Muhammad Irzam Liaqat[1], Qaiser Abbas[1], Shah Nawaz[1], Zaigham Zaheer[1], Marta Moscati[1], Yufang Hou[1], Muhammad Haris Khan[1], Salman Khan[1], Elisabeth Andre[1], and Markus Schedl[1]

[1]Affiliation not available

November 25, 2025

Multimodal Learning Under Imperfect Data Conditions: A Survey

MUHAMMAD IRZAM LIAQAT†, IMT School for Advance Studies, Italy
QAISER ABBAS†, College of Science and Engineering, Hamad Bin Khalifa University, Qatar
SHAH NAWAZ†, Institute of Computational Perception, Johannes Kepler University, Austria
ZAIGHAM ZAHEER, Mohamed bin Zayed University of Artificial Intelligence, UAE
MARTA MOSCATI, Institute of Computational Perception, Johannes Kepler University, Austria
YUFANG HOU, IT:U - Interdisciplinary Transformation University, Austria
MUHAMMAD HARIS KHAN, Mohamed bin Zayed University of Artificial Intelligence, UAE
SALMAN KHAN, Mohamed bin Zayed University of Artificial Intelligence, UAE
ELISABETH ANDRE, University of Augsburg, Germany
MARKUS SCHEDL, Institute of Computational Perception, Johannes Kepler University, Austria

Multimodal learning leverage multiple and diverse modalities such as images, text, or audio to enable contextual understanding and reliable decision-making. These methods have achieved state-of-the-art results by integrating multiple modalities in areas such as medical imaging, autonomous driving, and visual surveillance. However, the effectiveness of multimodal learning in real-world scenarios remains limited by practical challenges: data from some modalities may be missing, corrupted, or poorly aligned due to sensor failures, environmental noise, or bandwidth constraints. While prior surveys have proposed taxonomies on multimodal learning and strategies for handling missing or corrupted data, these perspectives are often treated in isolation. This separation overlooks the fact that data imperfections are interconnected, and effective multimodal learning requires a unified understanding across architectural design and modality reliability. To address this gap, we present comprehensive taxonomies that cover and integrate three major aspects: (1) architectural design for multimodal learning, (2) learning under missing modalities, and (3) learning under corrupted modalities. By framing these aspects together, our study highlights their interdependencies and facilitates a comprehensive understanding of multimodal learning. Furthermore, we discuss benchmark datasets and real-world applications through this taxonomic lens and outline open challenges and future directions for developing resilient methods. The complete list of related resources is available at: GitHub repository.

Additional Key Words and Phrases: Representation learning, Missing modalities, Corrupted modalities, Taxonomy

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 [311]. While traditional

† Equal contribution

Authors' Contact Information: Muhammad Irzam Liaqat†, irzam.liaqat@imtlucca.it, IMT School for Advance Studies, Lucca, Italy; Qaiser Abbas†, qaab89376@hbku.edu.qa, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar; Shah Nawaz†, shah.nawaz@jku.at, Institute of Computational Perception, Johannes Kepler University, Linz, Austria; Zaigham Zaheer, zaigham.zaheer@mbzuai.ac.ae, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE; Marta Moscati, marta.moscati@jku.at, Institute of Computational Perception, Johannes Kepler University, Linz, Austria; Yufang Hou, yufang.hou@it-u.at, IT:U - Interdisciplinary Transformation University, Linz, Austria; Muhammad Haris Khan, muhammad.haris@mbzuai.ac.ae, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE; Salman Khan, salman.khan@mbzuai.ac.ae, Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE; Elisabeth Andre, elisabeth.andre@uni-a.de, University of Augsburg, Augsburg, Germany; Markus Schedl, markus.schedl@jku.at, Institute of Computational Perception, Johannes Kepler University, Linz, Austria.

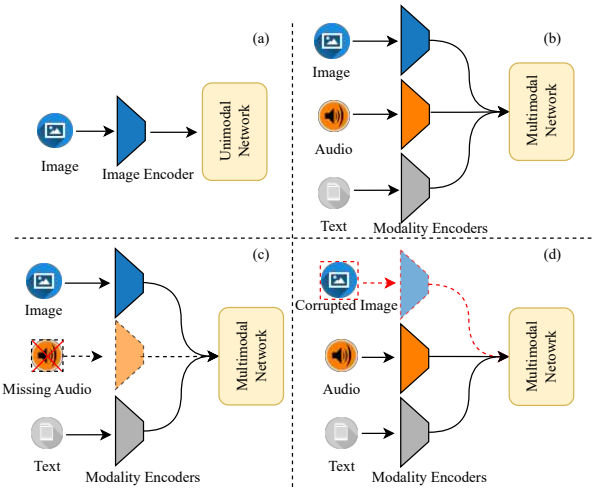

Fig. 1. Overview of (a) Unimodal (b) Multimodal Learning (c) with missing modalities (d) with corrupted modalities

unimodal methods are designed for individual independent modalities, MML has emerged to model and learn from heterogeneous data streams, notably enhancing performance across tasks [153]. Existing MML models have demonstrated remarkable success in domains ranging from medical imaging [148] and emotion recognition [324] to autonomous vehicles [44].

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 [153]. 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 [70]. This motivates us to focus on three interconnected challenges that are essential for deploying 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. Fig. 1 presents a high-level overview of these scenarios.

Architectural Design for Modality Integration: A fundamental challenge lies in designing multimodal architectures that can effectively learn representations across multiple data types [311]. 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 [109]. 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, multimodal learning is vulnerable when one or more modalities are missing during inference [285]. 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 [122], modality dropout training [160, 183], and conditional learning to improve model flexibility [75].

Resilience to Corrupted Modalities: Beyond simply missing modalities, a more realistic yet often overlooked challenge is handling corrupted modality problem (CMP) [323]. 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 [110]. 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.

While prior surveys on MML have addressed individual aspects such as architectural design [18], missing modality handling [285], or data incompleteness [323], they provide only partial views of the field. As shown in Table 1, existing works do not offer an integrated understanding of MML under imperfect data conditions, where modalities may be entirely missing or corrupted. Meanwhile, these challenges are interconnected, for instance, architectural design choices directly influence how networks respond to missing data [156]. Meanwhile, strategies for handling missing data often rely on assumptions about the quality of the remaining modalities [78]. Conversely, corrupted modalities can mimic missing data, blurring the line between the two [111]. However, existing surveys treat these aspects in isolation, leaving a gap in unified understanding. To address this gap, our survey presents structured taxonomies and critical analyses that integrate advances across missing and corrupted modality handling and multimodal system design. Our key contributions are as follows:

(1) We introduce taxonomies that span three interconnected dimensions: architectural design, missing modality handling, and robustness under corrupted modalities, offering a unified perspective lacking in prior surveys.
(2) We systematically organize our survey to outline the existing benchmark datasets, application areas, and methodological strengths and limitations within the three core dimensions.
(3) Building on our findings, we identify key open challenges and outline future research directions for robust MML.

By focusing on these aspects in a systematic manner, our work serves as a foundational reference for researchers, practitioners, and students interested in developing deployable multimodal systems that can operate reliably under the imperfect conditions commonly encountered in practical applications. The remainder of this study is organized as follows. Section 2 explains our strategy for literature search and selection. Section 3 provides an overview of existing MML paradigms and presents the architectural design principles and integration strategies across heterogeneous modalities. Section 4 introduces our taxonomy for MML under missing modality and discusses existing methods. Section 5 focuses on MML under corrupted modalities, presents the proposed taxonomy and summarizes recent progress

Table 1. Comparison of existing surveys. A check mark (✓) indicates that the survey provides a taxonomy for the given challenge, while a cross (✗) indicates that it does not provide a dedicated taxonomy even if the challenge is mentioned in passing. **MisM**: Missing modality taxonomy; **CorM**: Corrupted modality taxonomy.

Survey	Venue & Year	MML	MisM	CorM
Multimodal fusion for multimedia analysis: a survey [11]	Multimedia Systems, 2010	✓	✗	✗
Multimodal Machine Learning: A Survey and Taxonomy [18]	TPAMI, 2018	✓	✗	✗
Deep multimodal representation learning: A survey [72]	IEEE, 2019	✓	✗	✗
A survey on deep learning for multimodal data fusion [65]	Neural Computing, 2020	✓	✗	✗
A survey on deep multimodal learning for computer vision [22]	Visual Computer, 2022	✓	✗	✗
Multimodal learning with transformers: A survey [294]	TPAMI, 2023	✓	✗	✗
Deep multimodal learning with missing modality: A survey [285]	arXiv, 2024	✓	✓	✗
Multimodal fusion on low-quality data: A comprehensive survey [323]	arXiv, 2024	✓	✗	✓
Ours	—	✓	✓	✓

in handling noisy, misaligned, or degraded inputs. Finally, Section 6 highlights open challenges and promising research directions for robust MML.

2 Literature Search and Filtering Strategy

To construct a comprehensive survey, 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. 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 [1]. 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.

3 Multimodal Learning

This section offers a concise overview structured around two complementary perspectives. The first follows the taxonomy of Baltrušaitis et al. [18], 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 [294] 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.

Representation learning concerns how features from different modalities are modeled. *Joint representation* methods aim to map multiple modalities into a unified embedding space. Such representations have been widely explored in

[1] Complete list of resources available at: GitHub repository

multimodal classification and verification tasks. Model-agnostic fusion approaches such as *early fusion* concatenate input-level features before learning, as seen in multimodal emotion recognition with audio and visual signals [13]. *Late fusion* combines unimodal outputs, yielding robustness against noisy or missing modalities [54, 203]. *Hybrid strategies* integrate both levels, achieving success in tasks such as event detection and emotion classification [40, 196]. Beyond model-agnostic methods, model-assisted approaches introduce algorithmic guidance. *Kernel-based methods* such as multiple kernel learning have been applied in object classification, disease detection, and affect recognition [243]. *Probabilistic models* including dynamic Bayesian networks facilitate joint modeling of temporal sequences [102], 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 [80, 319]. In contrast, *coordinated representations* maintain modality-specific spaces and enforce consistency through cross-modal objectives. These include contrastive learning approaches [98] and distance-based methods [137].

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 [35]. Meanwhile, *generative* methods explicitly produce one modality from another. For instance, encoder–decoder architectures map videos to textual descriptions [16] or generating image captions [91, 122]. *Hybrid approaches* combine retrieval and generation, balancing diversity with accuracy [84, 179]. 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 [340]. *Explicit* alignment relies either on supervision, where paired annotations guide the alignment [33, 306] or on unsupervised learning, which exploits co-occurrence statistics or structural constraints [90, 289]. 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 [25]. *Late fusion* aggregates unimodal predictions, often through voting or weighted averaging, which can mitigate the impact of noisy modalities [54, 203]. *Hybrid fusion* strategies integrate intermediate-level representations with final predictions, offering robustness in tasks such as multimedia event detection and emotion recognition [40, 188]. *Model-assisted fusion* employs more principled learning approaches: *kernel-based methods* capture modality-specific similarities [243], *probabilistic models* handle temporal dependencies [102], and neural architectures learn complex integration functions across heterogeneous inputs [319].

Co-learning leverages one modality to enhance learning in another and involves iteratively updating classifiers across modalities to improve generalization [214]. *Transfer learning* transfers knowledge from a well-labeled modality to improve performance in a resource-scarce one, with applications in audio-visual recognition [87], disease classification [82], and action recognition [280]. *Conceptual grounding* maps symbolic linguistic concepts to perceptual modalities that aids semantic representation learning [116]. *Zero-shot learning* has also emerged as an important approach that enables models to generalize to unseen modality combinations using shared semantic embeddings [194]. In cross-lingual transliteration, *hybrid bridging* approaches use intermediate representations to facilitate knowledge transfer [269].

Multimodal Learning in the Era of Transformers: Transformers have revolutionized MML by leveraging unified frameworks for processing and integrating multiple modalities, representing a significant shift from traditional approaches that relied on modality-specific encoders and fusion mechanisms [294]. 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 [212] and DALL-E [218]. Modern architectures can be categorized into three paradigms [294]: *Single-stream Architectures* (e.g., UNITER [35], FLAVA [244], CLIPPO [259]) that process all modalities through shared transformer backbones; *Multi-stream Architectures* exemplified by CLIP's dual-encoder design and extended by ALIGN [101] and Florence [310]; and *Hybrid Architectures* like ALBEF [143], BLIP [141], and InstructBLIP [45] 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 [212] and extended by ALIGN [101], Chinese-CLIP [299], and domain-specific applications in medical imaging [330] and scientific literature [121], maximizes agreement between corresponding multimodal pairs. Masked Modeling leverages Masked Language Modeling (MLM) with Masked Region Modeling (MRM) as demonstrated by VinVL [320], OSCAR [147], BEiT [21], and MAE [81]. Generative Modeling encompasses cross-modal content generation, e.g., DALL-E [217, 218], Flamingo [4], and GPT-4V [200]. Applications span Vision-Language Models (e.g., BLIP-2 [140], VideoBERT [250]), Audio-Visual Models (e.g., AVBert [236], SpeechT5 [7], LAVISH [158], AV-HuBERT [235]), and Multimodal Large Language Models (e.g., LLaVA [161], Video-ChatGPT [180]), which represent the current frontier. Moreover, modern training strategies include small-scale specialized models like LayoutLM [295] and CLIP-Score [84], that focus on domain-specific tasks with computational efficiency. Similarly, Large-Scale Foundation Models including CLIP [212], DALL-E2 [217], and Flamingo [4] leverage massive datasets for general-purpose representations with emergent capabilities. Instruction-Tuned Models like InstructBLIP [45] and LLaVA [161] 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.

3.1 Summary

MML has evolved significantly from relying on hand-crafted features like SIFT and MFCC [168] to employing deep learning-based methods using CNNs or RNNs [85, 127, 191]. This shift has enabled more sophisticated multimodal representations and evolved beyond simple feature concatenation [44]. 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 [169, 212, 294]. 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 [140, 161]. However, the integration of modalities involves challenges such as managing inherent heterogeneity, and most critically, addressing scenarios with imperfect modalities [248]. The progression towards large-scale foundation [4, 212] and instruction-tuned models [45] promises more general-purpose, robust, and interpretable systems capable of reasoning across an ever-expanding set of tasks and domains. Laslty, 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 [33], Visual Genome [126], and LAION-5B [227], 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 2.

