Deep Multimodal Learning with Missing Modality: A Survey

Anonymous authors
Paper under double-blind review

Abstract

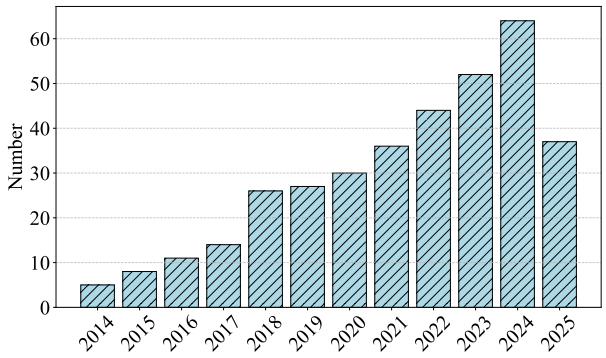
During multimodal model training and testing, certain data modalities may be absent due to sensor limitations, cost constraints, privacy concerns, or data loss, which can degrade performance. Multimodal learning techniques that explicitly account for missing modalities aim to improve robustness by enabling models to perform reliably even when certain inputs are unavailable. This survey presents the first comprehensive review of Multimodal Learning with Missing Modality (MLMM), with a focus on deep learning approaches. We outline the motivations and key distinctions between MLMM and conventional multimodal learning, provide a detailed analysis of existing methods, applications, and datasets, and conclude by highlighting open challenges and future research directions.

1 Introduction

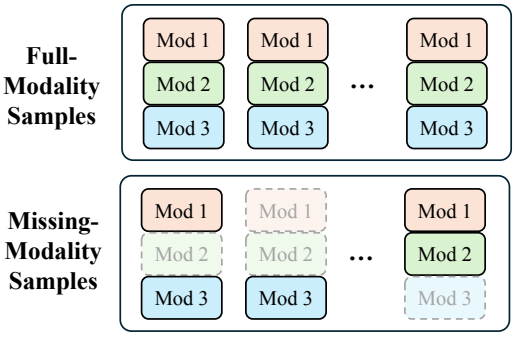
Multimodal learning has become a crucial field in Artificial Intelligence (AI). It focuses on jointly analyzing various data modalities, including visual, textual, auditory, and sensory information. This approach mirrors the human capacity to combine multiple senses for better understanding and interaction with the environment. Modern multimodal models leverage the robust generalization capabilities of deep learning to uncover complex patterns and relationships that single-modal systems might not detect. This capability is advancing work across multiple domains (Bayoudh et al., 2022; Hu et al., 2023; Streli et al., 2023). Recent surveys show that multimodal learning significantly boosts performance and enables more advanced AI applications (Baltrušaitis et al., 2018; Zhao et al., 2024).

However, real-world multimodal systems frequently encounter scenarios where certain data modalities are missing or incomplete. Such cases arise from sensor malfunctions, hardware limitations, privacy concerns, environmental interference, and data transmission failures. Interest in addressing this challenge has grown considerably in recent years, as reflected in the publication trends shown in figure 1a. For instance, in a three-modality setting, data samples can be classified as either full-modality (containing all three modalities) or missing-modality (lacking one or more). Missing modalities can occur at any stage, from data collection to deployment, and can substantially degrade model performance.

Early works in affective computing (Cohen et al., 2004; Sebe et al., 2005) reported that when collecting facial and audio data, some image or audio samples were not useful due to excessive microphone noise or camera obstructions. This has forced them to propose audio-visual models that can handle the missing modality to recognize human emotional states. In the field of medical AI, privacy concerns and the challenges of obtaining certain data modalities during surgeries or invasive treatments often lead to missing modalities in multimodal datasets (Zhang et al., 2023b; Li et al., 2025b). Similarly, in space exploration, NASA's Ingenuity Mars helicopter (Donaldson, 2024) faced a missing modality challenge when its inclinometer failed due to extreme temperature cycles on Mars. To address this issue, NASA applied a software patch that modified the initialization of navigation algorithms (Team, 2022). Beyond such domain-specific cases, structural or quantitative heterogeneity across sensor types, versions, or brands can also yield inconsistent modality availability. Combined with the unpredictability of real-world scenarios and the diversity of data sources, these factors make robustness to missing modalities an essential requirement for multimodal systems. Consequently, developing models that can effectively handle incomplete inputs has become a central research focus.



(a) Trends in Missing Modality Publications (2014-2025 August)



(b) Full-Modality Samples vs. Missing-Modality Samples

Figure 1: (a) The trend of papers published on deep multimodal learning with missing modality in the past 10 years. The number of publications has increased over time and has received widespread attention from the community. (b) Description of a three-modality scenario with full- and missing-modality samples. We abbreviate "Modality" as "Mod" in all figures of this paper and use dashed boxes with fading colors to represent missing modalities/modules.

In this survey, we focus on the challenge of handling incomplete inputs in multimodal learning, which we refer to as the **missing modality problem**. We term solutions to this challenge **Multimodal Learning with Missing Modality (MLMM)**. Unlike the conventional full-modality setting, where all **N** modalities are assumed to be consistently available, MLMM must operate when one or more modalities are absent during training or testing. The central goal of MLMM is to robustly exploit the available modalities while maintaining performance comparable to systems trained with complete modality information.

This survey reviews recent advances in MLMM and its applications across diverse domains, including information retrieval (Malitesta et al., 2024; Pipoli et al., 2025), remote sensing (Wei et al., 2023), robotic vision (Gunasekar et al., 2020; Park et al., 2025), medical diagnosis (Zhou et al., 2023; Shah et al., 2023; Azad et al., 2022; Zhu et al., 2025), sentiment analysis (Wagner et al., 2011; Wang et al., 2025), and multi-view clustering (Chao et al., 2021). We also introduce a fine-grained taxonomy of MLMM methodologies, application scenarios, and corresponding datasets.

Contributions: **(1)** A comprehensive survey of MLMM methods across diverse domains, accompanied by an extensive compilation of relevant datasets, highlighting the versatility of MLMM in addressing real-world challenges. **(2)** A novel, fine-grained taxonomy of MLMM methodologies, with a multi-faceted categorization framework based on multi-modal integration stages and missing modality recovery strategies. **(3)** An in-depth analysis of current MLMM approaches, their challenges, and future research directions, contextualized within the proposed taxonomical framework.

Paper Collection: In our literature search methodology, we primarily source papers from Google Scholar and major conferences/journals in AI, Machine Learning, Computer Vision, Natural Language Processing, Audio Signal Processing, Data Mining, Multimedia, Medical Imaging, and Remote Sensing. The collected papers are from, but not limited to, top-tier conferences (e.g., AAAI, IJCAI, NeurIPS, ICLR, ICML, CVPR, ICCV, ECCV, ACL, EMNLP, KDD, ACM MM, MICCAI, ICASSP) and journals (e.g., TPAMI, TIP, TMI, TMM, JMLR). Please refer to Section A in the Supplementary Materials for the full names of these conferences and journals. We have compiled a total of 354 significant papers from the period spanning 2012 to August 2025. Our search strategy involved using keywords such as "incomplete," "missing," "partial," "absent," and "imperfect," combined with terms like "multimodal learning," "deep learning," "representation learning," "multi-view learning," and "neural networks."

Survey Organization: Firstly, we explain the background and motivation of this survey in section 1. In section 2, we introduce our taxonomy and categorize existing deep MLMM methods from a methodological perspective, detailing them in two aspects and four types (figure 2). In the following section 3 and section 4,

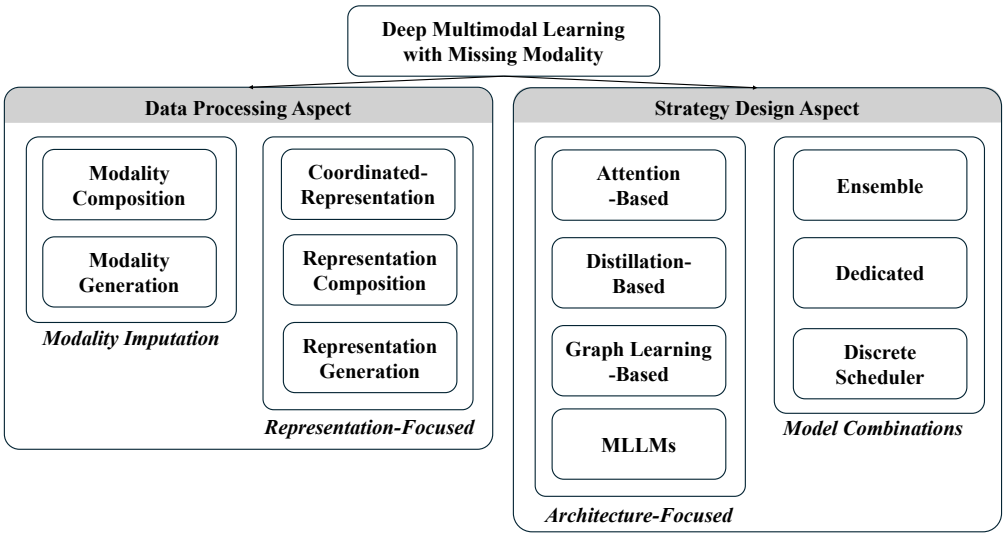


Figure 2: Our taxonomy of deep multimodal learning with missing modality methods. We categorize existing methods into two aspects: data processing and strategy design. **Data Processing**: we differentiate between modality imputation (handling at the modality data level) and representation-focused models (dealing with at the data representation level). **Strategy Design**: we distinguish between architecture-focused models (model architecture adjustments) and model combinations (combining multiple models externally). "MLLMs": multimodal large language models.

we introduce various methods from the aspects of model data processing and strategy design respectively. We then summarize current application scenarios and corresponding datasets used in this field in section 6. In section 7, we discuss unresolved challenges and future directions. Finally, we present our conclusions drawn from the exploration of deep MLMM in section 8.

2 Methodology Taxonomy: An Overview

We review current deep MLMM methods from two key aspects: data processing and strategy design, based on the contributions of existing work.

2.1 Data Processing Aspect

Methods that focus on exploring the data processing aspect of a model or framework can be divided into *Modality Imputation* and *Representation-Focused Models*, depending on whether the model's handling of missing modalities occurs at the modality data level or at the data representation level.

(1) Modality Imputation operates at the modality data level and fills in the missing information by compositing (Chen et al., 2020; Sun et al., 2024b; Parthasarathy & Sundaram, 2020; Yang et al., 2024; Campos et al., 2015) (modality composition methods) or generating (Chen et al., 2024a; Ma et al., 2021c; Arya & Saha, 2021; Wang et al., 2024c; Tran et al., 2017; Zhang et al., 2023c) absent modalities (modality generation methods) from available modalities. Those approaches are rooted in the idea that if a missing modality can be accurately imputed, downstream tasks can continue as if "full" modalities were available.

(2) Representation-Focused Models are designed to address missing modalities at the representation level. In some cases, the coordinated representation methods (Liang et al., 2019; Wang et al., 2020b; Ma et al., 2021b) impose some specific constraints to help align the representations of available modalities in the same semantic space, so that the model can be effectively trained even when facing missing modalities. Other representation-focused methods either generate the missing-modality representation using available data (Zheng et al., 2021; Li et al., 2024b; Hoffman et al., 2016; Ma et al., 2021c; Park et al., 2024) or

combine the representations of existing modalities (Zhi et al., 2024; Sun et al., 2024b; Zhou et al., 2021; Delgado-Escano et al., 2021; Zheng et al., 2024; Zhang et al., 2024) to fill in the gaps.

2.2 Strategy Design Aspect

Methods that explore the strategy design aspect are based on models that can dynamically adapt to different missing-modality cases during training and testing through flexible adjustments of the model architecture (internal model architecture adjustment) and the combination of multiple models (external model combinations). We name them *Architecture-Focused Models* and *Model Combinations*.

(1) Architecture-Focused Models address missing modalities by designing flexible model architectures that can adapt to varying numbers of available modalities during training or inference. A key technique here is based on attention mechanisms (Ge et al., 2023; Yao et al., 2024; Gong et al., 2023; Mordacq et al., 2024; Radosavovic et al., 2024; Jang et al., 2024; Qiu et al., 2023), which dynamically adjusts the modality fusion and processing, allowing the model to handle any number of input modalities. Another approach is based on knowledge distillation (Wang et al., 2020a; Saha et al., 2024; Zhang et al., 2023b; Chen et al., 2024b; Wang et al., 2023a; Shi et al., 2024; Wang et al., 2023c), where the model is trained to accommodate missing modalities by transferring knowledge from full-modality models to those operating with incomplete data or distilling between different branches inside the model. Additionally, graph learning-based methods (Zhao et al., 2022; Zhang et al., 2022; Lian et al., 2023; Malitesta et al., 2024) exploit the natural relationships between modalities, using graphs to dynamically fuse and process available modalities while compensating for missing ones. Finally, MLLMs (Zhan et al., 2024; Wu et al., 2023b; Li et al., 2023a) also play a crucial role in this category, as their ability to handle long contexts and act as feature processors enables them to accept and process representations from any number of modalities. These architectural strategies collectively allow models to maintain performance even when dealing with incomplete multimodal inputs.

(2) Model Combinations tackle missing modality problems by employing strategies that leverage multiple models or specialized training techniques. One approach is to use dedicated training strategies (Zeng et al., 2022; Chen et al., 2022; Xue & Marculescu, 2023) tailored for different modality cases, ensuring that each case is trained for optimal performance. Another approach involves ensemble methods (Hu et al., 2020; Wang et al., 2021; Lee et al., 2023b), where models trained on either partial/full sets of modalities are combined, allowing the system to select the most suitable model based on the available modalities to do joint predictions. Additionally, discrete scheduler methods (Wu et al., 2023a; Shen et al., 2024; Surís et al., 2023) can incorporate various downstream modules to flexibly process any number of modalities and handle specific tasks. These schedulers intelligently select and combine the outputs of multiple models or modules to manage missing-modality scenarios, offering a versatile solution for multimodal tasks.

Our taxonomy (figure 2) can reflect different aspects and levels of multimodal learning—ranging from modality data to data representations, architectural design, and model combinations—each providing a distinct way to approach the problem of missing modalities based on the task requirements and available resources.

3 Methodologies in Data Processing Aspect

3.1 Modality Imputation

Modality imputation refers to a technique used by MLMM methods that fills in missing-modality samples or generates missing modalities to complete the dataset with missing modalities by performing various transformations or operations on existing modalities. Modality imputation methods addressing the missing modality problem at the data modality level can be categorized into two types. (1) Modality composition methods use zero/random values or data copied from similar instances as the input for the missing modality data. The data produced via these methods to represent the missing data are then composited with the data from the available modalities to form a "full"-modality sample. (2) Modality generation methods generate the missing modality data using generative models such as Auto-Encoders (Hinton & Zemel, 1993), Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), or Diffusion Models (Ho et al., 2020).

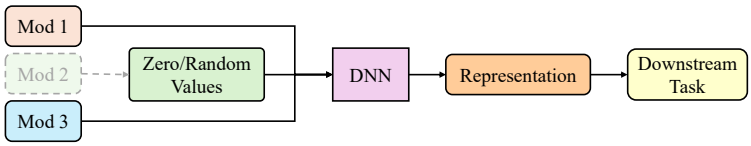


Figure 3: Zero/Random values composition methods. If we assume modality 2 is missing, then this modality will be replaced with zero/random values. "DNN" in all figures of this survey means different kinds of deep neural networks.

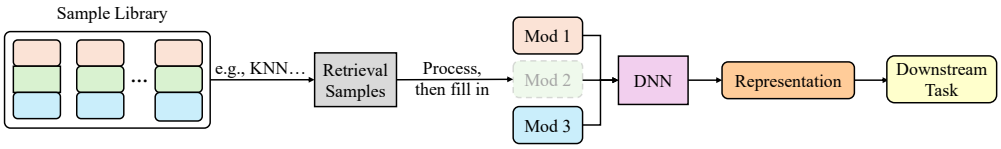


Figure 4: Retrieval-based modality composition methods search for one or more samples by randomly selecting or using simple retrieval algorithms like KNN, or its variants, from same-category samples that have the required missing modalities, and then compose them with the input missing-modality sample to form a "full"-modality sample.

