International Journal of Multimedia Information Retrieval (2025) 14:21
https://doi.org/10.1007/s13735-025-00370-y

REGULAR PAPER

Chameleon: A Multimodal Learning Framework Robust to Missing Modalities

Muhammad Irzam Liaqat[1] · Shah Nawaz[2] · Muhammad Zaigham Zaheer[3] · Muhammad Saad Saeed[4] · Hassan Sajjad[5] · Tom De Schepper[6] · Karthik Nandakumar[3] · Muhammad Haris Khan[3] · Ignazio Gallo[7] · Markus Schedl[2,8]

Received: 24 January 2025 / Revised: 21 April 2025 / Accepted: 25 April 2025 / Published online: 2 June 2025

Abstract
Multimodal learning has demonstrated remarkable performance improvements over unimodal architectures. However, multimodal learning methods often exhibit deteriorated performances if one or more modalities are missing. This may be attributed to the commonly used multi-branch design containing modality-specific components, making such approaches reliant on the availability of a complete set of modalities. In this work, we propose a robust multimodal learning framework, `Chameleon`, that adapts a common-space visual learning network to align all input modalities. To enable this, we present the unification of input modalities into one format by encoding any non-visual modality into visual representations thus making it robust to missing modalities. Extensive experiments are performed on multimodal classification task using four textual-visual (Hateful Memes, UPMC Food-101, MM-IMDb, and Ferramenta) and two audio-visual (avMNIST, VoxCeleb) datasets. `Chameleon` not only achieves superior performance when all modalities are present at train/test time but also demonstrates notable resilience in the case of missing modalities.

Keywords Multimodal learning · Vision and other modalities · Missing modalities

1 Introduction

In recent years, there has been a surge in the use of multimodal data for various applications. For instance, users combine text, image, audio, or video modalities to sell a product over an e-commerce platform or express views on social media platforms. Two or more of these modalities are often combined to solve different tasks such as multimodal

✉ Markus Schedl
markus.schedl@jku.at

Muhammad Irzam Liaqat
irzam.liaqat@imtlucca.it

Shah Nawaz
shah.nawaz@jku.at

Muhammad Zaigham Zaheer
Zaigham.Zaheer@mbzuai.ac.ae

Muhammad Saad Saeed
saad.saeed@uettaxila.edu.pk

Hassan Sajjad
hsajjad@dal.ca

Tom De Schepper
tom.deSchepper@imec.be

Karthik Nandakumar
karthik.nandakumar@mbzuai.ac.ae

Muhammad Haris Khan
muhammad.haris@mbzuai.ac.ae

Ignazio Gallo
ignazio.gallo@uninsubria.it

[1] IMT School for Advanced Studies of Lucca, Lucca, Italy

[2] Institute of Computational Perception, Johannes Kepler University, Linz, Austria

[3] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates

[4] Swarm Robotics Lab NCRA, University of Engineering and Technology, Taxila, Pakistan

[5] Dalhousie University, Halifax, Canada

[6] Interuniversity Microelectronics Centre (IMEC), Leuven, Belgium

[7] University of Insubria, Varese, Italy

[8] Linz Institute of Technology, AI Lab, JKU Linz, Austria

classification [1, 2], cross-modal retrieval [3], cross-modal verification [4], multimodal named entity recognition [5, 6], visual question answering [7, 8], image captioning [9], semantic relatedness [10], segmentation [11], and multimodal machine translation [12, 13]. In all these tasks, multimodal methods can experience scenarios where some modalities are missing, e.g., due to failures in data acquisition pipelines. Several researchers have recently concluded that multimodal learning is not *inherently* robust to missing modalities and can result in a significantly deteriorated performance when modalities are missing [14–16]. For example, as seen in Fig. 1, ViLT [17], a baseline multimodal Transformer demonstrates significant performance drops when the textual modality is missing at test time. The performance deterioration is attributed to the fundamental design that assumes the presence of all modalities concurrently for training, thus inadvertently developing a dependency on having a complete set of modalities to make a prediction. Therefore, if a modality is missed during inference, the performance deteriorates significantly [14]. Recently, Ma et al. [14] proposed a strategy to improve robustness of ViLT against missing modality by incorporating modal-incomplete data. As seen in Fig. 1, the performance on missing modality scenarios improves notably using this approach. Though it reduces the dependency of ViLT on having a complete set of modalities to make a prediction, it requires a data-centric fusion strategy to handle modal-incomplete data which is cumbersome to optimize [14].

To resolve this, it is pertinent to have an approach that may be trainable using the *complete set* of modalities but resilient to missing modalities. To this end, we propose `Chameleon`, a framework that adapts a common-space visual learning network to align all input modalities. This way, the network is able to process either of the two inputs, visual or non-visual, as well as both together, all as images. Therefore, when a modality is missing, `Chameleon` leverages the available modality to output the correct prediction scores. Moreover, mapping different modalities into a single shared embedding space not only simplifies the input interface but also allows harnessing architectures or training procedures designed for one domain into another domain. As seen in Fig. 1, `Chameleon` not only outperforms the baseline ViLT by a notable margin but also demonstrates better robustness than Ma et al. [14] without any data-centric optimization. To explore the general applicability of `Chameleon`, we incorporate various kinds of modalities (including, audio and text) for training the multimodal network by proposing an encoding scheme that transforms non-visual modalities into a visual format. We evaluate `Chameleon` on six benchmark datasets including four textual-visual datasets (i.e., UPMC Food-101 [18], Hateful Memes [2], MM-IMDb [19], and Ferramenta [20]) and two audio-visual datasets (i.e., avMNIST [21], VoxCeleb [22]). Our approach outperforms state-of-the-art (SOTA) multimodal methods on complete modalities and demonstrates superior robustness against missing modalities during training and testing. The key highlights of our work are as follows:

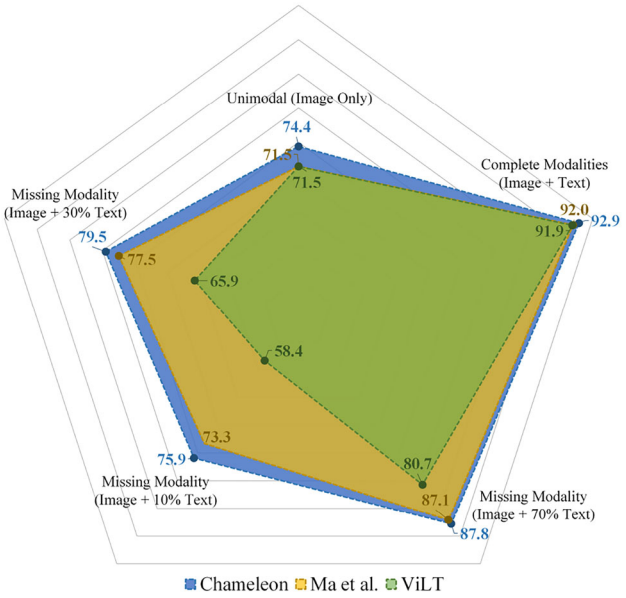


Fig. 1 Comparison of `Chameleon` with ViLT [17] & Ma et al. [14] on UPMC Food-101 dataset. `Chameleon` demonstrates better multimodal and unimodal performances and retains significantly superior performance when textual modality is missing at test time. Some drop in performance is generally expected when a modality is missed. However, for a multimodal method to be considered robust, the performance with missing modalities must be better than its independent unimodal performance.