Table 2. 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 [125]	✗	✓	✗	✓	Video Understanding
COCO Captions [33]	✗	✗	✓	✓	Computer Vision, Language
Flickr30K [306]	✗	✗	✓	✓	Vision–Language
MS COCO [157]	✗	✗	✓	✓	Object Recognition
MSR-VTT [293]	✗	✓	✗	✓	Video–Language Alignment
UPMC Food-101 [273]	✗	✗	✓	✓	Food Recognition
VATEX [275]	✗	✓	✗	✓	Multilingual Video Captioning
WebVid-10M [16]	✗	✓	✗	✓	Video–Text Pretraining
WIT [247]	✗	✗	✓	✓	Multilingual Multimodal
Medical Imaging and Specialized Domains					
MMIST-ccRCC [193]	✗	✗	✓	✓	Medical Imaging
OBELICS [130]	✗	✗	✓	✓	Document Understanding
Multimodal and Vision–Language Pretraining					
CC-12M [28]	✗	✗	✓	✓	Internet-scale Vision–Language
CLIP Benchmark [212]	✗	✗	✓	✓	Zero-Shot Learning
Conceptual Captions [233]	✗	✗	✓	✓	Vision–Language Pretraining
LAION-5B [228]	✗	✗	✓	✓	Large-scale Pretraining
LAION-5B [227]	✗	✗	✓	✓	Vision–Language Pretraining
UNITER eval. splits [35]	✗	✗	✓	✓	Vision–Language Pretraining
Sentiment and Emotion Analysis					
CMU-MOSEI [13]	✓	✓	✗	✓	Affective Computing
CMU-MOSI [313]	✓	✓	✗	✓	Affective Computing
Video Analysis and Audio–Visual Tasks					
AVA-ActiveSpeaker [223]	✓	✓	✗	✗	Video Analysis, AV Sync
HowTo100M [190]	✓	✓	✗	✓	Instructional Learning
XDViolence [284]	✓	✓	✗	✗	Violence Recognition
Visual Question Answering and Grounding					
CLEVR [108]	✗	✗	✓	✓	Reasoning
GQA [93]	✗	✗	✓	✓	Scene Understanding
MultiModalQA [254]	✗	✗	✓	✓	QA, Multimodal Reasoning
NLVR2 [249]	✗	✗	✓	✓	Visual Reasoning
VQA 2.0 [71]	✗	✗	✓	✓	Question Answering
Visual Genome [126]	✗	✗	✓	✓	Scene Understanding

4 Dissecting MML 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, [135, 172, 176] defined benchmarks for MMP with systematical drop of varying fractions of missing labels or modalities either during train or test stage. These studies demonstrate that existing MML architectures experience severe performance deterioration, at times under-performing their unimodal counterparts.

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, 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.

4.1 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

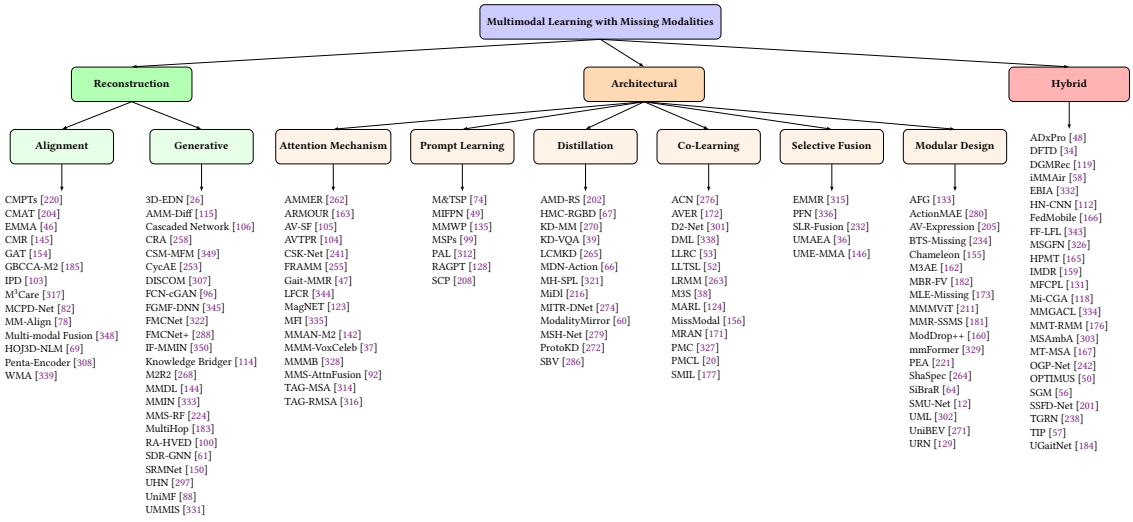

Fig. 2. Taxonomy of the existing works addressing multimodal learning (MML) with missing modality problem (MMP)

interactions and relationships during training while preserving the integrity of the original feature representations. Reconstruction approaches can be broadly classified into generative and alignment methods.

4.1.1 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 [26, 288, 297, 307, 322, 333, 350]. By capturing the statistical dependencies among modalities, generative methods are capable of generating semantically consistent and plausible representations of missing data [88, 96, 224, 253, 258, 331, 345, 349].

For instance, M2R2 [268] leveraged a data imputation strategy using common representation learning (CRL) for reconstructing the absent modality representations during inference. Similarly, Jeong et al. [100] 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 [115] integrated a diffusion-based generative model with an Image-Frequency Fusion Network for high-fidelity modality reconstruction. Ke et al. [114] explored the use of large multimodal models to generate, rank, and select the most appropriate reconstructions for missing modalities. John et al. [106] 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 [61] reconstructed missing modalities by aggregating spectral features through a spectral domain reasoning graph neural network. Li et al. [144] applied stacked autoencoders for effective modality imputation in traffic data scenarios, while Li et al. [150] 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 3a provides an overview of recent studies that leverage different generative methods.

Table 3. Comparison of methods for MMP (I)

(a) Generative methods for MMP

Method	Modalities	Application Area
3D-EDN [26]	Images, Genomics	Alzheimer Detection
AMM-Diff [115]	Multimodal MRI	MRI Synthesis
Cascaded Network [106]	Audio, Video, Thermal	Person Classification
CRA [258]	Images, Text	Object Recognition
CSM-MFM [349]	Multimodal MRI	Tumor Segmentation
CycAE [253]	Multimodal Images	Alzheimer Detection
DISCOM [307]	Images, Genomics	Alzheimer Prediction
FCN-cGAN [96]	MRI, Gene Expression	Medical Imaging
FGMF-DNN [345]	Multimodal MRI	Tumor Segmentation
FMCNet [322]	Images, Infrared	Person Re-Identification
FMCNet+ [288]	Images, Infrared	Person Re-Identification
IF-MMIN [350]	Audio, Video, Text	Emotion Recognition
Knowledge Bridger [114]	Video, Audio	Object Detection
M2R2 [268]	Audio, Video, Text	Emotion Recognition
MMDL [144]	Multisensor Data	Traffic Data Imputation
MMIN [333]	Audio, Video, Text	Emotion Recognition
MMS-RF [224]	Multisensor Data	Context Recognition
MultiHop [183]	Images, Text	Recommendation System
RA-HVED [100]	Multimodal MRI	Tumor Segmentation
SDR-GNN [61]	Video, Audio	Emotion Recognition
SRMNet [150]	Multimodal MRI	Tumor Segmentation
UHN [297]	Multimodal MRI	Medical Imaging
UniMF [88]	Audio, Video, Text	Sentiment Analysis
UMMIS [331]	Multimodal MRI	Medical Imaging

(b) Hybrid methods for MMP

Method	Modalities	Application Area
ADxPro [48]	MRI, PET, CSF, Biomarkers	Alzheimer Prediction
DFTD [34]	MRI, PET	Alzheimer Diagnosis
DGMRec [119]	Text, Images, Audio	Recommendation System
EBIA [332]	Text, Audio, Video	Sentiment Analysis
FedMobile [166]	Multimodal Sensor Data	Human Activity Detection
FF-LFL [343]	MRI, CT, PET	Brain Tumor Segmentation
HN-CNN [112]	Optical, DSM	Remote Sensing, Classification
HPMT [165]	Text, Images	Long Document Classification
IMDR [159]	Fundus, OCT	Ophthalmic Diagnosis
iMMAir [58]	Images, Sensor Data	Air Quality Prediction
MSGFN [326]	Text, Audio, Video	Sentiment Analysis
MFCPL [131]	Audio, Video, Text	Hate Speech Detection
Mi-CGA [118]	Text, Audio, Video	Emotion Recognition
MMGACL [334]	Text, Images	Recommendation System
MMT-RMM [176]	Images, Text	Hate Speech Detection
MSAmbA [303]	Audio, Video	Human Action Recognition
MT-MSA [167]	Text, Audio, Video	Sentiment Analysis
OGP-Net [242]	Images, Infrared	Scene Segmentation
OPTIMUS [50]	MRI, PET, CSF	Alzheimer Prediction
SGM [56]	Image, Audio, Signals	Emotion Recognition
SSFD-Net [201]	Palmprint, Palmvein	Biometric Recognition
TGRN [238]	Text, Audio, Video	Sentiment Analysis
TIP [57]	Tabular, Images	Myocardial Prediction
UGaitNet [184]	Optical Flow, Silhouette	Gait Recognition

(c) Co-learning methods for MMP

Method	Modalities	Application Area
ACN [276]	Multimodal MRI	Brain Tumor Segmentation
AVER [172]	Audio, Video	Emotion Recognition
D2-Net [301]	Multimodal MRI	Brain Tumor Segmentation
DML [338]	Multisensor Data	Weather Mapping
LLRC [53]	Images, Text	Face Recognition
LLTSL [52]	Images, Text	Face Recognition
LRMM [263]	Text, Images	Recommendation Systems
M3S [38]	Text, Audio, Video	Sentiment Analysis
MARL [124]	Multimodal MRI	Brain Tumor Segmentation
MissModal [156]	Text, Audio, Video	Sentiment Analysis
MRAN [171]	Text, Audio, Video	Sentiment Analysis
PMC [327]	Images, Text	Object Recognition
PMCL [20]	Multiple Sensors	Prototype Probing
SMIL [177]	Video, Audio, Text	Movie Genre Classification

(d) Distillation methods for MMP

Method	Modalities	Application Area
AMD-RS [202]	Multispectral/HSI	Remote Sensing Classification
HMC-RGBD [67]	Images, Depth	Video Action Recognition
KD-MM [270]	Multimodal MRI	Alzheimer's Diagnosis
KD-VQA [39]	Images, Text	Visual Question Answering
LCMKD [265]	Multimodal MRI	Brain Tumor Segmentation
MDN-Action [66]	Images, Depth, Skeleton	Human Action Recognition
MH-SPL [321]	Images, Text	Image Recognition
MiDl [216]	Video, Audio	Action Recognition
MITR-DNet [274]	Text, Audio, Video	Sentiment Intensity Analysis
ModalityMirror [60]	Audio, Video	Audio Classification
MSH-Net [279]	Multispectral/HSI	Remote Sensing Classification
ProtoKD [272]	Multimodal MRI	Brain Tumor Segmentation
SBV [286]	Audio, Video	Audiovisual Segmentation

4.1.2 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 [69, 82]. For instance, MM-Align [78] explored optimal transport theory with deep learning modules to effectively align modalities in a common latent space. M3Care [317] utilized auxiliary information derived from present modalities to assist the alignment process through the construction of similarity matrices. Leveraging Bayesian methods, Matsuura et al. [185] proposed a Bayesian Canonical Correlation Analysis (CCA) framework that projects multiple modalities into a shared 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 [46, 103, 145, 204, 348].

More recent innovations further expand the scope of alignment-based strategies. Liang et al. [154] introduced a graph attention network (GAT) using transfer learning to align heterogeneous data sources. Yu et al. [308] employed a

modality-dynamic encoder to capture tumor semantics across incomplete modalities. Reza et al. [220] applied cross-modal proxy tokens and alignment loss to guide the fusion process in the presence of missing data. Contrarily, Zhi et al. [339] incorporated Wasserstein distance into self-attention mechanisms to enhance modality fusion and alignment. Table 4e 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.

4.2 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.

4.2.1 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, 173, 205]. Table 4b highlights the key studies that introduced resilient architectural designs. For example, M3L [181] employed a semi-supervised framework that integrates modality masking to improve model generalization. Similarly, MMMVIT [211] leveraged Vision Transformers to learn a shared latent representation and extracted multi-scale features via intra-model fusion. Reza et al. [221] 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 [271] 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 [329] introduced a hybrid design with modality-specific encoders and inter-modal transformers, complemented by auxiliary regularizers that guide the learning of modality-invariant features. Similarly, UML [302] 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 [129, 234, 280], masked autoencoders for unsupervised feature reconstruction [162, 182], modular network designs with independently functioning sub-networks [12, 264], and dynamic co-learning mechanisms [160] that adaptively balance modality contributions.

4.2.2 Selective Fusion. Fusion methods aim to mitigate the impact of missing modalities by utilizing different techniques such as ensemble learning and weighted voting mechanisms [36, 146, 315]. Table 4d summarizes the studies that enable selective fusion dynamically to improve multimodal performance. For instance, Zheng et al. [336] proposed progressive

fusion that learns representations from single to multiple modalities. It involves taking the predictions from different modalities which are fused for decision making. Shao et al. [232] 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.

4.2.3 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 [53, 177], zero-shot learning [263], and conceptual grounding [124]. Moreover, as shown by Table 3c, these methods are leveraged to improve modular performance across a variety of application areas. In recent literature, Bao et al. [20] proposed a federated learning-based architecture that leverages a masking strategy to find, learn and compensate the missing information. Konwer et al. [124] proposed a meta-learning methodology in combination with conceptual grounding that enables the approximation of missing features from the learned shared latent space. MissModal [156] 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 [171] 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 [38], duel disentanglement [301], knowledge sharing [172, 338], progressive modality cooperation [327], adversarial co-training [276], and low-rank transfer [52].

4.2.4 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 [39].

Table 3d outlines recent studies that have explored various distillation strategies. For instance, Zhang et al. [321] proposed cross-branch supervision through a bi-directional distillation architecture, where one branch learns unimodal representations while the other performs MML. Wu et al. [286] 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 [272] transferred semantic knowledge from a complete-modality teacher to a student model that performs segmentation tasks under partial modality input. Cross-Model distillation techniques [265, 270] 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 [279], adversarial distillation [202], and hallucination-based distillation [66, 67].

4.2.5 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 [47, 241]. Table 4a outlines existing studies leveraging attention-based methods for modular robustness. In a recent study, MMAN-M2 [142] used a transformer-based encoder-decoder model with a multi-head mechanism for dynamic extraction and fusion of feature representations. Similarly, CTNet [105] 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.