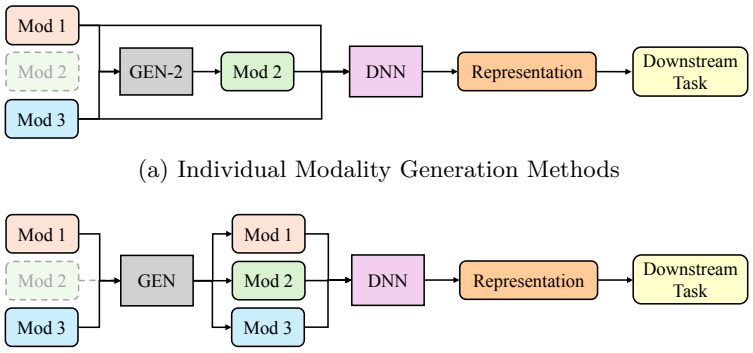
The generated data is then composited with the data from the available modalities to form a "full"-modality sample. We provide more details about these two methods in the next sub-sections.

3.1.1 Modality Composition Methods

Modality composition methods are widely employed for its simplicity and effectiveness to maintain the original dataset size. Zero/Random Values Composition Methods represent a type of modality composition method that replaces a missing modality with zeros or random values, as shown in figure 3. In recent research (Chen et al., 2020; Sun et al., 2024b; Liu et al., 2023a; Malitesta et al., 2024), they are often used as a baseline method for comparison with other more sophisticated methods. For the missing sequential data problem, such as missing frames in videos, the similar Frame-Zero method (Parthasarathy & Sundaram, 2020) was proposed to replace the missing frames. These methods are prevalent in typical multimodal learning procedures and can be used to balance and integrate information from different modalities when making predictions. This prevents the model from over-relying on dominant modalities available for each sample, enhancing its robustness by encouraging a more balanced integration of information across all available modalities.

Retrieval-Based Representation Composition Methods (figure 4) represent another type of modality composition method, which replaces the missing modality data by copying or averaging data from retrieved samples with the same classification. Some other methods randomly select a sample that has the same classification and required missing modality from other samples. The selected modality data is then composed with the missing-modality sample to form a full-modality sample for training. But those retrieval-based modality composition methods are not applicable to pixel-level tasks, such as segmentation, and are only suitable for simple tasks (e.g., classification) because they may lead to the overfitting of noisy data, if mismatched samples are combined. For example, Yang et al. (Yang et al., 2024) proposed Modal-mixup to complete the training datasets by randomly complementing same-category samples with missing modalities. However, such methods cannot solve the missing modality problem during the testing phase because they rely on known labels of training data samples. In some multimodal streaming data classification tasks, such as audio-visual expression recognition, video streams may experience frame drops due to network communication packet loss, etc. Frame-Repeat (Parthasarathy & Sundaram, 2020) was proposed to to make up for the missing frames by using past frames.

Other methods (Yang et al., 2024; Campos et al., 2015) also used K-Nearest Neighbors (KNN), or its variants, to retrieve the best-matched samples for composition. For these matched samples, they select the sample with the highest score or obtain the average values of these samples to supplement the missing modality



(a) Individual Modality Generation Methods

(b) Unified Modality Generation Methods

Figure 5: Description of two typical modality generation methods. We set modality 2 as the missing modality for examples and use other available modalities to generate modality 2. "GEN" in both figures represents modality generation networks. (a) We set up a modality-2 generator (GEN-2) from other modalities. (b) All modalities are input and generated together by using a single GEN.

data. Their experiments have shown that KNN-based methods generally perform better than the above methods, and can handle missing modality during testing. Nevertheless, such KNN retrieval-based modality composition methods often suffer from high computational complexity, sensitivity to imbalanced data, and significant memory overhead.

All methods above can complete datasets with missing modalities, but they reduce the diversity of the dataset because they may introduce duplicated training samples. This is especially problematic with datasets having a high rate of modality missing, where most samples have missing-modality data, as it increases the risk of overfitting to certain classes with few full-modality samples, if they pad missing-modality samples with duplicated ones.

3.1.2 Modality Generation Methods

With deep learning, synthesizing missing modalities has become more effective by leveraging powerful representation learning and generative models that can capture complex cross-modal relationships. Current methods for generating missing-modality data are divided into individual and unified generative methods.

Individual Modality Generation Methods train an individual generative model for each modality in case any modality is missing, as shown in figure 5a. Early works used Gaussian processes (Mario Christoudias et al., 2010) or Boltzmann machines (Sohn et al., 2014) to generate missing modalities from available data. With deep learning, models like Auto-Encoders (AEs) (Hinton & Zemel, 1993) can be used for missing modality generation. Li et al. (Li et al., 2014) used 3D-CNN to generate positron emission tomography (PET) data from magnetic resonance imaging (MRI) inputs. Chen et al. (Chen et al., 2024a) addressed missing modalities in MRI segmentation by training U-Net models to generate other two modalities from MRI. A recent work (Ma et al., 2021c) used AEs as one of the baselines to complete datasets by training one AE per modality. In domain adaptation, Zhang et al. (Zhang et al., 2021) proposed a U-Net-based Multi-Modality Data Generation module with domain adversarial learning to generate each missing modality by learning domain-invariant features.

GANs significantly improved image generation quality by using a generator to create realistic data and a discriminator to distinguish it from real data. Researchers therefore began replacing above models with GANs for modality generation. For example, GANs generate missing modalities using latent representations of existing ones in breast cancer prediction (Arya & Saha, 2021). In remote sensing, Bischke et al. (Bischke et al., 2018) used GANs to generate depth data, improving segmentation over RGB-only models. GANs were also used in robotic object recognition models to help generate missing RGB and depth imagesGunasekar et al. (2020). Recent studies (Ma et al., 2021c) show that GANs outperform AEs in generating more realistic missing modalities and can lead to better downstream-task model performance. Recently, the

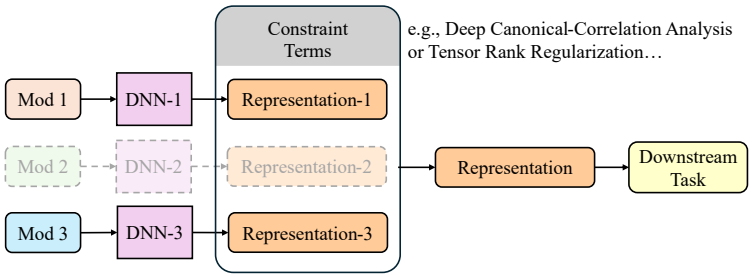


Figure 6: Illustration of coordinated-representation methods, which impose constraints to make the learned representations semantically consistent. Assuming that modality 2 is missing, these methods use constraints terms between the features of different modalities.

introduction of Diffusion models has further improved image generation quality. Wang et al. proposed the IMDer method (Wang et al., 2024c), which uses available modalities as conditions to make diffusion models generate missing modalities. Experiments showed it can reduce semantic ambiguity between recovered and missing modalities and achieves good generalization performance than previous works.

Unified Original Data Generative Methods train a unified model that can generate all modalities simultaneously (figure 5b). One representative model is Cascade AE (Tran et al., 2017), which stacks AEs to capture the differences between missing and existing modalities for generating all missing modalities. Recently, Zhang et al. (Zhang et al., 2023c), have attempted to use attention and max-pooling to integrate features of existing modalities, enabling modality-specific decoders to accept available modalities and generate other missing modalities. Experiments demonstrated that this method is more effective than using max-pooling alone (Chartsias et al., 2017) to integrate features from any number of modalities for generating all missing modalities together. Above methods for generating missing modalities can mitigate performance degradation to some extent. However, when facing a scenraio where one of the modalities is severely missing, training a generator that can produce high-quality missing modalities remains challenging. Additionally, the modality generation model significantly increases storage and computational requirements. As the number of modalities grows, the complexity of these generative models also increases, further complicating the training process and resource demands.

3.2 Representation-Focused Models

Representation-focused models address the missing modality problem at the representation level. We introduce such models by first presenting two coordinated-representation-based approaches that enhance the learning of more discriminative and robust representations by imposing specific constraints (figure 6). The next type of representation-focused methods that we discuss are the representation imputation methods, which can be categorized into representation composition and representation generation methods. Representation composition methods can borrow the solutions described in section 3.1.1 and operate at the representation level of the modalities or employs arithmetic operations (e.g., pooling) to fuse a dynamic number of modalities. Finally, we introduce the representation generation methods, which usually use small generative models to produce the representations of missing modalities.

3.2.1 Coordinated-Representation Methods

Coordinated-representation methods focus on introducing certain constraints between representations of different modalities to make the learned representations semantically consistent. They are usually divided into two categories, one based on regularization and the other driven by correlation. An example of regularization is Tensor Rank Regularization (TRR) (Liang et al., 2019). TRR achieves multimodal fusion by taking the outer products of different-modality tensors and taking the sum of all their outer products. In order to solve the high rank of multimodal representation tensors caused by imperfect modalities (missing or noisy) in time series, Paul et al. introduced tensor rank minimization to try to keep the multimodal representation low

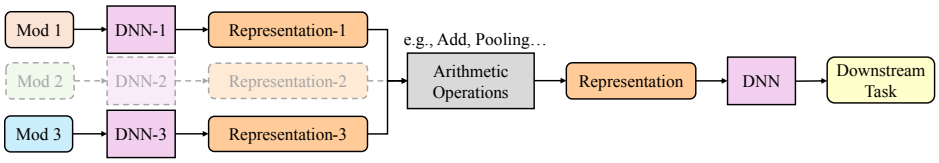


Figure 7: The general idea of the arithmetic operation-based representation composition methods. The representation of any number of modalities can be combined to a fixed dimension that is the same as the dimension required by the subsequent DNN layer.

rank to better express the tensors of true correlation and potential structure in multimodal data, thereby alleviating the imperfection of the input.

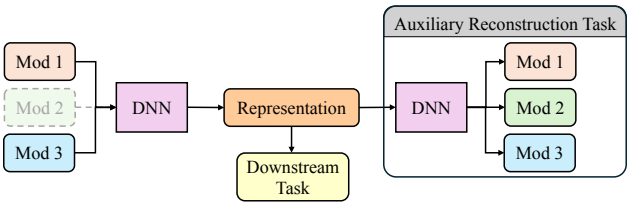
Some methods train models by coordinating the correlation between different modal features. For example, Wang et al. (Wang et al., 2020b) use a Deep Canonical Correlation Analysis (CCA) module to maximize feature associations of available modalities by using Canonical Correlation Coefficient, which enables training on incomplete datasets. Ma et al. (Ma et al., 2021b) proposed a Maximum likelihood function to characterize conditional distributions of full-modality and missing-modality samples during training. Other research efforts (Liu et al., 2021b) have added constraints based on the Hilbert-Schmidt Independence Criterion (HSIC), which helps models learn how to complete missing modality features by enforcing independence between irrelevant features while maximizing the dependence between relevant ones. These methods (Matsuura et al., 2018; Ma et al., 2021a; Li et al., 2022a) aid models in learning how to complete missing modality features or training on incomplete datasets by learning the similarity or correlation between features. A drawback of the above methods is that they perform well only when two or three modalities are used as input (Zhao et al., 2024), and their effectiveness tends to degrade as the number of modalities increases. Even when trained on a dataset with missing modalities, some methods still struggle to effectively address unseen missing modality combinations or cases during testing.

3.2.2 Representation Composition Methods

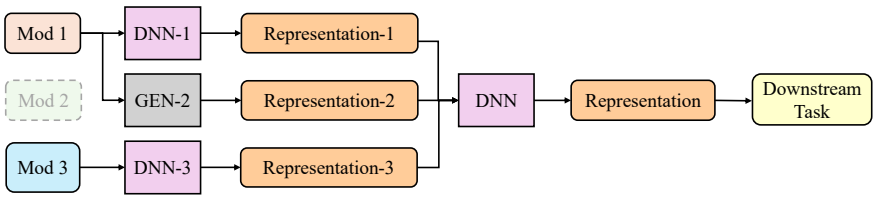
There are two types of representation composition methods, which are explained below. Retrieval-Based Representation Composition Methods attempt to recover the missing modality representation by retrieving the modality data from existing samples, similarly to the modality composition method in section 3.1.1. They typically use pre-trained feature extractors to generate features from available samples, storing them in a feature pool (Zhi et al., 2024; Sun et al., 2024b). Cosine similarity is then used to retrieve matched features for the input sample, with the average of these missing-modality features in top-K samples filling the representation of missing modalities. Additionally, some methods, such as Missing Modality Feature Generation (Wang et al., 2023b), replace missing modality features by averaging the representations of available modalities, assuming similar feature distributions across modalities.

Arithmetic Operation-Based Representation Composition Methods can flexibly fuse any number of modality representations through arithmetic operations such as pooling without learnable parameters (figure 7). Researchers (Zhou et al., 2021; Wang et al., 2023b) have fused features through operations like average/max pooling or addition, offering low computational complexity and efficiency. However, drawbacks include potential loss of important information. To address this, merge operations (Delgado-Escano et al., 2021) use a sign-max function, which selects the highest absolute value from feature vectors, yielding better results by preserving both positive and negative activation values. TMFormer (Zhang et al., 2024) introduces token merging based on cosine similarity. Similar approaches (Zheng et al., 2024) focus on selecting key vectors or merging tokens to handle missing modalities. A recent work (Sarkar et al., 2025) propose to perform weighted fusion of these pooled features to learn a fixed-size multimodal embedding

Therefore, those representation composition approaches not only allow the model to flexibly handle features from any number of modalities but also enable learning without introducing new learnable parameters. However, they are not good at capturing the inter-modality relationships through learning, as methods like



(a) Indirect-to-Task Representation Generation Methods



(b) Direct-to-Task Representation Generation Methods

Figure 8: Description of two typical representation generation methods. We assume modality 2 is missing here. (a) Indirect-to-task representation generation methods are supervised by the loss of two tasks, namely: auxiliary reconstruction and downstream tasks. Note that these two tasks sometimes can be trained separately, first training the reconstruction task, discarding the generator, and then training the downstream task, such as the training progress of ActionMAE Woo et al. (2023). While (b) direct-to-task representation generation methods directly use a modality generation network ("GEN-2" in figure) to generate representations of the missing modality 2 during training and inference without reconstructing the modality 2.

selecting the largest vector or the feature with the highest score is difficult to fully represent the characteristics of all modalities.

3.2.3 Representation Generation Methods

Compared to modality generation methods in section 3.1.2, representation generation methods can be integrated seamlessly into existing multimodal frameworks. Current methods fall into two categories: (1) Indirect-to-task representation generation methods (figure 8a) treat modality reconstruction as an auxiliary task during training, helping the model intrinsically generate missing modality representations for downstream tasks. Since the auxiliary task aids in representation generation during training but is dropped during inference of downstream task, it is termed "indirect-to-task." (2) Direct-to-task representation generation methods (figure 8b) train a small generative model to directly map available modality representations to the missing modalities. We provide more details about these two categories of methods below.

Indirect-to-Task Representation Generation Methods can indirectly generate missing modality representations and often employ the "encoder-decoder" architecture. During training, modality-specific encoders extract features from available modalities, reconstruction decoders reconstruct the missing modality, and the downstream module is supervised by downstream task loss for prediction. Those reconstruction decoders are discarded during inference, where predictions rely on both the generated missing modality representations and the existing ones. This approach typically employs Multi-Task Learning, training both downstream tasks and auxiliary tasks (modality reconstruction) simultaneously. Some methods, inspired by Masked Autoencoder (He et al., 2022), split reconstruction and downstream task training. Typically, in the training process of this type of methods, the features of the available modalities are usually input into the downstream task module and the reconstruction decoder of different modalities after being fused. However, some methods accept the outputs of different modality encoders into their respective reconstruction decoders. Based on this distinction, we divide this type of methods into before-fusion and after-fusion methods.

