sano, and R. Picard, "Multi-task, multi-kernel learning for estimating individual wellbeing," in *Proc. NIPS Workshop on Multimodal Machine Learning, Montreal, Quebec*, vol. 898, 2015, p. 3.

[46] X. Jiang, F. Wu, Y. Zhang, S. Tang, W. Lu, and Y. Zhuang, "The classification of multi-modal data with hidden conditional random field," *Pattern Recognition Letters*, vol. 51, pp. 63–69, 2015.

[47] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015, pp. 3128–3137.

[48] D. Kiela and S. Clark, "Multi-and cross-modal semantics beyond vision: Grounding in auditory perception," in *Proc. EMNLP*, 2015, pp. 2461–2470.

[49] R. Lebret, P. Pinheiro, and R. Collobert, "Phrase-based image captioning," in *ICML*, PMLR, 2015, pp. 2085–2094.

[50] Y. Li, S. Wang, Q. Tian, and X. Ding, "A survey of recent advances in visual feature detection," *Neurocomputing*, vol. 149, pp. 736–751, 2015.

[51] F. Liu, X. Xu, S. Qiu, C. Qing, and D. Tao, "Simple to complex transfer learning for action recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 949–960, 2015.

[52] B. McFee, C. Raffel, D. Liang, *et al.*, "Librosa: Audio and music signal analysis in python.," *SciPy*, vol. 2015, pp. 18–24, 2015.

[53] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Moddrop: Adaptive multi-modal gesture recognition," *IEEE TPAMI*, vol. 38, no. 8, pp. 1692–1706, 2015.

[54] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proc. IEEE international conference on computer vision*, 2015, pp. 2641–2649.

[55] M. Shao, Z. Ding, and Y. Fu, "Sparse low-rank fusion based deep features for missing modality face recognition," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, IEEE, vol. 1, 2015, pp. 1–6.

[56] M. Tapaswi, M. Bauml, and R. Stiefelhagen, "Book2movie: Aligning video scenes with book chapters," in *CVPR*, 2015, pp. 1827–1835.

[57] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," *arXiv:1511.06361*, 2015.

[58] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[59] X. Wang, D. Kumar, N. Thome, M. Cord, and F. Precioso, "Recipe recognition with large multimodal food dataset," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, IEEE, 2015, pp. 1–6.

[60] R. Xu, C. Xiong, W. Chen, and J. Corso, "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework," in *AAAI*, vol. 29, 2015.

[61] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30k: Multilingual english-german image descriptions," *arXiv preprint arXiv:1605.00459*, 2016.

[62] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, J. Su, K. Duh, and X. Carreras, Eds., Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 457–468. DOI: 10.18653/v1/D16-1044. [Online]. Available: https://aclanthology.org/D16-1044.

[63] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[65] A. E. Johnson, T. J. Pollard, L. Shen, *et al.*, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[66] B. Mahasseni and S. Todorovic, "Regularizing long short term memory with 3d human-skeleton sequences for action recognition," in *CVPR*, 2016, pp. 3054–3062.

[67] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *CVPR*, 2016, pp. 11–20.

[68] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *CVPR*, 2016, pp. 1010–1019.

[69] L. Specia, S. Frank, K. Sima'An, and D. Elliott, "A shared task on multimodal machine translation and crosslingual image description," in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 2016, pp. 543–553.

[70] G. Trigeorgis, M. A. Nicolaou, S. Zafeiriou, and B. W. Schuller, "Deep canonical time warping," in *CVPR*, 2016, pp. 5110–5118.

[71] G. Trigeorgis, F. Ringeval, R. Brueckner, *et al.*, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP*, IEEE, 2016, pp. 5200–5204.

[72] L. Wang, Y. Li, and S. Lazebnik, "Learning deep structure-preserving image-text embeddings," in *CVPR*, 2016, pp. 5005–5013.

[73] P. Wang, P. Nakov, and H. T. Ng, "Source language adaptation approaches for resource-poor machine translation," *Computational Linguistics*, vol. 42, no. 2, pp. 277–306, 2016.

[74] Y. Wang, C. Von Der Weth, Y. Zhang, K. H. Low, V. K. Singh, and M. Kankanhalli, "Concept based hybrid fusion of multimodal event signals," in *2016 IEEE International Symposium on Multimedia (ISM)*, IEEE, 2016, pp. 14–19.

[75] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *ICML*, PMLR, 2016, pp. 2397–2406.

[76] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *CVPR*, 2016, pp. 5288–5296.

[77] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv:1606.06259*, 2016.

[78] E. Alevizos, A. Skarlatidis, A. Artikis, and G. Paliouras, "Probabilistic complex event recognition: A survey," *ACM Computing Surveys (CSUR)*, vol. 50, no. 5, pp. 1–31, 2017.

[79] J. Arevalo, T. Solorio, M. Montes-y-Gómez, and F. A. González, "Gated multimodal units for information fusion," *arXiv:1702.01992*, 2017.

[80] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 6299–6308.

[81] I. Gallo, A. Calefati, and S. Nawaz, "Multimodal classification fusion in real-world scenarios," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, vol. 5, 2017, pp. 36–41.

[82] I. Gallo, S. Nawaz, and A. Calefati, "Semantic text encoding for text classification using convolutional neural networks," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, vol. 5, 2017, pp. 16–21.

[83] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, IEEE, 2017, pp. 776–780.

[84] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *CVPR*, 2017, pp. 6904–6913.

[85] Y. Gu, J. Chanussot, X. Jia, and J. A. Benediktsson, "Multiple kernel learning for hyperspectral image classification: A review," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 11, pp. 6547–6565, 2017.

[86] J. Johnson, B. Hariharan, L. Van Der Maaten, and ..., "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *CVPR*, 2017, pp. 2901–2910.

[87] D. Kiela and S. Clark, "Learning neural audio embeddings for grounding semantics in auditory perception," *Journal of Artificial Intelligence Research*, vol. 60, pp. 1003–1030, 2017.

[88] R. Krishna, K. Hata, F. Ren, J. C. Niebles, and L. Fei-Fei, "Dense-captioning events in videos," in *ICCV*, 2017, pp. 706–715.

[89] R. Krishna, Y. Zhu, O. Groth, J. Johnson, ..., and F.-F. Li, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, vol. 123, no. 1, pp. 32–73, 2017.

[90] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[91] V. Mitra, H. Franco, R. M. Stern, *et al.*, "Robust features in deep-learning-based speech recognition," *New Era for Robust Speech Recognition: Exploiting Deep Learning*, pp. 187–217, 2017.

[92] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *arXiv:1706.08612*, 2017.

[93] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[94] L. Tran, X. Liu, J. Zhou, and R. Jin, "Missing modalities imputation via cascaded residual autoencoder," in *CVPR*, 2017, pp. 1405–1414.

[95] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE TPAMI*, vol. 44, no. 12, pp. 8717–8727, 2018.

[96] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in *CVPR*, 2018, pp. 4971–4980.

[97] F. Alam, F. Ofli, and M. Imran, "Crisismmd: Multimodal twitter datasets from natural disasters," in *Proc. 12th International AAAI Conference on Web and Social Media (ICWSM)*, 2018, pp. 465–472.

[98] P. Anderson, X. He, C. Buehler, *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[99] A. A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *ACL*, 2018, pp. 2236–2246.

[100] S. Bakas, M. Reyes, A. Jakab, *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv:1811.02629*, 2018.

[101] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE TPAMI*, vol. 41, no. 2, pp. 423–443, 2018.

[102] L. Cai, Z. Wang, H. Gao, D. Shen, and S. Ji, "Deep adversarial learning for multi-modality missing data completion," in *Proc. 24th ACM SIGKDD*, 2018, pp. 1158–1166.

[103] C. Du, C. Du, H. Wang, *et al.*, "Semi-supervised deep generative modelling of incomplete multi-modality emotional data," in *ACM Multimedia*, 2018, pp. 108–116.

[104] I. Gallo, A. Calefati, S. Nawaz, and M. K. Janjua, "Image and encoded text fusion for multi-modal classification," in *2018 Digital Image Computing: Techniques and Applications (DICTA)*, IEEE, 2018, pp. 1–7.

[105] N. C. Garcia, P. Morerio, and V. Murino, "Modality distillation with multiple stream networks for action recognition," in *ECCV*, 2018, pp. 103–118.

[106] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *ECCV*, 2018, pp. 172–189.

[107] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Urban land cover classification with missing data modalities using deep convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 6, pp. 1758–1768, 2018.

[108] D. Kiela, E. Grave, A. Joulin, and T. Mikolov, "Efficient large-scale multi-modal classification," *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 5198–5204, 2018.

[109] V. Lostanlen, J. Salamon, M. Cartwright, *et al.*, "Per-channel energy normalization: Why and how," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 39–43, 2018.

[110] T. Matsuura, K. Saito, Y. Ushiku, and T. Harada, "Generalized bayesian canonical correlation analysis with missing modalities," in *ECCV*, 2018, pp. 0–0.

[111] S. Moon, L. Neves, and V. Carvalho, "Multimodal named entity recognition for short social media posts," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, M. Walker, H. Ji, and A. Stent, Eds., New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 852–860. DOI: 10.18653/v1/N18-1078. [Online]. Available: https://aclanthology.org/N18-1078/.

[112] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable pins: Cross-modal embeddings for person identity," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 71–88.

[113] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8427–8436.

[114] S. Nawaz, A. Calefati, M. K. Janjua, M. U. Anwaar, and I. Gallo, "Learning fused representations for large-scale multimodal classification," *IEEE Sensors Letters*, vol. 3, no. 1, pp. 1–4, 2018.