1. We propose `Chameleon`, a multimodal learning framework robust to missing modalities.
2. To eliminate the dependency of the learning network on complete set of modalities, we propose a novel encoding scheme that transforms non-visual modality into visual representations to carry out multimodal training using shared weights.
3. We evaluate our framework extensively on Convolutional Neural Networks (CNNs), Vision Transformers (ViT), and Adapters, demonstrating its general applicability across various networks. Notably, unlike previous study [14], we demonstrate that the robustness of Transformers against missing modalities can be improved significantly by using an appropriate learning framework such as `Chameleon`.
4. A wide range of experiments performed on four textual-visual and two audio-visual datasets under different multimodal, and missing modality settings during training and testing demonstrate the significance of `Chameleon` in application scenarios with missing modalities.

2 Related Work

2.1 Multimodal Learning and Missing Modality

Multiple modalities including texts and images often contain complementary information about a common subject. The goal of multimodal learning is to leverage complementary information across modalities to improve the performance of various machine learning tasks. Each task may be different from the other, however, the underlying objective remains the same: to learn joint representations across multiple modalities [23, 24]. Existing multimodal methods employ multi-branch networks to learn joint representations by minimizing the distance between the representations of multiple modalities [1, 4, 17, 19, 21, 25–29]. Such methods have achieved remarkable performance using modality-complete data [1, 2, 17, 19, 21, 30, 31].

However, multimodal methods suffer from performance deterioration if some modalities are absent either during training or testing [14, 16, 32–34]. Considering the importance of multimodal learning, recent years have seen a surge in studies handling missing modalities [14–16, 35–38]. For example, Ma et al. [14] improved the robustness of Transformer models via multi-task optimization. More recently, Lee et al. [16] introduced missing-modality-aware prompts which can be plugged into multimodal Transformers to handle missing modality. `Chameleon` also targets the missing modality problem. However, different from other approaches, it relies on encoding modalities to a single representation that enables training of a single modality framework. The reliance on a single modality representation resulted in a multimodal framework robust to missing modality.

2.2 Rendering Non-visual Modalities into Images

Recently, some researchers have explored the idea of transforming text information into a visual format to perform visual recognition task. For example, Gallo et al. [39] leveraged Word2Vec word embeddings to reconstruct the semantics associated with text as an image to train a visual network for text classification. Salesky et al. [40] rendered the raw text directly into visual format, divide the rendered image into overlapping slices, and produce representations with optical character recognition to train a machine translation model. Similarly, Rust et al. [41] proposed a Masked Autoencoding Visual Transformer to reconstruct the pixels in masked image patches to train a language model. More recently, Tschannen et al. [42] proposed CLIPPO that renders text information as images to train a pure pixel-based model performing multimodal learning. All approaches rendering text into visual format are, in essence, related to `Chameleon` as we also encode text into visual format. However, instead of the common approach of transforming raw text directly into images, we use embeddings to encode audio or textual information which is found to be effective in our experiments for the missing modality problem (Section 4.2.1). In addition, we acknowledge the existing approaches [22, 43, 44] that encode audio modality into visual format (e.g., spectrogram [22]) to train on a given task. However, we propose a generic encoding scheme instead of using a particular modality (like text in [42]) directly as input making our approach generalizable to any modality.

3 Methodology

`Chameleon` relies on the intuition that a common feature space across modalities enable learning of multimodal representations robust to missing modality. Each component of `Chameleon` is visualized in Fig. 2 and discussed subsequently in this section:

3.1 Problem Formulation and Overall Idea

Formally, given $\mathcal{D} = \{(x_i^a, x_i^b)\}_{i=1}^N$ is the training set where N is the number of pairs of modality a and modality b and x_i^a and x_i^b are the individual modality samples of the i^{th} instance respectively. Moreover, each pair (x_i^a, x_i^b) has a class label y_i. Typical existing multimodal methods take multiple modalities as input by using a multi-branch network $\mathcal{C}_m$ to perform the classification task:

$$\tilde{y}_i = \mathcal{C}_m(\{x^a, x^b\}, y_i) \tag{1}$$

Training such a multi-branch configuration inadvertently depends on modality-complete data to perform a given task. This results in a significant performance deterioration during inference in case of missing modality (as seen in Fig. 1). To alleviate this issue, we propose to encode non-visual modality as image and perform the multimodal task entirely in the visual domain, both during training and testing. In other words, a visual network is trained on multimodal data consisting of images and non-visual modality through a common interface of visual information. We hypothesize that unifying the modalities to an identical input format makes the model independent of modality-specific branches, thus eliminating the need of modality complete data and results in a framework robust to missing modalities. Therefore, in `Chameleon`, Eq. (1) takes the following form:

$$\tilde{y}_i = \mathcal{C}_v(\{\hat{x}^a, x^b\}, y_i), \tag{2}$$

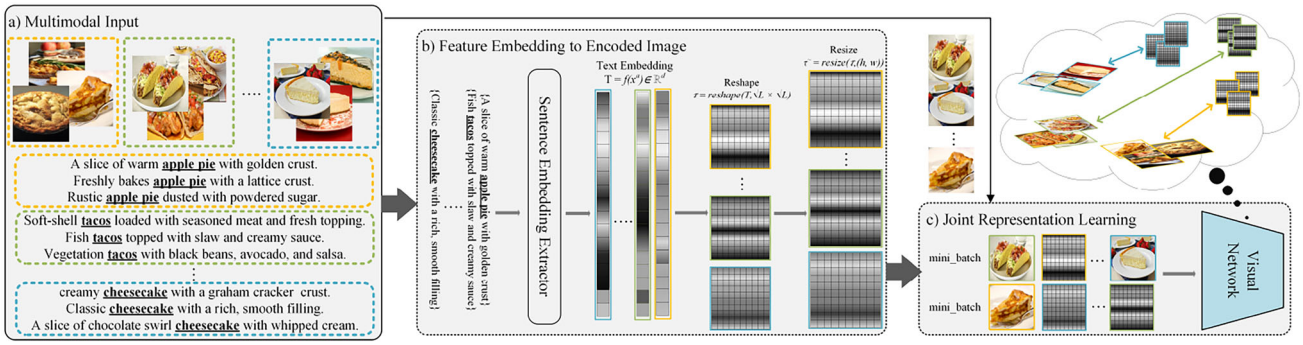


Fig. 2 Architecture of `Chameleon`. **(a)** Multimodal input consists of images and texts. **(b)** Embeddings of texts are encoded as images. **(c)** Visual and encoded text inputs are arbitrarily input to a weight-sharing visual network in a mini batch. Note that the figure represents the case of textual-visual modality pair. In the case of audio-visual pair, the same encoding is applied to the audio embedding.