Table 4. Comparison of methods for MMP (II)

(a) Attention-based methods for MMP

Method	Modalities	Application Area
AMMER [262]	Audio, Video, Text	Emotion Recognition
ARMOUR [163]	Text, Table	Mortality Prediction
AV-SF [105]	Audio, Video	Person Recognition
AVTPR [104]	Audio, Video	Person Recognition
CSK-Net [241]	Optical, Infrared	Semantic Segmentation
FRAMM [255]	Clinical Trials Data	Clinical Trial Site Selection
Gait-MMR [47]	Images, Silhouettes, Depth	Gait Recognition
LFCR [344]	Multimodal MRI	Brain Tumor Segmentation
MagNET [123]	Multimodal MRI	Brain Tumor Segmentation
MFI [335]	Multimodal MRI	Brain Tumor Segmentation
MMAN-M2 [142]	Audio, Video	Emotion Recognition
MMM-VoxCeleb [37]	Audio, Video	Speaker Identification
MMMB [328]	Images, Text	Image Aesthetic Prediction
MMS-AttnFusion [92]	Multimodal MRI	Brain Tumor Segmentation
TAG-MSA [314]	Video, Audio, Text	Sentiment Analysis
TAG-RMSA [316]	Video, Audio, Text	Sentiment Analysis

(b) Modular design methods for MMP

Method	Modalities	Application Area
AFG [133]	Audio, Video	Action Recognition
ActionMAE [280]	RGB, Depth, Infrared	Action Recognition
AV-Expression [205]	Audio, Video	Expression Recognition
BTS-Missing [234]	Multimodal MRI	Brain Tumor Segmentation
Chameleon [155]	Image, Text, Audio	Representation Learning
M3AE [162]	Multimodal MRI	Brain Tumor Segmentation
MBR-FV [182]	Facial Video	Biometrics Recognition
MLE-Missing [173]	Audio, Video	Emotion Recognition
MMMViT [211]	Multimodal MRI	Brain Tumor Segmentation
MMR-SSMS [181]	Images, Depth Imaging	Semantic Segmentation
ModDrop++ [160]	Multimodal MRI	Sclerosis Lesion Segmentation
mmFormer [329]	Multimodal MRI	Brain Tumor Segmentation
PEA [221]	Depth, Thermal, Images	Semantic Segmentation
ShaSpec [264]	Multimodal MRI	Brain Tumor Segmentation
SiBraR [64]	Image, Text, Audio	Recommendation System
SMU-Net [12]	Multimodal MRI	Brain Tumor Segmentation
UML [302]	Multimodal OCT	Visual Acuity Prediction
UniBEV [271]	LiDAR, Images	Object Detection
URN [129]	Multimodal MRI	Brain Tumor Segmentation

(c) Prompt learning methods for MMP

Method	Modalities	Application Area
M&TSP [74]	Text, Audio, Video	Emotion Recognition
MIFPN [49]	Multimodal MRI	Brain Tumor Segmentation
MMWP [135]	Images, Text	Food Recognition
MSPs [99]	Images, Audio	Movie Genre Classification
PAL [312]	Audio, Video, Text	Food Recognition
RAGPT [128]	Text, Images	Hate Speech Detection
SCP [208]	Images, Text	Hate Speech Detection

(d) Selective fusion methods for MMP

Method	Modalities	Application Area
EMMR [315]	Text, Audio, Video	Sentiment Analysis
PFN [336]	Video, Sensor	Person Re-identification
SLR-Fusion [232]	Images, Depth	Face Recognition
UMAEA [36]	Images, Text	Multimodal Entity Alignment
UME-MMA [146]	Audio, Video, Text	Audio–Video Event Localization

(e) Alignment methods for MMP

Method	Modalities	Application Area
CMPTs [220]	Images, Text	Robust Learning
CMAT [204]	Images, Skeleton	Human Action Recognition
CMR [145]	Multimodal Data	Cross-modal Retrieval
EMMA [46]	Text, Images	Object Retrieval
GAT [154]	MM Knowledge Graph	Representation Learning
GBCCA-M2 [185]	Images, Tabular	Image Retrieval
HOJ3D-NLM [69]	Video, Audio	Human Action Recognition
IPD [103]	Audio, Video, Text	Sentiment Intensity Analysis
M^3Care [317]	Multimodal EHRs	Ocular Disease Diagnosis
MCPD-Net [82]	Images, Accelerometer	Parkinson Classification
MM-Align [78]	Multimodal Sequences	Model Robustness
Multi-modal Fusion [348]	Images	Facial Recognition
Penta-Encoder [308]	Multimodal MRI	Brain Tumor Segmentation
WMA [339]	Video, Text, Audio	Hate Speech, Movie Genre

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. [262] enhanced model performance by introducing cross-model and self-attention in the underlying transformer architecture. Similar recent works include inter-model contrastive learning [163], masked cross model attention [255], and tag-assisted attention [314]. In addition, attention-based architectures are employed for sensor fusion [104], brain tumor segmentation [92, 335, 344, 346], sentiment analysis [37, 316], and image aesthetic prediction [328].

4.2.6 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 [99]. Table 4c outlines the existing literature that leverages prompt learning. For modality-aware prompts, Lee et al. [135] leveraged dynamically learnable training where the proposed model learns adaptive prompt embeddings to indicate

the availability of each modality. Pipoli et al. [208] 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. [312] introduced modality-specific and task-aware prompts 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. [128] 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. [74] 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.

4.3 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 [167] combined modality translation with fusion mechanisms in an encoder-decoder setup to enhance sentiment analysis performance under missing modality conditions. Zhou et al. [343] 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. [34] proposed imputation of missing modalities through combined distillation networks with a disentanglement module to capture both intra- and inter-modal characteristics. Table 3b highlights the work leveraging hybrid strategies such as model design integrated with dynamic fusion [176], self-supervised learning using attention and pre-trained encoders [184], and semi-supervised reconstruction-based co-learning frameworks [56, 112]. Furthermore, recent hybrid models continue this trend by combining diverse learning paradigms such as contrastive learning, disentanglement, proxy learning, and federated training. Dhivyaa et al. [48] and Zhang et al. [332] introduced hybrid models using VAEs, RNNs, and fuzzy-aware modules for healthcare and sentiment analysis. Kieu et al. [118] and Liu et al. [165] employed graph neural networks and hierarchical modeling to address missing modalities in emotion and document classification. More recent works such as Fan et al. [58], SSFD-Net [201], IMDR [159], and FedMobile [166] integrated diffusion-based disentanglement, proxy learning, and federated MML. Others such as Zhao et al. [334], Le et al. [131], and Du et al. [57] 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.

4.4 Summary

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 [61, 100, 114, 115, 268]. While powerful, these methods often entail high computational complexity and can be sensitive to noise or outliers in the training data [253, 333]. 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 [78, 185, 317]. Though computationally lighter, these methods can struggle with modality misalignment and semantic drift under high missingness. *Architectural* approaches avoid reconstructing missing modalities altogether and focus instead on adapting model architectures or learning dynamics to be inherently resilient. Techniques such as modality-specific architectural designs [181, 211], late fusion [232, 336], and co-learning frameworks [124, 156] 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 [216, 272, 321]. Attention-based mechanisms further enhance adaptability by dynamically weighting modalities based on relevance [142, 163]. On the other hand, prompt learning offers a lightweight and modular approach particularly suited to transformer-based architectures [135, 208]. 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 [34, 167, 343]. Recent advances further incorporate contrastive learning, federated learning, and graph-based reasoning to address complex real-world scenarios [58, 166, 292, 303]. While hybrid strategies offer improved robustness and generalization, they often involve increased architectural complexity and higher computational demands, which may hinder scalability [48, 332].

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 5a.

5 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 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 [323]. 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 [42, 164]. Empirical studies have further demonstrated that multimodal models are particularly sensitive to corrupted inputs, underscoring the necessity for robust MML methodologies [23, 30, 83, 186, 237, 257, 260]. 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

Table 5. Existing benchmark datasets for MML for MMP (right) and CMP (left). Modalities: Audio (A), Video (V), Text (T), Image (I). Corruption (C): Noise (N), Perturbations (P), Rephrasing (R), Variance (V). †One or more modalities are *synthetically* dropped or naturally absent in derived splits.

(a) Datasets for Missing Modality Problem

Dataset†	A	V	I	T	Application Area
Action and Activity Recognition					
Kinetics-Sounds [27]	✓	✓	✗	✗	Action Recognition
NTU RGB-D [230]	✗	✗	✓	✗	Activity Recognition
Event Detection					
AudioSet (inc.) [68]	✓	✓	✗	✗	Event Classification
General Classification					
MM-IMDB [10]	✗	✓	✓	✓	Multi-label Classification
Hate Speech Detection					
Hateful Memes [117]	✗	✗	✓	✓	Multimodal Classification
Medical AI (Diagnosis, Prognosis, Imaging)					
BraTS 2018/2020 [17]	✗	✗	✓	✗	Image segmentation
M3Care [317]	✗	✗	✗	✓	Prognosis Prediction
MIMIC-III [107]	✗	✗	✗	✓	Clinical Diagnosis
Multimodal Translation and ASR					
How2 [225]	✓	✓	✗	✓	Machine Translation
Retrieval					
COCO-Missing [277]	✗	✗	✓	✓	Image–text Retrieval
Flickr30K-Missing [277]	✗	✗	✓	✓	Image–text Retrieval
MIT-States [97]	✗	✗	✓	✓	Zero-shot Learning
Recipe1M+/Food-101 [273]	✗	✗	✓	✓	Cross-modal Retrieval
Sentiment and Emotion Classification					
CMU-MOSEI [13]	✓	✓	✗	✓	Emotion Recognition
IEMOCAP [24]	✓	✓	✗	✓	Emotion Recognition
MELD [209]	✓	✓	✗	✓	Emotion Recognition
MISA-MOSEI [80]	✓	✓	✗	✓	Sentiment Analysis
SMIL-MOSI [177]	✓	✓	✗	✓	Sentiment Analysis
Visual Question Answering					
VQA-Missing [240]	✗	✗	✓	✓	Robust Learning

(b) Datasets for Corrupt Modality Problem

Dataset	A	V	T	I	C	Application Area
Audio and Event Detection						
Common Voice-N [9]	✓	✗	✓	✗	N	Speech Recognition
CrisisMMD [3]	✗	✗	✓	✓	N	Crisis Recognition
DESED [188]	✓	✗	✗	✗	N	Event Detection
MIMII [210]	✓	✗	✗	✗	N	Anomaly Detection
Data Retrieval						
Flickr30K-N [143]	✗	✗	✓	✓	N	Cross-modal retrieval
MSR-VTT-N [170]	✗	✓	✓	✗	N	Video–text retrieval
Winoground [256]	✗	✗	✓	✓	N	Cross-modal reasoning
Image Classification and Captioning						
CIFAR-10-C [83]	✗	✗	✗	✓	N	Image classification
Cityscapes-C [189]	✗	✗	✗	✓	N	Semantic segmentation
COCO-C [189]	✗	✗	✗	✓	N	Object detection
COCO-Noisy [32]	✗	✗	✓	✓	N	Image–text alignment
ImageNet-C [83]	✗	✗	✗	✓	N	Image classification
ImageNet-P [83]	✗	✗	✗	✓	P	Image classification
PASCAL-C [189]	✗	✗	✗	✓	N	Object detection
Sentiment and Emotion Classification						
CMU-MOSI-R [215]	✓	✓	✓	✗	R	Sentiment Classification
M3ED [149]	✗	✓	✓	✗	V	Emotion Recognition
VoxCeleb-N [195]	✓	✓	✗	✗	N	Speaker Identification
Visual Question Answering						
R-Bench [138]	✗	✗	✓	✓	N	Multimodal Reasoning
VQA-CP [2]	✗	✗	✓	✓	R	Bias Evaluation
VQA-Rephrasings [229]	✗	✗	✓	✓	R	Linguistic Variation
VQA-Robust [95]	✗	✗	✓	✓	N	Noise-tolerant VQA

shown in Fig. 3. We also survey their application areas, outline corresponding benchmark datasets, and highlight open challenges in designing resilient multimodal systems.

5.1 Data Processing

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.

5.1.1 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 7a presents different denoising methods have been explored across diverse application domains. In medical imaging, RHViT [342] 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 [287] leveraged a denoising autoencoder for reconstruction of noisy sensor signals prior to multimodal representation pipeline. In cloud computing, Ikhlas et al. [94] targeted the noise in time series data by leveraging the proposed stacked denoising autoencoder with bidirectional gated recurrent units (BiGRUs). Similarly, Yin et al. [305] 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.

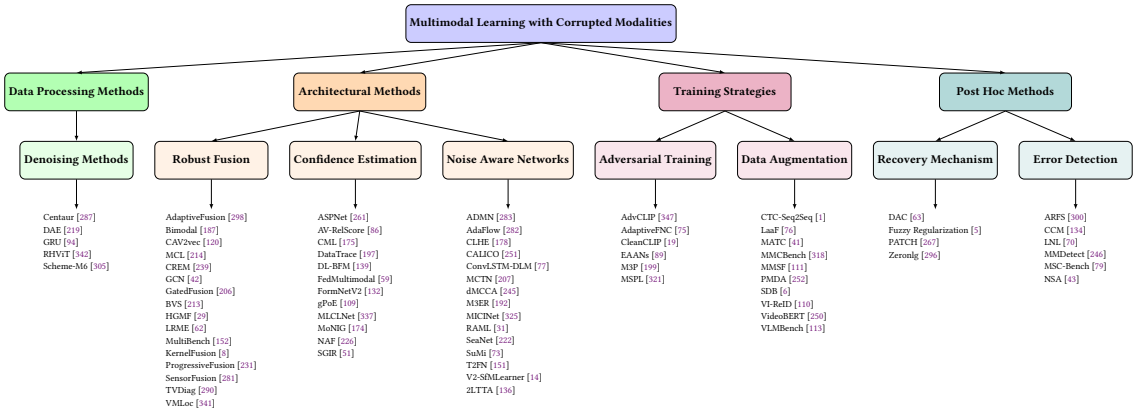

Fig. 3. Taxonomy of the existing works addressing multimodal learning (MML) with corrupted modality problem (CMP)

5.2 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 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.

5.2.1 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 6b highlights the existing studies that leverage novel design choices to withstand data corruption in multimodal networks. For instance, SuMi [73] 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. [325] proposed MICINet, which targets inter-class confusing information (ICI) by identifying and removing structured noise patterns across vision, audio, and text inputs. In network design, ADMN [283] proposed dynamic architectural adjustments in response to varying levels of Gaussian noise across RGB, depth, radar, audio, and WiFi inputs. Lei et al. [136] explored model parameter manipulation for test-time robustness through a two-level adaptation strategy. Leveraging contrastive learning, Ma et el. [178] proposed learning with noisy and sparse multimodal data in recommendation systems. Additional strategies including tensor rank regularization [151], cyclic translation between modalities [207], and cross-modal correlation learning [245] 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.