[115] D. Nguyen, K. Nguyen, S. Sridharan, D. Dean, and C. Fookes, "Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition," *Computer vision and image understanding*, vol. 174, pp. 33–42, 2018.

[116] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv:1810.02508*, 2018.

[117] A. Saeed, T. Ozcelebi, and J. Lukkien, "Synthesizing and reconstructing missing sensory modalities in behavioral context recognition," *Sensors*, vol. 18, no. 9, p. 2967, 2018.

[118] R. Sanabria, O. Caglayan, S. Palaskar, *et al.*, "How2: A large-scale dataset for multimodal language understanding," in *Proc.ViGIL*, 2018, pp. 1–4.

[119] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *ACL*, 2018, pp. 2556–2565.

[120] Z. Shou, H. Gao, L. Zhang, K. Miyazawa, and S.-F. Chang, "Autoloc: Weakly-supervised temporal action localization in untrimmed videos," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 154–171.

[121] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi, "A corpus for reasoning about natural language grounded in photographs," *arXiv:1811.00491*, 2018.

[122] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6479–6488.

[123] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie, "Centralnet: A multilayer approach for multimodal fusion," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[124] C. Wang, M. Niepert, and H. Li, "Lrmm: Learning to recommend with missing modalities," *arXiv:1808.06791*, 2018.

[125] J.-L. Yin, B.-H. Chen, and Y. Li, "Highly accurate image reconstruction for multimodal noise suppression using semisupervised learning on big data," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3045–3056, 2018.

[126] O. Arshad, I. Gallo, S. Nawaz, and A. Calefati, "Aiding intra-text representations with visual context for multimodal named entity recognition," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, 2019, pp. 337–342.

[127] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. [Online]. Available: https://aclanthology.org/N19-1423.

[128] N. C. Garcia, P. Morerio, and V. Murino, "Cross-modal learning by hallucinating missing modalities in rgb-d vision," in *Multimodal Scene Understanding*, Elsevier, 2019, pp. 383–401.

[129] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv:1903.12261*, 2019.

[130] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *CVPR*, 2019, pp. 6700–6709.

[131] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, and D. Testuggine, "Supervised multimodal bitransformers for classifying images and text," *arXiv preprint arXiv:1909.02950*, 2019.

[132] K. Lau, J. Adler, and J. Sjölund, "A unified representation network for segmentation with missing modalities," *arXiv:1908.06683*, 2019.

[133] H.-C. Lee, C.-Y. Lin, P.-C. Hsu, and W. H. Hsu, "Audio feature generation for missing modality problem in video action recognition," in *ICASSP*, IEEE, 2019, pp. 3956–3960.

[134] L. Li, M. Yatskar, D. Yin, C. Hsieh, and K. Chang, "A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[135] P. P. Liang, Z. Liu, Y.-H. H. Tsai, Q. Zhao, R. Salakhutdinov, and L.-P. Morency, "Learning representations from imperfect time series data via tensor rank regularization," *arXiv:1907.01011*, 2019.

[136] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS*, vol. 32, 2019.

[137] C. Michaelis, B. Mitzkus, R. Geirhos, *et al.*, "Benchmarking robustness in object detection: Autonomous driving when winter is coming," *arXiv:1907.07484*, 2019.

[138] A. Miech, D. Zhukov, J.-B. Alayrac, and *et al.*, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *ICCV*, 2019, pp. 2630–2640.

[139] S. Narayan, H. Cholakkal, F. S. Khan, and L. Shao, "3c-net: Category count and center loss for weakly-supervised action localization," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8679–8687.

[140] S. Nawaz, M. K. Janjua, I. Gallo, A. Mahmood, and A. Calefati, "Deep latent space learning for cross-modal mapping of audio and visual signals," in *2019 Digital Image Computing: Techniques and Applications (DICTA)*, IEEE, 2019, pp. 1–7.

[141] S. Nedelkoski, J. Cardoso, and O. Kao, "Anomaly detection from system tracing data using multimodal deep learning," in *CLOUD*, IEEE, 2019, pp. 179–186.

[142] S. Pande, A. Banerjee, S. Kumar, B. Banerjee, and S. Chaudhuri, "An adversarial approach to discriminative modality distillation for remote sensing image classification," in *CVPR*, 2019, pp. 0–0.

[143] N. Patel, A. Choromanska, P. Krishnamurthy, and F. Khorrami, "A deep learning gated architecture for ugv navigation robust to sensor failures," *Robotics and Autonomous Systems*, vol. 116, pp. 80–97, 2019.

[144] J.-M. Pérez-Rúa, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "Mfas: Multimodal fusion architecture search," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6966–6975.

[145] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *AAAI*, vol. 33, 2019, pp. 6892–6899.

[146] H. Purohit, R. Tanabe, T. Ichige, and *et al.*, "Mimii dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in *arXiv:1909.09347*, 2019.

[147] M. Shah, X. Chen, M. Rohrbach, and D. Parikh, "Cycle-consistency for robust visual question answering," in *CVPR*, 2019, pp. 6649–6658.

[148] Y. Shen and M. Gao, "Brain tumor segmentation on mri with missing modalities," in *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, Springer, 2019, pp. 417–428.

[149] K. Somandepalli, N. Kumar, R. Travadi, and S. Narayanan, "Multimodal representation learning using deep multiset canonical correlation," *arXiv:1904.01775*, 2019.

[150] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *IEEE/CVF ICCV*, 2019, pp. 7464–7473.

[151] X. Wang, J. Wu, J. Chen, and *et al.*, "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," in *ICCV*, 2019, pp. 4581–4591.

[152] Y. Wen, M. A. Ismail, W. Liu, B. Raj, and R. Singh, "Disjoint mapping network for cross-modal matching of voices and faces," in *7th International Conference on Learning Representations, ICLR 2019, USA, May 6-9, 2019*, 2019.

[153] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 5791–5795.

[154] Y. Zheng, Y. Li, and S. Wang, "Intention oriented image captions with guiding objects," in *CVPR*, 2019, pp. 8395–8404.

[155] R. Ardila, M. Branson, K. Davis, and *et al.*, "Common voice: A massively-multilingual speech corpus," *arXiv:1912.06670*, 2020.

[156] J. Chen and A. Zhang, "Hgmf: Heterogeneous graph-based fusion for multimodal data with incompleteness," in *Proc. 26th ACM SIGKDD*, 2020, pp. 1295–1305.

[157] Y.-C. Chen, L. Li, L. Yu, and *et al.*, "Uniter: Universal image-text representation learning," *ECCV*, pp. 104–120, 2020.

[158] Z. Chen, S. Wang, and Y. Qian, "Multi-modality matters: A performance leap on voxceleb.," in *Interspeech*, 2020, pp. 2252–2256.

[159] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168 865–168 878, 2020.

[160] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[161] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[162] I. Gallo, G. Ria, N. Landro, and R. La Grassa, "Image and text fusion for upmc food-101 using bert and cnns," in *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, IEEE, 2020, pp. 1–6.

[163] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *ACM MM*, 2020, pp. 1122–1131.

[164] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, 2020.

[165] D. Kiela, H. Firooz, A. Mohan, *et al.*, "The hateful memes challenge: Detecting hate speech in multimodal memes," *NeurIPS*, vol. 33, pp. 2611–2624, 2020.

[166] L. Li, B. Du, Y. Wang, L. Qin, and H. Tan, "Estimation of missing values in heterogeneous traffic data: Application of multimodal deep learning model," *Knowledge-Based Systems*, vol. 194, p. 105 592, 2020.

[167] X. Li, X. Yin, C. Li, *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *ECCV*, 2020, pp. 121–137.

[168] F. Ma, D. Meng, X. Dong, and Y. Yang, "Self-paced multi-view co-training," *Journal of Machine Learning Research*, vol. 21, no. 57, pp. 1–38, 2020.

[169] S. Maity, M. Abdel-Mottaleb, and S. S. Asfour, "Multimodal biometrics recognition from facial video with missing modalities using deep learning," *Journal of Information Processing Systems*, vol. 16, no. 1, pp. 6–29, 2020.

[170] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *AAAI*, vol. 34, 2020, pp. 1359–1367.

[171] B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire, and V. Chandran, "Automatic emotion recognition using temporal multimodal deep learning," *IEEE Access*, vol. 8, pp. 225 463–225 474, 2020.

[172] S. Parthasarathy and S. Sundaram, "Training strategies to handle missing modalities for audio-visual expression recognition," in *Proc. ICMI*, 2020, pp. 400–404.

[173] W. U. Rahman, M. K. Hasan, S. J. Lee, and *et al.*, "Integrating deep learning with logic fusion for information extraction," in *AAAI*, 2020, pp. 8912–8920.

[174] D. Roblek, K. Misiunas, M. Tagliasacchi, and P. Li, "Seanet: A multi-modal speech enhancement network," in *Interspeech*, 2020.

[175] J. Roth, S. Chaudhuri, O. Klejch, and *et al.*, "Ava-activespeaker: An audio-visual dataset for active speaker detection," in *ICASSP*, 2020, pp. 4492–4496.

[176] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," *Advances in neural information processing systems*, vol. 33, pp. 16 857–16 867, 2020.

[177] J. Tian, W. Cheung, N. Glaser, Y.-C. Liu, and Z. Kira, "Uno: Uncertainty-aware noisy-or multimodal fusion for unanticipated input degradation," in *ICRA*, IEEE, 2020, pp. 5716–5723.