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

3.2 Encoding Scheme

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

$$T = f(x^a) \in \mathbb{R}^d \tag{3}$$

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

$$\tau = \text{reshape}(T, \lceil \sqrt{d} \rceil \times \lceil \sqrt{d} \rceil) \tag{4}$$

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

$$\tilde{\tau} = \text{resize}\,(\tau, (h, w)) \tag{5}$$

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

3.3 Robustness to Missing Modalities

`Chameleon` inherently learns shared representations across modalities, aligning them into a common latent space. This results in a richer representation that is independent of the presence of both modalities, thus yielding a robust model in case one of the modality is missing. When a modality is missing, the available modality enables the use of the multimodal knowledge from the shared latent space to make a prediction. We observe this by visually comparing the activation maps [45] of cases taken from UPMC Food-101 for image-only unimodal training/testing and multimodal training with missing text modality testing in Fig. 3. In image-only unimodal training and testing (Fig. 3b), as expected, the model focuses on the distinct features of the object to predict the classification score. In the case of missing modality at test time, a multimodal architecture robust to missing modalities should ideally be able to retain the focus on the available modality and behave like a unimodal network. We observe this case in Fig. 3c for `Chameleon`, when the text modality is missing at test time. As seen, the model shifts its focus on the object itself and behaves comparably to the unimodal network. These visualizations highlight the internal working of `Chameleon` in successfully learning multimodal representations while demonstrating resilience against missing modalities. More examples are provided in Supplementary Section 2.

4 Experiments

Datasets We evaluate `Chameleon` on the multimodal classification task using four textual-visual datasets (UPMC Food-101 [18], Hateful Memes [46], MM-IMDb [21], and Ferramenta [20]) and two audio-visual (VoxCeleb [22] and avMNIST [21]) datasets. Table 1 summarizes statistics of the datasets.

Evaluation Metrics Following existing SOTA methods on complete and missing modalities [2, 14, 16, 18, 21], we report classification accuracy for UPMC Food-101, VoxCeleb, avMNIST and Ferramenta, area under the receiver

[1] While the description in Eq. 2 holds for arbitrary modality types, in our approach, one of the modalities is always visual.

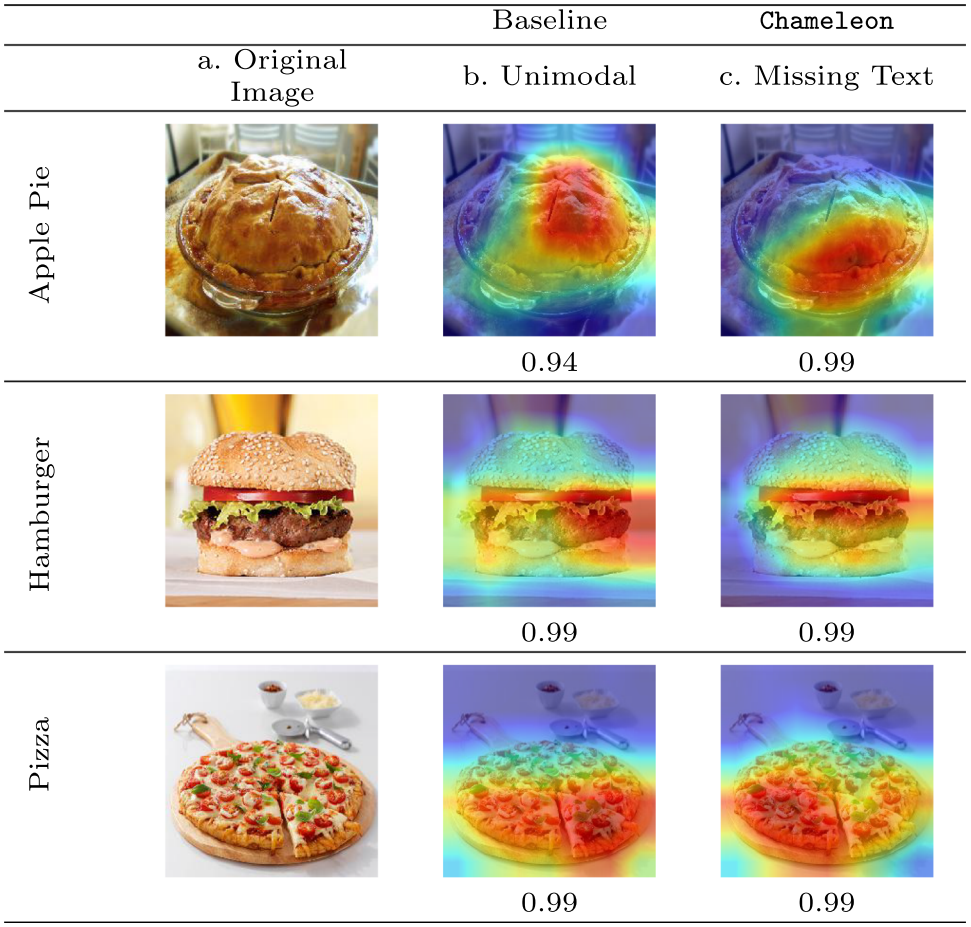


Fig. 3 (a) Orginal image. (b) Unimodal (image only) training and testing. (c) Multimodal training, unimodal testing (image only), i.e., 100% text missing. As seen, in unimodal image-only training (b), the model focuses on distinct features of the object. When the text modality (c) is missing during testing, the model focuses on the available modality to make correct final predictions demonstrating the success of our approach in training the multimodal system robust to missing modalities.

Table 1 Statistics of the datasets used in our experiments.

Dataset	Modalities	# of Classes	Train	Test
UPMC Food-101		101	67,988	22,716
Hateful Memes		2	8,500	2,000
MM-IMDb		23	15,552	7,799
Ferramenta		52	66,141	21,869
VoxCeleb		1251	145,265	8,251
avMNIST		10	1,125	375

operating characteristic (AUROC) for Hateful Memes and F1 macro-avaraged for MM-IMDb.

Implementation Details Unless specified otherwise, ViT (vit-base-patch16-224) is used as the default choice of visual learning network in `Chameleon`. For textual and audio modalities, we encode AnglE [47] and Ecapa-tdnn [48] embeddings. We utilize AnglE [47] for encoding textual modality into embeddings of size 1024 and Ecapa-tdnn [48] for encoding audio modality into embeddings of size 196. We provide further analysis on these choices in Sections 4.3.1, 4.3.3, & 4.3.5.

4.1 Evaluations Under Modality-complete Setting

To evaluate whether `Chameleon` is capable of learning from multiple data sources, we first experiment when complete modalities are present during training and testing. As seen in Table 2, `Chameleon` achieves SOTA performance on five of the six datasets. Specifically, `Chameleon` achieves classification performances of 92.9%, 73.1%, 51.5%, 96.9%, 99.5%, and 96.9%, respectively, on UPMC Food-101, Hateful memes, MM-IMDb, Ferramenta, avMNIST, and VoxCeleb. Overall, the results demonstrate that `Chameleon` is capable of realizing a performance comparable with that

Table 2 Comparison of Chameleon with SOTA multimodal methods on textual-visual (e.g., UPMC Food-101, Hateful Memes, MM-IMDb, Ferramenta) and audio-visual (e.g., avMNIST and VoxCeleb) datasets using modality-complete data. Best results are bold; second best are underlined.