Before Fusion methods receive the features from modality-specific encoders directly into each specific reconstruction decoder for missing modality reconstruction. For example, MGP-VAE (Hamghalam et al., 2021) uses output of the VAE encoder of each modality as the input of each reconstruction decoder and

the fused output for Gaussian Process prediction. Similarly, PFNet (Zheng et al., 2021) employs RGB and thermal-infrared modalities for pedestrian re-identification, feeding outputs of each encoder into corresponding decoders before fusing them together. Li et al. (Li et al., 2024b) adopted a similar approach for brain tumor segmentation, reconstructing each modality before fusing encoder outputs.

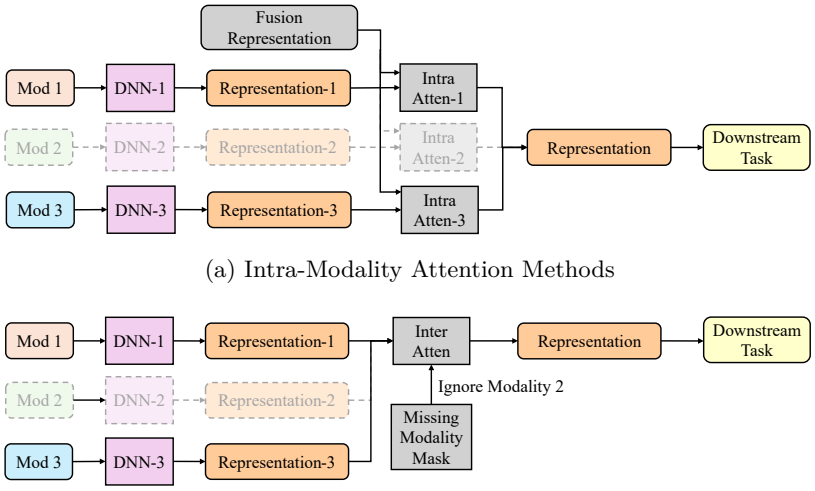
After Fusion methods use the fused features from modality-specific encoders as the input of reconstruction decoder. Tsai et al. (Tsai et al., 2018) introduced a Multimodal Factorization Model where modality-specific encoder-decoders reconstruct the missing modality, while the downstream task module employs a separate multimodal encoder-decoder. To recover the missing modality features, the outputs of the downstream task model's encoder are not only used for predictions of the downstream task decoder, but also fused with the outputs of each modality-specific encoder for inputting to the each corresponding reconstruction decoder. Chen et al. (Chen et al., 2019) and Li et al. (Li et al., 2025a) also decoupled the encoders for reconstruction and downstream tasks, showing that disentangling features helped eliminate irrelevant features. Jeong et al. (Jeong et al., 2022) proposed a simpler approach using multi-level skip connections in a U-Net architecture to fuse features for missing modality reconstruction, achieving better generalization in brain tumor segmentation than Chen et al. (Chen et al., 2019). Further, Zhou et al. (Zhou, 2023) and Sun et al. (Sun et al., 2024a) demonstrated effective cross-modal attention fusion techniques could help enhance reconstruction and segmentation. Most of these methods train reconstruction and downstream tasks concurrently. However, some approaches, such as ActionMAE (Woo et al., 2023) and M3AE (Liu et al., 2023a), first train reconstruction tasks and then fine-tune the pre-trained encoders for downstream tasks.

Although these methods achieve strong performance, they still require the missing modality data to be present during training to supervise the reconstruction decoders.

Direct-to-Task Representation Generation Methods usually use simple generative models to directly map information from avaiable modalities to the missing modality representation. The general concept is illustrated in figure 8b. Early work (Mario Christoudias et al., 2010) using Gaussian Processes to "hallucinate" missing modalities inspired Hoffman et al. (Hoffman et al., 2016) to propose a Hallucination Network (HN) that predicts missing depth modality features from RGB images. The HN aligns intermediate layer features with another network (trained on depth images) using an L2 loss, allowing RGB inputs to generate depth features to address missing modality challenges. This method has been extended to pose estimation (Choi et al., 2017) and land cover classification (Kampffmeyer et al., 2018), where the HN can be used to generate missing heat distribution and depth features.

In tasks like sentiment analysis, a translation module can map available-modality representations to the missing ones, with L2 loss applied to align outputs with those from networks trained on missing modalities (Huan et al., 2023). A notable recent work, SMIL (Ma et al., 2021c), employs Bayesian learning and meta-regularization to directly generate missing modality representations. SMIL has shown strong generalization on datasets with a high rate of missing modality samples. Subsequent work (Li et al., 2022d) addressed challenges such as distribution inconsistency, failure to capture specific modality information, and lack of correspondence between generated and existing modalities. Researchers have also explored cross-modal distribution transformation methods to align distributions of generated and existing modality representations, enhancing discriminative ability (Wang et al., 2023d). In audio-visual question answering (AVQA), Park et al. (Park et al., 2024) proposed a Relation-aware Missing Modal Generator (RMM) that consist of to generate pseudo features of missing modalities, improving robustness and accuracy. The RMM consists of a visual generator, auditory generator, and text generator, which generate missing auditory representations by analyzing the correlations between visual and textual modalities and utilizing learnable parameter vectors (slots) for reconstruction and synthesis of missing-modality features.

Recently, to address the potential absence of different modalities in sentiment analysis, Guo et al. (Guo et al., 2024) proposed a Missing Modality Generation Module for prompt learning. This module maps prompts of available modalities to prompts of missing modalities through a set of projection layers. Lin et al. (Lin et al., 2024) proposed Uncertainty Estimation Module which can identify useless tokens in different modality tokens to facilitate better use of U-Adapter-assisted pre-trained models to better utilize the tokens of available modalities when downstream tasks face missing modalities. Liao et al. (Liao et al., 2025) explore a variety of different conditions of sensor impairment are and the capabilities of various existing multimodal



(a) Intra-Modality Attention Methods

(b) Inter-Modality Attention Methods

Figure 9: Two typical attention fusion methods, assuming modality 2 as the missing modality. (a) The intra-modality attention ("Intra-Atten" in figure) of missing modality 2 can be skipped. Since each Intra-Atten of different modalities shares the fusion representation (like learnable query in Ge et al. (2023) or bottleneck tokens in Nagrani et al. (2021)), after calculating them one by one, the output representation can be regarded as the fusion of available modalities. (b) "Missing Modality Mask" is the custom mask that is generated according to the missing modality 2 and can help inter-modality attention ("Inter-Atten" in figure) ignore the missing modality 2. By setting the tokens of modality 2 to zero or negative infinity through masking, we can force the attention mechanism to ignore the missing modality 2.

segmentation models are evaluated. Park et al. (Park et al., 2025) use the Multi-modal Mixture of Expert, which not only enables the model to dynamically handle multi-modality, but also allows different branches to focus on their own modalities by sharing a query.

Compared to indirect methods, direct generation methods avoid the cumbersome reconstruction of modality data, using feature generators that are easier to train and integrate into multimodal models.

In summary, representation generation methods are limited by the imbalance of missing modalities in the training dataset or the limited number of full-modality samples, which may lead to training procedures that overfit to the existing modalities. In addition, indirect-to-task representation generation methods can access the missing modality data of the missing modality samples during training, which means that the training dataset is actually complete, so this type of method usually performs better than direct-to-task representation generation methods.

4 Methodologies in Strategy Design Aspect

4.1 Architecture-Focused Models

Different from above methods of handling missing modality problems at the modality or representation level, many researchers adjust the model training or testing architecture to adapt to the missing-modality cases. We divide them into four categories according to their core contributions in dealing with missing modalities: attention-based methods, distillation-based methods, graph learning-based methods, and multimodal large language models.

4.1.1 Attention Based Methods

In the self-attention mechanism (Vaswani et al., 2017), each input is linearly transformed to generate `Query`, `Key`, and `Value` vectors. Attention weights are computed by multiplying the query of each element with the

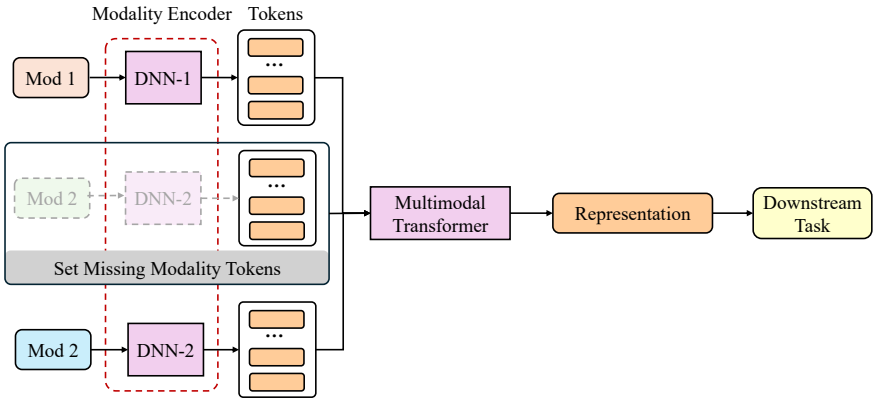


Figure 10: Description of general idea of joint representation learning methods, assuming Modality 2 is the missing modality. The missing modality 2 tokens can be replaced by specific masked tokens. The modality encoders (red dashed box) can be replaced by linear projection layers in some methods.

keys of others, followed by scaling and softmax to ensure the weights sum to 1. Finally, a weighted sum of the values generates the output. We classify attention-based MLMM methods into two categories. (1) Attention fusion methods put attention on modality fusion, integrating multimodal information, which do not rely on any specific model type and can be suitable for various model types as its input and output dimensions are the same. (2) Others are transformer-based methods, which stack attention layers to handle large-scale data with global information capture and parallelization. We provide more details of these two methods below.

Attention Fusion Methods have the powerful ability to capture key features and can be seen as plug-and-play modules. We categorize them into two types: intra- and inter-modality attention methods. Intra-Modality Attention Methods compute attention for each modality independently before fusing them, as shown in figure 9a. This approach focuses on relationships within a single modality, and the fusion between modalities is achieved by sharing partial information. For instance, in 3D detection, the BEV-Evolving Decoder (Ge et al., 2023) handles sensor failures by sharing same BEV-query with each modality-specific attention modules, allowing fusion of any number of modalities. Similarly, in clinical diagnosis, Lee et al. (Lee et al., 2023a) proposed modality-aware attention to perform intra-modality attention and predict decisions by using shared bottleneck fusion tokens (Nagrani et al., 2021). Inter-Modality Attention Methods, often based on masked attention, treat missing modality features as masked vectors (using zero or negative infinity values) to better capture dependencies across available modalities, as illustrated in figure 9b. Unlike conventional cross-modal attention, masked attention models share the same parameters across all embeddings, allowing flexible handling of missing modalities. For example, Qian et al. (Qian & Wang, 2023) designed an attention mask matrix to ignore missing modalities, improving model robustness. Similarly, DrFuseYao et al. (2024) decouples modalities into specific and shared representations, using the shared ones to replace missing modalities, with a custom mask matrix to help model ignore the specific representations of the missing modality.

Transformer Based Methods can be divided into two types: joint representation learning (JRL) and parameter efficient learning (PEL) according to full parameter training and a small amount of parameter fine-tuning. Since transformers have long context lengths to handle many feature tokens, the multimodal transformer can learn joint representations from any number of modality tokens (figure 10). Some methods first use modality encoders to extract feature tokens from available modalities and then feed them into a multimodal transformer. Gong et al. (Gong et al., 2023) introduced an egocentric multimodal task, proposing a transformer-based fusion module with a flexible number of modality tokens and a cross-modal contrast alignment loss to map features into a common space. Similarly, Mordacq et al. (Mordacq et al., 2024) leveraged Masked Multimodal Transformers, treating missing modalities as masked tokens for robust JRL. Ma et al. (Ma et al., 2022) proposed an optimal fusion strategy search strategy within the multimodal transformer to help find the best fusion method for different missing/full-modality representations. Radosavovic

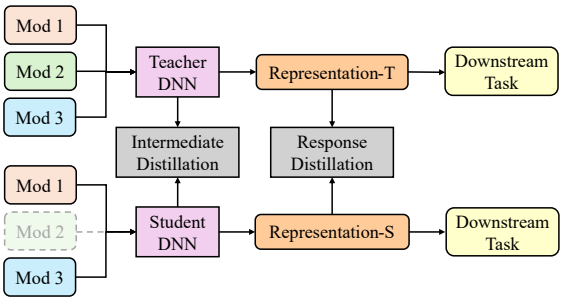


Figure 11: Description of representation distillation methods. Here we set modality 2 is missing. In order to distinguish the representations of teachers and students, we add "-T" and "-S" after their representations in the figure. We refer to intermediate distillation as using features from any combination of intermediate layers within models.

et al. (Radosavovic et al., 2024) introduced specially designed mask tokens for an autoregressive transformer to manage missing modalities, successfully deploying the model in real-world scenarios.

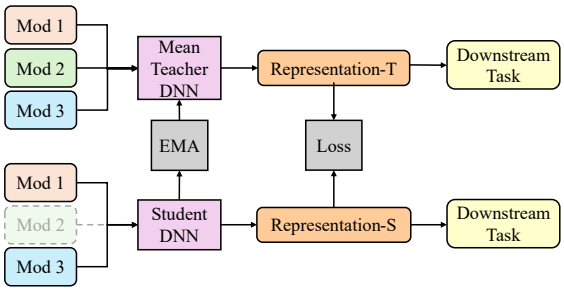
With the rise of pre-trained transformer models, PEL methods have been developed to fine-tune these models by training few parameters. Two common PEL methods for pre-trained models are prompt and adapter tuning. Initially used in natural language processing, prompt tuning optimizes input prompts while keeping model parameters fixed. It has since been extended to multimodal models. Jang et al. (Jang et al., 2024) introduced modality-specific prompts to address limitations in earlier methods, which merge when all modalities are present and allow the model to update all learnable prompts during training. Liu et al. (Liu et al., 2024b) further improved this by proposing Fourier Prompt, which uses fast Fourier transform to encode global spectral information of available modalities into learnable prompt tokens which can be used to supplement missing-modality features, enabling cross-attention with feature tokens to address missing modalities. Adapter tuning, on the other hand, involves inserting lightweight adapter layers (e.g., MLPs) into pre-trained models to adapt to new tasks without modifying the original parameters. Qiu et al. (Qiu et al., 2023) proposed a method that used a classifier to identify different missing-modality cases and used intermediate features of that classifier as the missing-modality prompt to cooperate with lightweight Adapters to address missing modality problems.

Although attention-based fusion mechanisms in above methods can effectively help deal with the missing modality problem in any framework, none of them cares about the missing modality that may contain important information required for prediction. In the transformer-based methods, the JRL methods are often limited by a large amount of computing resources and require relatively large datasets to achieve good performance. The PEL methods can achieve efficient fine-tuning, but their performance is still not comparable to that of JRL.

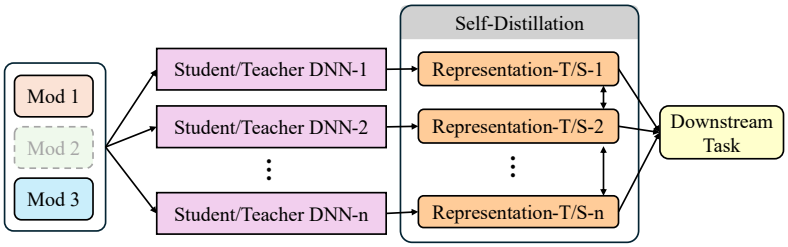
4.1.2 Distillation Based Methods

Knowledge distillation (Hinton et al., 2015) transfers knowledge from a teacher model to a student model. The teacher model, with access to more information, helps the student reconstruct missing modalities. Below, we categorize two types of distillation methods for addressing this problem.

