REDEEMing Modality Information Loss: Retrieval-Guided Conditional Generation for Severely Modality Missing Learning

Jian Lang
jian_lang@std.uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Rongpei Hong
rongpei.hong@std.uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Zhangtao Cheng
zhangtao.cheng@outlook.com
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Ting Zhong
zhongting@uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Yong Wang
wangyong@ipplus360.com
Aiwen Tech
Zhengzhou, Henan, China
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong, China

Fan Zhou*
fan.zhou@uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
Key Laboratory of Intelligent Digital Media Technology of Sichuan Province
Chengdu, Sichuan, China

Abstract

Traditional multimodal learning approaches often assume that all modalities are available during both the training and inference phases. However, this assumption is often impractical in real-world scenarios due to challenges such as sensor failures, data corruption, or privacy concerns. While recent efforts focus on enhancing the robustness of pre-trained Multimodal Transformers (MTs) under missing modality conditions, mainstream work in this field often overlook reconstructing the missing modalities and rely on static, sample-agnostic prompt-tuning techniques, undermining their efficacy in severe modality missing scenarios. To address these limitations, we propose **REDEEM**, a novel **RE**trieval-gui**DE**d conditional g**E**nerative fra**M**ework that largely alleviates the modality missing problems on pre-trained MTs. REDEEM consists of a new adaptive retrieval mechanism to identify relevant instances for both modality-complete and -incomplete samples. It then conditions on the remaining modalities and utilizes the retrieved data as experts to effectively recover the missing ones in modality-incomplete instances through a within-modal reconstruction manner. Finally, REDEEM generates sample-aware inter-modal prompts from the retrieved instances to guide MTs in tackling severe modality missing challenges. Comprehensive experiments on three diverse multimodal classification benchmarks demonstrate that REDEEM significantly outperforms competitive baselines.

*Corresponding author.

KDD '25, August 3–7, 2025, Toronto, ON, Canada

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3737101

CCS Concepts

• **Computing methodologies** → **Computer vision**; • **Information systems** → **Multimedia and multimodal retrieval**.

Keywords

Incomplete multimodal learning, multimodal transformer, retrieval augmentation, mixture of experts, prompt tuning

ACM Reference Format:
Jian Lang, Rongpei Hong, Zhangtao Cheng, Ting Zhong, Yong Wang, and Fan Zhou. 2025. REDEEMing Modality Information Loss: Retrieval-Guided Conditional Generation for Severely Modality Missing Learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3711896.3737101

1 Introduction

Multimodal learning has garnered significant attention in both research and industry communities across various domains, including medical service [13, 14], autonomous driving [22, 34], and malicious content detection [4, 48]. Traditional multimodal learning approaches often implicitly assume that all modalities are available during training and inference. However, real-world applications often face the challenge of modality missing due to sensor failures, data corruption, or privacy concerns [21, 26]. These missing modalities can substantially curtail the performance and robustness of traditional multimodal models, even those Multimodal Transformers (MTs) [1, 29, 30] pre-trained on large corpora.

Recently, as the MTs have become the popular and dominant method in multimodal learning across various tasks, such as text-image retrieval and video generation [10, 31, 49], researchers have focused on enhancing the robustness of pre-trained MTs under modality missing conditions [12, 18, 25]. For instance, AMTR [25] adopts multi-task optimization and policy-based fusion in MTs to tackle modality-missing challenges, while prevailing methods MAPs [18] and MSPs [12] leverage prompt tuning to improve the

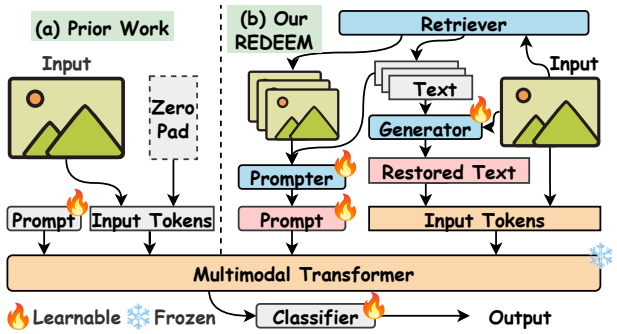


Figure 1: Prior MT-based methods (Zero Pad & Static Prompting) *vs.* REDEEM (Within-Modal Conditional Completion & Sample-Aware Inter-Modal Prompting) in image-only case.

pre-trained MTs' performance in such conditions without fine-tuning the entire framework.