UPMC Food-101 [18]	
Method	Acc.
Wang et al. [18]	85.1
Fused Rep. [49]	85.7
CLIPPO [42]	91.2
CentralNet [21]	91.5
ViLT [17]	91.9
Ma et al. [14]	92.0
MMBT [50]	92.1
BL [51]	<u>92.5</u>
Chameleon	**92.9** ± **0.3**
Hateful Memes [2]	
Method	AUROC
MMBT-G [50]	67.3
ViLT [17]	70.2
Clippo [42]	70.7
Ma et al. [14]	71.8
MMBT-R [50]	72.2
Visual BERT [52]	73.2
ViLBERT [28]	**73.4**
Chameleon	<u>73.1</u> ± 0.6
MM-IMDb [21]	
Method	F1 Macro
CBGP [1]	52.9
CBP [53]	53.2
GMU [19]	53.9
Lee et al. [16]	54.0
ViLT [17]	55.3
MFAS [54]	<u>55.7</u>
CentralNet [21]	**56.1**
Chameleon	51.5 ± 0.5
Ferramenta [20]	
Method	Acc.
Ferramenta [20]	92.9
Fused Rep. [49]	94.8
IeTF [55]	95.2
Two-Branch [26]	96.2
MHFNet [56]	<u>96.5</u>
Chameleon	**96.9** ± **0.4**
avMNIST [21]	
Method	Acc.
Moddrop [57]	94.8
Gated Multimodal Unit [19]	94.1
CentralNet [21]	95.0
SMIL [15]	98.2
Chameleon	**99.5** ± **0.2**
VoxCeleb [22]	
Method	Acc.
FOP [26]	92.9
SBNet. [58]	94.8
Chameleon	**96.9** ± **0.4**

of existing SOTA multimodal methods when learning from multiple modalities.

4.2 Evaluations Under Modality-missing Setting

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

4.2.1 Missing Modalities During Testing

Table 3 compares Chameleon with other methods: multimodal Transformers (ViLT [17], Ma et al. [14], CLIPPO [42]), and large vision-language model (LLaVA [59]) for varying amounts of missing modality on textual-visual datasets including UPMC Food-101, Hateful Memes, MM-IMDb, and Ferramenta during testing. As seen, Chameleon outperforms compared methods on UPMC Food-101, Hateful Memes, and Ferramenta demonstrating resilience against missing modalities. For example, in the case when only 10% of text modality is available on the UPMC Food-101 dataset, Chameleon demonstrates an accuracy of 75.9%. In comparison, ViLT [17], Ma et al. [14], CLIPPO [42] and LLaVA demonstrate performances of 58.4%, 73.3%, 74.1%, and 75.4%, respectively. Similar trends are noticeable on Hateful Memes and Ferramenta datasets. In the case of MM-IMDb dataset, though the performance of Chameleon is lower when all modalities are available, it outperforms ViLT, CLIPPO and LLaVa when a modality is severely missing. Overall, considering all four textual-visual datasets presenting 24 scenarios of complete and missing modality, Chameleon outperforms all compared methods in 18 while second best in 2 scenarios. Another interesting observation is seen when results of Chameleon are compared to CLIPPO which can be considered a baseline that renders text directly as an image. The superior performance of Chameleon demonstrates that directly rendering text as input may not be an optimal choice compared to our proposed feature encoding approach. Furthermore, we compare results with a large-

Table 3 Comparison of Chameleon with different levels of available modality at test time using textual and visual datasets (e.g., UPMC-Food-101, Hateful Memes, MM-IMDb, and Ferramenta). Comparison is provided with multimodal Transformers (ViLT [17]*, Ma et al. [14], and CLIPPO [42]). *ViLT values are taken from Ma et al. [14]. Boldface and underline denote, respectively, the best and second-best results. † indicates our implementation.

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

scale vision-language model (LLaVA), which converts image features to textual tokens. Though the methodology is different compared to Chameleon, the essence of converting one modality into the other is similar. Comparing results directly with CLIPPO and LLaVA demonstrates the effectiveness of the design choices in Chameleon. Furthermore, Fig. 4 compares Chameleon with other methods: FOP [26] and SBNet [58] for varying amounts of missing modality on audio-visual dataset (VoxCeleb). As seen, Chameleon significantly outperforms compared methods demonstrating resilience against missing modalities.

The overall persistent robustness to missing modality of Chameleon compared to existing methods across various challenging datasets and scenarios is attributed to proposed framework that enables the training of all input modalities in a common learning space. This way, if one modality is missing during testing, Chameleon shifts its focus to the available modality, thus yielding higher performance leveraging multimodal knowledge. More insights on this are provided in Fig. 3 where Grad-CAM visualizations clearly show the shifting of focus from multimodal data to the available modality for predictions. More visual examples are provided in the supplementary material.

4.2.2 Missing Modalities During Training

Derived from the motivation that modality can be missing in any data sample, Lee et al. [16] recently extended the evaluation protocol on textual-visual datasets by introducing scenarios of missing modality during training and testing, i.e., 30% of one modality is available against 100% of the other modality at train and test time. Table 4 compares Chameleon with existing methods ViLT [17], Lee et al. [16], and CLIPPO [42] using this protocol on UPMC Food-101, Hateful Memes, and MM-IMDb datasets. In most scenarios, Chameleon demonstrates comparable or

Table 4 Comparison of Chameleon with ViLT [17]*, Lee et al. [16], and CLIPPO [42] on UPMC Food-101 [18], Hateful Memes [2], and MM-IMDb [19] under different missing modality settings during training and testing. *ViLT values are taken from Lee et al. [16]. Boldface and underline denote, respectively, the best and second best results. † indicates our implementation.

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

Table 5 Performance analysis of Chameleon with various visual networks using UPMC-Food-101, and Hateful Memes datasets.

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

better performance than existing methods. Notably, compared to CLIPPO, Chameleon demonstrates robustness to missing modalities during training and testing thus reiterating the importance of our encoding scheme. Furthermore, Fig. 5 compares Chameleon with other methods: SMIL [15], Autoencoder (AE) [60], and Generative Adversarial Network (GAN) for varying amounts of missing modality on audio-visual (avMNIST). As seen, Chameleon significantly outperforms compared methods on avMNIST demonstrating resilience against missing modalities.