Representation-Based Distillation Methods transfer rich representations from the teacher model to help the student capture and reconstruct missing modality features. We classify them based on whether they use logits or intermediate features. figure 11 illustrates this type of method. Response distillation methods focus on transferring teacher models' logits to students, helping it mimic probability distributions. Wang et al. (Wang et al., 2020a) trained modality-specific teachers for missing modalities, then used their soft labels to guide a multimodal student. Hafner and Ban (Hafner & Ban, 2023) employed logits from a teacher model trained with optical data to supervise a reconstruction network for approximating missing optical features from radar data. Pramit et al. (Saha et al., 2024) proposed a Modality-Aware Distillation, leveraging both



(a) Mean Teacher Distillation Methods



(b) Self Distillation Methods

Figure 12: Description of two process-based distillation methods. (a) "EMA" means exponential moving average. To demonstrate the general architecture of Mean Teacher Distillation, we set the loss on the logits representation. (b) Self-distillation methods for missing modality problems involve the use of models that act as teachers and students, where these models use their own soft labels/representations from each branch to refine themselves during training.

global and local teacher knowledge in a federated learning setting to help student models to learn how to handle missing modalities.

Intermediate distillation methods align intermediate features between teacher and student models. Shen et al. (Shen & Gao, 2019) used Domain Adversarial Similarity Loss to align intermediate layers of the teacher and student, improving segmentation in missing modality settings. Zhang et al. (Zhang et al., 2023b) applied intermediate distillation in endometriosis diagnosis by distilling features from a TVUS-trained teacher to a student using MRI data.

Process-Based Distillation Methods focus on overall distillation strategies, like Mean Teacher Distillation (MTD) (Tarvainen & Valpola, 2017) and self-distillationZhang et al. (2019). These methods emphasize procedural learning over direct representation transfer. MTD enhances stability by using the exponential moving average of the student model's parameters as a teacher (figure 12a). Chen et al. (Chen et al., 2024b) applied this to missing modality sentiment analysis, treating missing samples as augmented data. Li et al. (Li et al., 2022c) used MTD for Lidar-Radar segmentation, improving robustness against missing modalities.

Self-distillation helps a model improve by learning from its own soft representations (figure 12b). Wang et al. proposed the ShaSpec (Wang et al., 2023a), which utilizes distillation between modality-shared branches and modality-specific branches of available modalities. Based on ShaSpec, researchers proposed Meta-learned Cross-modal Knowledge Distillation (Wang et al., 2024a) to further weigh the importance of different available modalities to improve performance. Ramazanova et al. (Ramazanova et al., 2024) used mutual information and self-distillation for egocentric tasks, making predictions invariant to missing modalities. Shi et al. proposed PASSION (Shi et al., 2024), a self-distillation method designed to use the multi-modal branch to help other uni-modal branches of available modalities to improve multimodal medical segmentation performance with missing modality.

Hybrid Distillation Methods combine various distillation approaches to improve student performance. For medical segmentation, Yang et al. (Yang et al., 2022) distilled teacher model knowledge (including logits and

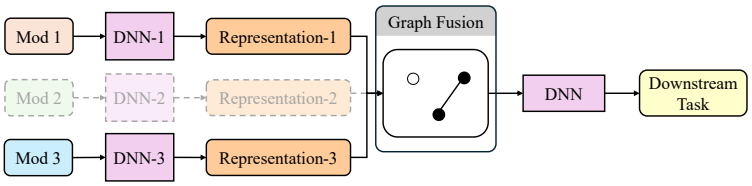


Figure 13: Illustration of graph fusion methods. We set modality 2 (white circle in the graph) as the missing modalities. Features from modality 1 and 3 can be aggregated via this graph fusion method and fused features maintains the same dimension.

intermediate features) at every decoder layer, outperforming ACNetWang et al. (2021). Wang et al. (Wang et al., 2023c) introduced ProtoKD, which captures inter-class feature relationships for improved segmentation under missing modality conditions. Recently, CorrKD (Li et al., 2024a) leverages contrastive distillation and prototype learning to enhance performance in uncertain missing modality cases.

Aforementioned methods address the missing modality problem and achieve good generalization during testing. However, except for some intermediate and self-distillation methods, most assume the complete dataset is available for training (means teacher can access full-modality samples during training), with missing modalities encountered only during testing. Therefore, most distillation methods are unsuitable for handling incomplete training datasets.

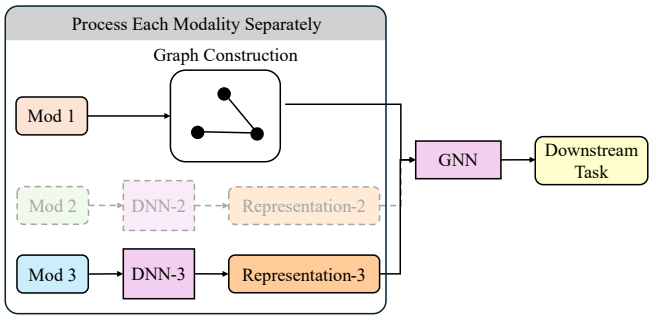
4.1.3 Graph Learning Based Methods

Graph-learning based methods leverage relationships between nodes and edges in graph-structured data for representation learning and prediction. We categorize approaches for addressing the missing modality problem into two main types: graph fusion and graph neural network (GNN) methods.

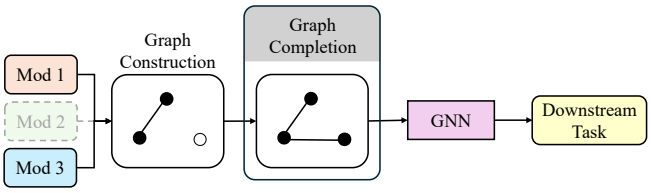
Graph Fusion Methods integrate multimodal data using a graph structure (figure 13), making them adaptable to various networks. For example, Angelou et al. (Angelou et al., 2019) proposed a method mapping each modality to a common space using graph techniques to preserve distances and internal structures. Chen et al. (Chen & Zhang, 2020) introduced HGMF, which builds complex relational networks using hyper edges that dynamically connect available modalities. Zhao et al. (Zhao et al., 2022) developed Modality-Adaptive Feature Interaction for Brain Tumor Segmentation, adjusting feature interactions across modalities based on binary existence codes. More recently, Yang et al. (Yang et al., 2023a) proposed a graph attention based Fusion Block to adaptively fuse multimodal features, using attention-based message passing to share information between modalities. These fusion methods can be flexibly inserted into any network for integrating multiple modalities.

Graph Neural Network Methods directly encode multimodal information into a graph structure, with GNNs used to learn and fuse this information. Early approaches (Wang et al., 2015) employed Laplacian graphs to connect complete and incomplete samples. Individual GNN methods (figure 14a), such as DESAlign (Wang et al., 2024b), extract features using neural networks or GNNs and fuse them for prediction. Unified GNN methods (figure 14b) first complete the graph and then use GNNs for prediction, such as in Zhang et al.'s M3Care (Zhang et al., 2022), which uses adaptive weights to integrate information from similar patients. Lian et al. (Lian et al., 2023) proposed the Graph Completion Network, which reconstructs missing modalities by mapping features back to the input space. In recommendation systems, FeatPropMalitesta et al. (2024) propagates known multimodal features to infer missing ones, while MUSE (Wu et al., 2024) represents patient-modality relationships as bipartite graphs and learns unified patient representations. In knowledge graphs, Chen et al. (Chen et al., 2023) introduced Entity-Level Modality Alignment to dynamically assign lower weights to missing/uncertain modalities, reducing the risk of misguidance in learning.

The methods above can leverage the graph structure to better capture relationships both within/between modalities and samples. However, these methods tend to have lower efficiency, scalability, and higher development complexity compared to other kind of approaches.



(a) Individual GNN Methods



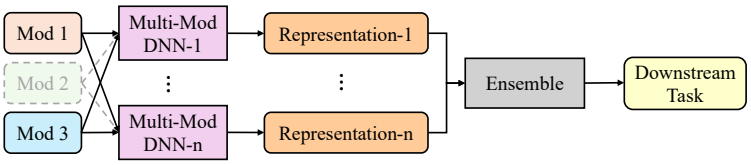
(b) Unified GNN Methods

Figure 14: Illustration of graph neural network (GNN) based methods. We set modality 2 as the missing modalities. (a) Each modality first passes through DNN or builds its graph, and then they are uniformly processed using GNNs. (b) This method first complete the graph from missing-modality samples via approaches like FeatProp Malitesta et al. (2024), then process it by GNNs.

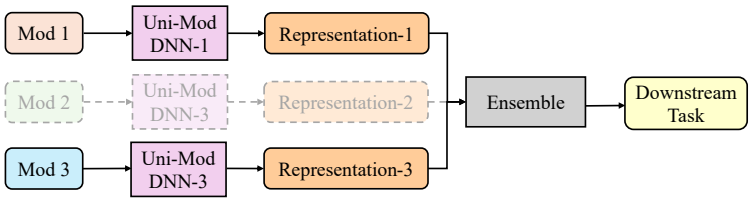
4.1.4 Multimodal Large Language Models (LLM)

The impressive transformative power of LLMs, like ChatGPT (Bubeck et al., 2023), can be explained by their impressive generalization capabilities across many tasks. However, our understanding of the world depends not only on language, but also on other data modalities, like vision and audio. This has led researchers to explore MLLMs, designed to handle diverse user inputs across modalities, including cases with missing modalities, leveraging the flexibility of Transformers. Their architectures are similar to that shown in figure 10. Due to the idiosyncrasies of MLLMs, we distinguish them from normal transformer-based joint representation learning methods in section 4.1.1 and introduce them below. In some current MLLM architectures, LLMs act as feature processors, integrating feature tokens from different modality-specific encoders and passing the output to task-/modality-specific decoders. This enables LLM to not only capture rich inter-modal dependencies, but also naturally carry the ability to handle any number of modalities, that is, the ability to solve the missing modality problem.

Most MLLMs employ transformer-based modality encoders, such as CLIP, ImageBind (Girdhar et al., 2023), and LanguageBind (Zhu et al., 2023), which encode multimodal inputs into a unified representation space. Examples include BLIP-2 (Li et al., 2023b), which bridges the modality gap using a lightweight Querying Transformer, and LLaVA (Liu et al., 2024a), which enhances visual-language understanding with GPT-4-generated instruction data (Bubeck et al., 2023). These models are optimized for tasks like Visual Question Answering, Dialogue, and Captioning. Recent advancements extend output generation to multiple modalities, such as images. AnyGPT (Zhan et al., 2024) and NExT-GPT (Wu et al., 2023b) unify modalities like text, speech, and images using discrete representations and multimodal projection adaptors, enabling seamless multimodal interaction. CoDi (Tang et al., 2024) introduces Composable Multimodal Conditioning, allowing arbitrary modality generation through weighted feature summation. There are also some other methods that discard the modality encoders and use linear transformation directly, such as Fuyu (Bavishi et al., 2023) and OtterHD (Li et al., 2023a). Although MLLMs can flexibly handle any number of modalities, they have many disadvantages, such as the inconsistent multi-modal positional encoding, training difficulty, and high GPU resource requirements. Additionally, no specific MLLM benchmarks on missing modality problems have been proposed.



(a) Multimodal Model Ensemble Methods



(b) Unimodal Model Ensemble Methods

Figure 15: Two general architectures for ensemble methods. (a) The multimodal model ensemble methods contain n three-modal DNNs and produce final predictions based on the outputs of all n DNNs. (b) Each modality is processed by a unimodal DNN in unimodal model ensemble methods, where final predictions are produced by aggregating the outputs of all accessible unimodal DNNs.

4.2 Model Combinations

Model combinations aim to use selected models for downstream tasks. These methods can be categorized into ensemble, dedicated training, and discrete scheduler methods. Ensemble methods combine predictions from multiple selected models through different types of aggregation methods, such as voting, weighted averaging, and similar approaches to improve the accuracy and stability. Dedicated training methods allocate different sub-tasks (e.g., different missing modality cases) to specialized individual models, focusing on specific sub-tasks or sub-datasets. In the discrete scheduler methods, users can use natural language instructions to enable the LLMs to autonomously select the appropriate model based on types of modalities and downstream tasks. We provide more details about these methods below.

4.2.1 Ensemble Methods

As detailed in the next paragraphs, ensemble learning methods allow flexibility in supporting different numbers of expert models to combine their predictions, as shown in figure 15. Multimodal Model Ensemble Methods: Early work (Wagner et al., 2011) in sentiment analysis employed ensemble learning to handle missing modalities, averaging predictions from uni-modal models. With deep learning advancements, the Ensemble-based Missing Modality Reconstruction network (Zeng et al., 2022) was introduced, leveraging weighted judgments from multiple full-modality models when generated missing modality features exhibit semantic inconsistencies. This type of multimodal model ensemble method, depicted in figure 15a, integrates various full-modality models to aid decision-making.

Unimodal Model Ensemble Methods: The general architecture of this method is shown in figure 15b, where each modality is processed by a uni-modal model, and only the available modalities contribute to decision-making. In multimodal medical image diagnosis, early studies found uniformly weighted methods performed better than weighted averaging and voting approaches (Yuan et al., 2012). In multimodal object detection, Chen et al. (Chen et al., 2022) proposed a probabilistic ensemble method. This method does not require training and can flexibly handle missing modalities through probabilistic marginalization, demonstrating high efficiency in experiments. Li et al. (Li et al., 2023c) recently proposed Uni-Modal Ensemble with Missing Modality Adaptation, training models per modality and performing late-fusion training. Other approaches (Miech et al., 2018) calculate feature-based weights to fuse modalities, where weights reflect feature importance for final predictions.

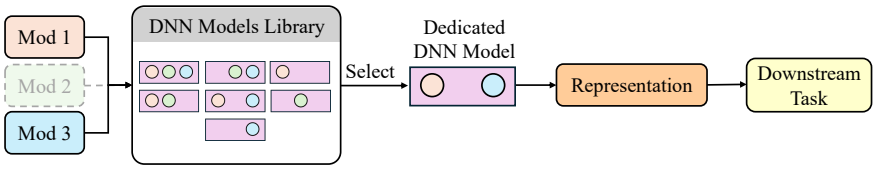


Figure 16: General idea of dedicated training methods. Assuming that modality 2 is missing, dedicated methods select the sample-suited model from DNN models library for training and testing. In the figure, the three modalities can form a total of seven models with different modality combinations. For easy understanding, we use colored circles in purple rectangles to represent the modalities that the model can handle. In order to adapt to the missing modality 2 (light green circles), dedicated methods usually select the DNN model (purple rectangle) that can handle modality 1 (light orange circles) and modality 3 (light blue circles) for training and testing.

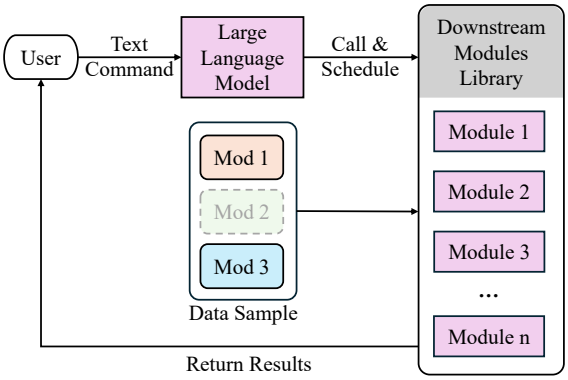


Figure 17: General procedure of discrete scheduler methods. The user types text instructions to make an LLM call and schedule the selected downstream task modules to process data samples and return results to the user. As long as there are sufficiently diverse downstream task modules, it is possible to handle the same downstream tasks with different numbers of modalities.