[178] Q. Wang, L. Zhan, P. Thompson, and J. Zhou, "Multimodal learning with incomplete modalities by knowledge distillation," in *Proc. 26th ACM SIGKDD*, 2020, pp. 1828–1838.

[179] Z. Wang, Z. Wan, and X. Wan, "Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis," in *Proc. of The Web Conference 2020*, 2020, pp. 2514–2520.

[180] P. Wu, J. Liu, Y. Shi, *et al.*, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, Springer, 2020, pp. 322–339.

[181] P. Wu, j. Liu, Y. Shi, *et al.*, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *ECCV*, 2020.

[182] W. Xia, Y. Yang, and J.-H. Xue, "Unsupervised multi-domain multimodal image-to-image translation with explicit domain-constrained disentanglement," *Neural Networks*, vol. 131, pp. 50–63, 2020.

[183] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "Layoutlm: Pre-training of text and layout for document image understanding," in *Proc. 26th ACM SIGKDD*, 2020, pp. 1192–1200.

[184] H. Yang, T. Wang, and L. Yin, "Adaptive multimodal fusion for facial action units recognition," in *ACM Multimedia*, 2020, pp. 2982–2990.

[185] K. You, Z. Kou, M. Long, and J. Wang, "Co-tuning for transfer learning," *NeurIPS*, vol. 33, pp. 17 236–17 246, 2020.

[186] G. Yu, Q. Li, D. Shen, and Y. Liu, "Optimal sparse linear prediction for block-missing multi-modality data without imputation," *Journal of the American Statistical Association*, vol. 115, no. 531, pp. 1406–1419, 2020.

[187] M. Z. Zaheer, A. Mahmood, M. Astrid, and S.-I. Lee, "Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, Springer, 2020, pp. 358–376.

[188] M. Z. Zaheer, A. Mahmood, H. Shin, and S.-I. Lee, "A self-reasoning framework for anomaly detection using video-level labels," *IEEE Signal Processing Letters*, vol. 27, pp. 1705–1709, 2020.

[189] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, vol. 59, pp. 103–126, 2020.

[190] H. Zhong, J. Chen, C. Shen, H. Zhang, J. Huang, and X.-S. Hua, "Self-adaptive neural module transformer for visual question answering," *IEEE Transactions on Multimedia*, vol. 23, pp. 1264–1273, 2020.

[191] T. Zhou, S. Canu, P. Vera, and S. Ruan, "Brain tumor segmentation with missing modalities via latent multi-source correlation representation," in *MICCAI*, Springer, 2020, pp. 533–541.

[192] H. Akbari, L. Yuan, R. Qian, *et al.*, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," *NeurIPS*, vol. 34, pp. 24 206–24 221, 2021.

[193] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *ICCV*, 2021, pp. 1728–1738.

[194] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv:2106.08254*, 2021.

[195] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *CVPR*, 2021, pp. 3558–3568.

[196] M. Chen, Z. Wang, and F. Zheng, "Benchmarks for corruption invariant person re-identification," *arXiv:2111.00880*, 2021.

[197] S. Chen, Q. Jin, P. Wang, and Q. Wu, "Learning with noisy correspondence for cross-modal matching," in *NeurIPS*, 2021, pp. 29 406–29 419.

[198] J. W. Cho, D.-J. Kim, J. Choi, Y. Jung, and I. S. Kweon, "Dealing with missing modalities in the visual question answer-difference prediction task through knowledge distillation," in *CVPR*, 2021, pp. 1592–1601.

[199] M. Cross and V. Radu, "An immune inspired algorithm for fault tolerant enhanced multimodal machine learning," in *BIBM*, IEEE, 2021, pp. 2194–2200.

[200] R. Delgado-Escano, F. M. Castro, N. Guil, V. Kalogeiton, and M. J. Marin-Jimenez, "Multimodal gait recognition under missing modalities," in *ICIP*, IEEE, 2021, pp. 3003–3007.

[201] A. Gharahdaghi, F. Razzazi, and A. Amini, "A non-linear mapping representing human action recognition under missing modality problem in video data," *Measurement*, vol. 186, p. 110 123, 2021.

[202] F. Heidarivincheh, R. McConville, C. Morgan, *et al.*, "Multimodal classification of parkinson's disease in home environments with resiliency to missing modalities," *Sensors*, vol. 21, no. 12, p. 4133, 2021.

[203] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," in *Proc. 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 7514–7528.

[204] F. Huang, A. Jolfaei, and A. K. Bashir, "Robust multimodal representation learning with evolutionary adversarial attention networks," *IEEE Transactions on Evolutionary Computation*, vol. 25, no. 5, pp. 856–868, 2021.

[205] Y. Huang, H. Xue, B. Liu, and Y. Lu, "Unifying multimodal transformer for bi-directional image and text generation," in *Proc. ACM MM*, 2021, pp. 1138–1147.

[206] M. Islam, N. Wijethilake, and H. Ren, "Glioblastoma multiforme prognosis: Mri missing modality generation, segmentation and radiogenomic survival prediction," *Computerized Medical Imaging and Graphics*, vol. 91, p. 101 906, 2021.

[207] C. Jia, Y. Yang, Y. Xia, *et al.*, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021, pp. 4904–4916.

[208] D. Kiela, H. Firooz, A. Mohan, *et al.*, "The hateful memes challenge: Competition report," in *NeurIPS 2020 Competition and Demonstration Track*, PMLR, 2021, pp. 344–360.

[209] J.-H. Kim, D.-H. Kim, S. Yi, and T. Lee, "Semi-orthogonal embedding for efficient unsupervised anomaly segmentation," *arXiv preprint arXiv:2105.14737*, 2021.

[210] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International Conference on Machine Learning*, PMLR, 2021, pp. 5583–5594.

[211] J. Y. Koh, D. Fried, and R. Mottaghi, "Zero-shot image-to-text generation for visual-semantic arithmetic," in *CVPR*, 2021, pp. 5670–5679.

[212] M. A. Lee, M. Tan, Y. Zhu, and J. Bohg, "Detect, reject, correct: Crossmodal compensation of corrupted sensors," in *ICRA*, IEEE, 2021, pp. 909–916.

[213] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *NeurIPS*, vol. 34, 2021, pp. 9694–9705.

[214] P. P. Liang, Y. Lyu, X. Fan, *et al.*, "Multibench: Multiscale benchmarks for multimodal representation learning," *NeurIPS*, vol. 2021, no. DB1, p. 1, 2021.

[215] H. Luo, L. Ji, M. Zhong, and *et al.*, "Clip4clip: An empirical study of clip for end-to-end video clip retrieval," *arXiv:2104.08860*, 2021.

[216] F. Ma, S.-L. Huang, and L. Zhang, "An efficient approach for audio-visual emotion recognition with missing labels and missing modalities," in *2021 IEEE international conference on multimedia and Expo (ICME)*, IEEE, 2021, pp. 1–6.

[217] F. Ma, X. Xu, S.-L. Huang, and L. Zhang, "Maximum likelihood estimation for multimodal learning with missing modality," *arXiv:2108.10513*, 2021.

[218] H. Ma, Z. Han, C. Zhang, H. Fu, J. T. Zhou, and Q. Hu, "Trustworthy multimodal regression with mixture of normal-inverse gamma distributions," *NeurIPS*, vol. 34, pp. 6881–6893, 2021.

[219] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "Smil: Multimodal learning with severely missing modality," in *AAAI*, vol. 35, 2021, pp. 2302–2310.

[220] Z. Ma, F. Ma, B. Sun, and S. Li, "Hybrid mutimodal fusion for dimensional emotion recognition," in *Proc. 2nd on Multimodal Sentiment Analysis Challenge*, 2021, pp. 29–36.

[221] M. J. Marín-Jiménez, F. M. Castro, R. Delgado-Escaño, V. Kalogeiton, and N. Guil, "Ugaitnet: Multimodal gait recognition with missing input modalities," *IEEE TIFS*, vol. 16, pp. 5452–5462, 2021.

[222] A. Mesaros, T. Heittola, and T. Virtanen, "Detection and classification of acoustic scenes and events: Outcome of the dcase 2019 challenge," *IEEE/ACM TASLP*, vol. 29, pp. 2453–2471, 2021.

[223] S. Nawaz, M. S. Saeed, P. Morerio, *et al.*, "Cross-modal speaker verification and recognition: A multilingual perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1682–1691.

[224] M. Ni, H. Huang, L. Su, *et al.*, "M3p: Learning universal representations via multitask multilingual multimodal pre-training," in *CVPR*, 2021, pp. 3977–3986.

[225] H. Ning, X. Zheng, X. Lu, and Y. Yuan, "Disentangled representation learning for cross-modal biometric matching," *IEEE Transactions on Multimedia*, vol. 24, pp. 1763–1774, 2021.

[226] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 2887–2905, 2021.

[227] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[228] A. Radford, J. W. Kim, C. Hallacy, and *et al.*, "Learning transferable visual models from natural language supervision," *ICML*, pp. 8748–8763, 2021.

[229] A. Ramesh, M. Pavlov, G. Goh, *et al.*, "Zero-shot text-to-image generation," in *ICML*, 2021, pp. 8821–8831.

[230] E. Salesky, D. Etter, and M. Post, "Robust open-vocabulary translation from visual text representations," *arXiv preprint arXiv:2104.08211*, 2021.

[231] L. Sarı, K. Singh, J. Zhou, L. Torresani, N. Singhal, and Y. Saraf, "A multi-view approach to audio-visual speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6194–6198.

[232] H. Sato, T. Ochiai, K. Kinoshita, M. Delcroix, T. Nakatani, and S. Araki, "Multimodal attention fusion for target speaker extraction," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2021, pp. 778–784.