4.3 Analysis and Discussion

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

4.3.1 Is Chameleon Agnostic to the Visual Network?

Chameleon can generally adapt any visual network for training. Investigating the extent to which the visual clas-

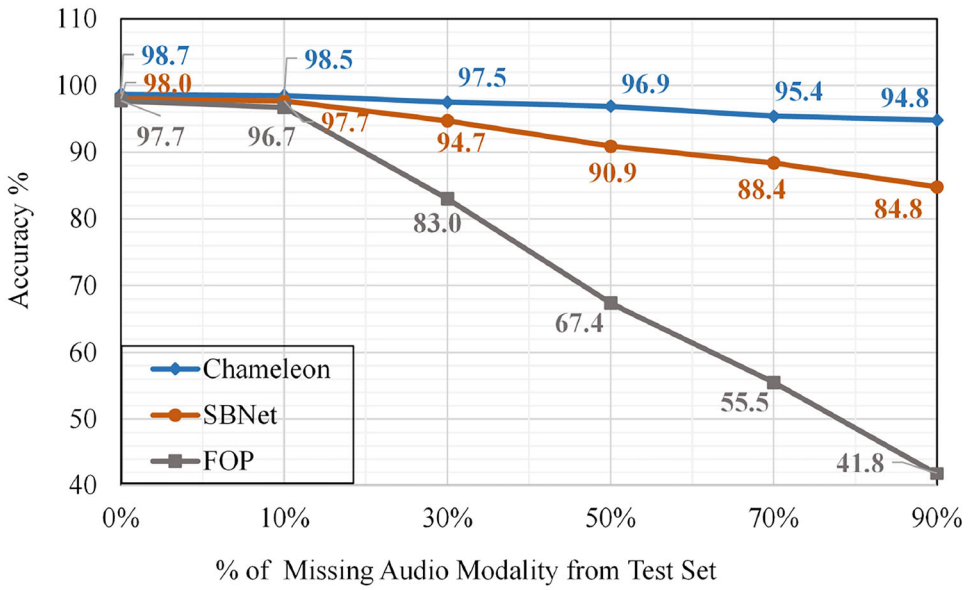


Fig. 4 Comparison of Chameleon with FOP [26], and SBNet [58] on audio-visual dataset (e.g., VoxCeleb) under different levels of missing modality at test time.

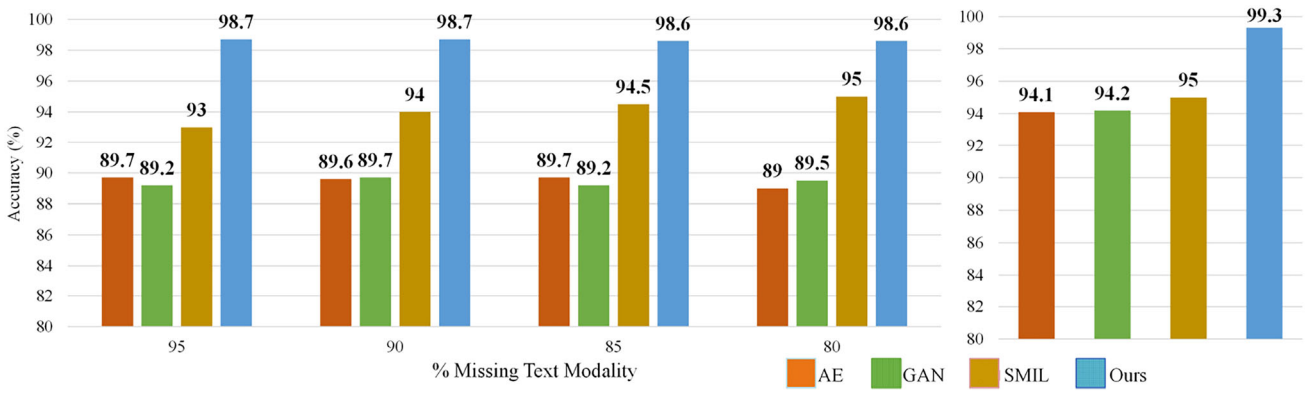
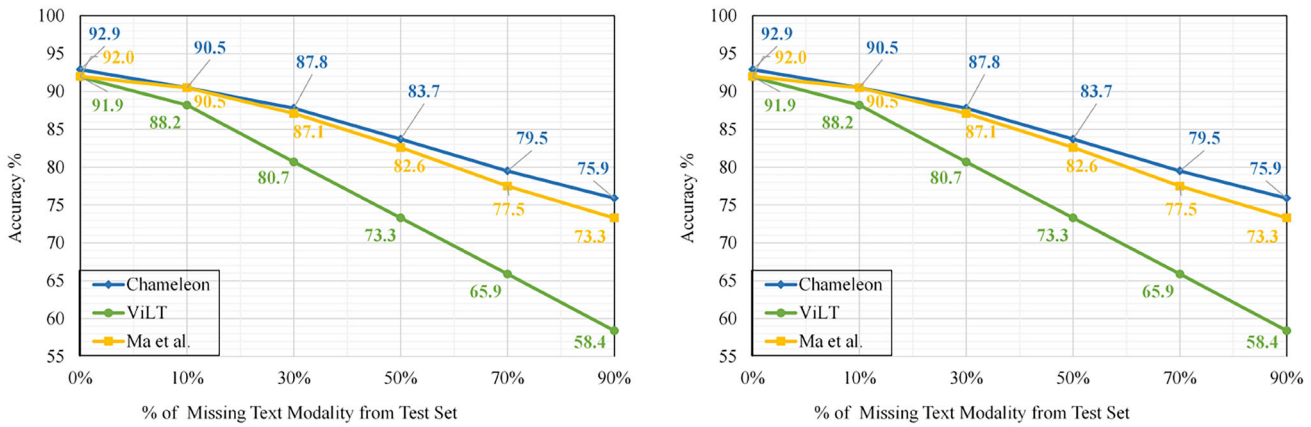


Fig. 5 Comparison of `Chameleon` with SMIL [15], Autoencoder (AE) [60], and Generative Adversarial Network (GAN) on avMNIST; an audio-visual dataset. (Left) training with 100% Image + n% Audio and testing with Image Only. (Right) training with 100% Image + 20% Audio and testing with Image + Audio.



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

Fig. 6 Comparisons of `Chameleon` with ViLT [17] and Ma et al. [14] on various levels of missing textual modality during testing on textual-visual datasets (e.g., UPMC Food-101 and Hateful Memes). A smaller drop in performance by our approach in most cases signifies its effectiveness towards training multimodal Transformer resilient to missing modalities without dataset-centric fusion strategies.

sification network influences performance, we conduct a series of experiments, using CNN (ResNet-101) [61], ViT [62], and Adapter [63]. The results of these experiments are summarized in Table 5 on a complete set of modalities. While adapting ResNet-101 yields 89.5% accuracy and 70.9% AUROC on UPMC Food-101 and Hateful Memes respectively, using Adapter [63] boosts the performance to 90.1% and 71.3%. Finally, using ViT yields the best performance of 92.9% and 73.9% on corresponding datasets. While the trend conforms with the existing literature comparing ResNets, ViTs, and Adapters, the overall competitive performance demonstrates that `Chameleon` is agnostic to the vision network.

Table 6 Cross-modal verification results on unseen-unheard configurations of `Chameleon` and existing state-of-the-art methods. Best results are highlighted in bold text whereas the second best are underlined.