5.2.2 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

Table 6. Comparison of methods for CMP (I)

(a) Robust fusion methods for CMP

Method	Modalities	Application Area
AFusion [298]	Images, Sensor	Emotion Detection
Bimodal [187]	Speech, Facial Images	Smart Surveillance
CAV2vec [120]	Audio, Video	Speech Recognition
MCL [214]	Gas Sensor, Images	Industry 5.0
CREM [239]	Optical Imagery, DSMs	Remote Sensing
GCN [42]	Images, Text	Scene Recognition
GatedFusion [206]	Images, LiDAR	Object Detection
BVS [213]	Images, Sensor	Biometric Security
HGMF [29]	Images, Audio, Video	Data Mining
LRME [62]	Sensor, Time-Series	Industry 5.0
MultiBench [152]	Text, Images, Video, Audio	HCI, Healthcare
KernelFusion [8]	Image, Text, Audio, Video	Content Analysis
PFusion [231]	Text, Images, Audio, Sensor	Sentiment Analysis
SensorFusion [281]	LiDAR, Images	Autonomous Driving
TVDiag [290]	Logs, Metrics, Traces	Failure Diagnosis
VMLoc [341]	RGB Images, Depth Maps	Robotics

(b) Noise-Aware methods for CMP

Method	Modalities	Application Area
ADMN [283]	Images, Audio, Sensor	Object Localization
AdaFlow [282]	Camera, LiDAR, GPS	Mobile Sensing
CLHE [178]	Text, tabular data	E-commerce
CALICO [251]	Camera, LiDAR	3D Object Detection
ConvLSTM-DLM [77]	Images	Medical Imaging
MCTN [207]	Text, Video, Audio	Sentiment Analysis
dMCCA [245]	Images, Text	Multimodal Learning
M3ER [192]	Images, Text, Audio	Emotion Recognition
MICINet [325]	Images, Text, Audio	Security
RAML [31]	Video, Text, Audio	Emotion Recognition
SeaNet [222]	Audio, Accelerometer	Speech Enhancement
SuMi [73]	Images, Text, Audio	Test-Time Adaptation
T2FN [151]	Text, Video, Audio	Speech Recognition
V2-SfMLearner [14]	Images, Vibration	Robotics
2LTTA [136]	Images, Text, Audio	Test-Time Adaptation

(c) Augmentation-based methods for CMP

Method	Modalities	Application Area
CTC-Seq2Seq [1]	Audio, Video	Speech Recognition
LaaF [76]	MRI, Clinical Data	Medical Imaging
MATC [41]	Text, Audio	Emotion Recognition
MMCBench [318]	Images, Text, Audio	Multimodal Classification
MMSF [111]	Images, Infrared	Smart Surveillance
PMDA [252]	Images, Infrared	Smart Surveillance
SDB [6]	Images, Sensor	Industry 5.0
VI-ReID [110]	Images, Infrared	Surveillance Systems
VideoBERT [250]	Audio, Video	Action Recognition
VLMBench [113]	Text, Images	Long Form Generation

(d) Confidence estimation methods for CMP

Method	Modalities	Application Area
ASPNet [261]	Video, Sensor	Action Segmentation
AV-RelScore [86]	Images, Audio	Speech Recognition
CML [175]	Video, Text, Audio	Classification
DataTrace [197]	Text, Sensor	Anomaly Detection
DL-BFM [139]	Text, Audio, Sensor	Health Monitoring
FedMultimodal [59]	Images, Text, Audio	Activity Recognition
FormNetV2 [132]	Text, Layout, Image	Document Analysis
gPoE [109]	Images, Sensor	Video Segmentation
MLCLNet [337]	Images, Sensor	Medical AI
MoNIG [174]	Images, Text	Sentiment Analysis
NAF [226]	Video, Audio	Speech Extraction
SGIR [51]	Images, Text, Audio	Classification

prediction. Table 6d summarizes the recent studies leveraging diverse strategies for confidence estimation for CMP. For instance, Li et al. [139] 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. [337] proposed MLCLNet for multi-level confidence estimation to assess both feature and label reliability. Similarly, Ma et al. [174] 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) [175] introduced a regularization method that mitigates compromised predictions in the presence of corrupted or missing modalities. FedMultimodal [59] explored federated learning with structured and unstructured noise and emphasized on the need for calibrated confidence signals across distributed and unreliable input sources. Hong et al. [86] proposed AV-RelScore, a reliability scoring module for audio-visual speech recognition (AVSR). Similarly, Ding et al. [51] proposed Star Graph Interaction for confidence-based weighting via centralized hubs and private pathways to maintain robustness under noisy conditions. In document understanding, FormNetV2 [132] 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.

5.2.3 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 6a

highlights the recent works that leverage fusion for CMP. For instance, Kim et al. [120] 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. [239] 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. [206] 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, Fu et al. [62] applied a low-rank multi-manifold embedding approach for monitoring incomplete sensor readings Rahate et al. [214] 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. [29] 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. [298] proposed Adaptive Multimodal Fusion (AMF) that dynamically adjusts the architectural weights to handle modality-specific occlusions and improve expression recognition. Similarly, Raghavendra et al. [213] 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.

5.3 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.

5.3.1 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 [250]. For speech recognition, time masking and additive noise improve model robustness to environmental corruptions [1]. In multimodal sentiment analysis, partial masking of audio or textual inputs has been employed to force the model to rely on complementary visual signals [41].

Recent studies have expanded corruption-based data augmentation to increasingly complex and realistic multimodal settings as shown in Table 6c. Kaplan et al. [113] 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 systematic textual noise. Zhang et al. [318] 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. [6] leveraged augmentation techniques specifically tailored to realistic sensor failures. In medical imaging, Hager et al. [76] 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.

Table 7. Comparison of methods for CMP (II)

(a) Denoising methods for CMP

Method	Modalities	Application Area
Centaur [287]	Sensors	Activity Recognition
DAE [219]	Images, acoustic bathymetry	Autonomous Vehicles
GRU [94]	vCPU, memory, storage	Resource Forecasting
RHViT [342]	3D multimodal MRI	Medical Imaging
Scheme-M6 [305]	Images	Video Classification

(b) Adversarial training methods for CMP

Method	Modalities	Application Area
AdvCLIP [347]	Image, Text	Contrastive Learning
AdaptiveFNC [75]	Image, Video	Video Classification
CleanCLIP [19]	Image, Text	Adversarial Learning
EAANs [89]	Image, Video, Text	Multimodal Classification
M3P [199]	Text, Image	Captioning, Retrieval
MSPL [321]	Images	Image Classification

(c) Error detection methods for CMP

Method	Modalities	Application Area
ARFS [300]	Images, Audio, Text	Multimodal Learning
CCM [134]	Images, Depth, Sensor	Autonomous Vehicles
LNL [70]	Images, Text	Language Understanding
MMDetect [246]	Images, Text	LLM Evaluation
MSC-Bench [79]	Camera, LiDAR	3D Object Detection
NSA [43]	Images	Activity Recognition

(d) Recovery based methods for CMP

Method	Modalities	Application Area
DAC [63]	Images, 3D Point Clouds	Cross-Modal Retrieval
MMFF [5]	Image, Time Series, Sensor	Industrial Automation
PATCH [267]	Sensor, Video, Infrared, Thermal	Autonomous Driving
Zeronlg [296]	Images, Videos, Text	Video Captioning

5.3.2 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 7b 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. [89] 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* [19] specifically targets data poisoning attacks in multimodal contrastive learning, where poisoned images are introduced to implant backdoors in vision-language models. Similarly, *AdvCLIP* [347] leveraged artificially curated adversarial examples that target both image and text modalities. Similarly, Gupta et al. [75] 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.

5.4 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.

5.4.1 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 7c presents the recent studies that have explored various post-hoc error detection techniques for CMP. For instance, Gou et al. [70] 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. [134] proposed a "Detect, Reject, Correct" framework for detecting corrupted sensor modalities by evaluating reconstruction inconsistencies. Cross et al. [43] 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.

5.4.2 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 7d, 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. [63] leveraged a divide-and-conquer method that mitigates the effects of noisy labels in 2D-3D retrieval via credibility modeling. Similarly, PATCH [267] 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. [5] leveraged fuzzy regularization to handle CMP. Yang et al. [296] 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. However, 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.

5.5 Summary

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 [94, 287] 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 [73, 175, 198]. These methods demonstrate strong performance in critical domains such as medical imaging [342] and autonomous driving [283], where runtime resilience is essential. Nonetheless, they introduce higher architectural complexity and require accurate noise modeling which remains a key challenge particularly in multimodal 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 [1, 111], while adversarial training strategies [19, 347] 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 [70, 134, 267], 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 [23, 318] hinders consistent evaluation across tasks and domains.

Lastly, several datasets have been developed to evaluate multimodal robustness under corruption and noise. Datasets such as ImageNet-C [83], COCO-C [189], and PASCAL-C [189] introduce systematic perturbations to assess model stability against visual degradation. These resources, summarized in Table 5b, enable consistent assessment of corruption-aware multimodal networks.

6 Challenges & Future Research

In this section, we synthesize the research gaps and outline concrete, evidence-based future directions.

6.1 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 [292], often suffer from computational bloat. For instance, AMM-Diff [115] and OPTIMUS [50] 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 [303], cross-domain pretraining on heterogeneous multimodal datasets could improve transferability. Finally, benchmarking models under varying computational cost would provide clearer cost-benefit trade-offs.

6.2 Modality Drift and Misalignment:

Han et al. [78] and Zhang et al. [325] are examples of alignment-based models that show performance deterioration when modalities are asynchronous or semantically divergent. Similarly, many architectures such as M3Care [317] 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 [73]. Similarly, interested researchers can investigate meta-learning strategies that can adapt alignment functions dynamically under the contextual shifts in semantic meaning.

6.3 Adversarial Robustness:

While adversarial robustness has been addressed in contrastive learning frameworks including CleanCLIP [19] and AdvCLIP [347], it is often confined to vision-language pairs and fail under open-set corruptions or cross-modal perturbations. Zhang et al. [318] 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 [75].

6.4 Standardized Benchmarks and Taxonomies:

Many studies use custom or synthetic corruption setups, making cross-comparison difficult. While recent work such as Beemelmanns et al. [23], Zhang et al. [318], Dong et al. [55], and Yu et al. [309] 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.

6.5 Interpretability and Explainability:

Despite exceptional success across multiple application areas such as Cross-modal Information Retrieval [278], Multimodal Emotion Recognition [40], Sentiment Analysis [304], and Genere Classification [176], 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 [266], compensated for [291], or hallucinated [15].

6.6 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 [86] 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.

7 Conclusion

In this survey, 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 present the open issues, benchmark datasets and future directions to enhance robustness to compromised modalities. In essence, this survey serves as a valuable resource for researchers aiming to deepen their understanding and strengthen the robustness of multimodal learning.

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. *IEEE TPAMI* 44, 12 (2018), 8717–8727.

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In *CVPR*. 4971–4980.

[3] Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. CrisisMMD: Multimodal Twitter datasets from natural disasters. In *Proc. 12th International AAAI Conference on Web and Social Media (ICWSM)*. 465–472.

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *NeurIPS* 35 (2022), 23716–23736.

[5] Diyar Altinses and Andreas Schwung. 2023. Deep Multimodal Fusion with Corrupted Spatio-Temporal Data Using Fuzzy Regularization. In *IECON*. IEEE, 1–7.

[6] Diyar Altinses and Andreas Schwung. 2023. Multimodal Synthetic Dataset Balancing: A Framework for Realistic and Balanced Training Data Generation in Industrial Settings. In *IECON*. IEEE, 1–7.

[7] Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. 2022. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. *Proc.ACL* (2022), 5723–5738.

[8] Hrishikesh Aradhye and Chitra Dorai. 2002. New kernels for analyzing multimodal data in multimedia using kernel machines. In *Proceedings. IEEE International Conference on Multimedia and Expo*, Vol. 2. IEEE, 37–40.

[9] Rosana Ardila, Megan Branson, Kelly Davis, and *et al.* 2020. Common Voice: A massively-multilingual speech corpus. *arXiv:1912.06670* (2020).

[10] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. *arXiv:1702.01992* (2017).

[11] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems* 16 (2010), 345–379.

[12] Reza Azad, Nika Khosravi, and Dorit Merhof. 2022. SMU-Net: Style matching U-Net for brain tumor segmentation with missing modalities. In *International Conference on Medical Imaging with Deep Learning*. PMLR, 48–62.

[13] Amir Ali Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *ACL*. 2236–2246.

[14] Long Bai, Beilei Cui, Liangyu Wang, Yanheng Li, Shilong Yao, Sishen Yuan, Yanan Wu, Yang Zhang, Max Q-H Meng, Zhen Li, et al. 2025. V 2-SfMLearner: Learning Monocular Depth and Ego-motion for Multimodal Wireless Capsule Endoscopy. *IEEE T-ASE* (2025).

[15] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv:2404.18930* (2024).

[16] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in Time: A joint video and image encoder for end-to-end retrieval. In *ICCV*. 1728–1738.

[17] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv:1811.02629* (2018).

[18] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI* 41, 2 (2018), 423–443.

[19] Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. 2023. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. In *IEEE/CVF ICCV*. 112–123.

[20] Guangyin Bao, Qi Zhang, Duoqian Miao, Zixuan Gong, Liang Hu, Ke Liu, Yang Liu, and Chongyang Shi. 2023. Multimodal federated learning with missing modality via prototype mask and contrast. *arXiv:2312.13508* (2023).

[21] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. BEiT: BERT pre-training of image transformers. *arXiv:2106.08254* (2021).

[22] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. 2022. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer* 38, 8 (2022), 2939–2970.

[23] Till Beemelmanns, Quan Zhang, Christian Geller, and Lutz Eckstein. 2024. Multicorrupt: A multi-modal robustness dataset and benchmark of lidar-camera fusion for 3d object detection. In *IVS*. IEEE, 3255–3261.

[24] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, and *et al.* 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources & Evaluation* 42, 4 (2008), 335–359.

[25] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proc. ICMI*. 205–211.

[26] Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. 2018. Deep adversarial learning for multi-modality missing data completion. In *Proc. 24th ACM SIGKDD*. 1158–1166.

[27] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*. 6299–6308.

[28] Soravit Changpinyo, Parthasarathi Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*. 3558–3568.

[29] Jiayi Chen and Aidong Zhang. 2020. Hgmf: heterogeneous graph-based fusion for multimodal data with incompleteness. In *Proc. 26th ACM SIGKDD*. 1295–1305.

[30] Minghui Chen, Zhiqiang Wang, and Feng Zheng. 2021. Benchmarks for Corruption Invariant Person Re-identification. *arXiv:2111.00880* (2021).

[31] Mengxi Chen, Jiangchao Yao, Linyu Xing, Yu Wang, Ya Zhang, and Yanfeng Wang. 2023. Redundancy-adaptive multimodal learning for imperfect data. *arXiv:2310.14496* (2023).

[32] Shizhe Chen, Qin Jin, Pan Wang, and Qi Wu. 2021. Learning with noisy correspondence for cross-modal matching. In *NeurIPS*. 29406–29419.

[33] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv:1504.00325* (2015).

[34] Yuanyuan Chen, Yongsheng Pan, Yong Xia, and Yixuan Yuan. 2023. Disentangle first, then distill: A unified framework for missing modality imputation and Alzheimer's disease diagnosis. *IEEE Transactions on Medical Imaging* (2023).

[35] Yen-Chun Chen, Linjie Li, Licheng Yu, and *et al.* 2020. UNITER: UNiversal Image-TExt Representation Learning. *ECCV* (2020), 104–120.

[36] Zhuo Chen, Lingbing Guo, Yin Fang, Yichi Zhang, Jiaoyan Chen, Jeff Z Pan, Yangning Li, Huajun Chen, and Wen Zhang. 2023. Rethinking uncertainly missing and ambiguous visual modality in multi-modal entity alignment. In *International Semantic Web Conference*. Springer, 121–139.

[37] Zhengyang Chen, Shuai Wang, and Yanmin Qian. 2020. Multi-Modality Matters: A Performance Leap on VoxCeleb.. In *Interspeech*. 2252–2256.

[38] Haozhe Chi, Minghua Yang, Junhao Zhu, Guanhong Wang, and Gaoang Wang. 2022. Missing modality meets meta sampling (M3S): An efficient universal approach for multimodal sentiment analysis with missing modality. *arXiv:2210.03428* (2022).