[233] B. Shi, W.-N. Hsu, and A. Mohamed, "Robust self-supervised audio-visual speech recognition," *arXiv:2201.01763*, 2021.

[234] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork, "Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning," *arXiv:2103.01913*, 2021.

[235] W. Sun, F. Ma, Y. Li, S.-L. Huang, S. Ni, and L. Zhang, "Semi-supervised multimodal image translation for missing modality imputation," in *ICASSP*, IEEE, 2021, pp. 4320–4324.

[236] A. Talmor, O. Yoran, D. Lahav, and *et al.*, "Multimodalqa: Complex question answering over text, tables and images," *arXiv:2104.06039*, 2021.

[237] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, and G. Carneiro, "Weakly-supervised video anomaly detection with robust temporal feature magnitude learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4975–4986.

[238] T. Wang, L. Zhang, and W. Hu, "Bridging deep and multiple kernel learning: A review," *Information Fusion*, vol. 67, pp. 3–13, 2021.

[239] Y. Wang, Y. Zhang, Y. Liu, *et al.*, "Acn: Adversarial co-training network for brain tumor segmentation with missing modalities," in *MICCAI*, Springer, 2021, pp. 410–420.

[240] B. Xin, J. Huang, Y. Zhou, J. Lu, and X. Wang, "Interpretation on deep multimodal fusion for diagnostic classification," in *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2021, pp. 1–8.

[241] K. Yang, W.-Y. Lin, M. Barman, F. Condessa, and Z. Kolter, "Defending multimodal fusion models against single-source adversaries," in *CVPR*, 2021, pp. 3340–3349.

[242] C. Yi, S. Yang, H. Li, Y. Tan, and A. C. Kot, "Benchmarking the robustness of spatial-temporal models against corruptions," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, J. Vanschoren and S. Yeung, Eds., 2021.

[243] L. Yuan, D. Chen, Y.-L. Chen, *et al.*, "Florence: A new foundation model for computer vision," *arXiv:2111.11432*, 2021.

[244] P. Zhang, X. Li, X. Hu, *et al.*, "Vinvl: Revisiting visual representations in vision-language models," in *CVPR*, 2021, pp. 5579–5588.

[245] W. Zhang, D. Xu, J. Zhang, and W. Ouyang, "Progressive modality cooperation for multi-modality domain adaptation," *IEEE Transactions on Image Processing*, vol. 30, pp. 3293–3306, 2021.

[246] J. Zhao, R. Li, and Q. Jin, "Missing modality imagination network for emotion recognition with uncertain missing modalities," in *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 2608–2618.

[247] A. Zheng, M. Hu, B. Jiang, Y. Huang, Y. Yan, and B. Luo, "Adversarial-metric learning for audio-visual cross-modal matching," *IEEE Transactions on Multimedia*, vol. 24, pp. 338–351, 2021.

[248] A. Zheng, Z. Wang, Z. Chen, C. Li, and J. Tang, "Robust multi-modality person re-identification," in *AAAI*, vol. 35, 2021, pp. 3529–3537.

[249] Z. Zheng, A. Ma, L. Zhang, and Y. Zhong, "Deep multisensor learning for missing-modality all-weather mapping," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 174, pp. 254–264, 2021.

[250] K. Zhou, C. Chen, B. Wang, M. R. U. Saputra, N. Trigoni, and A. Markham, "Vmloc: Variational fusion for learning-based multimodal camera localization," in *AAAI*, vol. 35, 2021, pp. 6165–6173.

[251] T. Zhou, S. Canu, P. Vera, and S. Ruan, "Feature-enhanced generation and multi-modality fusion based deep neural network for brain tumor segmentation with missing mr modalities," *Neurocomputing*, vol. 466, pp. 102–112, 2021.

[252] T. Zhou, S. Canu, P. Vera, and S. Ruan, "Latent correlation representation learning for brain tumor segmentation with missing mri modalities," *IEEE Transactions on Image Processing*, vol. 30, pp. 4263–4274, 2021.

[253] Y. Zhu, S. Wang, R. Lin, Y. Hu, and Q. Chen, "Brain tumor segmentation for missing modalities by supplementing missing features," in *ICCCBDA*, IEEE, 2021, pp. 652–656.

[254] J.-B. Alayrac, J. Donahue, P. Luc, *et al.*, "Flamingo: A visual language model for few-shot learning," *NeurIPS*, vol. 35, pp. 23 716–23 736, 2022.

[255] J. Ao, R. Wang, L. Zhou, *et al.*, "Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing," *Proc. 60th Annual Meeting of the Association for Computational Linguistics*, pp. 5723–5738, 2022.

[256] R. Azad, N. Khosravi, and D. Merhof, "Smu-net: Style matching u-net for brain tumor segmentation with missing modalities," in *International Conference on Medical Imaging with Deep Learning*, PMLR, 2022, pp. 48–62.

[257] H. Chi, M. Yang, J. Zhu, G. Wang, and G. Wang, "Missing modality meets meta sampling (m3s): An efficient universal approach for multimodal sentiment analysis with missing modality," *arXiv:2210.03428*, 2022.

[258] M. Hammad, L. Tawalbeh, A. M. Iliyasu, *et al.*, "Efficient multimodal deep-learning-based covid-19 diagnostic system for noisy and corrupted images," *Journal of King Saud University-Science*, vol. 34, no. 3, p. 101 898, 2022.

[259] W. Han, H. Chen, M.-Y. Kan, and S. Poria, "Mm-align: Learning optimal transport-based alignment dynamics for fast and accurate inference on missing modality sequences," *arXiv:2210.12798*, 2022.

[260] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022, pp. 16 000–16 009.

[261] Z. Huang, L. Lin, P. Cheng, L. Peng, and X. Tang, "Multi-modal brain tumor segmentation via missing modality synthesis and modality-level attention fusion," *arXiv:2203.04586*, 2022.

[262] H. Ikhlasse, D. Benjamin, C. Vincent, and M. Hicham, "Multimodal cloud resources utilization forecasting using a bidirectional gated recurrent unit predictor based on a power efficient stacked denoising autoencoders," *Alexandria Engineering Journal*, vol. 61, no. 12, pp. 11 565–11 577, 2022.

[263] S. Jeong, H. Cho, J. Kwon, and H. Park, "Region-of-interest attentive heteromodal variational encoder-decoder for segmentation with missing modalities," in *Proc. Asian Conference on Computer Vision*, 2022, pp. 3707–3723.

[264] V. John and Y. Kawanishi, "A multimodal sensor fusion framework robust to missing modalities for person recognition," in *ACM Multimedia*, 2022, pp. 1–5.

[265] A. Joshi, N. Gupta, J. Shah, B. Bhattarai, A. Modi, and D. Stoyanov, "Generalized product-of-experts for learning multimodal representations in noisy environments," in *Proc. ICMI*, 2022, pp. 83–93.

[266] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022, pp. 12 888–12 900.

[267] M. Li, S.-L. Huang, and L. Zhang, "A general framework for incomplete cross-modal retrieval with missing labels and missing modalities," in *ICASSP*, IEEE, 2022, pp. 4763–4767.

[268] Y. Li, A. W. Yu, T. Meng, *et al.*, "Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection," in *CVPR*, 2022, pp. 17 182–17 191.

[269] Y. Li, X. Zhang, Y. Wang, *et al.*, "M3ed: Multi-modal multi-scene emotional dataset for affective computing," *IEEE Transactions on Affective Computing*, 2022.

[270] H. Liu, Y. Fan, H. Li, *et al.*, "Moddrop++: A dynamic filter network with intra-subject co-training for multiple sclerosis lesion segmentation with missing modalities," in *MICCAI*, Springer, 2022, pp. 444–453.

[271] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?" In *CVPR*, 2022, pp. 18 177–18 186.

[272] M. F. Naeem, Y. Xian, L. V. Gool, and F. Tombari, "I2dformer: Learning image to document attention for zero-shot image classification," *NeurIPS*, vol. 35, pp. 12 283–12 294, 2022.

[273] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv:2204.06125*, 2022.

[274] P. Rust, J. F. Lotz, E. Bugliarello, E. Salesky, M. de Lhoneux, and D. Elliott, "Language modelling with pixels," *arXiv preprint arXiv:2207.06991*, 2022.

[275] M. S. Saeed, M. H. Khan, S. Nawaz, M. H. Yousaf, and A. Del Bue, "Fusion and orthogonal projection for improved face-voice association," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7057–7061.

[276] C. Schuhmann, R. Beaumont, C. Vencu, *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *arXiv:2210.08402*, 2022.

[277] C. Schuhmann, R. Beaumont, R. Vencu, and *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *NeurIPS*, vol. 35, pp. 25 278–25 294, 2022.

[278] S. Shankar, L. Thompson, and M. Fiterau, "Progressive fusion for multimodal integration," *arXiv:2209.00302*, 2022.

[279] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *arXiv:2201.02184*, 2022.

[280] B. Shi, W.-N. Hsu, and A. Mohamed, "Robust self-supervised audio-visual speech recognition," *arXiv:2201.01763*, 2022.

[281] N. Shvetsova, B. Chen, A. Rouditchenko, and *et al.*, "Everything at once – multi-modal fusion transformer for video retrieval," in *CVPR*, 2022, pp. 20 020–20 029.

[282] A. Singh, R. Hu, V. Goswami, *et al.*, "Flava: A foundational language and vision alignment model," in *CVPR*, 2022, pp. 15 638–15 650.

[283] T. Thrush, R. Jiang, M. Bartolo, *et al.*, "Winoground: Probing vision and language models for visio-linguistic compositionality," in *CVPR*, 2022, pp. 5238–5248.