Hybird Methods: Dynamic Multimodal Fusion (DynMM) (Xue & Marculescu, 2023) uses gating mechanisms to select uni/multi-modal models dynamically. Recently, Cha et al. (Cha et al., 2024) proposed Proximity-based Modality Ensemble (PME) to use a cross-attention mechanism with an attention bias to integrate box predictions from different modalities. PME can adaptively combine box features from uni-/multi-model models and reduce noise in multi-modal decoding.

4.2.2 Dedicated Training Methods

Dedicated training methods assign different tasks to specialized models. We show the general idea of those methods in figure 16. KDNet (Hu et al., 2020) was the first dedicated method proposed to handle different combinations of missing modalities. It treats uni-modality specific models as student models, learning multimodal knowledge from the features and logits of a multimodal teacher model. Those trained uni-modals can be used for different missing modality cases. Following KDNet, ACNet (Wang et al., 2021) also utilizes this distillation method but introduces adversarial co-training, further improving the performance. Lee et al. (Lee et al., 2023b) proposed missing-modality-aware prompts to address missing modality problems based on the prompt learning by using input- and attention- level prompts for each kind of missing modality case. This method only needed 1% of the model parameters to be fine-tuned for downstream tasks. Similarly, some researchers (Reza et al., 2023) introduce different adapter layers for each missing modality case.

4.2.3 Discrete Scheduler Methods

In discrete scheduler methods (figure 17), LLMs act as controllers, determining the execution order of different discrete steps broken down from the major task/instruction. While the LLM does not directly process multimodal data, it interprets language instructions and orchestrates task execution across uni- and multimodal modules. This structured yet flexible approach is particularly effective for outputs requiring sequential tasks, enabling the system to handle any number of modalities and naturally addressing missing modality problems. For example, Visual ChatGPT (Wu et al., 2023a) integrates multiple foundation models with ChatGPT to enable interaction through text or images, allowing users to pose complex visual questions, provide editing instructions, and receive feedback within a multi-step collaboration framework. HuggingGPT (Shen et al., 2024) is an LLM-driven agent that manages and coordinates various AI models from Hugging Face. It leverages LLMs for task planning, model selection, and summarization to address complex multimodal tasks. ViperGPT (Surís et al., 2023) combines visual and language tasks into modular subprograms, generating and executing Python code for complex visual queries without additional training to achieve effective outputs. There are other similar discrete scheduler approaches, such as MM-REACT (Yang et al., 2023b) and LLaVA-Plus (Liu et al., 2023c).

Some aforementioned dedicated and ensemble methods can flexibly handle the missing modality problem without additional training, but most of them require more model storage space, which is not feasible for many resource-constrained devices. For instance, DynMM requires storing various uni- and multi-modality models. As the number of modalities increases, the required number of models also rises. Also, unimodal model ensemble methods struggle to adequately model the complex inter-modality relationships to make final predictions. Although the some dedicated fine-tuning methods do not require too much time and consumption of resources such as GPU, it is still difficult to compare with full parameter training. In addition, discrete scheduler methods can solve a variety of tasks when there are sufficient types of downstream modules, but they usually require LLMs to respond quickly and understand human instructions accurately in real world scenarios.

5 Methodology Discussion

Table 1: Comparison of deep multimodal learning with missing modality methods from four types and two aspects.

Aspect	Type	Method	Pros and Cons
Data Processing	**Modality Imputation**	Modality Composition	Pros: Simple and effective for data augmentation. Cons: Unsuitable for pixel-level tasks (Composition); Modality number ↑, generative model number or training complexity ↑.
		Modality Generation	
	Representation Focused	Coordinated-Representation	Pros: Flexible for novel modalities with better generalization. Cons: Hard to balance constraints and handle dataset imbalance.
		Representation Composition	
		Representation Generation	
Strategy Design	**Architecture Focused**	Attention-based	Pros: Attention methods scale well; Distillation and graph learning are effective. Cons: Most attention methods demand high computational resources; Distillation and graph learning struggle with incomplete datasets and training complexity.
		Distillation-based	
		Graph Learning-based	
		Multimodal Large Language Model	
	Model Combinations	Ensemble	Pros: Effective for specific tasks. Cons: Grows complex with more modalities; Suffers from modality imbalance; Demands significant storage space.
		Dedicated	
		Discrete Scheduler	

In the above section 3 and section 4, we divide the existing MLMM methods into four types from the aspects of data processing and strategy design: modality imputation, representation-focused, architecture-focused, and model combinations. We also further subdivide these four types into twelve categories, exploring a fine-grained methodology taxonomy proposed by us. Table 1 summarizes the overall pros and cons of these methods. Generative and distillation methods are the most common approaches; they are easy to implement and deliver strong performance. With the rise of Transformers, Transformers methods have become more

popular due to their larger receptive fields and parallelism. However, indirect-to-task generation methods and most distillation methods (except for some intermediate (Zhang et al., 2023b) and self-distillation methods (Wang et al., 2023a; Shi et al., 2024)) are currently unable to handle incomplete training datasets (means cannot access missing modality data during training). Below we provide a concise analysis of these twelve methods from two aspects and four types.

5.1 Data Processing Aspect

(1) Modality Imputation: Modality composition methods operate directly on the input modality data level to address missing modality problems by combining existing data samples or filling missing data. However, they are typically not good at pixel-level downstream tasks and heavily rely on available modalities. On the other hand, modality generation methods employ generative models to synthesize missing modalities, mitigating the performance degradation caused by missing data. But these methods are often limited by the availability and number of full-modality samples and the increased training complexity due to the additional modalities. They also require extra storage for generative models.

(2) Representation Focused: Coordinated-representation methods allow novel modalities to be introduced by simply adding corresponding branches. However, as the number of constraints and modalities increases, balancing them becomes challenging. Correlation-driven methods often fail to handle missing modality samples in the test environment, as they are typically designed for pre-training with incomplete datasets. Similar to modality composition methods, representation composition methods attempt to recover missing modality representations by combining available-modality representations. Since representations typically carry more generalized information than modality data, these methods tend to yield better results. Representation generation methods further improve this process by generating representations based on relationships between modalities, allowing reconstruction of missing modality representations. Many studies have confirmed the effectiveness of this approach in handling missing modalities. However, if the dataset contains missing modality samples, indirect-to-task representation generation methods that aim to recover missing modality representations through reconstruction of modality data are not feasible. When the dataset has severely imbalanced modality combinations, generative methods may also fail, becoming overly reliant on existing modalities.

5.2 Strategy Design Aspect

(1) Architecture Focused: Some methods focus on designing model training or inference architecture to alleviate the performance degradation caused by missing modalities. Attention-based methods are valued by researchers because they can effectively capture the relationships between various modalities, are scalable in terms of dataset size, and are highly parallelizable. Single attention mechanisms used for modality fusion have the advantage of being plug-and-play. The drawbacks of transformer-based methods in this category are also relatively obvious: training multimodal transformers from scratch requires excessive training time and many GPU resources. Recently popular PEL methods can effectively mitigate this drawback, but the generalization performance of this method is still not as well as full-parameter training or tuning. Although MLLMs can handle an arbitrary number of modalities with flexibility, they are constrained by training complexity and require substantial computational resources.

Distillation-based methods are relatively easy to implement, with student models learning how to reconstruct missing modality representations and inter-modal relationships from teacher models. Since the teacher model typically receives full-modality samples as input, it can simulate any missing modality cases during training, ensuring strong performance. However, most distillation-based methods are limited to the requirement of complete datasets and are inapplicable to datasets with inherent missing modalities. To our knowledge, only one intermediate distillation method (Zhang et al., 2023b) has attempted to input mismatched samples of the same class into the teacher and some self-distillation methods (Wang et al., 2023a; Shi et al., 2024) use representations of available modality branches for distillation purposes.

Graph learning-based methods capture intra- and inter-modal relationships more effectively, but as the number of modalities increases, the development complexity and inefficiency of these methods become more pronounced. They are also currently unsuitable for large-scale datasets.

(2) Model Combinations: These methods select models for prediction. Ensemble and dedicated training methods can be affected by imbalances in modality combinations within the dataset. As the number of modalities grows, the complexity and models to be trained also increase, especially for dedicated methods. Discrete scheduler methods requires an LLM to coordinate, but if callable modules are insufficient, downstream tasks may not be completed. Additionally, inference speed is limited by LLMs and modules. Model combinations methods also require significant model storage.

5.3 Recovery and Non-Recovery Methods & Some Statistics

Table 2: Another common taxonomy of recovery and non-recovery methods in multimodal learning is based on three stages: early, intermediate, and late stage, as seen in the conventional taxonomy. In this context, recovery and non-recovery methods refer to how missing modalities are handled at each stage.

Methods	Early	Intermediate	Late	Hybrid
Recovery	Modality Composition Yang et al. (2024) Modality Generation Wang et al. (2024c)	Representation Composition Wang et al. (2023b) Representation Generation Woo et al. (2023) Distillation-based Shen & Gao (2019) Graph Learning-based Malitesta et al. (2024)	Representation Generation Ma et al. (2021c) Distillation-based Li et al. (2022c) Coordinated-Representation Ma et al. (2021b)	Distillation-based Sikdar et al. (2023)
Non-Recovery	Dedicated Wang et al. (2021) Discrete Scheduler Wu et al. (2023a)	Representation Composition Delgado-Escano et al. (2021) Attention-based Ge et al. (2023) Graph Learning-based Zhao et al. (2022) Multimodal Large Language Model Wu et al. (2023b)	Coordinated-Representation Liang et al. (2019) Distillation-based Shi et al. (2024) Ensemble Chen et al. (2022)	Not Applicable

We have further classified MLMMs using the taxonomy in Table 1 to facilitate researchers to better distinguish MLMMs. This classification is based on the three-stage (early, intermediate, and late) classic multimodal learning taxonomy. Table 2 leverages this classic taxonomy of MLMMs and specifies whether they recover missing modalities or not. We also list reference papers for different methods. In our analysis of 315 papers, we found that 75.5% of the works focus on recovering missing modality information, while only 24.5% explore inference without modality recovery. Among the recovery methods, early-stage and intermediate-stage modality recovery methods account for 20.3% and 45.8%, respectively, with late-stage and multi-stage recovery methods accounting for 4.7%, each. For non-recovery methods, early, intermediate, and late-stage modality fusion methods represent 4.2%, 14.1%, and 6.3%, respectively. We were not able to find any non-recovery methods that combine multiple stages. We can observe that methods on recovering missing modality features in the intermediate stage account for the largest proportion. We think this is because recovering features, as opposed to the raw data, can avoid much noise and bias while providing more modality-specific/shared information. Additionally, compared to later-stage features, intermediate-stage features tend to be more enriched.

Limitations: Due to the variety of settings in most missing-modality works, it is difficult to compare performances of different methods across datasets. This represents a research gap that we encourage the community to address.

6 Applications and Datasets

The collection of multimodal datasets is often labor-intensive and costly. In certain specific application directions, issues such as user privacy concerns, sensor malfunctions on data collection devices, and other factors can result in datasets with missing modalities. In severe cases, up to 90% of the samples may have missing modalities, making it challenging for conventional MLFM to achieve good performance. This has given rise to the task of MLMM. Since the factors causing incomplete datasets usually stem from different application directions, we introduce the following datasets based on the applications which is common in MLMM tasks: Sentiment Analysis, Medical Diagnosis, Retrieval/Captioning, Remote Sensing, Robotic Vision, and others. We categorize those datasets according to the applications and data type in Table 3.

6.1 Sentiment Analysis

Typically, the goal of sentiment analysis is to classify the current emotional state by combining information from multiple modalities such as textual, auditory, and RGB information. This field has garnered significant attention due to its promising applications in various specific scenarios, including market research, health

Table 3: Common datasets used by deep MLMM methods divided by main applications and modality types. "Vision" includes data from visual sensors such as RGB images, Depth, Infrared, LiDAR, Radar, Event, and Optical Flow. "Bio-Sensors" includes data from sensors like: electrophysiological and respiratory sensors. Other modalities include audio, text, CT scans, MRI, and skeleton.

Applications	Modality Types	Common and Typical Datasets
Sentiment Analysis	Vision+Audio	eNTERFACE'05 Martin et al. (2006), RAVDESS Livingstone & Russo (2018), CREMA-D Cao et al. (2014)
	Vision+Text	eBDtheque Guérin et al. (2013), DCM Nguyen et al. (2018), Manga 109 Matsui et al. (2017)
	Vision+Bio-Sensors	Ulm-TSST Stappen et al. (2021a;b), RECOLA Ringeval et al. (2013), SEED Zheng & Lu (2015)
	Vision+Audio+Text	CMU-MOSI Zadeh et al. (2016), CMU-MOSEI Zadeh et al. (2018), IEMOCAP Busso et al. (2008),MSP-IMPROV Busso et al. (2016), MELD Poria et al. (2018), CH-SIMI Yu et al. (2020), ICT-MMMO Wöllmer et al. (2013), YouTube Morency et al. (2011)
Medical Diagnosis	MRI, CT Scans	BRATS-series Menze et al. (2014); Myronenko (2019), ADNI Jack Jr et al. (2008), IXI [https://brain-development.org/ixi-dataset]
	Bio-Sensors	StressID Chaptoukaev et al. (2024), PhysioNet Kemp et al. (2000)
	Electronic Health Record	MIMIC-CXR Johnson et al. (2019), MIMIC-IV Johnson et al. (2023), NCH Lee et al. (2022), BCNB Xu et al. (2021), ODIR Li et al. (2021)
Information Retrieval	Vision+Text	Amazon Series Datasets Lakkaraju et al. (2013); McAuley et al. (2015); He & McAuley (2016); PromptCloud (2017), MIR-Flickr25K Duin, NUS-WIDE Chua et al. (July 8-10, 2009)
	Vision+Audio+Text	MSR-VTT Xu et al. (2016), TikTok Lin et al. (2023)
Remote Sensing	Vision	WHU-OPT-SAR Li et al. (2022b), DFC2020 Yokoya et al. (2020), MSAW Shermeyer et al. (2020), Trento Rasti et al. (2017), Houston Piacentini (2023a), Berlin Hong et al. (2021), GRSS Debes et al. (2014)
Robotic Vision	Vision	RegDB Ye et al. (2021), SYSU-MM01 Wu et al. (2017), UWA3DII Rahmani et al. (2016), Delivery Zhang et al. (2023a), NuScenes Caesar et al. (2020), MVSS Ji et al. (2023)
	Vision+Audio	Ego4D-AR Grauman et al. (2022), Epic-Sounds Huh et al. (2023), Epic-Kitchens Damen et al. (2022), Speaking Faces John & Kawanishi (2022)
	Vision+Skeleton/Inertial Data	MMG-Ego4D Gong et al. (2023), Northwestern-UCLA Wang et al. (2014), NTU60 Shahroudy et al. (2016)

monitoring, and advertising. In the early stages of audio-visual sentiment analysis research, researchers discovered that facial detection algorithm in data collection devices sometimes failed to capture faces (e.g., due to occlusion) or the recorded audio is too noisy to use (Cohen et al., 2004; Sebe et al., 2005), resulting in some samples that only contained audio or image. Consequently, researchers began exploring methods to enable sentiment analysis models to continue functioning effectively when facing missing-modality samples. Currently, the datasets for applying sentiment analysis can be roughly divided into two categories. One type of dataset (Martin et al., 2006; Livingstone & Russo, 2018; Cao et al., 2014; Stappen et al., 2021a;b; Ringeval et al., 2013; Zheng & Lu, 2015; Zadeh et al., 2018; Busso et al., 2008; 2016; Poria et al., 2018; Yu et al., 2020; Wöllmer et al., 2013; Morency et al., 2011) involves using video, audio, text, and biosensors to determine human emotional states in real life or in movies. The other type (Guérin et al., 2013; Nguyen et al., 2018; Matsui et al., 2017) involves using text and images to assess emotional states in contexts such as comic books. Please refer to Table 3 for details.