[39] Jae Won Cho, Dong-Jin Kim, Jinsoo Choi, Yunjae Jung, and In So Kweon. 2021. Dealing with missing modalities in the visual question answer-difference prediction task through knowledge distillation. In *CVPR*. 1592–1601.

[40] Yucel Cimtay, Erhan Ekmekcioglu, and Seyma Caglar-Ozhan. 2020. Cross-subject multimodal emotion recognition based on hybrid fusion. *IEEE Access* 8 (2020), 168865–168878.

[41] Dror Cohen, Ido Rosenberger, Moshe Butman, and Kfir Bar. 2023. Masking important information to assess the robustness of a multimodal classifier for emotion recognition. *Frontiers in Artificial Intelligence* 6 (2023), 1091443.

[42] Willams de Lima Costa, Raul Ismayilov, Nicola Strisciuglio, and Estefania Talavera Martinez. 2024. Indoor scene recognition from images under visual corruptions. *arXiv:2408.13029* (2024).

[43] Mattias Cross and Valentin Radu. 2021. An immune inspired algorithm for fault tolerant enhanced multimodal machine learning. In *BIBM*. IEEE, 2194–2200.

[44] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. 2024. A survey on multimodal large language models for autonomous driving. In *WACV*. 958–979.

[45] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *arXiv:2305.06500* (2023).

[46] Kasra Darvish, Edward Raff, Francis Ferraro, Cynthia Matuszek, et al. 2023. Multimodal Language Learning for Object Retrieval in Low Data Regimes in the Face of Missing Modalities. *Transactions on Machine Learning Research* (2023).

[47] Ruben Delgado-Escano, Francisco M Castro, Nicolás Guil, Vicky Kalogeiton, and Manuel J Marin-Jimenez. 2021. Multimodal gait recognition under missing modalities. In *ICIP*. IEEE, 3003–3007.

[48] SP Dhivyaa, Duy-Phuong Dao, Hyung-Jeong Yang, and Jahae Kim. 2025. Adaptive Cross-Modal Representation Learning for Heterogeneous Data Types in Alzheimer Disease Progression Prediction with Missing Time Point and Modalities. In *ICPR*. Springer, 267–282.

[49] Yueqin Diao, Huihui Fang, Hanyi Yu, Fan Li, and Yanwu Xu. 2025. Multimodal invariant feature prompt network for brain tumor segmentation with missing modalities. *Neurocomputing* 616 (2025), 128847.

[50] Christelle Schneuwly Diaz, Duy-Thanh Vu, Julien Bodelet, Duy-Cat Can, Guillaume Blanc, Haiting Jiang, Lin Yao, Pantaleo, et al. 2025. OPTIMUS: Predicting Multivariate Outcomes in Alzheimer's Disease Using Multi-modal Data amidst Missing Values. *arXiv:2503.11282* (2025).

[51] Zhiwei Ding, Guilin Lan, Yanzhi Song, and Zhouwang Yang. 2023. SGIR: Star Graph-Based Interaction for Efficient and Robust Multimodal Representation. *IEEE Transactions on Multimedia* 26 (2023), 4217–4229.

[52] Zhengming Ding, Shao Ming, and Yun Fu. 2014. Latent low-rank transfer subspace learning for missing modality recognition. In *AAAI*, Vol. 28.

[53] Zhengming Ding, Ming Shao, and Yun Fu. 2015. Missing modality transfer learning via latent low-rank constraint. *IEEE transactions on Image Processing* 24, 11 (2015), 4322–4334.

[54] Chhavi Dixit and Shashank Mouli Satapathy. 2024. Deep CNN with late fusion for real time multimodal emotion recognition. *Expert Systems with Applications* (2024).

[55] Yinpeng Dong, Caixin Kang, Jinlai Zhang, Zijian Zhu, Yikai Wang, Xiao Yang, Hang Su, Xingxing Wei, and Jun Zhu. 2023. Benchmarking Robustness of 3D Object Detection to Common Corruptions. In *CVPR*. 1022–1032.

[56] Changde Du, Changying Du, Hao Wang, Jinpeng Li, Wei-Long Zheng, Bao-Liang Lu, and Huiguang He. 2018. Semi-supervised deep generative modelling of incomplete multi-modality emotional data. In *ACM Multimedia*. 108–116.

[57] Siyi Du, Shaoming Zheng, Yinsong Wang, Wenjia Bai, Declan P O'Regan, and Chen Qin. 2024. TIP: Tabular-image pre-training for multimodal classification with incomplete data. In *ECCV*. Springer, 478–496.

[58] Jinxiao Fan, Mengshi Qi, Liang Liu, and Huadong Ma. 2025. Diffusion-driven Incomplete Multimodal Learning for Air Quality Prediction. *ACM Transactions on Internet of Things* 6, 1 (2025), 1–24.

[59] Tiantian Feng, Digbalay Bose, Tuo Zhang, Rajat Hebbar, Anil Ramakrishna, Rahul Gupta, Mi Zhang, Salman Avestimehr, and Shrikanth Narayanan. 2023. Fedmultimodal: A benchmark for multimodal federated learning. In *Proc.ACM SIGKDD*. 4035–4045.

[60] Tiantian Feng, Tuo Zhang, Salman Avestimehr, and Shrikanth Narayanan. 2025. 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. 78–83.

[61] Fangze Fu, Wei Ai, Fan Yang, Yuntao Shou, Tao Meng, and Keqin Li. 2025. SDR-GNN: Spectral Domain Reconstruction Graph Neural Network for incomplete multimodal learning in conversational emotion recognition. *Knowledge-Based Systems* 309 (2025), 112825.

[62] Yuanjian Fu, Jinliang Ding, and Xue Xu. 2023. Low-rank multimanifold embedding learning for multimode process monitoring. *IEEE Transactions on Industrial Informatics* 20, 3 (2023), 3468–3477.

[63] Chaofan Gan, Yuanpeng Tu, Yuxi Li, and Weiyao Lin. 2024. DAC: 2D-3D Retrieval with Noisy Labels via Divide-and-Conquer Alignment and Correction. In *ACM Multimedia*. 4217–4226.

[64] Christian Ganhör, Marta Moscati, Anna Hausberger, Shah Nawaz, and Markus Schedl. 2024. A multimodal single-branch embedding network for recommendation in cold-start and missing modality scenarios. In *Proc. 18th ACM Conference on Recommender Systems*. 380–390.

[65] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. 2020. A survey on deep learning for multimodal data fusion. *Neural Computation* 32, 5 (2020), 829–864.

[66] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. 2018. Modality distillation with multiple stream networks for action recognition. In *ECCV*. 103–118.

[67] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. 2019. Cross-modal Learning by Hallucinating Missing Modalities in RGB-D Vision. In *Multimodal Scene Understanding*. Elsevier, 383–401.

[68] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *ICASSP*. IEEE, 776–780.

[69] Aidin Gharahdaghi, Farbod Razzazi, and Arash Amini. 2021. A non-linear mapping representing human action recognition under missing modality problem in video data. *Measurement* 186 (2021), 110123.

[70] Yunhao Gou, Hansi Yang, Zhili Liu, Kai Chen, Yihan Zeng, Lanqing Hong, Zhenguo Li, Qun Liu, James T Kwok, and Yu Zhang. 2025. Corrupted but Not Broken: Rethinking the Impact of Corrupted Data in Visual Instruction Tuning. *arXiv:2502.12635* (2025).

[71] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*. 6904–6913.

[72] Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep multimodal representation learning: A survey. *Ieee Access* 7 (2019), 63373–63394.

[73] Zirun Guo and Tao Jin. 2025. Smoothing the shift: Towards stable test-time adaptation under complex multimodal noises. *arXiv:2503.02616* (2025).

[74] Zirun Guo, Shulei Wang, Wang Lin, Weicai Yan, Yangyang Wu, and Tao Jin. 2025. Efficient prompting for continual adaptation to missing modalities. *arXiv:2503.00528* (2025).

[75] Rohit Gupta, Naveed Akhtar, Ajmal Mian, and Mubarak Shah. 2023. Contrastive self-supervised learning leads to higher adversarial susceptibility. In *AAAI*, Vol. 37. 14838–14846.

[76] Paul Hager, Martin J Menten, and Daniel Rueckert. 2023. Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In *CVPR*. 23924–23935.

[77] Mohamed Hammad, Lo'ai Tawalbeh, Abdullah M Iliyasu, Ahmed Sedik, Fathi E Abd El-Samie, Monagi H Alkinani, and Ahmed A Abd El-Latif. 2022. Efficient multimodal deep-learning-based COVID-19 diagnostic system for noisy and corrupted images. *Journal of King Saud University-Science* 34, 3 (2022), 101898.

[78] Wei Han, Hui Chen, Min-Yen Kan, and Soujanya Poria. 2022. Mm-align: Learning optimal transport-based alignment dynamics for fast and accurate inference on missing modality sequences. *arXiv:2210.12798* (2022).

[79] Xiaoshuai Hao, Guanqun Liu, Yuting Zhao, Yuheng Ji, Mengchuan Wei, Haimei Zhao, Lingdong Kong, Rong Yin, and Yu Liu. 2025. MSC-Bench: Benchmarking and Analyzing Multi-Sensor Corruption for Driving Perception. *arXiv:2501.01037* (2025).

[80] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-Invariant and-Specific Representations for Multimodal Sentiment Analysis. In *ACM MM*. 1122–1131.

[81] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *CVPR*. 16000–16009.

[82] Farnoosh Heidarivincheh, Ryan McConville, Catherine Morgan, Roisin McNaney, Alessandro Masullo, Majid Mirmehdi, Alan L Whone, and Ian Craddock. 2021. Multimodal classification of parkinson's disease in home environments with resiliency to missing modalities. *Sensors* 21, 12 (2021), 4133.

[83] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions. *arXiv:1903.12261* (2019).

[84] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing*. 7514–7528.

[85] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine* 29, 6 (2012), 82–97.

[86] Joanna Hong, Minsu Kim, Jeongsoo Choi, and Yong Man Ro. 2023. Watch or listen: Robust audio-visual speech recognition with visual corruption modeling and reliability scoring. In *CVPR*. 18783–18794.

[87] Sirshendu Hore and Tanmay Bhattacharya. 2024. Audio-visual expression-based emotion recognition model for neglected people in real-time: a late-fusion approach. *Multimedia Tools and Applications* (2024), 1–39.

[88] Ruohong Huan, Guowei Zhong, Peng Chen, and Ronghua Liang. 2023. Unimf: a unified multimodal framework for multimodal sentiment analysis in missing modalities and unaligned multimodal sequences. *IEEE Transactions on Multimedia* (2023).

[89] Feiran Huang, Alireza Jolfaei, and Ali Kashif Bashir. 2021. Robust multimodal representation learning with evolutionary adversarial attention networks. *IEEE Transactions on Evolutionary Computation* 25, 5 (2021), 856–868.

[90] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *ECCV*. 172–189.

[91] Yupan Huang, Hongwei Xue, Bei Liu, and Yutong Lu. 2021. Unifying multimodal transformer for bi-directional image and text generation. In *Proc. ACM MM*. 1138–1147.

[92] Ziqi Huang, Li Lin, Pujin Cheng, Linkai Peng, and Xiaoying Tang. 2022. Multi-modal brain tumor segmentation via missing modality synthesis and modality-level attention fusion. *arXiv:2203.04586* (2022).

[93] Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*. 6700–6709.

[94] Hamzaoui Ikhlasse, Duthil Benjamin, Courboulay Vincent, and Medromi Hicham. 2022. Multimodal cloud resources utilization forecasting using a Bidirectional Gated Recurrent Unit predictor based on a power efficient Stacked denoising Autoencoders. *Alexandria Engineering Journal* 61, 12 (2022), 11565–11577.

[95] Md Farhan Ishmam, Ishmam Tashdeed, Talukder Asir Saadat, Md Hamjajul Ashmafee, Abu Raihan Mostofa Kamal, and Md Azam Hossain. 2025. Visual robustness benchmark for visual question answering (vqa). In *WACV*. IEEE, 6623–6633.

[96] Mobarakol Islam, Navodini Wijethilake, and Hongliang Ren. 2021. Glioblastoma multiforme prognosis: MRI missing modality generation, segmentation and radiogenomic survival prediction. *Computerized Medical Imaging and Graphics* 91 (2021), 101906.

[97] Phillip Isola, Jun-Yan Lim, and Edward Adelson. 2015. Discovering states and transformations in image collections. In *CVPR*. 1383–1391.

[98] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *Technologies* 9, 1 (2020), 2.

[99] Jaehyuk Jang, Yooseung Wang, and Changick Kim. 2024. Towards Robust Multimodal Prompting with Missing Modalities. In *ICASSP*. IEEE, 8070–8074.

[100] Seungwan Jeong, Hwanho Cho, Junmo Kwon, and Hyunjin Park. 2022. Region-of-interest attentive heteromodal variational encoder-decoder for segmentation with missing modalities. In *Proc. Asian Conference on Computer Vision*. 3707–3723.

[101] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*. 4904–4916.

[102] Xinyang Jiang, Fei Wu, Yin Zhang, Siliang Tang, Weiming Lu, and Yueting Zhuang. 2015. The classification of multi-modal data with hidden conditional random field. *Pattern Recognition Letters* 51 (2015), 63–69.

[103] Tao Jin, Xize Cheng, Linjun Li, Wang Lin, Ye Wang, and Zhou Zhao. 2023. Rethinking Missing Modality Learning from a Decoding Perspective. In *ACM Multimedia*. 4431–4439.

[104] Vijay John and Yasutomo Kawanishi. 2022. A Multimodal Sensor Fusion Framework Robust to Missing Modalities for Person Recognition. In *ACM Multimedia*. 1–5.

[105] Vijay John and Yasutomo Kawanishi. 2023. Audio-Visual Sensor Fusion Framework using Person Attributes Robust to Missing Visual Modality for Person Recognition. In *International Conference on Multimedia Modeling*. Springer, 523–535.

[106] Vijay John and Yasutomo Kawanishi. 2025. Multimodal Cascaded Framework with Multimodal Latent Loss Functions Robust to Missing Modalities. *ACM Transactions on Multimedia Computing, Communications and Applications* (2025).

[107] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

[108] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, and ... 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*. 2901–2910.

[109] Abhinav Joshi, Naman Gupta, Jinang Shah, Binod Bhattarai, Ashutosh Modi, and Danail Stoyanov. 2022. Generalized product-of-experts for learning multimodal representations in noisy environments. In *Proc. ICMI*. 83–93.

[110] Arthur Josi, Mahdi Alehdaghi, Rafael MO Cruz, and Eric Granger. 2023. Multimodal data augmentation for visual-infrared person ReID with corrupted data. In *WACV*. 32–41.

[111] Arthur Josi, Mahdi Alehdaghi, Rafael MO Cruz, and Eric Granger. 2025. Fusion for visual-infrared person ReID in real-world surveillance using corrupted multimodal data. *International Journal of Computer Vision* (2025), 1–22.

[112] Michael Kampffmeyer, Arnt-Børre Salberg, and Robert Jenssen. 2018. 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* 11, 6 (2018), 1758–1768.