[284] N. Wang, H. Cao, J. Zhao, R. Chen, D. Yan, and J. Zhang, "M2r2: Missing-modality robust emotion recognition framework with iterative data augmentation," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 5, pp. 1305–1316, 2022.

[285] S. Wang, A. Politis, A. Mesaros, and T. Virtanen, "Self-supervised learning of audio representations from audio-visual data using spatial alignment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1467–1479, 2022.

[286] Z. Wang, X. Liu, H. Li, and *et al.*, "Multi-modal learning with missing modality via shared-specific feature modelling," in *CVPR*, 2022, pp. 8875–8884.

[287] P. Wu, X. Liu, and J. Liu, "Weakly supervised audio-visual violence detection," *IEEE Transactions on Multimedia*, 2022.

[288] J. Yang, H. Duan, R. Socher, C. Xiong, L. Fei-Fei, and H. Wang, "Chinese clip: Contrastive vision-language pretraining in chinese," *arXiv:2211.01335*, 2022.

[289] Q. Yang, X. Guo, Z. Chen, P. Y. Woo, and Y. Yuan, "D 2-net: Dual disentanglement network for brain tumor segmentation with missing modalities," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2953–2964, 2022.

[290] J. Ye, J. Zhou, J. Tian, *et al.*, "Sentiment-aware multimodal pretraining for multimodal sentiment analysis," *Knowledge-Based Systems*, vol. 258, p. 110 021, 2022.

[291] J. Yu, J. Liu, Y. Cheng, R. Feng, and Y. Zhang, "Modality-aware contrastive instance learning with self-distillation for weakly-supervised audio-visual violence detection," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 6278–6287.

[292] L. Zaadnoordijk, T. R. Besold, and R. Cusack, "Lessons from infant learning for unsupervised machine learning," *Nature Machine Intelligence*, vol. 4, no. 6, pp. 510–520, 2022.

[293] M. Z. Zaheer, A. Mahmood, M. H. Khan, M. Segu, F. Yu, and S.-I. Lee, "Generative cooperative learning for unsupervised video anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14 744–14 754.

[294] M. Z. Zaheer, J.-H. Lee, A. Mahmood, M. Astrid, and S.-I. Lee, "Stabilizing adversarially learned one-class novelty detection using pseudo anomalies," *IEEE Transactions on Image Processing*, vol. 31, pp. 5963–5975, 2022.

[295] J. Zeng, T. Liu, and J. Zhou, "Tag-assisted multimodal sentiment analysis under uncertain missing modalities," in *Proc. 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 1545–1554.

[296] J. Zeng, J. Zhou, and T. Liu, "Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities," in *Proc. 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 2924–2934.

[297] J. Zeng, J. Zhou, and T. Liu, "Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities," *IEEE Transactions on Multimedia*, vol. 25, pp. 6301–6314, 2022.

[298] C. Zhang, X. Chu, L. Ma, *et al.*, "M3care: Learning with missing modalities in multimodal healthcare data," in *Proc. 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 2418–2428.

[299] Q. Zhang, C. Lai, J. Liu, N. Huang, and J. Han, "Fmcnet: Feature-level modality compensation for visible-infrared person re-identification," in *CVPR*, 2022, pp. 7349–7358.

[300] X. Zhang, Q. Song, and G. Liu, "Multimodal image aesthetic prediction with missing modality," *Mathematics*, vol. 10, no. 13, p. 2312, 2022.

[301] Y. Zhang, N. He, J. Yang, *et al.*, "Mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation," in *MICCAI*, Springer, 2022, pp. 107–117.

[302] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Machine Learning for Healthcare Conference*, 2022, pp. 2–25.

[303] Z. Zhao, H. Yang, and J. Sun, "Modality-adaptive feature interaction for brain tumor segmentation with missing modalities," in *MICCAI*, Springer, 2022, pp. 183–192.

[304] D. Altinses and A. Schwung, "Deep multimodal fusion with corrupted spatio-temporal data using fuzzy regularization," in *IECON*, IEEE, 2023, pp. 1–7.

[305] D. Altinses and A. Schwung, "Multimodal synthetic dataset balancing: A framework for realistic and balanced training data generation in industrial settings," in *IECON*, IEEE, 2023, pp. 1–7.

[306] B. van Amsterdam, A. Kadkhodamohammadi, I. Luengo, and D. Stoyanov, "Aspnet: Action segmentation with shared-private representation of multiple data sources," in *CVPR*, 2023, pp. 2384–2393.

[307] H. Bansal, N. Singhi, Y. Yang, F. Yin, A. Grover, and K.-W. Chang, "Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning," in *IEEE/CVF ICCV*, 2023, pp. 112–123.

[308] G. Bao, Q. Zhang, D. Miao, *et al.*, "Multimodal federated learning with missing modality via prototype mask and contrast," *arXiv:2312.13508*, 2023.

[309] M. Chen, J. Yao, L. Xing, Y. Wang, Y. Zhang, and Y. Wang, "Redundancy-adaptive multimodal learning for imperfect data," *arXiv:2310.14496*, 2023.

[310] Y. Chen, Y. Pan, Y. Xia, and Y. Yuan, "Disentangle first, then distill: A unified framework for missing modality imputation and alzheimer's disease diagnosis," *IEEE Transactions on Medical Imaging*, 2023.

[311] Z. Chen, L. Guo, Y. Fang, *et al.*, "Rethinking uncertainly missing and ambiguous visual modality in multi-modal entity alignment," in *International Semantic Web Conference*, Springer, 2023, pp. 121–139.

[312] M. Cho, M. Kim, S. Hwang, C. Park, K. Lee, and S. Lee, "Look around for anomalies: Weakly-supervised anomaly detection via context-motion relational learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 12 137–12 146.

[313] D. Cohen, I. Rosenberger, M. Butman, and K. Bar, "Masking important information to assess the robustness of a multimodal classifier for emotion recognition," *Frontiers in Artificial Intelligence*, vol. 6, p. 1 091 443, 2023.

[314] W. Dai, J. Li, D. Li, *et al.*, "Instructblip: Towards general-purpose vision-language models with instruction tuning," *arXiv:2305.06500*, 2023.

[315] K. Darvish, E. Raff, F. Ferraro, C. Matuszek, *et al.*, "Multimodal language learning for object retrieval in low data regimes in the face of missing modalities," *Transactions on Machine Learning Research*, 2023.

[316] Z. Ding, G. Lan, Y. Song, and Z. Yang, "Sgir: Star graph-based interaction for efficient and robust multimodal representation," *IEEE Transactions on Multimedia*, vol. 26, pp. 4217–4229, 2023.

[317] Y. Dong, C. Kang, J. Zhang, *et al.*, "Benchmarking robustness of 3d object detection to common corruptions," in *CVPR*, Jun. 2023, pp. 1022–1032.

[318] T. Feng, D. Bose, T. Zhang, *et al.*, "Fedmultimodal: A benchmark for multimodal federated learning," in *Proc. 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 4035–4045.

[319] Y. Fu, J. Ding, and X. Xu, "Low-rank multimanifold embedding learning for multimode process monitoring," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 3, pp. 3468–3477, 2023.

[320] R. Gupta, N. Akhtar, A. Mian, and M. Shah, "Contrastive self-supervised learning leads to higher adversarial susceptibility," in *AAAI*, vol. 37, 2023, pp. 14 838–14 846.

[321] P. Hager, M. J. Menten, and D. Rueckert, "Best of both worlds: Multimodal contrastive learning with tabular and imaging data," in *CVPR*, 2023, pp. 23 924–23 935.

[322] J. Hong, M. Kim, J. Choi, and Y. M. Ro, "Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring," in *CVPR*, 2023, pp. 18 783–18 794.

[323] R. Huan, G. Zhong, P. Chen, and R. Liang, "Unimf: A unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences," *IEEE Transactions on Multimedia*, 2023.

[324] S. Jabeen, X. Li, M. S. Amin, O. Bourahla, S. Li, and A. Jabbar, "A review on methods and applications in multimodal deep learning," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 2s, pp. 1–41, 2023.

[325] T. Jin, X. Cheng, L. Li, W. Lin, Y. Wang, and Z. Zhao, "Rethinking missing modality learning from a decoding perspective," in *ACM Multimedia*, 2023, pp. 4431–4439.

[326] V. John and Y. Kawanishi, "Audio-visual sensor fusion framework using person attributes robust to missing visual modality for person recognition," in *International Conference on Multimedia Modeling*, Springer, 2023, pp. 523–535.

[327] A. Josi, M. Alehdaghi, R. M. O. Cruz, and E. Granger, "Multimodal data augmentation for visual-infrared person reid with corrupted data," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, Jan. 2023, pp. 32–41.

[328] A. Josi, M. Alehdaghi, R. M. Cruz, and E. Granger, "Multimodal data augmentation for visual-infrared person reid with corrupted data," in *WACV*, 2023, pp. 32–41.

[329] L. Kong, Y. Liu, X. Li, *et al.*, "Robo3d: Towards robust and reliable 3d perception against corruptions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 994–20 006.

[330] A. Konwer, C. Chen, and P. Prasanna, "Magnet: Modality-agnostic network for brain tumor segmentation and characterization with missing modalities," in *International Workshop on Machine Learning in Medical Imaging*, Springer, 2023, pp. 361–371.

[331] A. Konwer, X. Hu, J. Bae, X. Xu, C. Chen, and P. Prasanna, "Enhancing modality-agnostic representations via meta-learning for brain tumor segmentation," in *IEEE/CVF ICCV*, 2023, pp. 21 415–21 425.