6.2 Medical Diagnosis

Medical diagnosis requires comprehensive judgments from various modalities such as medical history, physical examination and imaging data, which is exactly what multimodal learning is good at, many researchers developed multimodal intelligent medical diagnosis systems. The current datasets for multimodal medical diagnosis that contain data with missing modality problems are focused on on eight areas: Neuroimaging and Brain Disorders (Menze et al., 2014; Myronenko, 2019; Young et al., 2017; Jack Jr et al., 2008; Lesjak et al., 2018), Cardiovascular (Liu et al., 2016; Moody & Mark, 2001), Cancer Detection (Arya & Saha, 2020; Tomczak et al., 2015), Women Health Analysis (Zhang et al., 2023b; Xu et al., 2021), Ophthalmology (Li et al., 2021; Zhang et al., 2022), Sleep Disorders (Lee et al., 2022; Marcus et al., 2013), Clinical Predictions (Johnson et al., 2016; 2023; 2019), and Biomedical Analysis (Chaptoukaev et al., 2024; Schmidt et al., 2018; Kemp et al., 2000). We have categorized some commonly used and representative datasets from list above by modality type, as shown in Table 3.

6.3 Information Retrieval

Information retrieval is a technology that automatically retrieves relevant content or data based on queries, historical behavior, preferences, and attribute data. It has received widespread attention due to its success on various platforms such as search engines and social media. It improves user satisfaction and information access experience through algorithmic analysis and prediction of content that users may find relevant. Multimodal learning makes modern retrieval systems possible because it can effectively process and analyze multiple types of information, including text, images, audio, etc. Due to user privacy concerns and data sparsity issues, some researchers have begun exploring MLMM approaches in retrieval systems. Many of these studies leverage datasets (Lakkaraju et al., 2013; McAuley et al., 2015; He & McAuley, 2016; PromptCloud, 2017; Duin; Chua et al., July 8-10, 2009; Xu et al., 2016; Lin et al., 2023) from websites such as Amazon and TikTok for their research. We list those datasets based on modality type in Table 3.

6.4 Remote Sensing

By leveraging multimodal learning, we can integrate various types of visual data, such as Synthetic Aperture Radar data and multi-/hyper-spectral data. As a result, various datasets have emerged (Yokoya et al., 2020; Shermeyer et al., 2020; Rasti et al., 2017; Piacentini, 2023a; Hong et al., 2021; Debes et al., 2014). Those datasets enable the assessment and analysis of different environmental conditions, resources, and disasters on earth from satellites or aircraft. Such capabilities can significantly aid in environmental protection, resource management, and disaster response. In practice, remote sensing tasks, such as multimodal land cover classification tasks, optical modalities may be unavailable due to cloud cover or sensor damage. Therefore, it is also necessary to address the missing modality problem for remote sensing problems.

6.5 Robotic Vision

Robotic vision is the field that enables robots to perceive and understand their environment through visual sensors by acquiring and processing image data. This includes tasks such as object and facial recognition, environment modeling, navigation, and behavior understanding, allowing robots to autonomously perform complex operations and interact with humans. Typically, the modality combination formula in robotic vision tasks can be expressed as RGB+X, where X can include, but are not limited to, LiDAR, radar, infrared sensors, depth sensors and auditory sensors. Both RGB and other sensor data can be missing for various reasons. Below, we introduce five common tasks in robotic vision concerning the missing modality problems. (1) Multimodal Segmentation: This task aims to input multimodal data and use segmentation masks to locate objects of interest. In the context of missing modalities, this task is common in datasets for autonomous vehicles, where sensors might fail due to damage or adverse weather conditions. Typical data combinations include RGB+(Depth/Optical Flow/LiDAR/Radar/Infrared/Event data) (Takumi et al., 2017; Liu et al., 2023b; Caesar et al., 2020; Zhang et al., 2023a). Other datasets focus on indoor scene segmentation (Silberman et al., 2012; Ha et al., 2017) (RGB, Depth, Thermal) and material segmentation (Liang et al., 2022) (RGB, Angle of Linear Polarization, Degree of Linear Polarization, Near-Infrared images). (2) Multimodal Detection: Similar to multimodal segmentation, multimodal detection aims to locate objects of interest using bounding boxes. Commonly used multimodal detection combinations in the context of missing modalities include: RGB with Depth (Silberman et al., 2012; Lai et al., 2011) and RGB with Thermal (Hwang et al., 2015; kaggle, 2020). (3) Multimodal Activity Recognition: This task can input data from the aforementioned visual sensors and also use audio cues to recognize human activities. Common datasets (Grauman et al., 2022; Huh et al., 2023; Damen et al., 2022; Gong et al., 2023; Wang et al., 2014; Shahroudy et al., 2016) for this task are combinations of RGB with Audio, Depth, Thermal, Inertial and skeleton data. (4) Multimodal Person Re-Identification: This task aims to use depth and thermal (Ye et al., 2021; Wu et al., 2017) information to robustly identify individuals under varying lighting conditions, such as at night. (5) Multimodal Face Anti-Spoofing: This task utilizes multiple sensors (such as RGB, infrared, and depth cameras) to detect and prevent spoofing in facial recognition systems (John & Kawanishi, 2022; Zhang et al., 2020; Liu et al., 2021a). Its purpose is to enhance the system's ability to identify fake faces (such as photos, videos, masks, etc.) by comprehensively analyzing different modalities, thereby increasing the security and reliability of facial recognition systems.

Other Applications: There are many other areas where MLMM methods are being explored. For example, conventional audio-visual classification (Vielzeuf et al., 2018); audio-visual question answering (Park et al., 2024); multimodal large language models in visual-dialogue (Bubeck et al., 2023), captioning (Liu et al., 2023c); hand pose estimation using depth images and heat distributions (Tompson et al., 2014; Choi et al., 2017); knowledge graph completion utilizing multimodal data with missing modalities (Liu et al., 2019); multimodal time series prediction, such as stock prediction (Luo et al., 2024) and air quality forecasting (Zheng et al., 2015); multimodal gesture generation on how to create natural animations (Liu et al., 2022); multimodal analysis of single-cell data (Mimitou et al., 2021) in biology.

Discussion: Current MLMM research primarily focuses on the areas of sentiment analysis, medical diagnosis, remote sensing analysis, retrieval/captioning, and robotic vision. Among these, a significant portion of research is dedicated to addressing the missing modality problem in video, text, and audio-based sentiment analysis, MRI segmentation and clinical prediction in medical diagnosis, and multi-sensor semantic segmen-

tation for autonomous vehicles. In contrast, research efforts in MLMM for streaming data and scientific fields are relatively limited. In addition, the majority of the publicly available datasets mentioned above are complete in terms of modalities, and naturally occurring datasets with missing modalities are rare. As a result, these tasks are often evaluated by considering all possible combinations of missing modalities based on the existing types of modalities in the dataset, and then performing training and testing on the performance of the missing modalities, supplemented by different missing modality rates. In the works we reviewed, 38% controlled for varying degrees of missing modality rates and tested accordingly, while the remaining 62% used random modality missing rates for training and testing. When categorizing missing modality rates into mild (<30%), moderate (30%-70%), and severe (>70%) levels, the proportions are 72.6%, 82.2%, and 69.9%, respectively. It is worth noting that a single study may perform validation across multiple missing modality rate levels. Additionally, we found that 36.4% of the works trained on incomplete training datasets (cannot access missing modality data during training), while 63.6% actually focused on testing with missing modalities. We can see that there is still small number of research on incomplete training datasets, so it is important to emphasise the need for more research to this field.

7 Open Issues and Future Research Directions

7.1 Accurate Missing Modality/Representation Data Generation

Some researchers (Ma et al., 2021c) have shown that recovering, reconstructing or generating missing modalities or their representations can help improve the performance of the model when encountering missing-modality samples. However, these generated modalities or representations usually have artifacts or hallucinations, which may be caused by the generative model itself or its training datasets. Therefore, exploring how to generate accurate and unbiased missing modalities or their representations will be the important research direction in the future.

7.2 Recovery or Non-Recovery Methods ?

In MLMM, *"to recover or not to recover the missing modality"*, that is the question. According to our observations, most current approaches to handling MLMM can be broadly categorized into two types: one focusing on recovering the missing modal information so that the multimodal model can continue functioning (Wang et al., 2023a; Ma et al., 2021c), and the other attempting to make predictions using only the data modalities available (Hu et al., 2020; Zhang et al., 2024). However, some studies (Yao et al., 2024) have shown that the recovered modality information might be ignored or dominated by the existing modality information due to the imbalance missing rates of different modalities in the training set, leading the model to still rely on the available modalities when making predictions. Moreover, Lin et al. (Lin et al., 2024) have found that in situations where crucial modality information is missing, the model might be unduly influenced by less important existing modality information, resulting in incorrect outcomes. In such cases, attempting to recover missing modality information can potentially alleviate this problem. Consequently, determining under what circumstances recovery or non-recovery methods are more beneficial for model predictions, as well as how to measure whether the recovered modality information is not dominated by other modality information or is unbiased, remains a significant challenge. Addressing these issues is essential for advancing the effectiveness and reliability of deep multimodal learning with missing modality methods.

7.3 Benchmarking and Evaluations for Missing Modality Problems

Benchmarking and fair evaluation play crucial roles in guiding the field. Many works are trained/tested under different missing modality settings, which makes it difficult for researchers to find works with the same settings for comparison. Also, the recent rise of large pre-trained models, such as GPT-4, has given researchers hope for achieving Artificial General Intelligence. Consequently, more large models (Liu et al., 2024a) are integrating visual, auditory, and other modal information to realize the vision of Multimodal Large Language Models. Although some research has been done based on those models, detailed analysis and benchmarks are urgently needed to evaluate the performance of MLLMs on missing modality problems. We call on the community to build a work similar to MultiBench (Liang et al., 2021), covering common

datasets mentioned in section 6 and settings of missing modality problems, to help researchers conduct more systematic research.

7.4 Method Efficiency

Current MLMM methods often overlook the need for more lightweight and efficient methods. For instance, some model combinations methods (Hu et al., 2020; Xue & Marculescu, 2023) require training an independent model for each modality combination. Similarly, methods (Chen et al., 2024a; Zhang et al., 2023c) that aim to recover missing modality information typically involve using a distinct model for each kind of missing information or a large unified model to generate all modalities. Although these approaches generally perform well, they are often too heavy. In the real world, many multimodal models need to be deployed on resource-constrained devices, such as space or disaster-response robots. These devices cannot accommodate high-performance GPUs and are prone to sensor damage, which can be difficult/costly to repair. Therefore, there is an urgent need to explore efficient and lightweight solutions for MLMM that can operate effectively under these constraints.

7.5 Multimodal Streaming Data with Missing Modality

Currently, only a small part of research focuses on handling missing modalities in multimodal streaming data. An example is sentiment analysis (Lin & Hu, 2023). Long sequences of multimodal data, such as RGB+X video and multimodal time series, are common in real world and essential for tasks like anomaly detection and video tracking, especially on robots. Therefore, MLMM needs to address missing modality problems in streaming data to advance this field further.

7.6 Multimodal Reinforcement Learning with Missing Modality

Multimodal reinforcement learning leverages information from different sensory modalities, enabling agents to learn effective strategies in environments. This approach has wide applications in areas such as robotic grasping, drone control, and driving decisions. However, these agent-based tasks often face practical challenges, as real-world scenarios frequently have sensor failures or restricted access to certain modality data, compromising the robustness of the algorithms. Currently, only a limited number of papers (Vasco et al., 2021; Lee et al., 2021; Awan et al., 2022) aim to address the problem of missing data or modalities in reinforcement learning. Therefore, we hope the community can focus more on the practically significant missing modality problem in reinforcement learning.

7.7 Multimodal AI with Missing Modality for Natural Science

In scientific fields such as drug prediction (Deng et al., 2020) and materials science (Muroga et al., 2023), multimodal learning plays a key role by integrating diverse data types like molecular structures, genomic sequences, and spectral images. This integration can enhance predictive accuracy and uncover new insights. However, implementing multimodal learning in these domains is challenging due to data access restrictions, high acquisition costs, and the incompatibility of heterogeneous data. These issues often result in missing modalities, hindering the potential of multimodal models. Despite the critical need, there has been limited research (He et al., 2024) on MLMM in scientific applications. Addressing this gap is essential for advancing AI-driven discoveries, requiring new methods to handle incomplete multimodal datasets and create robust models capable of learning despite missing data.

8 Conclusion

In this survey, we present the first comprehensive survey of Deep Multimodal Learning with Missing Modality. We begin with a brief introduction to the motivation of the missing modality problem and the real-world reasons that underscore its significance. Following this, we summarize the current advances based on our fine-grained taxonomy and review the applications and relevant datasets. Finally, we discuss the existing challenges and potential future directions in this field. Although more and more researchers are involved in

studying the problem of missing modality, we are also concerned about some urgent issues that need to be addressed, such as a unified benchmark for testing (e.g., multimodal large language models) and the need for a wider range of applications (e.g., natural science). With our comprehensive and detailed approach, we hope that this survey will inspire more researchers to explore missing modality problems.

References

Michalis Angelou, Vassilis Solachidis, Nicholas Vretos, and Petros Daras. Graph-based multimodal fusion with metric learning for multimodal classification. *Pattern Recognition*, 95:296–307, 2019.

Nikhilanand Arya and Sriparna Saha. Multi-modal classification for human breast cancer prognosis prediction: proposal of deep-learning based stacked ensemble model. *IEEE/ACM transactions on computational biology and bioinformatics*, 19(2):1032–1041, 2020.

Nikhilanand Arya and Sriparna Saha. Generative incomplete multi-view prognosis predictor for breast cancer: Gimpp. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(4):2252–2263, 2021.

Saqib Ejaz Awan, Mohammed Bennamoun, Ferdous Sohel, Frank Sanfilippo, and Girish Dwivedi. A reinforcement learning-based approach for imputing missing data. *Neural Computing and Applications*, 34 (12):9701–9716, 2022.

Reza Azad, Nika Khosravi, Mohammad Dehghanmanshadi, Julien Cohen-Adad, and Dorit Merhof. Medical image segmentation on mri images with missing modalities: A review. *arXiv preprint arXiv:2203.06217*, 2022.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. URL `https://www.adept.ai/blog/fuyu-8b`.

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

Benjamin Bischke, Patrick Helber, Florian Koenig, Damian Borth, and Andreas Dengel. Overcoming missing and incomplete modalities with generative adversarial networks for building footprint segmentation. In *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, pp. 1–6. IEEE, 2018.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80, 2016.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Sergio Campos, Luis Pizarro, Carlos Valle, Katherine R Gray, Daniel Rueckert, and Héctor Allende. Evaluating imputation techniques for missing data in adni: a patient classification study. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 20th Iberoamerican Congress, CIARP 2015, Montevideo, Uruguay, November 9-12, 2015, Proceedings 20*, pp. 3–10. Springer, 2015.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4): 377–390, 2014.

Juhan Cha, Minseok Joo, Jihwan Park, Sanghyeok Lee, Injae Kim, and Hyunwoo J. Kim. Robust multimodal 3d object detection via modality-agnostic decoding and proximity-based modality ensemble, 2024. URL `https://arxiv.org/abs/2407.19156`.

Guoqing Chao, Shiliang Sun, and Jinbo Bi. A survey on multiview clustering. *IEEE Transactions on Artificial Intelligence*, 2(2):146–168, 2021. doi: 10.1109/TAI.2021.3065894.

Hava Chaptoukaev, Valeriya Strizhkova, Michele Panariello, Bianca Dalpaos, Aglind Reka, Valeria Manera, Susanne Thümmler, Esma Ismailova, Massimiliano Todisco, Maria A Zuluaga, et al. Stressid: a multimodal dataset for stress identification. *Advances in Neural Information Processing Systems*, 36, 2024.