Method	EER ↓ Unseen-Unheard	AUC ↑
DIMNet [64]	*24.9*	–
Learnable Pins [25]	29.6	78.5
MAV-Celeb [65]	29.0	78.9
Single Stream Net. [44]	29.5	78.8
Multi-view [66]	28.0	–
AML [67]	–	80.6
DRL [68]	25.0	**84.6**
FOP [26]	*24.9*	83.5
SBNet. [58]	25.7	82.4
`Chameleon`	**24.0**	*83.6*

Table 7 Comparison of different embeddings to encode textual modality under different levels of missing modality at test time.

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

4.3.2 Is Chameleon Suitable for Other Multimodal Task?

We conduct experiments to perform cross-modal verification task with `Chameleon` to establish face-voice association using VoxCeleb [22] dataset. We compare the performance of `Chameleon` against several SOTA face-voice association methods including DIMNet [64], Learnable Pins [25], MAV-Celeb [65], Single Stream Network [44], Multi-view [66], Adversarial Metric learning [67] (AML), Disentangled Representation Learning [68] (DRL), FOP [26], and Single Branch Network [58] (SBNet) Table 6 shows the results of cross-modal verification task on unseen-unheard configuration. We observe that `Chameleon` outperforms the other SOTA methods in terms of EER metric by achieving the EER scores of 24.0% whereas the existing best FOP [26] obtains the EER score of 24.9%. In terms of AUC metric, `Chameleon` obtains the favorably better score of 83.6% as compared to the DRL [68] that achieves the score of 84.6%.

4.3.3 Embedding Extractors

`Chameleon` can generally encode any audio or textual embedding for training. In Table 7, we compare the performance of various textual embeddings (e.g., AngIE [47], CLIP [29], MPNet [69], Bert [70] and Doc2Vec [71]) on UPMC Food-101 and Hateful Memes datasets under different levels of missing modality. Though, best performance is obtained with AngIE embeddings, other representations such as CLIP and MPNet also produces competitive results on complete as well as various levels of missing modalities.

4.3.4 Are Dataset-specific Transformers Necessary?

Ma et al. [14] have observed that multimodal Transformers are not only sensitive to missing modalities but different design choices, such as fusion strategy, should be tailored to the dataset. They further conclude that it may not be possible to design a general multimodal Transformer architecture to be used across datasets. In contrast, with our proposed design of encoding audio or text modality into visual format and then training a visual network such as ViT, `Chameleon` becomes independent of dataset-centric components. To evaluate this, we train two separate ViTs on the UPMC Food-101 and Hateful Memes datasets. Each model is then evaluated on different levels of missing textual modality at test time. Fig. 6 shows the corresponding results and compares the performances with ViLT [17] and Ma et al. [14]. As Ma et al. have proposed dataset-optimal fusion strategies, their approach demonstrates robustness to missing modality compared to ViLT. `Chameleon`, on the other hand demonstrates superior resilience against missing modality while training on similar ViTs without any changes across datasets. As seen, `Chameleon` enables the training of dataset-agnostic transformers capable of handling severe missing modalities.

Table 8 Comparison of word- and sentence-level embeddings using UPMC-Food-101, and Hateful Memes datasets.

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

4.3.5 Comparison Between Word-level and Sentence-level Embeddings

`Chameleon` utilizes non-visual-modality embeddings that are subsequently encoded into a common input modality. This enables any non-visual modality (e.g., audio or text) to be encoded into a visual format to train the network. However, in the case of text, another possibility of extracting embeddings would be to utilize word-level encoders (see Supplementary Section 3 for more details on word-level embedding). Table 8 compares word and sentence-level embedding performance on UPMC Food-101 and Hateful Memes datasets. As seen, the performance on both word and sentence embeddings are comparable, either one of them can be used depending on the task and application.

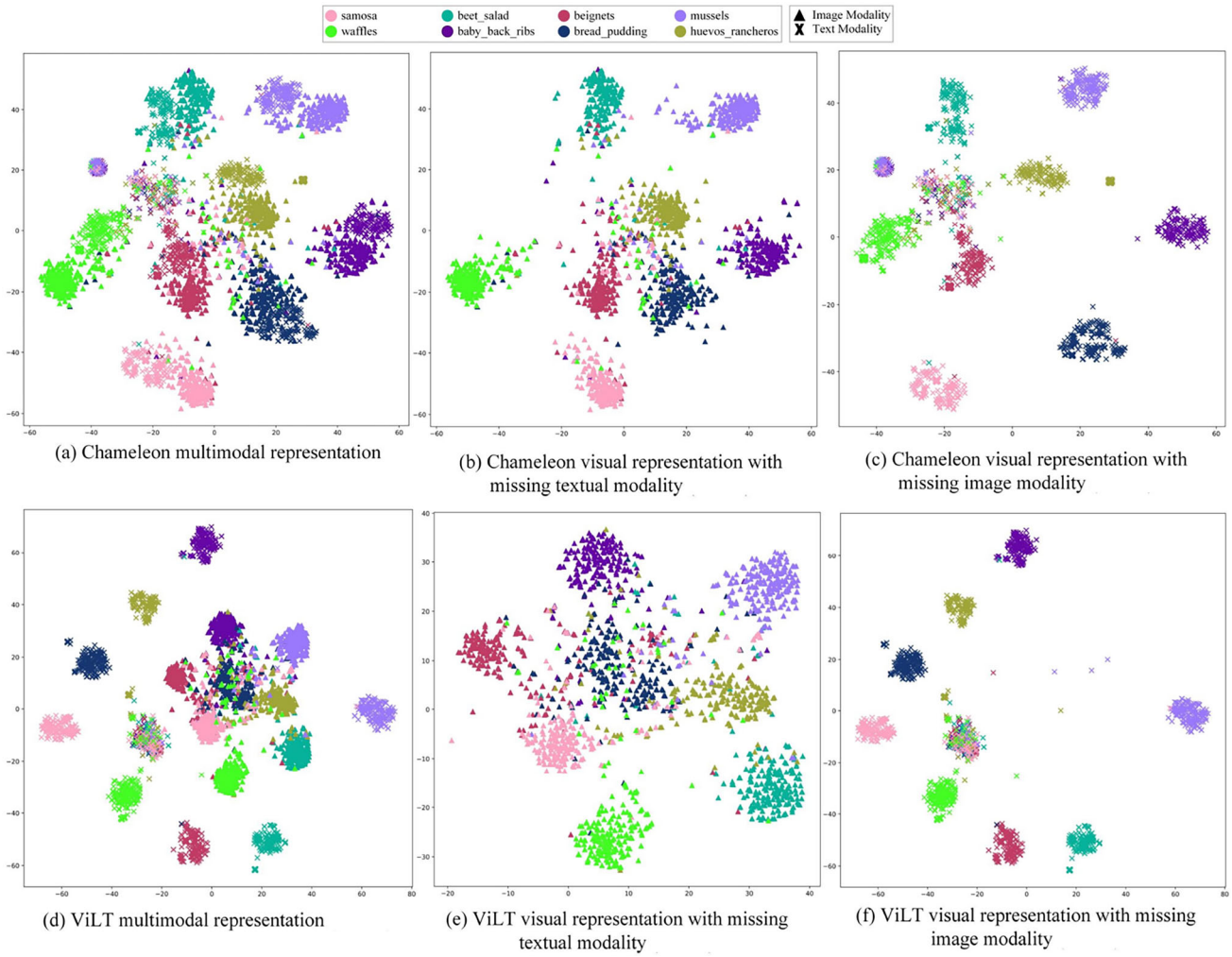


Fig. 7 t-SNE visualizations of the embedding space of `Chameleon` **(a – c)** and ViLT **(d - f)** along with accuracy on test set of UPMC Food-101. Compared to ViLT, `Chameleon` not only enhances the classification boundaries when complete modalities are available at test time but also retains these boundaries when the textual or visual modality is completely missing during test time. Note that classes are selected randomly from the test set.