[332] H. Laurençon, L. Saulnier, L. Tronchon, *et al.*, "Obelics: An open web-scale filtered dataset of interleaved image-text documents," *arXiv:2306.16527*, 2023.

[333] C.-Y. Lee, C.-L. Li, H. Zhang, *et al.*, "Formnetv2: Multimodal graph contrastive learning for form document information extraction," *arXiv:2305.02549*, 2023.

[334] Y.-L. Lee, Y.-H. Tsai, W.-C. Chiu, and C.-Y. Lee, "Multimodal prompting with missing modalities for visual recognition," in *CVPR*, 2023, pp. 14 943–14 952.

[335] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*, 2023, pp. 19 730–19 742.

[336] S. Li, C. Du, Y. Zhao, Y. Huang, and H. Zhao, "What makes for robust multi-modal models in the face of missing modalities?" *arXiv:2310.06383*, 2023.

[337] X. Li and J. Li, "Angle-optimized text embeddings," *arXiv preprint arXiv:2309.12871*, 2023.

[338] R. Lin and H. Hu, "Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1686–1702, 2023.

[339] Y.-B. Lin, Y.-H. Tsai, K.-W. Cheng, and H.-Y. Lee, "Lavish: Language-vision instruction tuning with home videos," *arXiv:2305.14837*, 2023.

[340] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023.

[341] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *arXiv:2304.08485*, 2023.

[342] H. Liu, D. Wei, D. Lu, J. Sun, L. Wang, and Y. Zheng, "M3ae: Multimodal representation learning for brain tumor segmentation with missing modalities," in *AAAI*, vol. 37, 2023, pp. 1657–1665.

[343] J. Liu, D. Capurro, A. Nguyen, and K. Verspoor, "Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities," *Journal of Biomedical Informatics*, vol. 145, p. 104 466, 2023.

[344] Y. Liu, D. Yang, Y. Wang, *et al.*, "Generalized video anomaly event detection: Systematic taxonomy and comparison of deep models," *ACM Computing Surveys*, 2023.

[345] W. Luo, M. Xu, and H. Lai, "Multimodal reconstruct and align net for missing modality problem in sentiment analysis," in *International Conference on Multimedia Modeling*, Springer, 2023, pp. 411–422.

[346] H. Ma, Q. Zhang, C. Zhang, *et al.*, "Calibrating multimodal learning," in *ICML*, PMLR, 2023, pp. 23 429–23 450.

[347] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," *arXiv:2306.05424*, 2023.

[348] B. McKinzie, V. Shankar, J. Y. Cheng, Y. Yang, J. Shlens, and A. T. Toshev, "Robustness in multimodal learning under train-test modality mismatch," in *ICML*, PMLR, 2023, pp. 24 291–24 303.

[349] X. Ning, X. Wang, S. Xu, *et al.*, "A review of research on co-training," *Concurrency and computation: practice and experience*, vol. 35, no. 18, e6276, 2023.

[350] OpenAI, "Gpt-4 technical report," *arXiv:2303.08774*, 2023.

[351] Y. Park, S. Woo, S. Lee, M. A. Nugroho, and C. Kim, "Cross-modal alignment and translation for missing modality action recognition," *Computer Vision and Image Understanding*, vol. 236, p. 103 805, 2023.

[352] A. Rahate, S. Mandaokar, P. Chandel, R. Walambe, S. Ramanna, and K. Kotecha, "Employing multimodal co-learning to evaluate the robustness of sensor fusion for industry 5.0 tasks," *Soft Computing*, vol. 27, no. 7, pp. 4139–4155, 2023.

[353] M. S. Saeed, S. Nawaz, M. H. Khan, *et al.*, "Single-branch network for multimodal training," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[354] A. Sikdar, J. Teotia, and S. Sundaram, "Contrastive learning-based spectral knowledge distillation for multi-modality and missing modality scenarios in semantic segmentation," *arXiv:2312.02240*, 2023.

[355] J. Sun, H. Zheng, Q. Zhang, A. Prakash, Z. M. Mao, and C. Xiao, "Calico: Self-supervised camera-lidar contrastive pre-training for bev perception," *arXiv:2306.00349*, 2023.

[356] Y. Tian, G. Jian, J. Wang, *et al.*, "A revised approach to orthodontic treatment monitoring from oralscan video," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 12, pp. 5827–5836, 2023.

[357] M. Tschannen, B. Mustafa, and N. Houlsby, "Clippo: Image-and-language understanding from pixels only," in *CVPR*, 2023, pp. 11 006–11 017.

[358] J. Vazquez-Rodriguez, G. Lefebvre, J. Cumin, and J. L. Crowley, "Accommodating missing modalities in time-continuous multimodal emotion recognition," in *ACII*, IEEE, 2023, pp. 1–8.

[359] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro, "Multi-modal learning with missing modality via shared-specific feature modelling," in *CVPR*, 2023, pp. 15 878–15 887.

[360] H. Wang, C. Ma, J. Zhang, *et al.*, "Learnable cross-modal knowledge distillation for multi-modal learning with missing modality," in *MICCAI*, Springer, 2023, pp. 216–226.

[361] J. Wang, L. Mou, L. Ma, T. Huang, and W. Gao, "Amsa: Adaptive multimodal learning for sentiment analysis," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 3s, pp. 1–21, 2023.

[362] J. Wang, G. Wang, X. Zhang, *et al.*, "Patch: A plug-in framework of non-blocking inference for distributed multimodal system," *Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 3, pp. 1–24, 2023.

[363] S. Wang, Z. Yan, D. Zhang, H. Wei, Z. Li, and R. Li, "Prototype knowledge distillation for medical segmentation with missing modality," in *ICASSP*, IEEE, 2023, pp. 1–5.

[364] S. Wei, Y. Luo, X. Ma, P. Ren, and C. Luo, "Msh-net: Modality-shared hallucination with joint adaptation distillation for remote sensing image classification using missing modalities," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[365] S. Woo, S. Lee, Y. Park, M. A. Nugroho, and C. Kim, "Towards good practices for missing modality robust action recognition," in *AAAI*, vol. 37, 2023, pp. 2776–2784.

[366] M. K. Wozniak, V. Kårefj¨ard, M. Thiel, and P. Jensfelt, "Toward a robust sensor fusion step for 3d object detection on corrupted data," *IEEE Robotics and automation letters*, vol. 8, no. 11, pp. 7018–7025, 2023.

[367] S. Xie, L. Kong, W. Zhang, *et al.*, "Benchmarking bird's eye view detection robustness to real-world corruptions," in *International Conference on Learning Representations 2023 Workshop on Scene Representations for Autonomous Driving*, 2023.

[368] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE TPAMI*, vol. 45, no. 10, pp. 12 113–12 132, 2023.

[369] H. Yang, J. Sun, and Z. Xu, "Learning unified hyper-network for multi-modal mr image synthesis and tumor segmentation with missing modalities," *IEEE Transactions on Medical Imaging*, 2023.

[370] K. Yu, T. Tao, H. Xie, *et al.*, "Benchmarking the robustness of lidar-camera fusion for 3d object detection," in *CVPR*, 2023, pp. 3188–3198.

[371] T. Yue, Y. Li, J. Qin, and Z. Hu, "Multi-modal hierarchical fusion network for fine-grained paper classification," *Multimedia Tools and Applications*, pp. 1–17, 2023.

[372] P.-F. Zhang and Z. H. Huang, "Multi-head siamese prototype learning against both data and label corruption," in *ACM Multimedia*, 2023, pp. 1–7.

[373] X. Zheng, C. Tang, Z. Wan, C. Hu, and W. Zhang, "Multi-level confidence learning for trustworthy multimodal classification," in *AAAI*, vol. 37, 2023, pp. 11 381–11 389.

[374] Q. Zhou, H. Zou, F. Luo, and Y. Qiu, "Rhvit: A robust hierarchical transformer for 3d multimodal brain tumor segmentation using biased masked image modeling pre-training," in *BIBM*, IEEE, 2023, pp. 1784–1791.

[375] T. Zhou, "Feature fusion and latent feature learning guided brain tumor segmentation and missing modality recovery network," *Pattern Recognition*, vol. 141, p. 109 665, 2023.

[376] Z. Zhou, S. Hu, M. Li, H. Zhang, Y. Zhang, and H. Jin, "Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning," in *ACM Multimedia*, 2023, pp. 6311–6320.

[377] Y. Zhu, X. Sun, and X. Zhou, "Exploiting multi-modal fusion for robust face representation learning with missing modality," in *International Conference on Artificial Neural Networks*, Springer, 2023, pp. 283–294.

[378] H. Zuo, R. Liu, J. Zhao, G. Gao, and H. Li, "Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities," in *ICASSP*, IEEE, 2023, pp. 1–5.

[379] Z. Bai, P. Wang, T. Xiao, *et al.*, "Hallucination of multimodal large language models: A survey," *arXiv:2404.18930*, 2024.

[380] T. Beemelmanns, Q. Zhang, C. Geller, and L. Eckstein, "Multicorrupt: A multi-modal robustness dataset and benchmark of lidar-camera fusion for 3d object detection," in *IVS*, IEEE, 2024, pp. 3255–3261.

[381] W. d. L. Costa, R. Ismayilov, N. Strisciuglio, and E. T. Martinez, "Indoor scene recognition from images under visual corruptions," *arXiv:2408.13029*, 2024.

[382] C. Cui, Y. Ma, X. Cao, *et al.*, "A survey on multimodal large language models for autonomous driving," in *WACV*, 2024, pp. 958–979.