Agisilaos Chartsias, Thomas Joyce, Mario Valerio Giuffrida, and Sotirios A Tsaftaris. Multimodal mr synthesis via modality-invariant latent representation. *IEEE transactions on medical imaging*, 37(3):803–814, 2017.

Cheng Chen, Qi Dou, Yueming Jin, Hao Chen, Jing Qin, and Pheng-Ann Heng. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pp. 447–456. Springer, 2019.

Jiayi Chen and Aidong Zhang. Hgmf: heterogeneous graph-based fusion for multimodal data with incompleteness. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1295–1305, 2020.

Qianqian Chen, Jiadong Zhang, Runqi Meng, Lei Zhou, Zhenhui Li, Qianjin Feng, and Dinggang Shen. Modality-specific information disentanglement from multi-parametric mri for breast tumor segmentation and computer-aided diagnosis. *IEEE Transactions on Medical Imaging*, 2024a.

Sishuo Chen, Lei Li, Shuhuai Ren, Rundong Gao, Yuanxin Liu, Xiaohan Bi, Xu Sun, and Lu Hou. Towards multimodal video paragraph captioning models robust to missing modality. *arXiv preprint arXiv:2403.19221*, 2024b.

Yi-Ting Chen, Jinghao Shi, Zelin Ye, Christoph Mertz, Deva Ramanan, and Shu Kong. Multimodal object detection via probabilistic ensembling. In *European Conference on Computer Vision*, pp. 139–158. Springer, 2022.

Zhengyang Chen, Shuai Wang, and Yanmin Qian. Multi-modality matters: A performance leap on voxceleb. In *INTERSPEECH*, pp. 2252–2256, 2020.

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

Chiho Choi, Sangpil Kim, and Karthik Ramani. Learning hand articulations by hallucinating heat distribution. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3104–3113, 2017.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*, Santorini, Greece., July 8-10, 2009.

Ira Cohen, Fabio Gagliardi Cozman, Nicu Sebe, Marcelo Cesar Cirelo, and Thomas S Huang. Semisupervised learning of classifiers: Theory, algorithms, and their application to human-computer interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(12):1553–1566, 2004.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pp. 1–23, 2022.

Christian Debes, Andreas Merentitis, Roel Heremans, Jürgen Hahn, Nikolaos Frangiadakis, Tim van Kasteren, Wenzhi Liao, Rik Bellens, Aleksandra Pižurica, Sidharta Gautama, et al. Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(6):2405–2418, 2014.

Ruben Delgado-Escano, Francisco M Castro, Nicolás Guil, Vicky Kalogeiton, and Manuel J Marin-Jimenez. Multimodal gait recognition under missing modalities. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 3003–3007. IEEE, 2021.

Yifan Deng, Xinran Xu, Yang Qiu, Jingbo Xia, Wen Zhang, and Shichao Liu. A multimodal deep learning framework for predicting drug–drug interaction events. *Bioinformatics*, 36(15):4316–4322, 2020.

Abbey A. Donaldson. After Three Years on Mars, NASA's Ingenuity Helicopter Mission Ends - NASA — nasa.gov. `https://www.nasa.gov/news-release/after-three-years-on-mars-nasas-ingenuity-helicopter-mission-ends/`, Jan 2024. [Accessed 13-05-2024].

Robert Duin. Multiple Features. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5HC70.

Chongjian Ge, Junsong Chen, Enze Xie, Zhongdao Wang, Lanqing Hong, Huchuan Lu, Zhenguo Li, and Ping Luo. Metabev: Solving sensor failures for 3d detection and map segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8721–8731, 2023.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Xinyu Gong, Sreyas Mohan, Naina Dhingra, Jean-Charles Bazin, Yilei Li, Zhangyang Wang, and Rakesh Ranjan. Mmg-ego4d: multimodal generalization in egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6481–6491, 2023.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022.

Clément Guérin, Christophe Rigaud, Antoine Mercier, Farid Ammar-Boudjelal, Karell Bertet, Alain Bouju, Jean-Christophe Burie, Georges Louis, Jean-Marc Ogier, and Arnaud Revel. ebdtheque: a representative database of comics. In *2013 12th International Conference on Document Analysis and Recognition*, pp. 1145–1149. IEEE, 2013.

Kausic Gunasekar, Qiang Qiu, and Yezhou Yang. Low to high dimensional modality hallucination using aggregated fields of view. *IEEE Robotics and Automation Letters*, 5(2):1983–1990, 2020.

Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition, 2024. URL `https://arxiv.org/abs/2407.05374`.

Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

Sebastian Hafner and Yifang Ban. Multi-modal deep learning for multi-temporal urban mapping with a partly missing optical modality. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 6843–6846. IEEE, 2023.

Mohammad Hamghalam, Alejandro F Frangi, Baiying Lei, and Amber L Simpson. Modality completion via gaussian process prior variational autoencoders for multi-modal glioma segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*, pp. 442–452. Springer, 2021.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pp. 507–517, 2016.

Zhen He, Shuofeng Hu, Yaowen Chen, Sijing An, Jiahao Zhou, Runyan Liu, Junfeng Shi, Jing Wang, Guohua Dong, Jinhui Shi, et al. Mosaic integration and knowledge transfer of single-cell multimodal data with midas. *Nature Biotechnology*, pp. 1–12, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and helmholtz free energy. *Advances in neural information processing systems*, 6, 1993.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 826–834, 2016.

Danfeng Hong, Jingliang Hu, Jing Yao, Jocelyn Chanussot, and Xiao Xiang Zhu. Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178:68–80, 2021.

Minhao Hu, Matthis Maillard, Ya Zhang, Tommaso Ciceri, Giammarco La Barbera, Isabelle Bloch, and Pietro Gori. Knowledge distillation from multi-modal to mono-modal segmentation networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pp. 772–781. Springer, 2020.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17853–17862, 2023.

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

Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. Epic-sounds: A large-scale dataset of actions that sound. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1037–1045, 2015.

Clifford R Jack Jr, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 27(4):685–691, 2008.

Jaehyuk Jang, Yooseung Wang, and Changick Kim. Towards robust multimodal prompting with missing modalities. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8070–8074. IEEE, 2024.

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

Wei Ji, Jingjing Li, Cheng Bian, Zongwei Zhou, Jiaying Zhao, Alan L Yuille, and Li Cheng. Multispectral video semantic segmentation: A benchmark dataset and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1094–1104, 2023.

Vijay John and Yasutomo Kawanishi. A multimodal sensor fusion framework robust to missing modalities for person recognition. In *Proceedings of the 4th ACM International Conference on Multimedia in Asia*, pp. 1–5, 2022.

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. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

kaggle. kaggle flir thermal, 2020. URL `https://www.kaggle.com/datasets/deepnewbie/flir-thermal-images-dataset`.

Michael Kampffmeyer, Arnt-Børre Salberg, and Robert Jenssen. Urban land cover classification with missing data modalities using deep convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(6):1758–1768, 2018.

Bob Kemp, Aeilko H Zwinderman, Bert Tuk, Hilbert AC Kamphuisen, and Josefien JL Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *2011 IEEE international conference on robotics and automation*, pp. 1817–1824. IEEE, 2011.

Himabindu Lakkaraju, Julian McAuley, and Jure Leskovec. What's in a name? understanding the interplay between titles, content, and communities in social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pp. 311–320, 2013.

Harlin Lee, Boyue Li, Shelly DeForte, Mark L Splaingard, Yungui Huang, Yuejie Chi, and Simon L Linwood. A large collection of real-world pediatric sleep studies. *Scientific Data*, 9(1):421, 2022.

Kwanhyung Lee, Soojeong Lee, Sangchul Hahn, Heejung Hyun, Edward Choi, Byungeun Ahn, and Joohyung Lee. Learning missing modal electronic health records with unified multi-modal data embedding and modality-aware attention. In *Machine Learning for Healthcare Conference*, pp. 423–442. PMLR, 2023a.

Michelle A Lee, Matthew Tan, Yuke Zhu, and Jeannette Bohg. Detect, reject, correct: Crossmodal compensation of corrupted sensors. In *2021 IEEE international conference on robotics and automation (ICRA)*, pp. 909–916. IEEE, 2021.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14943–14952, 2023b.

Žiga Lesjak, Alfiia Galimzianova, Aleš Koren, Matej Lukin, Franjo Pernuš, Boštjan Likar, and Žiga Špiclin. A novel public mr image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. *Neuroinformatics*, 16:51–63, 2018.

Bo Li, Peiyuan Zhang, Jingkang Yang, Yuanhan Zhang, Fanyi Pu, and Ziwei Liu. Otterhd: A high-resolution multi-modality model. *arXiv preprint arXiv:2311.04219*, 2023a.

Haitao Li, Ziyu Li, Yiheng Mao, Zhengyao Ding, and Zhengxing Huang. Dc-seg: Disentangled contrastive learning for brain tumor segmentation with missing modalities. *arXiv preprint arXiv:2505.11921*, 2025a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12458–12468, 2024a.

Mingyang Li, Shao-Lun Huang, and Lin Zhang. A general framework for incomplete cross-modal retrieval with missing labels and missing modalities. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4763–4767. IEEE, 2022a.

Ning Li, Tao Li, Chunyu Hu, Kai Wang, and Hong Kang. A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection. In *Benchmarking, Measuring, and Optimizing: Third BenchCouncil International Symposium, Bench 2020, Virtual Event, November 15–16, 2020, Revised Selected Papers 3*, pp. 177–193. Springer, 2021.

Rongjian Li, Wenlu Zhang, Heung-Il Suk, Li Wang, Jiang Li, Dinggang Shen, and Shuiwang Ji. Deep learning based imaging data completion for improved brain disease diagnosis. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part III 17*, pp. 305–312. Springer, 2014.

Sijie Li, Chen Chen, and Jungong Han. Simmlm: A simple framework for multi-modal learning with missing modality. *arXiv preprint arXiv:2507.19264*, 2025b.

Siting Li, Chenzhuang Du, Yue Zhao, Yu Huang, and Hang Zhao. What makes for robust multi-modal models in the face of missing modalities? *arXiv preprint arXiv:2310.06383*, 2023c.

Xue Li, Guo Zhang, Hao Cui, Shasha Hou, Shunyao Wang, Xin Li, Yujia Chen, Zhijiang Li, and Li Zhang. Mcanet: A joint semantic segmentation framework of optical and sar images for land use classification. *International Journal of Applied Earth Observation and Geoinformation*, 106:102638, 2022b.

Yu-Jhe Li, Jinhyung Park, Matthew O'Toole, and Kris Kitani. Modality-agnostic learning for radar-lidar fusion in vehicle detection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 918–927, 2022c.

Yulin Li, Tianzhu Zhang, Xiang Liu, Qi Tian, Yongdong Zhang, and Feng Wu. Visible-infrared person re-identification with modality-specific memory network. *IEEE Transactions on Image Processing*, 31: 7165–7178, 2022d.

Zhiyuan Li, Yafei Zhang, Huafeng Li, Yi Chai, and Yushi Yang. Deformation-aware and reconstruction-driven multimodal representation learning for brain tumor segmentation with missing modalities. *Biomedical Signal Processing and Control*, 91:106012, 2024b.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence*, 2023.

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

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. *Advances in neural information processing systems*, 2021(DB1):1, 2021.

Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19800–19808, 2022.

Chenfei Liao, Kaiyu Lei, Xu Zheng, Junha Moon, Zhixiong Wang, Yixuan Wang, Danda Pani Paudel, Luc Van Gool, and Xuming Hu. Benchmarking multi-modal semantic segmentation under sensor failures: Missing and noisy modality robustness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1576–1586, 2025.

Ronghao Lin and Haifeng Hu. Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702, 2023.

Xun Lin, Shuai Wang, Rizhao Cai, Yizhong Liu, Ying Fu, Zitong Yu, Wenzhong Tang, and Alex Kot. Suppress and rebalance: Towards generalized multi-modal face anti-spoofing. *arXiv preprint arXiv:2402.19298*, 2024.

Zhenghong Lin, Yanchao Tan, Yunfei Zhan, Weiming Liu, Fan Wang, Chaochao Chen, Shiping Wang, and Carl Yang. Contrastive intra-and inter-modality generation for enhancing incomplete multimedia recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 6234–6242, 2023.

Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z Li. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1179–1187, 2021a.

Chengyu Liu, David Springer, Qiao Li, Benjamin Moody, Ricardo Abad Juan, Francisco J Chorro, Francisco Castells, José Millet Roig, Ikaro Silva, Alistair EW Johnson, et al. An open access database for the evaluation of heart sound algorithms. *Physiological measurement*, 2016.

Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*, pp. 612–630. Springer, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. M3ae: multimodal representation learning for brain tumor segmentation with missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1657–1665, 2023a.

Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8115–8124, 2023b.

Ruiping Liu, Jiaming Zhang, Kunyu Peng, Yufan Chen, Ke Cao, Junwei Zheng, M Saquib Sarfraz, Kailun Yang, and Rainer Stiefelhagen. Fourier prompt tuning for modality-incomplete scene segmentation. *arXiv preprint arXiv:2401.16923*, 2024b.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. Llava-plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*, 2023c.

Yanbei Liu, Lianxi Fan, Changqing Zhang, Tao Zhou, Zhitao Xiao, Lei Geng, and Dinggang Shen. Incomplete multi-modal representation learning for alzheimer's disease diagnosis. *Medical Image Analysis*, 69:101953, 2021b.

Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. Mmkg: multi-modal knowledge graphs. In *The Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2–6, 2019, Proceedings 16*, pp. 459–474. Springer, 2019.

Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

Zihong Luo, Haochen Xue, Mingyu Jin, Chengzhi Liu, Zile Huang, Chong Zhang, and Shuliang Zhao. Mc-dbn: A deep belief network-based model for modality completion. *arXiv preprint arXiv:2402.09782*, 2024.

Fei Ma, Shao-Lun Huang, and Lin Zhang. An efficient approach for audio-visual emotion recognition with missing labels and missing modalities. In *2021 IEEE international conference on multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2021a.

Fei Ma, Xiangxiang Xu, Shao-Lun Huang, and Lin Zhang. Maximum likelihood estimation for multimodal learning with missing modality. *arXiv preprint arXiv:2108.10513*, 2021b.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2302–2310, 2021c.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18177–18186, 2022.

Daniele Malitesta, Emanuele Rossi, Claudio Pomo, Fragkiskos D Malliaros, and Tommaso Di Noia. Dealing with missing modalities in multimodal recommendation: a feature propagation-based approach. *arXiv preprint arXiv:2403.19841*, 2024.

Carole L Marcus, Reneé H Moore, Carol L Rosen, Bruno Giordani, Susan L Garetz, H Gerry Taylor, Ron B Mitchell, Raouf Amin, Eliot S Katz, Raanan Arens, et al. A randomized trial of adenotonsillectomy for childhood sleep apnea. *New England Journal of Medicine*, 2013.

C Mario Christoudias, Raquel Urtasun, Mathieu Salzmann, and Trevor Darrell. Learning to recognize objects from unseen modalities. In *Computer Vision–ECCV 2010*, pp. 677–691. Springer, 2010.

Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. The enterface'05 audio-visual emotion database. In *22nd international conference on data engineering workshops (ICDEW'06)*, pp. 8–8. IEEE, 2006.

Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia tools and applications*, 76: 21811–21838, 2017.

Toshihiko Matsuura, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Generalized bayesian canonical correlation analysis with missing modalities. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pp. 43–52, 2015.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018.

Eleni P Mimitou, Caleb A Lareau, Kelvin Y Chen, Andre L Zorzetto-Fernandes, Yuhan Hao, Yusuke Takeshima, Wendy Luo, Tse-Shun Huang, Bertrand Z Yeung, Efthymia Papalexi, et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nature biotechnology*, 39(10):1246–1258, 2021.