4.3.6 Embedding Space Analysis

We plot results of t-SNE projections to take a peek into the embedding space of `Chameleon` trained on UPMC Food-101 dataset and provide comparisons with ViLT [17] in Fig. 7. Comparisons are provided under the following three settings: Fig. 7a & d) complete modalities during training and testing, b & e) complete modalities during training but 100% missing textual modality during testing, and c & f) complete modalities during training but 100% missing visual modality during testing. In the case of multimodal training/testing, compared to ViLT, `Chameleon` is able to project modalities belonging to the same classes closer to each other while demonstrating inter-class separability (Fig. 7a & d). This highlights the success of `Chameleon` in training a robust multimodal learning method. Finally, in the case of multimodal training but 100% missing modality during testing (Fig. 7b, c, e & f), although some distortions are noticeable, compared to ViLT, `Chameleon` successfully retains the overall separability of the classes under such severe missing modality case. This demonstrates the resilience of `Chameleon` towards missing modalities.

5 Conclusion

We presented a multimodal learning framework that encodes embeddings of non-visual modality into visual format to learn modality-independent representations of the input modalities. The common input format facilitates joint representation learning by sharing weights across multiple modalities resulting in multimodal learning robust to missing

modalities. Extensive experiments are performed on UPMC Food-101, Hateful Memes, MM-IMDb, Ferramenta, avMNIST, and VoxCeleb. The proposed method is thoroughly evaluated on complete modalities as well as missing modalities during training and testing. The experimental results indicate that the proposed framework obtains superior performance when a complete set of modalities is available. In the case of missing modalities, the performance deterioration is noticeably smaller than that of the existing multimodal learning methods, indicating significant robustness of `Chameleon`.

Acknowledgements This research was funded in whole or in part by the Austrian Science Fund (FWF): https://doi.org/10.55776/COE12, https://doi.org/10.55776/DFH23, https://doi.org/10.55776/P36413.

Funding Open access funding provided by Johannes Kepler University Linz.

References

1. Kiela D, Grave E, Joulin A, Mikolov T (2018). Efficient large-scale multi-modal classification. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 5198–5204
2. Kiela D, Firooz H, Mohan A, Goswami V, Singh A, Ringshia P, Testuggine D (2020) The hateful memes challenge: Detecting hate speech in multimodal memes. Advances in Neural Information Processing Systems 33:2611–2624
3. Wang L, Li Y, Lazebnik S (2016). Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5005–5013
4. Nagrani A, Albanie S, Zisserman A (2018). Seeing voices and hearing faces: Cross-modal biometric matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8427–8436
5. Moon S, Neves L, Carvalho V (2018). Multimodal named entity recognition for short social media posts. In: Walker, M., Ji, H., Stent, A. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 852–860. Association for Computational Linguistics, New Orleans, Louisiana . https://doi.org/10.18653/v1/N18-1078 . https://aclanthology.org/N18-1078/
6. Arshad O, Gallo I, Nawaz S, Calefati A (2019). Aiding intra-text representations with visual context for multimodal named entity recognition. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 337–342
7. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L(2018). Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6077–6086
8. Fukui A, Park D.H, Yang D, Rohrbach A, Darrell T, Rohrbach M (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv:1606.01847
9. Vinyals O, Toshev A, Bengio S, Erhan D (2015). Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164
10. Kiela D, Bottou L (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 36–45
11. Tian Y, Jian G, Wang J, Chen H, Pan L, Xu Z, Li J, Wang R (2023) A revised approach to orthodontic treatment monitoring from oralscan video. IEEE Journal of Biomedical and Health Informatics 27(12):5827–5836
12. Specia L, Frank S, Sima'An K, Elliott D (2016). A shared task on multimodal machine translation and crosslingual image description. In: Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, pp. 543–553
13. Elliott D, Frank S, Sima'an K, Specia L (2016). Multi30k: Multilingual english-german image descriptions. arXiv:1605.00459
14. Ma M, Ren J, Zhao L, Testuggine D, Peng X (2022). Are multimodal transformers robust to missing modality? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18177–18186
15. Ma M, Ren J, Zhao L, Tulyakov S, Wu C, Peng X (2021). Smil: Multimodal learning with severely missing modality. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 2302–2310
16. Lee Y.-L, Tsai Y.-H, Chiu W.-C, Lee C.-Y (2023). Multimodal prompting with missing modalities for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14943–14952
17. Kim W, Son B, Kim I (2021). Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning, pp. 5583–5594
18. Wang X, Kumar D, Thome N, Cord M, Precioso F (2015). Recipe recognition with large multimodal food dataset. In: 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 1–6
19. Arevalo J, Solorio T, Montes-y-Gómez M, González F.A (2017). Gated multimodal units for information fusion. arXiv:1702.01992
20. Gallo I, Calefati A, Nawaz S (2017) Multimodal classification fusion in real-world scenarios. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 5, pp. 36–41
21. Vielzeuf V, Lechervy A, Pateux S, Jurie F (2018). Centralnet: a multilayer approach for multimodal fusion. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp. 0–0
22. Nagrani A, Chung J.S, Zisserman A (2017). Voxceleb: a large-scale speaker identification dataset. In: INTERSPEECH
23. Baltrušaitis T, Ahuja C, Morency L-P (2018) Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence 41(2):423–443
24. Xu P, Zhu X, Clifton DA (2023) Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(10):12113–12132