[383] C. Dixit and S. M. Satapathy, "Deep cnn with late fusion for real time multimodal emotion recognition," *Expert Systems with Applications*, 2024.

[384] S. Du, S. Zheng, Y. Wang, W. Bai, D. P. O'Regan, and C. Qin, "Tip: Tabular-image pre-training for multimodal classification with incomplete data," in *ECCV*, Springer, 2024, pp. 478–496.

[385] C. Gan, Y. Tu, Y. Li, and W. Lin, "Dac: 2d-3d retrieval with noisy labels via divide-and-conquer alignment and correction," in *ACM Multimedia*, 2024, pp. 4217–4226.

[386] C. Ganhör, M. Moscati, A. Hausberger, S. Nawaz, and M. Schedl, "A multimodal single-branch embedding network for recommendation in cold-start and missing modality scenarios," in *Proc. 18th ACM Conference on Recommender Systems*, 2024, pp. 380–390.

[387] A. Ghadiya, P. Kar, V. Chudasama, and P. Wasnik, "Cross-modal fusion and attention mechanism for weakly supervised video anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1965–1974.

[388] S. Hore and T. Bhattacharya, "Audio-visual expression-based emotion recognition model for neglected people in real-time: A late-fusion approach," *Multimedia Tools and Applications*, pp. 1–39, 2024.

[389] J. Jang, Y. Wang, and C. Kim, "Towards robust multimodal prompting with missing modalities," in *ICASSP*, IEEE, 2024, pp. 8070–8074.

[390] D. Z. Kaplan, A. Roger, M. Osman, and I. Rish, "The effect of data corruption on multimodal long form responses," in *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

[391] J. Y. Koh, D. Fried, and R. R. Salakhutdinov, "Generating images with multimodal language models," *NeurIPS*, vol. 36, 2024.

[392] A. Al-Lahham, N. Tastan, M. Z. Zaheer, and K. Nandakumar, "A coarse-to-fine pseudo-labeling (c2fpl) framework for unsupervised video anomaly detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 6793–6802.

[393] A. Al-lahham, M. Z. Zaheer, N. Tastan, and K. Nandakumar, "Collaborative learning of anomalies with privacy (clap) for unsupervised video anomaly detection: A new baseline," *arXiv preprint arXiv:2404.00847*, 2024.

[394] J. Lei and F. Pernkopf, "Two-level test-time adaptation in multimodal learning," in *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

[395] J. Li, L. Li, R. Sun, G. Yuan, S. Wang, and S. Sun, "Mman-m2: Multiple multi-head attentions network based on encoder with missing modalities," *Pattern Recognition Letters*, vol. 177, pp. 110–120, 2024.

[396] Y. Li, M. E. H. Daho, P.-H. Conze, *et al.*, "A review of deep learning-based information fusion techniques for multimodal medical image classification," *Computers in Biology and Medicine*, p. 108 635, 2024.

[397] Z. Li, Y. Zhang, H. Li, Y. Chai, and Y. Yang, "Deformation-aware and reconstruction-driven multimodal representation learning for brain tumor segmentation with missing modalities," *BSPC*, vol. 91, p. 106 012, 2024.

[398] P. P. Liang, A. Zadeh, and L.-P. Morency, "Foundations & trends in multimodal machine learning: Principles, challenges, and open questions," *ACM Computing Surveys*, vol. 56, no. 10, pp. 1–42, 2024.

[399] Y. Liang, "Multimodal knowledge graph embedding with missing data integration," *IEEE Transactions on Computational Social Systems*, 2024.

[400] Q. Liu and W. Ma, "Navigating data corruption in machine learning: Balancing quality, quantity, and imputation strategies," *arXiv:2412.18296*, 2024.

[401] Z. Liu, B. Zhou, D. Chu, Y. Sun, and L. Meng, "Modality translation-based multimodal sentiment analysis under uncertain missing modalities," *Information Fusion*, vol. 101, p. 101 973, 2024.

[402] Y. Ma, X. Liu, Y. Wei, Z. Tao, X. Wang, and T.-S. Chua, "Leveraging multimodal features and item-level user feedback for bundle construction," in *Proc. 17th ACM International Conference on Web Search and Data Mining*, 2024, pp. 510–519.

[403] H. Maheshwari, Y.-C. Liu, and Z. Kira, "Missing modality robustness in semi-supervised multi-modal semantic segmentation," in *WACV*, 2024, pp. 1020–1030.

[404] D. Malitesta, E. Rossi, C. Pomo, T. Di Noia, and F. D. Malliaros, "Do we really need to drop items with missing modalities in multimodal recommendation?" In *Proc. CIKM*, 2024, pp. 3943–3948.

[405] I. E. Marouf, E. Tartaglione, and S. Lathuilière, "Mini but mighty: Finetuning vits with mini adapters," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 1732–1741.

[406] T. Mota, M. R. Verdelho, D. J. Araújo, A. Bissoto, C. Santiago, and C. Barata, "Mmist-ccrcc: A real world medical dataset for the development of multi-modal systems," in *CVPR*, 2024, pp. 2395–2403.

[407] C. Qiu, Y. Song, Y. Liu, *et al.*, "Mmmvit: Multiscale multimodal vision transformer for brain tumor segmentation with missing modalities," *BSPC*, vol. 90, p. 105 827, 2024.

[408] M. K. Reza, A. Prater-Bennette, and M. S. Asif, "Robust multimodal learning with missing modalities via parameter-efficient adaptation," *IEEE TPAMI*, 2024.

[409] W. Shi, W. Qin, and Z. Yun, "Learning rich multimodal representation for robust land cover classification in fog," *IEEE Sensors Journal*, 2024.

[410] D. Song, S. Lai, S. Chen, L. Sun, and B. Wang, "Both text and images leaked! a systematic analysis of multimodal llm data contamination," *arXiv:2411.03823*, 2024.

[411] R. Sun, L. Chen, L. Zhang, R. Xie, and J. Gao, "Robust visible-infrared person re-identification based on polymorphic mask and wavelet graph convolutional network," *IEEE TIFS*, vol. 19, pp. 2800–2813, 2024.

[412] B. Theodorou, L. Glass, C. Xiao, and J. Sun, "Framm: Fair ranking with missing modalities for clinical trial site selection," *Patterns*, vol. 5, no. 3, 2024.

[413] S. Wang, H. Caesar, L. Nan, and J. F. Kooij, "Unibev: Multi-modal 3d object detection with uniform bev encoders for robustness against missing sensor modalities," in *IVS*, IEEE, 2024, pp. 2776–2783.

[414] F. Wu, S. Liu, B. Guo, X. Li, Y. Gao, and Z. Yu, "Adaflow: Non-blocking inference with heterogeneous multi-modal mobile sensor data," in *CSCAIoT*, IEEE, 2024, pp. 8–9.

[415] R. Wu, H. Wang, H.-T. Chen, and G. Carneiro, "Deep multimodal learning with missing modality: A survey," *arXiv:2409.07825*, 2024.

[416] R. Wu, H. Wang, F. Dayoub, and H.-T. Chen, "Segment beyond view: Handling partially missing modality for audio-visual semantic segmentation," in *AAAI*, vol. 38, 2024, pp. 6100–6108.

[417] S. Xaviar, X. Yang, and O. Ardakanian, "Centaur: Robust multimodal fusion for human activity recognition," *IEEE Sensors Journal*, 2024.

[418] R. Xi, N. Huang, C. Lai, Q. Zhang, and J. Han, "Fmcnet +: Feature-level modality compensation for visible-infrared person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[419] S. Xie, J. Wang, H. He, *et al.*, "Tvdiag: A task-oriented and view-invariant failure diagnosis framework with multimodal data," *arXiv:2407.19711*, 2024.

[420] B. Yang, F. Liu, Y. Zou, X. Wu, Y. Wang, and D. A. Clifton, "Zeronlg: Aligning and autoencoding domains for zero-shot multimodal and multilingual natural language generation," *IEEE TPAMI*, 2024.

[421] G. Yu, L. Wang, Q. Chen, *et al.*, "Penta-encoder with medical transformer for incomplete multimodal learning of brain tumor segmentation," in *2024 IEEE 17th International Conference on Signal Processing (ICSP)*, IEEE, 2024, pp. 642–646.

[422] J. Zhang, T. Pang, C. Du, Y. Ren, B. Li, and M. Lin, "Benchmarking large multimodal models against common corruptions," *arXiv:2401.11943*, 2024.

[423] Q. Zhang, Y. Wei, Z. Han, *et al.*, "Multimodal fusion on low-quality data: A comprehensive survey," *arXiv:2404.18947*, 2024.

[424] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, and X. Zhao, "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects," *Expert Systems with Applications*, vol. 237, p. 121 692, 2024.

[425] Y. Zhang, C. Peng, Q. Wang, D. Song, K. Li, and S. K. Zhou, "Unified multi-modal image synthesis for missing modality imputation," *IEEE Transactions on Medical Imaging*, 2024.

[426] X. Zhou, X. Peng, H. Wen, *et al.*, "Learning weakly supervised audio-visual violence detection in hyperbolic space," *Image and Vision Computing*, vol. 151, p. 105 286, 2024.

[427] L. Bai, B. Cui, L. Wang, *et al.*, "V 2-sfmlearner: Learning monocular depth and ego-motion for multimodal wireless capsule endoscopy," *IEEE T-ASE*, 2025.

[428] S. Dhivyaa, D.-P. Dao, H.-J. Yang, and J. Kim, "Adaptive cross-modal representation learning for heterogeneous data types in alzheimer disease progression prediction with missing time point and modalities," in *ICPR*, Springer, 2025, pp. 267–282.