[113] Daniel Z Kaplan, Alexis Roger, Mohamed Osman, and Irina Rish. 2024. The Effect of Data Corruption on Multimodal Long Form Responses. In *ICML 2024 Workshop on Foundation Models in the Wild*.

[114] Guanzhou Ke, Shengfeng He, Xiao Li Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Yi Xie, and HeXing Su. 2025. Knowledge Bridger: Towards Training-free Missing Multi-modality Completion. *arXiv:2502.19834* (2025).

[115] Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, Pierre Vera, and Su Ruan. 2025. AMM-Diff: Adaptive Multi-Modality Diffusion Network for Missing Modality Imputation. *arXiv:2501.12840* (2025).

[116] Douwe Kiela and Stephen Clark. 2017. Learning neural audio embeddings for grounding semantics in auditory perception. *Journal of Artificial Intelligence Research* 60 (2017), 1003–1030.

[117] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *NeurIPS* 33 (2020), 2611–2624.

[118] Hai-Dang Kieu, Quang-Thuy Ha, Xuan-Hieu Phan, Duc-Trong Le, et al. 2025. Mi-CGA: Cross-modal Graph Attention Network for robust emotion recognition in the presence of incomplete modalities. *Neurocomputing* 623 (2025), 129342.

[119] Jiwan Kim, Hongseok Kang, Sein Kim, Kibum Kim, and Chanyoung Park. 2025. Disentangling and Generating Modalities for Recommendation in Missing Modality Scenarios. *arXiv:2504.16352* (2025).

[120] Sungnyun Kim, Sungwoo Cho, Sangmin Bae, Kangwook Jang, and Se-Young Yun. 2025. Multi-Task Corrupted Prediction for Learning Robust Audio-Visual Speech Representation. In *The Thirteenth International Conference on Learning Representations*.

[121] Joosung Yoon Koh, Daniel Fried, and Roozbeh Mottaghi. 2021. Zero-shot image-to-text generation for visual-semantic arithmetic. In *CVPR*. 5670–5679.

[122] Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2024. Generating images with multimodal language models. *NeurIPS* 36 (2024).

[123] Aishik Konwer, Chao Chen, and Prateek Prasanna. 2023. MagNET: Modality-Agnostic Network for Brain Tumor Segmentation and Characterization with Missing Modalities. In *International Workshop on Machine Learning in Medical Imaging*. Springer, 361–371.

[124] Aishik Konwer, Xiaoling Hu, Joseph Bae, Xuan Xu, Chao Chen, and Prateek Prasanna. 2023. Enhancing modality-agnostic representations via meta-learning for brain tumor segmentation. In *IEEE/CVF ICCV*. 21415–21425.

[125] Ranjay Krishna, Kenji Hata, Frederic Ren, Juan Carlos Niebles, and Li Fei-Fei. 2017. Dense-captioning events in videos. In *ICCV*. 706–715.

[126] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, ..., and Fei-Fei Li. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV* 123, 1 (2017), 32–73.

[127] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.

[128] Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. 2025. Retrieval-Augmented Dynamic Prompt Tuning for Incomplete Multimodal Learning. *arXiv:2501.01120* (2025).

[129] Kenneth Lau, Jonas Adler, and Jens Sjölund. 2019. A unified representation network for segmentation with missing modalities. *arXiv:1908.06683* (2019).

[130] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. 2023. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. *arXiv:2306.16527* (2023).

[131] Huy Q Le, Chu Myaet Thwal, Yu Qiao, Ye Lin Tun, Minh NH Nguyen, Eui-Nam Huh, and Choong Seon Hong. 2025. Cross-modal prototype based multimodal federated learning under severely missing modality. *Information Fusion* (2025), 103219.

[132] Chen-Yu Lee, Chun-Liang Li, Hao Zhang, Timothy Dozat, Vincent Perot, Guolong Su, Xiang Zhang, Kihyuk Sohn, Nikolai Glushnev, Renshen Wang, et al. 2023. Formnetv2: Multimodal graph contrastive learning for form document information extraction. *arXiv:2305.02549* (2023).

[133] Hu-Cheng Lee, Chih-Yu Lin, Pin-Chun Hsu, and Winston H Hsu. 2019. Audio feature generation for missing modality problem in video action recognition. In *ICASSP*. IEEE, 3956–3960.

[134] Michelle A Lee, Matthew Tan, Yuke Zhu, and Jeannette Bohg. 2021. Detect, reject, correct: Crossmodal compensation of corrupted sensors. In *ICRA*. IEEE, 909–916.

[135] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. 2023. Multimodal prompting with missing modalities for visual recognition. In *CVPR*. 14943–14952.

[136] Jixiang Lei and Franz Pernkopf. 2024. Two-level test-time adaptation in multimodal learning. In *ICML 2024 Workshop on Foundation Models in the Wild*.

[137] Baoli Li and Liping Han. 2013. Distance weighted cosine similarity measure for text classification. In *IDEAL*. Springer, 611–618.

[138] Chunyi Li, Jianbo Zhang, Zicheng Zhang, Haoning Wu, Yuan Tian, Wei Sun, Guo Lu, Xiongkuo Min, Xiaohong Liu, Weisi Lin, et al. 2025. R-Bench: Are your Large Multimodal Model Robust to Real-world Corruptions? *IEEE Journal of Selected Topics in Signal Processing* (2025).

[139] Jialin Li, Muhammad Azeem Akbar, Syed Hassan Shah, Zhi Wang, and Jing Yang. 2025. Deep Learning-Driven Behavioral Modeling in IoST for Mental Health Monitoring and Intervention. *IEEE Transactions on Computational Social Systems* (2025).

[140] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*. 19730–19742.

[141] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*. 12888–12900.

[142] Jiayao Li, Li Li, Ruizhi Sun, Gang Yuan, Shufan Wang, and Shulin Sun. 2024. MMAN-M2: Multiple multi-head attentions network based on encoder with missing modalities. *Pattern Recognition Letters* 177 (2024), 110–120.

[143] Junnan Li, Ramprasaath Selvaraju, Anirudh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, Vol. 34. 9694–9705.

[144] Linchao Li, Bowen Du, Yonggang Wang, Lingqiao Qin, and Huachun Tan. 2020. Estimation of missing values in heterogeneous traffic data: Application of multimodal deep learning model. *Knowledge-Based Systems* 194 (2020), 105592.

[145] Mingyang Li, Shao-Lun Huang, and Lin Zhang. 2022. A general framework for incomplete cross-modal retrieval with missing labels and missing modalities. In *ICASSP*. IEEE, 4763–4767.

[146] Siting Li, Chenzhuang Du, Yue Zhao, Yu Huang, and Hang Zhao. 2023. What Makes for Robust Multi-Modal Models in the Face of Missing Modalities? *arXiv:2310.06383* (2023).

[147] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. OSCAR: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*. 121–137.

[148] Yihao Li, Mostafa El Habib Daho, Pierre-Henri Conze, Rachid Zeghlache, Hugo Le Boité, Ramin Tadayoni, Béatrice Cochener, Mathieu Lamard, and Gwenolé Quellec. 2024. A review of deep learning-based information fusion techniques for multimodal medical image classification. *Computers in Biology and Medicine* (2024), 108635.

[149] Yixuan Li, Xinyu Zhang, Yifan Wang, Yuhang Hu, Zhen Li, Yifei Liu, Zhen Wang, Yadong Wang, and Guoying Zhang. 2022. M3ED: Multi-modal multi-scene emotional dataset for affective computing. *IEEE Transactions on Affective Computing* (2022).

[150] Zhiyuan Li, Yafei Zhang, Huafeng Li, Yi Chai, and Yushi Yang. 2024. Deformation-aware and reconstruction-driven multimodal representation learning for brain tumor segmentation with missing modalities. *BSPC* 91 (2024), 106012.

[151] Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2019. Learning representations from imperfect time series data via tensor rank regularization. *arXiv:1907.01011* (2019).

[152] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. 2021. Multibench: Multiscale benchmarks for multimodal representation learning. *NeurIPS* 2021, DB1 (2021), 1.

[153] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2024. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *Comput. Surveys* 56, 10 (2024), 1–42.

[154] Yuan Liang. 2024. Multimodal Knowledge Graph Embedding With Missing Data Integration. *IEEE Transactions on Computational Social Systems* (2024).

[155] Muhammad Irzam Liaqat, Shah Nawaz, Muhammad Zaigham Zaheer, Muhammad Saad Saeed, Hassan Sajjad, Tom De Schepper, Karthik Nandakumar, Muhammad Haris Khan, Ignazio Gallo, and Markus Schedl. 2025. Chameleon: A Multimodal Learning Framework Robust to Missing Modalities. *International Journal of Multimedia Information Retrieval* 14, 2 (2025), 21.

[156] Ronghao Lin and Haifeng Hu. 2023. MissModal: Increasing Robustness to Missing Modality in Multimodal Sentiment Analysis. *Transactions ACL* 11 (2023), 1686–1702.

[157] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. *ECCV* (2014), 740–755.

[158] Yung-Bo Lin, Yi-Hsuan Tsai, Kuan-Wei Cheng, and Hung-Yi Lee. 2023. LAVISH: Language-vision instruction tuning with home videos. *arXiv:2305.14837* (2023).

[159] Chengzhi Liu, Zile Huang, Zhe Chen, Feilong Tang, Yu Tian, Zhongxing Xu, Zihong Luo, Yalin Zheng, and Yanda Meng. 2025. Incomplete modality disentangled representation for ophthalmic disease grading and diagnosis. In *AAAI*, Vol. 39. 5361–5369.

[160] Han Liu, Yubo Fan, Hao Li, Jiacheng Wang, Dewei Hu, Can Cui, Ho Hin Lee, Huahong Zhang, and Ipek Oguz. 2022. Moddrop++: A dynamic filter network with intra-subject co-training for multiple sclerosis lesion segmentation with missing modalities. In *MICCAI*. Springer, 444–453.

[161] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv:2304.08485* (2023).

[162] Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. 2023. M3AE: multimodal representation learning for brain tumor segmentation with missing modalities. In *AAAI*, Vol. 37. 1657–1665.

[163] Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2023. Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities. *Journal of Biomedical Informatics* 145 (2023), 104466.

[164] Qi Liu and Wanjing Ma. 2024. Navigating Data Corruption in Machine Learning: Balancing Quality, Quantity, and Imputation Strategies. *arXiv:2412.18296* (2024).

[165] Tengfei Liu, Yongli Hu, Mingjie Li, Junfei Yi, Xiaojun Chang, Junbin Gao, and Baocai Yin. 2025. Tackling Real-world Complexity: Hierarchical Modeling and Dynamic Prompting for Multimodal Long Document Classification. *IEEE Transactions CSVT* (2025).

[166] Yi Liu, Cong Wang, and Xingliang Yuan. 2025. FedMobile: Enabling Knowledge Contribution-aware Multi-modal Federated Learning with Incomplete Modalities. In *Proc. ACM on Web Conference 2025*. 2775–2786.

[167] Zhizhong Liu, Bin Zhou, Dianhui Chu, Yuhang Sun, and Lingqiang Meng. 2024. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion* 101 (2024), 101973.

[168] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60 (2004), 91–110.

[169] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, Vol. 32.

[170] Huaishao Luo, Lin Ji, Mengchen Zhong, and *et al.* 2021. CLIP4CLIP: An empirical study of CLIP for end-to-end video clip retrieval. *arXiv:2104.08860* (2021).

[171] Wei Luo, Mengying Xu, and Hanjiang Lai. 2023. Multimodal reconstruct and align net for missing modality problem in sentiment analysis. In *International Conference on Multimedia Modeling*. Springer, 411–422.

[172] Fei Ma, Shao-Lun Huang, and Lin Zhang. 2021. 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, 1–6.

[173] Fei Ma, Xiangxiang Xu, Shao-Lun Huang, and Lin Zhang. 2021. Maximum likelihood estimation for multimodal learning with missing modality. *arXiv:2108.10513* (2021).

[174] Huan Ma, Zongbo Han, Changqing Zhang, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. 2021. Trustworthy multimodal regression with mixture of normal-inverse gamma distributions. *NeurIPS* 34 (2021), 6881–6893.

[175] Huan Ma, Qingyang Zhang, Changqing Zhang, Bingzhe Wu, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. 2023. Calibrating multimodal learning. In *ICML*. PMLR, 23429–23450.

[176] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality?. In *CVPR*. 18177–18186.

[177] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. 2021. Smil: Multimodal learning with severely missing modality. In *AAAI*, Vol. 35. 2302–2310.

[178] Yunshan Ma, Xiaohao Liu, Yinwei Wei, Zhulin Tao, Xiang Wang, and Tat-Seng Chua. 2024. Leveraging multimodal features and item-level user feedback for bundle construction. In *Proc. 17th ACM International Conference on Web Search and Data Mining*. 510–519.

[179] Ziyu Ma, Fuyan Ma, Bin Sun, and Shutao Li. 2021. Hybrid mutimodal fusion for dimensional emotion recognition. In *Proc. 2nd on Multimodal Sentiment Analysis Challenge*. 29–36.

[180] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-ChatGPT: Towards detailed video understanding via large vision and language models. *arXiv:2306.05424* (2023).

[181] Harsh Maheshwari, Yen-Cheng Liu, and Zsolt Kira. 2024. Missing modality robustness in semi-supervised multi-modal semantic segmentation. In *WACV*. 1020–1030.

[182] Sayan Maity, Mohamed Abdel-Mottaleb, and Shihab S Asfour. 2020. Multimodal biometrics recognition from facial video with missing modalities using deep learning. *Journal of Information Processing Systems* 16, 1 (2020), 6–29.

[183] Daniele Malitesta, Emanuele Rossi, Claudio Pomo, Tommaso Di Noia, and Fragkiskos D Malliaros. 2024. Do We Really Need to Drop Items with Missing Modalities in Multimodal Recommendation?. In *Proc. CIKM*. 3943–3948.

[184] Manuel J Marín-Jiménez, Francisco M Castro, Rubén Delgado-Escaño, Vicky Kalogeiton, and Nicolas Guil. 2021. UGaitNet: multimodal gait recognition with missing input modalities. *IEEE TIFS* 16 (2021), 5452–5462.

[185] Toshihiko Matsuura, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Generalized bayesian canonical correlation analysis with missing modalities. In *ECCV*. 0–0.

[186] Brandon McKinzie, Vaishaal Shankar, Joseph Yitan Cheng, Yinfei Yang, Jonathon Shlens, and Alexander T Toshev. 2023. Robustness in multimodal learning under train-test modality mismatch. In *ICML*. PMLR, 24291–24303.

[187] Niall McLaughlin, Ji Ming, and Danny Crookes. 2013. Robust multimodal person identification with limited training data. *IEEE Transactions on Human-Machine Systems* 43, 2 (2013), 214–224.

[188] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2021. Detection and classification of acoustic scenes and events: Outcome of the DCASE 2019 challenge. *IEEE/ACM TASLP* 29 (2021), 2453–2471.

[189] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. 2019. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv:1907.07484* (2019).

[190] Antoine Miech, Dmitry Zhukov, Jean-Baptiste Alayrac, and *et al.* 2019. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*. 2630–2640.