25. Nagrani A, Albanie S, Zisserman A (2018). Learnable pins: Cross-modal embeddings for person identity. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 71–88
26. Saeed M.S, Khan M.H, Nawaz S, Yousaf M.H, Del Bue A (2022). Fusion and orthogonal projection for improved face-voice association. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7057–7061
27. Kim C, Shin H.V, Oh T.-H, Kaspar A, Elgharib M, Matusik W (2018). On learning associations of faces and voices. In: Asian Conference on Computer Vision, pp. 276–292 . Springer
28. Lu J, Batra D, Parikh D, Lee S (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32**
29. Radford A, Kim J.W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al (2021). Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 . PMLR
30. He X, Peng Y (2017). Fine-grained image classification via combining vision and language. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5994–6002
31. Yang F, Peng X, Ghosh G, Shilon R, Ma H, Moore E, Predovic G (2019). Exploring deep multimodal fusion of text and photo for hate speech classification. In: Proceedings of the Third Workshop on Abusive Language Online, pp. 11–18
32. Zhang X, Song Q, Liu G (2022) Multimodal image aesthetic prediction with missing modality. Mathematics 10(13):2312
33. Suo Q, Zhong W, Ma F, Yuan Y, Gao J, Zhang A (2019). Metric learning on healthcare data with incomplete modalities. In: IJCAI, pp. 3534–3540
34. Li M, Yang D, Liu Y, Wang S, Chen J, Wang S, Wei J, Jiang Y, Xu Q, Hou X, et al (2024). Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. arXiv:2411.02793
35. Zhang C, Chu X, Ma L, Zhu Y, Wang Y, Wang J, Zhao J (2022) M3care: Learning with missing modalities in multimodal healthcare data. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 2418–2428
36. Wang N, Cao H, Zhao J, Chen R, Yan D, Zhang J (2022). M2r2: Missing-modality robust emotion recognition framework with iterative data augmentation. IEEE Transactions on Artificial Intelligence
37. Wang H, Chen Y, Ma C, Avery J, Hull L, Carneiro G (2023). Multimodal learning with missing modality via shared-specific feature modelling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15878–15887
38. Lan G, Du Y, Yang Z (2024) Robust multimodal representation under uncertain missing modalities. ACM Transactions on Multimedia Computing, Communications and Applications
39. Gallo I, Nawaz S, Calefati A (2017). Semantic text encoding for text classification using convolutional neural networks. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 5, pp. 16–21
40. Salesky E, Etter D, Post M (2021). Robust open-vocabulary translation from visual text representations. arXiv:2104.08211
41. Rust P, Lotz J.F, Bugliarello E, Salesky E, Lhoneux M, Elliott D (2022). Language modelling with pixels. arXiv:2207.06991
42. Tschannen M, Mustafa B, Houlsby N (2023). Clippo: Image-and-language understanding from pixels only. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11006–11017
43. Xie W, Nagrani A, Chung J.S, Zisserman A (2019). Utterance-level aggregation for speaker recognition in the wild. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5791–5795
44. Nawaz S, Janjua M.K, Gallo I, Mahmood A, Calefati A (2019). Deep latent space learning for cross-modal mapping of audio and visual signals. In: 2019 Digital Image Computing: Techniques and Applications (DICTA), pp. 1–7
45. Selvaraju R.R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626
46. Kiela D, Firooz H, Mohan A, Goswami V, Singh A, Fitzpatrick C.A, Bull P, Lipstein G, Nelli T, Zhu R, et al (2021). The hateful memes challenge: competition report. In: NeurIPS 2020 Competition and Demonstration Track, pp. 344–360
47. Li X, Li J (2023). Angle-optimized text embeddings. arXiv:2309.12871
48. Desplanques B, Thienpondt J, Demuynck K (2020). Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. arXiv:2005.07143
49. Nawaz S, Calefati A, Janjua MK, Anwaar MU, Gallo I (2018) Learning fused representations for large-scale multimodal classification. IEEE Sensors Letters 3(1):1–4
50. Kiela D, Bhooshan S, Firooz H, Perez E, Testuggine D (2019). Supervised multimodal bitransformers for classifying images and text. arXiv preprint arXiv:1909.02950
51. Gallo I, Ria G, Landro N, La Grassa R (2020). Image and text fusion for upmc food-101 using bert and cnns. In: 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ), pp. 1–6 . IEEE
52. Li L, Yatskar M, Yin D, Hsieh C, Chang K (2019). A simple and performant baseline for vision and language. arXiv:1908.03557
53. Fukui A, Park D.H, Yang D, Rohrbach A, Darrell T, Rohrbach M (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Su, J., Duh, K., Carreras, X. (eds.) Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 457–468. Association for Computational Linguistics, Austin, Texas . https://doi.org/10.18653/v1/D16-1044 . https://aclanthology.org/D16-1044
54. Pérez-Rúa J.-M, Vielzeuf V, Pateux S, Baccouche M, Jurie F (2019). Mfas: Multimodal fusion architecture search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6966–6975
55. Gallo I, Calefati A, Nawaz S, Janjua M.K (2018). Image and encoded text fusion for multi-modal classification. In: 2018 Digital Image Computing: Techniques and Applications (DICTA), pp. 1–7
56. Yue T, Li Y, Qin J, Hu Z (2023). Multi-modal hierarchical fusion network for fine-grained paper classification. Multimedia Tools and Applications, 1–17
57. Neverova N, Wolf C, Taylor G, Nebout F (2015) Moddrop: adaptive multi-modal gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(8):1692–1706
58. Saeed M.S, Nawaz S, Khan M.H, Zaheer M.Z, Nandakumar K, Yousaf M.H, Mahmood A (2023). Single-branch network for multimodal training. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5
59. Liu H, Li C, Wu Q, Lee Y.J (2023). Visual instruction tuning. In: NeurIPS
60. Lee H.-C, Lin C.-Y, Hsu P.-C, Hsu W.H (2019). Audio feature generation for missing modality problem in video action recognition. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3956–3960
61. He K, Zhang X, Ren S, Sun J (2016). Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778

62. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929
63. Marouf I.E, Tartaglione E, Lathuilière S (2024). Mini but mighty: Finetuning vits with mini adapters. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1732–1741
64. Wen Y, Ismail M.A, Liu W, Raj B, Singh R (2019). Disjoint mapping network for cross-modal matching of voices and faces. In: 7th International Conference on Learning Representations, ICLR 2019, USA, May 6-9, 2019
65. Nawaz S, Saeed M.S, Morerio P, Mahmood A, Gallo I, Yousaf M.H, Del Bue A (2021). Cross-modal speaker verification and recognition: A multilingual perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1682–1691
66. Sarı L, Singh K, Zhou J, Torresani L, Singhal N, Saraf Y (2021). A multi-view approach to audio-visual speaker verification. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6194–6198 . IEEE
67. Zheng A, Hu M, Jiang B, Huang Y, Yan Y, Luo B (2021) Adversarial-metric learning for audio-visual cross-modal matching. IEEE Transactions on Multimedia 24:338–351
68. Ning H, Zheng X, Lu X, Yuan Y (2021) Disentangled representation learning for cross-modal biometric matching. IEEE Transactions on Multimedia 24:1763–1774
69. Song K, Tan X, Qin T, Lu J, Liu T-Y (2020) Mpnet: Masked and permuted pre-training for language understanding. Advances in neural information processing systems 33:16857–16867
70. Devlin J, Chang M.-W, Lee K, Toutanova K (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota . https://doi.org/10.18653/v1/N19-1423 https://aclanthology.org/N19-1423
71. Mikolov T, Sutskever I, Chen K, Corrado G.S, Dean J (2013). Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119