[429] Y. Diao, H. Fang, H. Yu, F. Li, and Y. Xu, "Multimodal invariant feature prompt network for brain tumor segmentation with missing modalities," *Neurocomputing*, vol. 616, p. 128 847, 2025.

[430] C. S. Diaz, D.-T. Vu, J. Bodelet, *et al.*, "Optimus: Predicting multivariate outcomes in alzheimer's disease using multi-modal data amidst missing values," *arXiv:2503.11282*, 2025.

[431] J. Fan, M. Qi, L. Liu, and H. Ma, "Diffusion-driven incomplete multimodal learning for air quality prediction," *ACM Transactions on Internet of Things*, vol. 6, no. 1, pp. 1–24, 2025.

[432] T. Feng, T. Zhang, S. Avestimehr, and S. Narayanan, "Modalitymirror: Enhancing audio classification in modality heterogeneity federated learning via multimodal distillation," in *Proc. 35th Workshop on Network and Operating System Support for Digital Audio and Video*, 2025, pp. 78–83.

[433] F. Fu, W. Ai, F. Yang, Y. Shou, T. Meng, and K. Li, "Sdr-gnn: Spectral domain reconstruction graph neural network for incomplete multimodal learning in conversational emotion recognition," *Knowledge-Based Systems*, vol. 309, p. 112 825, 2025.

[434] Y. Gou, H. Yang, Z. Liu, *et al.*, "Corrupted but not broken: Rethinking the impact of corrupted data in visual instruction tuning," *arXiv:2502.12635*, 2025.

[435] Z. Guo and T. Jin, "Smoothing the shift: Towards stable test-time adaptation under complex multimodal noises," *arXiv:2503.02616*, 2025.

[436] Z. Guo, S. Wang, W. Lin, W. Yan, Y. Wu, and T. Jin, "Efficient prompting for continual adaptation to missing modalities," *arXiv:2503.00528*, 2025.

[437] X. Hao, G. Liu, Y. Zhao, *et al.*, "Msc-bench: Benchmarking and analyzing multi-sensor corruption for driving perception," *arXiv:2501.01037*, 2025.

[438] M. F. Ishmam, I. Tashdeed, T. A. Saadat, M. H. Ashmafee, A. R. M. Kamal, and M. A. Hossain, "Visual robustness benchmark for visual question answering (vqa)," in *WACV*, IEEE, 2025, pp. 6623–6633.

[439] V. John and Y. Kawanishi, "Multimodal cascaded framework with multimodal latent loss functions robust to missing modalities," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025.

[440] A. Josi, M. Alehdaghi, R. M. Cruz, and E. Granger, "Fusion for visual-infrared person reid in real-world surveillance using corrupted multimodal data," *International Journal of Computer Vision*, pp. 1–22, 2025.

[441] G. Ke, S. He, X. L. Wang, *et al.*, "Knowledge bridger: Towards training-free missing multi-modality completion," *arXiv:2502.19834*, 2025.

[442] A. Kebaili, J. Lapuyade-Lahorgue, P. Vera, and S. Ruan, "Amm-diff: Adaptive multi-modality diffusion network for missing modality imputation," *arXiv:2501.12840*, 2025.

[443] H.-D. Kieu, Q.-T. Ha, X.-H. Phan, D.-T. Le, *et al.*, "Mi-cga: Cross-modal graph attention network for robust emotion recognition in the presence of incomplete modalities," *Neurocomputing*, vol. 623, p. 129 342, 2025.

[444] J. Kim, H. Kang, S. Kim, K. Kim, and C. Park, "Disentangling and generating modalities for recommendation in missing modality scenarios," *arXiv:2504.16352*, 2025.

[445] S. Kim, S. Cho, S. Bae, K. Jang, and S.-Y. Yun, "Multi-task corrupted prediction for learning robust audio-visual speech representation," in *The Thirteenth International Conference on Learning Representations*, 2025.

[446] J. Lang, Z. Cheng, T. Zhong, and F. Zhou, "Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning," *arXiv:2501.01120*, 2025.

[447] H. Q. Le, C. M. Thwal, Y. Qiao, *et al.*, "Cross-modal prototype based multimodal federated learning under severely missing modality," *Information Fusion*, p. 103 219, 2025.

[448] C. Li, J. Zhang, Z. Zhang, *et al.*, "R-bench: Are your large multimodal model robust to real-world corruptions?" *IEEE Journal of Selected Topics in Signal Processing*, 2025.

[449] J. Li, M. A. Akbar, S. H. Shah, Z. Wang, and J. Yang, "Deep learning-driven behavioral modeling in iost for mental health monitoring and intervention," *IEEE Transactions on Computational Social Systems*, 2025.

[450] M. I. Liaqat, S. Nawaz, M. Z. Zaheer, *et al.*, "Chameleon: A multimodal learning framework robust to missing modalities," *International Journal of Multimedia Information Retrieval*, vol. 14, no. 2, p. 21, 2025.

[451] C. Liu, Z. Huang, Z. Chen, *et al.*, "Incomplete modality disentangled representation for ophthalmic disease grading and diagnosis," in *AAAI*, vol. 39, 2025, pp. 5361–5369.

[452] T. Liu, Y. Hu, M. Li, *et al.*, "Tackling real-world complexity: Hierarchical modeling and dynamic prompting for multimodal long document classification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[453] Y. Liu, C. Wang, and X. Yuan, "Fedmobile: Enabling knowledge contribution-aware multi-modal federated learning with incomplete modalities," in *Proc. ACM on Web Conference 2025*, 2025, pp. 2775–2786.

[454] Z. Pan, J. Xu, S. Jiang, and J. Wang, "Ssfd-net: Shared-specific feature disentanglement network for multimodal biometric recognition with missing modality," *Digital Signal Processing*, vol. 159, p. 105 003, 2025.

[455] V. Pipoli, F. Bolelli, S. Sarto, *et al.*, "Semantically conditioned prompts for visual recognition under missing modality scenarios," in *WACV*, 2025.

[456] M. Ramazanova, A. Pardo, B. Ghanem, and M. Alfarra, "Test-time adaptation for combating missing modalities in egocentric videos," in *The Thirteenth International Conference on Learning Representations*, 2025.

[457] M. K. Reza, A. Patil, M. Solh, and M. S. Asif, "Robust multimodal learning via cross-modal proxy tokens," *arXiv:2501.17823*, 2025.

[458] P. Shi, M. Hu, S. Nakagawa, X. Zheng, X. Shi, and F. Ren, "Text-guided reconstruction network for sentiment analysis with uncertain missing modalities," *IEEE Transactions on Affective Computing*, 2025.

[459] A. Sikdar, J. Teotia, and S. Sundaram, "Ogp-net: Optical guidance meets pixel-level contrastive distillation for robust multimodal and missing modality segmentation," in *AAAI*, vol. 39, 2025, pp. 6922–6930.

[460] M. Usama, S. A. Asim, S. B. Ali, S. T. Wasim, and U. B. Mansoor, "Analysing the robustness of vision-language-models to common corruptions," *arXiv:2504.13690*, 2025.

[461] X. Wang, L. Wang, Y. Yu, and X. Jiao, "Modality-invariant bidirectional temporal representation distillation network for missing multimodal sentiment analysis," *arXiv:2501.05474*, 2025.

[462] J. Wu, K. Yang, L. Kaplan, and M. Srivastava, "Admn: A layer-wise adaptive multimodal network for dynamic input noise and compute resources," *arXiv:2502.07862*, 2025.

[463] W. Xiong, T. Wang, X. Chen, *et al.*, "Disentanglement and codebook learning-induced feature match network to diagnose neurodegenerative diseases on incomplete multimodal data," *Pattern Recognition*, vol. 165, p. 111 597, 2025.

[464] T. Yang, Q. Zhou, H. Zou, H. Jiang, and Y. Wang, "Uml: A unified multimodal learning framework for cataract postoperative visual acuity prediction with uncertain missing modalities," in *ICASSP*, IEEE, 2025, pp. 1–5.

[465] Z. Yang, R. Jiang, X. Fu, W. Xi, and J. Zhao, "Open-modality latent modality interaction maximization for audio-visual learning," in *ICASSP*, IEEE, 2025, pp. 1–5.

[466] Y. Yuan, Z. Li, and B. Zhao, "A survey of multimodal learning: Methods, applications, and future," *ACM Computing Surveys*, 2025.

[467] X. Yue, Y. Chen, X. Zhang, *et al.*, "Pal: Prompting analytic learning with missing modality for multi-modal class-incremental learning," *arXiv:2501.09352*, 2025.

[468] T. Zhang, B. Song, Z. Zhang, and Y. Zhang, "Multimodal sentiment analysis based on multi-stage graph fusion networks under random missing modality conditions," *IET Image Processing*, vol. 19, no. 1, e13310, 2025.

[469] T. Zhang, S. Shen, and C. Chen, "Micinet: Multi-level inter-class confusing information removal for reliable multimodal classification," *arXiv:2502.19674*, 2025.

[470] Y. Zhang, D. Xie, D. Luo, and B. Sun, "Emotional boundaries and intensity aware model for incomplete multimodal sentiment analysis," *Digital Signal Processing*, p. 105 023, 2025.

[471] W. Zhao, K. Yang, P. Ding, C. Na, and W. Li, "Graph attention contrastive learning with missing modality for multimodal recommendation," *Knowledge-Based Systems*, p. 113 035, 2025.

[472] Z. Zhi, Y. Sun, *et al.*, "Wasserstein modality alignment makes your multimodal transformer more robust," *Transactions on Machine Learning Research*, 2025.