George B Moody and Roger G Mark. The impact of the mit-bih arrhythmia database. *IEEE engineering in medicine and biology magazine*, 20(3):45–50, 2001.

Julie Mordacq, Leo Milecki, Maria Vakalopoulou, Steve Oudot, and Vicky Kalogeiton. Adapt: Multimodal learning for detecting physiological changes under missing modalities. In *Medical Imaging with Deep Learning*, 2024.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pp. 169–176, 2011.

Shun Muroga, Yasuaki Miki, and Kenji Hata. A comprehensive and versatile multimodal deep-learning approach for predicting diverse properties of advanced materials. *Advanced Science*, 10(24):2302508, 2023.

Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, pp. 311–320. Springer, 2019.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213, 2021.

Nhu-Van Nguyen, Christophe Rigaud, and Jean-Christophe Burie. Digital comics image indexing based on deep learning. *Journal of Imaging*, 4(7):89, 2018.

Konyul Park, Yecheol Kim, Daehun Kim, and Jun Won Choi. Resilient sensor fusion under adverse sensor failures via multi-modal expert fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6720–6729, 2025.

Kyu Ri Park, Hong Joo Lee, and Jung Uk Kim. Learning trimodal relation for audio-visual question answering with missing modality, 2024. URL `https://arxiv.org/abs/2407.16171`.

Srinivas Parthasarathy and Shiva Sundaram. Training strategies to handle missing modalities for audio-visual expression recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pp. 400–404, 2020.

Claudio Persello; Saurabh Prasad; Gemine Vivone; Vincent Lonjou ; Frédéric Bretar ; Raquel Rodriguez-Suquet ; Pauline Guntzburger ; Vincent Poulain ; Jacqueline Le Moigne; Benjamin Smith ; Sujay Kumar ; Thomas Huang ; Sophie Ricci ; Thanh Huy Nguyen ; Andrea Piacentini. 2024 ieee grss data fusion contest - flood rapid mapping. *IEEE Dataport*, 2023a. doi: 10.21227/73zj-4303. URL `https://dx.doi.org/10.21227/73zj-4303`.

Vittorio Pipoli, Alessia Saporita, Federico Bolelli, Marcella Cornia, Lorenzo Baraldi, Costantino Grana, Rita Cucchiara, Elisa Ficarra, et al. Missrag: Addressing the missing modality challenge in multimodal large language models. In *Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision*, 2025.

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

PromptCloud. Toy products on amazon, Sep 2017. URL `https://www.kaggle.com/datasets/PromptCloudHQ/toy-products-on-amazon`.

Shuwei Qian and Chongjun Wang. Com: Contrastive masked-attention model for incomplete multimodal learning. *Neural Networks*, 162:443–455, 2023.

Yansheng Qiu, Ziyuan Zhao, Hongdou Yao, Delin Chen, and Zheng Wang. Modal-aware visual prompting for incomplete multi-modal brain tumor segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3228–3239, 2023.

Ilija Radosavovic, Bike Zhang, Baifeng Shi, Jathushan Rajasegaran, Sarthak Kamat, Trevor Darrell, Koushil Sreenath, and Jitendra Malik. Humanoid locomotion as next token prediction. *arXiv preprint arXiv:2402.19469*, 2024.

Hossein Rahmani, Arif Mahmood, Du Huynh, and Ajmal Mian. Histogram of oriented principal components for cross-view action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(12): 2430–2443, 2016.

Merey Ramazanova, Alejandro Pardo, Bernard Ghanem, and Motasem Alfarra. Combating missing modalities in egocentric videos at test time. *arXiv preprint arXiv:2404.15161*, 2024.

Behnood Rasti, Pedram Ghamisi, and Richard Gloaguen. Hyperspectral and lidar fusion using extinction profiles and total variation component analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3997–4007, 2017.

Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *arXiv preprint arXiv:2310.03986*, 2023.

Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pp. 1–8. IEEE, 2013.

Pramit Saha, Divyanshu Mishra, Felix Wagner, Konstantinos Kamnitsas, and J Alison Noble. Examining modality incongruity in multimodal federated learning for medical vision and language-based disease detection. *arXiv preprint arXiv:2402.05294*, 2024.

Sayan Deb Sarkar, Ondrej Miksik, Marc Pollefeys, Daniel Barath, and Iro Armeni. Crossover: 3d scene cross-modal alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8985–8994, 2025.

Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pp. 400–408, 2018.

Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. Multimodal approaches for emotion recognition: a survey. In *Internet Imaging VI*, volume 5670, pp. 56–67. SPIE, 2005.

Deep Shah, Amit Barve, Brijesh Vala, and Jay Gandhi. A survey on brain tumor segmentation with missing mri modalities. In *International Conference on Information Technology*, pp. 299–308. Springer, 2023.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010–1019, 2016.

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

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36, 2024.

Jacob Shermeyer, Daniel Hogan, Jason Brown, Adam Van Etten, Nicholas Weir, Fabio Pacifici, Ronny Hansch, Alexei Bastidas, Scott Soenen, Todd Bacastow, et al. Spacenet 6: Multi-sensor all weather mapping dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 196–197, 2020.

Junjie Shi, Caozhi Shang, Zhaobin Sun, Li Yu, Xin Yang, and Zengqiang Yan. Passion: Towards effective incomplete multi-modal medical image segmentation with imbalanced missing rates, 2024. URL `https://arxiv.org/abs/2407.14796`.

Aniruddh Sikdar, Jayant Teotia, and Suresh Sundaram. Contrastive learning-based spectral knowledge distillation for multi-modality and missing modality scenarios in semantic segmentation. *arXiv preprint arXiv:2312.02240*, 2023.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pp. 746–760. Springer, 2012.

Kihyuk Sohn, Wenling Shang, and Honglak Lee. Improved multimodal deep learning with variation of information. *Advances in neural information processing systems*, 27, 2014.

Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertolli, Eva-Maria Messner, Erik Cambria, Guoying Zhao, and Björn W Schuller. The muse 2021 multimodal sentiment analysis challenge: sentiment, emotion, physiological-emotion, and stress. In *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pp. 5–14. 2021a.

Lukas Stappen, Eva-Maria Meßner, Erik Cambria, Guoying Zhao, and Björn W Schuller. Muse 2021 challenge: Multimodal emotion, sentiment, physiological-emotion, and stress detection. In *Proceedings of the 29th ACM International conference on multimedia*, pp. 5706–5707, 2021b.

Paul Streli, Rayan Armani, Yi Fei Cheng, and Christian Holz. Hoov: Hand out-of-view tracking for proprioceptive interaction using inertial sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2023.

Jun Sun, Xinxin Zhang, Shoukang Han, Yu-Ping Ruan, and Taihao Li. Redcore: Relative advantage aware cross-modal representation learning for missing modalities with imbalanced missing rates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15173–15182, 2024a.

Yuhang Sun, Zhizhong Liu, Quan Z Sheng, Dianhui Chu, Jian Yu, and Hongxiang Sun. Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, pp. 102454, 2024b.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11888–11898, 2023.

Karasawa Takumi, Kohei Watanabe, Qishen Ha, Antonio Tejero-De-Pablos, Yoshitaka Ushiku, and Tatsuya Harada. Multispectral object detection for autonomous vehicles. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pp. 35–43, 2017.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36, 2024.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

NASA Science Editorial Team. Keeping Our Sense of Direction: Dealing With a Dead Sensor - NASA Science — science.nasa.gov. `https://science.nasa.gov/missions/mars-2020-perseverance/ingenuity-helicopter/keeping-our-sense-of-direction-dealing-with-a-dead-sensor/`, JUN 2022.

Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015.

Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, 33(5):1–10, 2014.

Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1405–1414, 2017.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018.

Miguel Vasco, Hang Yin, Francisco S Melo, and Ana Paiva. How to sense the world: Leveraging hierarchy in multimodal perception for robust reinforcement learning agents. *arXiv preprint arXiv:2110.03608*, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

Johannes Wagner, Elisabeth Andre, Florian Lingenfelser, and Jonghwa Kim. Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Transactions on Affective Computing*, 2(4): 206–218, 2011.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal 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, 2023a.

Hu Wang, Congbo Ma, Jianpeng Zhang, Yuan Zhang, Jodie Avery, Louise Hull, and Gustavo Carneiro. Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 216–226. Springer, 2023b.

Hu Wang, Congbo Ma, Yuyuan Liu, Yuanhong Chen, Yu Tian, Jodie Avery, Louise Hull, and Gustavo Carneiro. Enhancing multi-modal learning: Meta-learned cross-modal knowledge distillation for handling missing modalities. *arXiv preprint arXiv:2405.07155*, 2024a.

Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2649–2656, 2014.

Qi Wang, Liang Zhan, Paul Thompson, and Jiayu Zhou. Multimodal learning with incomplete modalities by knowledge distillation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1828–1838, 2020a.

Qianqian Wang, Huanhuan Lian, Gan Sun, Quanxue Gao, and Licheng Jiao. Icmsc: Incomplete cross-modal subspace clustering. *IEEE Transactions on Image Processing*, 30:305–317, 2020b.

Qifan Wang, Luo Si, and Bin Shen. Learning to hash on partial multi-modal data. In *24th International Joint Conference on Artificial Intelligence*, 2015.

Shuai Wang, Zipei Yan, Daoan Zhang, Haining Wei, Zhongsen Li, and Rui Li. Prototype knowledge distillation for medical segmentation with missing modality. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023c.

Yixin Wang, Yang Zhang, Yang Liu, Zihao Lin, Jiang Tian, Cheng Zhong, Zhongchao Shi, Jianping Fan, and Zhiqiang He. Acn: adversarial co-training network for brain tumor segmentation with missing modalities. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*, pp. 410–420. Springer, 2021.

Yiyu Wang, Haifang Jian, Jian Zhuang, Huimin Guo, and Yan Leng. Sslmm: Semi-supervised learning with missing modalities for multimodal sentiment analysis. *Information Fusion*, 120:103058, 2025.

Yuanyi Wang, Haifeng Sun, Jiabo Wang, Jingyu Wang, Wei Tang, Qi Qi, Shaoling Sun, and Jianxin Liao. Towards semantic consistency: Dirichlet energy driven robust multi-modal entity alignment. *arXiv preprint arXiv:2401.17859*, 2024b.

Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22025–22034, 2023d.

Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36, 2024c.

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

Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53, 2013.

Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2776–2784, 2023.

Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pp. 5380–5389, 2017.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023a.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv:2309.05519*, 2023b.

Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. Multimodal patient representation learning with missing modalities and labels. In *The Twelfth International Conference on Learning Representations*, 2024.

Feng Xu, Chuang Zhu, Wenqi Tang, Ying Wang, Yu Zhang, Jie Li, Hongchuan Jiang, Zhongyue Shi, Jun Liu, and Mulan Jin. Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Frontiers in oncology*, 11:759007, 2021.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Zihui Xue and Radu Marculescu. Dynamic multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2575–2584, 2023.

Heran Yang, Jian Sun, and Zongben Xu. Learning unified hyper-network for multi-modal mr image synthesis and tumor segmentation with missing modalities. *IEEE Transactions on Medical Imaging*, 2023a.

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

Yanwu Yang, Hairui Chen, Zhikai Chang, Yang Xiang, Chenfei Ye, and Ting Ma. Incomplete learning of multi-modal connectome for brain disorder diagnosis via modal-mixup and deep supervision. In *Medical Imaging With Deep Learning*, pp. 1006–1018. PMLR, 2024.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023b.

Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16416–16424, 2024.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

Naoto Yokoya, Pedram Ghamisi, Ronny Haensch, and Michael Schmitt. 2020 ieee grss data fusion contest: Global land cover mapping with weak supervision [technical committees]. *IEEE Geoscience and Remote Sensing Magazine*, 8(1):154–157, 2020.

Jeffrey T Young, Yundi Shi, Marc Niethammer, Michael Grauer, Christopher L Coe, Gabriele R Lubach, Bradley Davis, Francois Budin, Rebecca C Knickmeyer, Andrew L Alexander, et al. The unc-wisconsin rhesus macaque neurodevelopment database: a structural mri and dti database of early postnatal development. *Frontiers in neuroscience*, 11:29, 2017.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 3718–3727, 2020.

Lei Yuan, Yalin Wang, Paul M Thompson, Vaibhav A Narayan, Jieping Ye, Alzheimer's Disease Neuroimaging Initiative, et al. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *NeuroImage*, 61(3):622–632, 2012.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

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

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.

Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. 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, 2022.

Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1136–1147, 2023a.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3713–3722, 2019.

Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z Li. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):182–193, 2020.

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

Yuan Zhang, Hu Wang, David Butler, Minh-Son To, Jodie Avery, M Louise Hull, and Gustavo Carneiro. Distilling missing modality knowledge from ultrasound for endometriosis diagnosis with magnetic resonance images. In *2023 IEEE 20th International Symposium on Biomedical Imaging*, 2023b.

Yue Zhang, Chengtao Peng, Qiuli Wang, Dan Song, Kaiyan Li, and S Kevin Zhou. Unified multi-modal image synthesis for missing modality imputation. *arXiv preprint arXiv:2304.05340*, 2023c.

Zheyu Zhang, Gang Yang, Yueyi Zhang, Huanjing Yue, Aiping Liu, Yunwei Ou, Jian Gong, and Xiaoyan Sun. Tmformer: Token merging transformer for brain tumor segmentation with missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM Computing Surveys*, 56(9):1–36, 2024.

Zechen Zhao, Heran Yang, and Jian Sun. Modality-adaptive feature interaction for brain tumor segmentation with missing modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 183–192. Springer, 2022.

Aihua Zheng, Zi Wang, Zihan Chen, Chenglong Li, and Jin Tang. Robust multi-modality person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 3529–3537, 2021.

Wei-Long Zheng and Bao-Liang Lu. Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Transactions on autonomous mental development*, 7(3):162–175, 2015.

Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Learning modality-agnostic representation for semantic segmentation from any modalities, 2024. URL `https://arxiv.org/abs/2407.11351`.

Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li. Forecasting fine-grained air quality based on big data. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 2267–2276, 2015.

Zhuo Zhi, Ziquan Liu, Moe Elbadawi, Adam Daneshmend, Mine Orlu, Abdul Basit, Andreas Demosthenous, and Miguel Rodrigues. Borrowing treasures from neighbors: In-context learning for multimodal learning with missing modalities and data scarcity. *arXiv:2403.09428*, 2024.

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

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

Tongxue Zhou, Su Ruan, and Haigen Hu. A literature survey of mr-based brain tumor segmentation with missing modalities. *Computerized Medical Imaging and Graphics*, 104:102167, 2023.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023.

Xiaoguang Zhu, Lianlong Sun, Yang Liu, Pengyi Jiang, Uma Srivatsa, Nipavan Chiamvimonvat, and Vladimir Filkov. Causal debiasing medical multimodal representation learning with missing modalities. *arXiv preprint arXiv:2509.05615*, 2025.

A Paper Collection

The following shows the full names of the mentioned conference and journal abbreviations in the "Paper Collection" paragraph of Section Introduction. The collected papers come from, but are not limited to, the following conferences (such as the Association for the Advancement of Artificial Intelligence (AAAI), International Joint Conference on Artificial Intelligence (IJCAI), Conference on Neural Information Processing Systems (NeurIPS), International Conference on Learning Representations (ICLR), International Conference on Machine Learning (ICML), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV), Annual Meeting of the Association for Computational Linguistics (ACL), Conference on Empirical Methods in Natural Language Processing (EMNLP), ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), ACM International Conference on Multimedia (ACM MM), International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), etc.) and journals (such as IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), IEEE Transactions on Image Processing (TIP), IEEE Transactions on Medical Imaging (TMI), IEEE Transactions on Multimedia (TMM), Journal of Machine Learning Research (JMLR), etc.).