[191] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *NeurIPS* 26 (2013).

[192] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *AAAI*, Vol. 34. 1359–1367.

[193] Tiago Mota, M Rita Verdelho, Diogo J Araújo, Alceu Bissoto, Carlos Santiago, and Catarina Barata. 2024. MMIST-ccRCC: A Real World Medical Dataset for the Development of Multi-Modal Systems. In *CVPR*. 2395–2403.

[194] Muhammad Ferjad Naeem, Yongqin Xian, Luc V Gool, and Federico Tombari. 2022. I2dformer: Learning image to document attention for zero-shot image classification. *NeurIPS* 35 (2022), 12283–12294.

[195] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: A large-scale speaker identification dataset. *arXiv:1706.08612* (2017).

[196] Bahareh Nakisa, Mohammad Naim Rastgoo, Andry Rakotonirainy, Frederic Maire, and Vinod Chandran. 2020. Automatic emotion recognition using temporal multimodal deep learning. *IEEE Access* 8 (2020), 225463–225474.

[197] Sasho Nedelkoski, Jorge Cardoso, and Odej Kao. 2019. Anomaly detection from system tracing data using multimodal deep learning. In *CLOUD*. IEEE, 179–186.

[198] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. 2015. Moddrop: adaptive multi-modal gesture recognition. *IEEE TPAMI* 38, 8 (2015), 1692–1706.

[199] Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. 2021. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *CVPR*. 3977–3986.

[200] OpenAI. 2023. GPT-4 technical report. *arXiv:2303.08774* (2023).

[201] Zaiyu Pan, Jiameng Xu, Shuangtian Jiang, and Jun Wang. 2025. SSFD-Net: Shared-Specific Feature Disentanglement Network for Multimodal Biometric Recognition with Missing Modality. *Digital Signal Processing* 159 (2025), 105003.

[202] Shivam Pande, Avinandan Banerjee, Saurabh Kumar, Biplab Banerjee, and Subhasis Chaudhuri. 2019. An adversarial approach to discriminative modality distillation for remote sensing image classification. In *CVPR*. 0–0.

[203] Yagya Raj Pandeya and Joonwhoan Lee. 2021. Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimedia Tools and Applications* 80, 2 (2021), 2887–2905.

[204] Yeonju Park, Sangmin Woo, Sumin Lee, Muhammad Adi Nugroho, and Changick Kim. 2023. Cross-modal alignment and translation for missing modality action recognition. *Computer Vision and Image Understanding* 236 (2023), 103805.

[205] Srinivas Parthasarathy and Shiva Sundaram. 2020. Training strategies to handle missing modalities for audio-visual expression recognition. In *Proc. ICMI*. 400–404.

[206] Naman Patel, Anna Choromanska, Prashanth Krishnamurthy, and Farshad Khorrami. 2019. A deep learning gated architecture for UGV navigation robust to sensor failures. *Robotics and Autonomous Systems* 116 (2019), 80–97.

[207] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *AAAI*, Vol. 33. 6892–6899.

[208] Vittorio Pipoli, Federico Bolelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, Costantino Grana, Rita Cucchiara, Elisa Ficarra, et al. 2025. Semantically Conditioned Prompts for Visual Recognition under Missing Modality Scenarios. In *WACV*.

[209] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv:1810.02508* (2018).

[210] Harsh Purohit, Ryo Tanabe, Toshiya Ichige, and *et al.* 2019. MIMII Dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. In *arXiv:1909.09347*.

[211] Chengjian Qiu, Yuqing Song, Yi Liu, Yan Zhu, Kai Han, Victor S Sheng, and Zhe Liu. 2024. MMMViT: Multiscale multimodal vision transformer for brain tumor segmentation with missing modalities. *BSPC* 90 (2024), 105827.

[212] Alec Radford, Jong Wook Kim, Craig Hallacy, and *et al.* 2021. Learning transferable visual models from natural language supervision. *ICML* (2021), 8748–8763.

[213] Ramachandra Raghavendra, Mohammad Imran, Ashok Rao, and G Hemantha Kumar. 2010. Multimodal biometrics: Analysis of handvein & palmprint combination used for person verification. In *ICETET*. IEEE, 526–530.

[214] Anil Rahate, Shruti Mandaokar, Pulkit Chandel, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. 2023. Employing multimodal co-learning to evaluate the robustness of sensor fusion for industry 5.0 tasks. *Soft Computing* 27, 7 (2023), 4139–4155.

[215] Wasi Uddin Rahman, Md Kamrul Hasan, Se Jin Lee, and *et al.* 2020. Integrating deep learning with logic fusion for information extraction. In *AAAI*. 8912–8920.

[216] Merey Ramazanova, Alejandro Pardo, Bernard Ghanem, and Motasem Alfarra. 2025. Test-Time Adaptation for Combating Missing Modalities in Egocentric Videos. In *The Thirteenth International Conference on Learning Representations*.

[217] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *arXiv:2204.06125* (2022).

[218] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *ICML*. 8821–8831.

[219] Dushyant Rao, Mark De Deuge, Navid Nourani-Vatani, Bertrand Douillard, Stefan B Williams, and Oscar Pizarro. 2014. Multimodal learning for autonomous underwater vehicles from visual and bathymetric data. In *ICRA*. IEEE, 3819–3825.

[220] Md Kaykobad Reza, Ameya Patil, Mashhour Solh, and M. Salman Asif. 2025. Robust Multimodal Learning via Cross-Modal Proxy Tokens. *arXiv:2501.17823* (2025).

[221] Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. 2024. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE TPAMI* (2024).

[222] Dominik Roblek, Karolis Misiunas, Marco Tagliasacchi, and Pen Li. 2020. Seanet: A multi-modal speech enhancement network. In *Interspeech*.

[223] Joseph Roth, Sourish Chaudhuri, Ondřej Klejch, and *et al.* 2020. AVA-ActiveSpeaker: An audio-visual dataset for active speaker detection. In *ICASSP*. 4492–4496.

[224] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. 2018. Synthesizing and reconstructing missing sensory modalities in behavioral context recognition. *Sensors* 18, 9 (2018), 2967.

[225] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loic Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. In *Proc.ViGIL*. 1–4.

[226] Hiroshi Sato, Tsubasa Ochiai, Keisuke Kinoshita, Marc Delcroix, Tomohiro Nakatani, and Shoko Araki. 2021. Multimodal attention fusion for target speaker extraction. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 778–784.

[227] Christoph Schuhmann, Romain Beaumont, Cade Vencu, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *arXiv:2210.08402* (2022).

[228] Christoph Schuhmann, Romain Beaumont, Richard Vencu, and *et al.* 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *NeurIPS* 35 (2022), 25278–25294.

[229] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *CVPR*. 6649–6658.

[230] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*. 1010–1019.

[231] Shiv Shankar, Laure Thompson, and Madalina Fiterau. 2022. Progressive fusion for multimodal integration. *arXiv:2209.00302* (2022).

[232] Ming Shao, Zhengming Ding, and Yun Fu. 2015. 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)*, Vol. 1. IEEE, 1–6.

[233] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*. 2556–2565.

[234] Yan Shen and Mingchen Gao. 2019. 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, 417–428.

[235] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv:2201.02184* (2022).

[236] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. 2021. Robust self-supervised audio-visual speech recognition. *arXiv:2201.01763* (2021).

[237] Bowen Shi, Wei-Ning Hsu, and Abdelrahman Mohamed. 2022. Robust self-supervised audio-visual speech recognition. *arXiv:2201.01763* (2022).

[238] Piao Shi, Min Hu, Satoshi Nakagawa, Xiangming Zheng, Xuefeng Shi, and Fuji Ren. 2025. Text-guided Reconstruction Network for Sentiment Analysis with Uncertain Missing Modalities. *IEEE Transactions on Affective Computing* (2025).

[239] Weipeng Shi, Wenhu Qin, and Zhonghua Yun. 2024. Learning rich multimodal representation for robust land cover classification in fog. *IEEE Sensors Journal* (2024).

[240] Nadezhda Shvetsova, Boyang Chen, Andrew Rouditchenko, and *et al.* 2022. Everything at once – Multi-modal fusion transformer for video retrieval. In *CVPR*. 20020–20029.

[241] Aniruddh Sikdar, Jayant Teotia, and Suresh Sundaram. 2023. Contrastive Learning-Based Spectral Knowledge Distillation for Multi-Modality and Missing Modality Scenarios in Semantic Segmentation. *arXiv:2312.02240* (2023).

[242] Aniruddh Sikdar, Jayant Teotia, and Suresh Sundaram. 2025. OGP-Net: Optical Guidance Meets Pixel-Level Contrastive Distillation for Robust Multi-Modal and Missing Modality Segmentation. In *AAAI*, Vol. 39. 6922–6930.

[243] Karan Sikka, Karmen Dykstra, Suchitra Sathyanarayana, Gwen Littlewort, and Marian Bartlett. 2013. Multiple kernel learning for emotion recognition in the wild. In *Proc. ACM ICMI*. 517–524.

[244] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *CVPR*. 15638–15650.

[245] Krishna Somandepalli, Naveen Kumar, Ruchir Travadi, and Shrikanth Narayanan. 2019. Multimodal representation learning using deep multiset canonical correlation. *arXiv:1904.01775* (2019).

[246] Dingjie Song, Sicheng Lai, Shunian Chen, Lichao Sun, and Benyou Wang. 2024. Both Text and Images Leaked! A Systematic Analysis of Multimodal LLM Data Contamination. *arXiv:2411.03823* (2024).

[247] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning. *arXiv:2103.01913* (2021).

[248] Nitish Srivastava and Russ R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. *NeurIPS* 25 (2012).

[249] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv:1811.00491* (2018).

[250] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *IEEE/CVF ICCV*. 7464–7473.

[251] Jiachen Sun, Haizhong Zheng, Qingzhao Zhang, Atul Prakash, Z Morley Mao, and Chaowei Xiao. 2023. CALICO: Self-supervised camera-liDAR contrastive pre-training for BEV perception. *arXiv:2306.00349* (2023).

[252] Rui Sun, Long Chen, Lei Zhang, Ruirui Xie, and Jun Gao. 2024. Robust visible-infrared person re-identification based on polymorphic mask and wavelet graph convolutional network. *IEEE TIFS* 19 (2024), 2800–2813.

[253] Wangbin Sun, Fei Ma, Yang Li, Shao-Lun Huang, Shiguang Ni, and Lin Zhang. 2021. Semi-supervised multimodal image translation for missing modality imputation. In *ICASSP*. IEEE, 4320–4324.

[254] Alon Talmor, Ori Yoran, Dana Lahav, and *et al.* 2021. MultiModalQA: Complex question answering over text, tables and images. *arXiv:2104.06039* (2021).

[255] Brandon Theodorou, Lucas Glass, Cao Xiao, and Jimeng Sun. 2024. FRAMM: Fair ranking with missing modalities for clinical trial site selection. *Patterns* 5, 3 (2024).

[256] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*. 5238–5248.

[257] Junjiao Tian, Wesley Cheung, Nathaniel Glaser, Yen-Cheng Liu, and Zsolt Kira. 2020. Uno: Uncertainty-aware noisy-or multimodal fusion for unanticipated input degradation. In *ICRA*. IEEE, 5716–5723.

[258] Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. 2017. Missing modalities imputation via cascaded residual autoencoder. In *CVPR*. 1405–1414.

[259] Michael Tschannen, Basil Mustafa, and Neil Houlsby. 2023. Clippo: Image-and-language understanding from pixels only. In *CVPR*. 11006–11017.

[260] Muhammad Usama, Syeda Aisha Asim, Syed Bilal Ali, Syed Talal Wasim, and Umair Bin Mansoor. 2025. Analysing the Robustness of Vision-Language-Models to Common Corruptions. *arXiv:2504.13690* (2025).

[261] Beatrice van Amsterdam, Abdolrahim Kadkhodamohammadi, Imanol Luengo, and Danail Stoyanov. 2023. Aspnet: Action segmentation with shared-private representation of multiple data sources. In *CVPR*. 2384–2393.

[262] Juan Vazquez-Rodriguez, Grégoire Lefebvre, Julien Cumin, and James L Crowley. 2023. Accommodating Missing Modalities in Time-Continuous Multimodal Emotion Recognition. In *ACII*. IEEE, 1–8.

[263] Cheng Wang, Mathias Niepert, and Hui Li. 2018. LRMM: Learning to recommend with missing modalities. *arXiv:1808.06791* (2018).

[264] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. 2023. Multi-modal learning with missing modality via shared-specific feature modelling. In *CVPR*. 15878–15887.

[265] Hu Wang, Congbo Ma, Jianpeng Zhang, Yuan Zhang, Jodie Avery, Louise Hull, and Gustavo Carneiro. 2023. Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In *MICCAI*. Springer, 216–226.

[266] Jingyao Wang, Luntian Mou, Lei Ma, Tiejun Huang, and Wen Gao. 2023. AMSA: Adaptive multimodal learning for sentiment analysis. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 3s (2023), 1–21.

[267] Juexing Wang, Guangjing Wang, Xiao Zhang, Li Liu, Huacheng Zeng, Li Xiao, Zhichao Cao, Lin Gu, and Tianxing Li. 2023. Patch: A plug-in framework of non-blocking inference for distributed multimodal system. *Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–24.

[268] Ning Wang, Hui Cao, Jun Zhao, Ruilin Chen, Dapeng Yan, and Jie Zhang. 2022. M2R2: Missing-Modality Robust emotion Recognition framework with iterative data augmentation. *IEEE Transactions on Artificial Intelligence* 4, 5 (2022), 1305–1316.

[269] Pidong Wang, Preslav Nakov, and Hwee Tou Ng. 2016. Source language adaptation approaches for resource-poor machine translation. *Computational Linguistics* 42, 2 (2016), 277–306.

[270] Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. 2020. Multimodal learning with incomplete modalities by knowledge distillation. In *Proc. 26th ACM SIGKDD*. 1828–1838.

[271] Shiming Wang, Holger Caesar, Liangliang Nan, and Julian FP Kooij. 2024. Unibev: Multi-modal 3d object detection with uniform bev encoders for robustness against missing sensor modalities. In *IVS*. IEEE, 2776–2783.

[272] Shuai Wang, Zipei Yan, Daoan Zhang, Haining Wei, Zhongsen Li, and Rui Li. 2023. Prototype knowledge distillation for medical segmentation with missing modality. In *ICASSP*. IEEE, 1–5.

[273] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. 2015. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 1–6.

[274] Xincheng Wang, Liejun Wang, Yinfeng Yu, and Xinxin Jiao. 2025. Modality-Invariant Bidirectional Temporal Representation Distillation Network for Missing Multimodal Sentiment Analysis. *arXiv:2501.05474* (2025).

[275] Xin Wang, Jia Wu, Junkun Chen, and *et al.* 2019. VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*. 4581–4591.

[276] Yixin Wang, Yang Zhang, Yang Liu, Zihao Lin, Jiang Tian, Cheng Zhong, Zhongchao Shi, Jianping Fan, and Zhiqiang He. 2021. ACN: adversarial co-training network for brain tumor segmentation with missing modalities. In *MICCAI*. Springer, 410–420.

[277] Zechen Wang, Xin Liu, Hang Li, and *et al.* 2022. Multi-modal Learning with Missing Modality via Shared-Specific Feature Modelling. In *CVPR*. 8875–8884.

[278] Zilong Wang, Zhaohong Wan, and Xiaojun Wan. 2020. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proc. of The Web Conference 2020*. 2514–2520.

[279] Shicai Wei, Yang Luo, Xiaoguang Ma, Peng Ren, and Chunbo Luo. 2023. MSH-Net: Modality-shared hallucination with joint adaptation distillation for remote sensing image classification using missing modalities. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–15.

[280] Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. 2023. Towards good practices for missing modality robust action recognition. In *AAAI*, Vol. 37. 2776–2784.

[281] Maciej K Wozniak, Viktor Kårefj"ard, Marko Thiel, and Patric Jensfelt. 2023. Toward a robust sensor fusion step for 3D object detection on corrupted data. *IEEE Robotics and automation letters* 8, 11 (2023), 7018–7025.

[282] Fengmin Wu, Sicong Liu, Bin Guo, Xiaocheng Li, Yuan Gao, and Zhiwen Yu. 2024. AdaFlow: Non-blocking Inference with Heterogeneous Multi-modal Mobile Sensor Data. In *CSCAIoT*. IEEE, 8–9.

[283] Jason Wu, Kang Yang, Lance Kaplan, and Mani Srivastava. 2025. ADMN: A Layer-Wise Adaptive Multimodal Network for Dynamic Input Noise and Compute Resources. *arXiv:2502.07862* (2025).

[284] Peng Wu, jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. 2020. Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision. In *ECCV*.

[285] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. 2024. Deep multimodal learning with missing modality: A survey. *arXiv:2409.07825* (2024).

[286] Renjie Wu, Hu Wang, Feras Dayoub, and Hsiang-Ting Chen. 2024. Segment Beyond View: Handling Partially Missing Modality for Audio-Visual Semantic Segmentation. In *AAAI*, Vol. 38. 6100–6108.

[287] Sanju Xaviar, Xin Yang, and Omid Ardakanian. 2024. Centaur: Robust Multimodal Fusion for Human Activity Recognition. *IEEE Sensors Journal* (2024).

[288] Ruida Xi, Nianchang Huang, Changzhou Lai, Qiang Zhang, and Jungong Han. 2024. FMCNet +: Feature-Level Modality Compensation for Visible-Infrared Person Re-Identification. *IEEE Transactions on Neural Networks and Learning Systems* (2024).

[289] Weihao Xia, Yujiu Yang, and Jing-Hao Xue. 2020. Unsupervised multi-domain multimodal image-to-image translation with explicit domain-constrained disentanglement. *Neural Networks* 131 (2020), 50–63.

[290] Shuaiyu Xie, Jian Wang, Hanbin He, Zhihao Wang, Yuqi Zhao, Neng Zhang, and Bing Li. 2024. TVDiag: A Task-oriented and View-invariant Failure Diagnosis Framework with Multimodal Data. *arXiv:2407.19711* (2024).

[291] Bowen Xin, Jing Huang, Yun Zhou, Jie Lu, and Xiuying Wang. 2021. Interpretation on deep multimodal fusion for diagnostic classification. In *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[292] Wei Xiong, Tao Wang, Xiumei Chen, Yue Zhang, Wencong Zhang, Qianjin Feng, Meiyan Huang, Alzheimer's Disease Neuroimaging Initiative, et al. 2025. Disentanglement and codebook learning-induced feature match network to diagnose neurodegenerative diseases on incomplete multimodal data. *Pattern Recognition* 165 (2025), 111597.

[293] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*. 5288–5296.

[294] Peng Xu, Xiatian Zhu, and David A Clifton. 2023. Multimodal learning with transformers: A survey. *IEEE TPAMI* 45, 10 (2023), 12113–12132.

[295] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In *Proc. 26th ACM SIGKDD*. 1192–1200.

[296] Bang Yang, Fenglin Liu, Yuexian Zou, Xian Wu, Yaowei Wang, and David A Clifton. 2024. Zeronlg: Aligning and autoencoding domains for zero-shot multimodal and multilingual natural language generation. *IEEE TPAMI* (2024).

[297] Heran Yang, Jian Sun, and Zongben Xu. 2023. Learning unified hyper-network for multi-modal MR image synthesis and tumor segmentation with missing modalities. *IEEE Transactions on Medical Imaging* (2023).

[298] Huiyuan Yang, Taoyue Wang, and Lijun Yin. 2020. Adaptive multimodal fusion for facial action units recognition. In *ACM Multimedia*. 2982–2990.

[299] Jianwei Yang, Hao Duan, Richard Socher, Caiming Xiong, Li Fei-Fei, and Hailin Wang. 2022. Chinese CLIP: Contrastive vision-language pretraining in Chinese. *arXiv:2211.01335* (2022).

[300] Karren Yang, Wan-Yi Lin, Manash Barman, Filipe Condessa, and Zico Kolter. 2021. Defending multimodal fusion models against single-source adversaries. In *CVPR*. 3340–3349.

[301] Qiushi Yang, Xiaoqing Guo, Zhen Chen, Peter YM Woo, and Yixuan Yuan. 2022. D 2-Net: Dual disentanglement network for brain tumor segmentation with missing modalities. *IEEE Transactions on Medical Imaging* 41, 10 (2022), 2953–2964.

[302] Tongyu Yang, Qian Zhou, Hua Zou, Haifeng Jiang, and Yong Wang. 2025. UML: A Unified Multimodal Learning Framework for Cataract Postoperative Visual Acuity Prediction with Uncertain Missing Modalities. In *ICASSP*. IEEE, 1–5.

[303] Zhao Yang, Rui Jiang, Xiao Fu, Wei Xi, and Jizhong Zhao. 2025. Open-Modality Latent Modality Interaction Maximization for Audio-Visual Learning. In *ICASSP*. IEEE, 1–5.

[304] Junjie Ye, Jie Zhou, Junfeng Tian, Rui Wang, Jingyi Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022. Sentiment-aware multimodal pre-training for multimodal sentiment analysis. *Knowledge-Based Systems* 258 (2022), 110021.

[305] Jia-Li Yin, Bo-Hao Chen, and Ying Li. 2018. Highly accurate image reconstruction for multimodal noise suppression using semisupervised learning on big data. *IEEE Transactions on Multimedia* 20, 11 (2018), 3045–3056.

[306] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *TACL*, Vol. 2. 67–78.

[307] Guan Yu, Quefeng Li, Dinggang Shen, and Yufeng Liu. 2020. Optimal sparse linear prediction for block-missing multi-modality data without imputation. *J. Amer. Statist. Assoc.* 115, 531 (2020), 1406–1419.

[308] Guohui Yu, Lihui Wang, Qijian Chen, Li Wang, Kun Tang, Xinyu Cheng, and Yuemin Zhu. 2024. Penta-Encoder with Medical Transformer for Incomplete Multimodal Learning of Brain Tumor Segmentation. In *2024 IEEE 17th International Conference on Signal Processing (ICSP)*. IEEE, 642–646.

[309] Kaicheng Yu, Tang Tao, Hongwei Xie, Zhiwei Lin, Tingting Liang, Bing Wang, Peng Chen, Dayang Hao, Yongtao Wang, and Xiaodan Liang. 2023. Benchmarking the robustness of lidar-camera fusion for 3d object detection. In *CVPR*. 3188–3198.

[310] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Chunyuan Li, Linjie Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv:2111.11432* (2021).

[311] Yuan Yuan, Zhaojian Li, and Bin Zhao. 2025. A Survey of Multimodal Learning: Methods, Applications, and Future. *Comput. Surveys* (2025).

[312] Xianghu Yue, Yiming Chen, Xueyi Zhang, Xiaoxue Gao, Mengling Feng, Mingrui Lao, Huiping Zhuang, and Haizhou Li. 2025. PAL: Prompting Analytic Learning with Missing Modality for Multi-Modal Class-Incremental Learning. *arXiv:2501.09352* (2025).

[313] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. *arXiv:1606.06259* (2016).

[314] Jiandian Zeng, Tianyi Liu, and Jiantao Zhou. 2022. Tag-assisted multimodal sentiment analysis under uncertain missing modalities. In *Proc. 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1545–1554.

[315] Jiandian Zeng, Jiantao Zhou, and Tianyi Liu. 2022. Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities. In *Proc. 2022 Conference on Empirical Methods in Natural Language Processing*. 2924–2934.

[316] Jiandian Zeng, Jiantao Zhou, and Tianyi Liu. 2022. Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities. *IEEE Transactions on Multimedia* 25 (2022), 6301–6314.

[317] Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. 2022. M3care: Learning with missing modalities in multimodal healthcare data. In *Proc. ACM SIGKDD*. 2418–2428.

[318] Jiawei Zhang, Tianyu Pang, Chao Du, Yi Ren, Bo Li, and Min Lin. 2024. Benchmarking large multimodal models against common corruptions. *arXiv:2401.11943* (2024).

[319] Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. 2020. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion* 59 (2020), 103–126.

[320] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting visual representations in vision-language models. In *CVPR*. 5579–5588.

[321] Peng-Fei Zhang and Zi Helen Huang. 2023. Multi-head siamese prototype learning against both data and label corruption. In *ACM Multimedia*. 1–7.

[322] Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. 2022. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *CVPR*. 7349–7358.

[323] Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, et al. 2024. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv:2404.18947* (2024).

[324] Shiqing Zhang, Yijiao Yang, Chen Chen, Xingnan Zhang, Qingming Leng, and Xiaoming Zhao. 2024. 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* 237 (2024), 121692.

[325] Tong Zhang, Shu Shen, and CL Chen. 2025. MICINet: Multi-Level Inter-Class Confusing Information Removal for Reliable Multimodal Classification. *arXiv:2502.19674* (2025).

[326] Ting Zhang, Bin Song, Zhiyong Zhang, and Yajuan Zhang. 2025. Multimodal sentiment analysis based on multi-stage graph fusion networks under random missing modality conditions. *IET Image Processing* 19, 1 (2025), e13310.

[327] Weichen Zhang, Dong Xu, Jing Zhang, and Wanli Ouyang. 2021. Progressive modality cooperation for multi-modality domain adaptation. *IEEE Transactions on Image Processing* 30 (2021), 3293–3306.

[328] Xiaodan Zhang, Qiao Song, and Gang Liu. 2022. Multimodal Image Aesthetic Prediction with Missing Modality. *Mathematics* 10, 13 (2022), 2312.

[329] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. 2022. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *MICCAI*. Springer, 107–117.

[330] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2022. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*. 2–25.

[331] Yue Zhang, Chengtao Peng, Qiuli Wang, Dan Song, Kaiyan Li, and S Kevin Zhou. 2024. Unified multi-modal image synthesis for missing modality imputation. *IEEE Transactions on Medical Imaging* (2024).

[332] Yuqing Zhang, Dongliang Xie, Dawei Luo, and Baosheng Sun. 2025. Emotional Boundaries and Intensity Aware Model for Incomplete Multimodal Sentiment Analysis. *Digital Signal Processing* (2025), 105023.

[333] Jinming Zhao, Ruichen Li, and Qin Jin. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proc. ACL*. 2608–2618.

[334] Wenqian Zhao, Kai Yang, Peijin Ding, Ce Na, and Wen Li. 2025. Graph attention contrastive learning with missing modality for multimodal recommendation. *Knowledge-Based Systems* (2025), 113035.

[335] Zechen Zhao, Heran Yang, and Jian Sun. 2022. Modality-adaptive feature interaction for brain tumor segmentation with missing modalities. In *MICCAI*. Springer, 183–192.

[336] Aihua Zheng, Zi Wang, Zihan Chen, Chenglong Li, and Jin Tang. 2021. Robust multi-modality person re-identification. In *AAAI*, Vol. 35. 3529–3537.

[337] Xiao Zheng, Chang Tang, Zhiguo Wan, Chengyu Hu, and Wei Zhang. 2023. Multi-level confidence learning for trustworthy multimodal classification. In *AAAI*, Vol. 37. 11381–11389.

[338] Zhuo Zheng, Ailong Ma, Liangpei Zhang, and Yanfei Zhong. 2021. Deep multisensor learning for missing-modality all-weather mapping. *ISPRS Journal of Photogrammetry and Remote Sensing* 174 (2021), 254–264.

[339] Zhuo Zhi, Yuxuan Sun, et al. 2025. Wasserstein Modality Alignment Makes Your Multimodal Transformer More Robust. *Transactions on Machine Learning Research* (2025).

[340] Huasong Zhong, Jingyuan Chen, Chen Shen, Hanwang Zhang, Jianqiang Huang, and Xian-Sheng Hua. 2020. Self-adaptive neural module transformer for visual question answering. *IEEE Transactions on Multimedia* 23 (2020), 1264–1273.

[341] Kaichen Zhou, Changhao Chen, Bing Wang, Muhamad Risqi U Saputra, Niki Trigoni, and Andrew Markham. 2021. Vmloc: Variational fusion for learning-based multimodal camera localization. In *AAAI*, Vol. 35. 6165–6173.

[342] Qian Zhou, Hua Zou, Fei Luo, and Yishi Qiu. 2023. RHViT: A Robust Hierarchical Transformer for 3D Multimodal Brain Tumor Segmentation Using Biased Masked Image Modeling Pre-training. In *BIBM*. IEEE, 1784–1791.

[343] Tongxue Zhou. 2023. Feature fusion and latent feature learning guided brain tumor segmentation and missing modality recovery network. *Pattern Recognition* 141 (2023), 109665.

[344] Tongxue Zhou, Stéphane Canu, Pierre Vera, and Su Ruan. 2020. Brain tumor segmentation with missing modalities via latent multi-source correlation representation. In *MICCAI*. Springer, 533–541.

[345] Tongxue Zhou, Stéphane Canu, Pierre Vera, and Su Ruan. 2021. Feature-enhanced generation and multi-modality fusion based deep neural network for brain tumor segmentation with missing MR modalities. *Neurocomputing* 466 (2021), 102–112.

[346] Tongxue Zhou, Stéphane Canu, Pierre Vera, and Su Ruan. 2021. Latent correlation representation learning for brain tumor segmentation with missing MRI modalities. *IEEE Transactions on Image Processing* 30 (2021), 4263–4274.

[347] Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. 2023. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *ACM Multimedia*. 6311–6320.

[348] Yizhe Zhu, Xin Sun, and Xi Zhou. 2023. Exploiting Multi-modal Fusion for Robust Face Representation Learning with Missing Modality. In *International Conference on Artificial Neural Networks*. Springer, 283–294.

[349] Yian Zhu, Shaoyu Wang, Runlong Lin, Yun Hu, and Qiang Chen. 2021. Brain tumor segmentation for missing modalities by supplementing missing features. In *ICCCBDA*. IEEE, 652–656.

[350] Haolin Zuo, Rui Liu, Jinming Zhao, Guanglai Gao, and Haizhou Li. 2023. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. In *ICASSP*. IEEE, 1–5.

Received 15 November 2025