However, these mainstream MT-based methods often fill missing modalities by padding dummy values (e.g., zeros) without any reconstruction and rely on static, sample-agnostic prompts, as shown in Figure 1(a). As a result, there are two major **limitations** that lie in these methods: **(i)** Since pre-trained MTs lack exposure to modality missing scenarios during pre-training, they may struggle to interpret dummy values that easily introduce noise and instability [8, 25]. **(ii)** Static, sample-agnostic prompts fall short in adapting to every sample, offering limited applicability particularly in datasets containing a mix of modality-complete and -incomplete instances [5, 46]. The increased variability in such datasets amplifies distributional diversity, further restricting these static prompts to guide pre-trained MTs in addressing modality missing challenges.

To address these issues, we propose **REDEEM**, a novel **RE**trieval-gui**DE**d conditional g**E**nerative fra**M**ework, as illustrated in Figure 1(b). Inspired by the notion that modality-missing results in information loss, REDEEM enhances informational completeness by incorporating relevant data through retrieval mechanisms. This retrieval-guided conditional generation paradigm facilitates both the conditional completion of missing modalities and generation of inter-modal dynamic prompts for specific samples, enhancing the robustness of MTs without fine-tuning the whole framework.

To address issue **(i)**, we propose to recover missing content by leveraging both the remaining modalities and retrieved information, ensuring more robust multimodal inputs for pre-trained MTs. To achieve this, we first design a new Missing Self-Adaptive Retriever that dynamically adjusts its retrieval strategy based on specific modality missing scenarios and leverages the available modalities to perform effective within-modal retrieval. Subsequently, unlike traditional reconstruction methods [26, 41], which often generate the missing modalities merely through the available ones and thus introduce modality heterogeneity issues [37, 39], we propose a novel Conditional Mixture of Experts Generator for within-modal reconstruction. Specifically, for a missing modality, the generator identifies a set of within-modal experts from the retrieved data with the same modality as the missing one. A carefully designed router then conditions on the interplay between the experts and the remaining modalities, adaptively modulating the contribution of each expert in the reconstruction process and aggregating these within-modal experts to accurately recover the missing content.

Drawing the inspiration of few-shot learning in MTs [28, 47], which leverages a few contextual examples to adapt MTs to specific situations, we propose a fresh Sample-Aware Inter-Modal Prompter to address issue **(ii)**. This prompter generates instance-specific, dynamic prompts by extracting informative cross-modal patterns from retrieved modality-complete samples (i.e., image-text pairs). By excavating the inter-modal relationships embedded in these target-relevant instances, the prompter provides MTs with fine-grained cross-modal cues, enabling MTs to understand the correspondence between modalities and robustly handle both complete and incomplete data. Our contributions are summarized as follows:

- We propose REDEEM, a novel framework that pioneers a retrieval-guided conditional generation paradigm for both missing modality recovery and prompt generation. It can effectively enhance the performance and robustness of MTs under severe modality missing challenges.
- We design a new Conditional Mixture of Experts Generator to realize within-modal reconstruction. This generator leverages the remaining modalities as conditions to guide the within-modal experts – retrieved data corresponding to the same modality as the missing one – to effectively recover the missing modalities.
- We develop a fresh Sample-Aware Inter-Modal Prompter that extracts inter-modal relationships from target-relevant modality-complete data, yielding dynamic prompts. These tailored prompts explicitly deliver sample-specific and cross-modal cues, largely enhance the robustness of pre-trained MTs in tackling severe modality missing scenarios.

Extensive experiments on three diverse benchmarks demonstrate that REDEEM outperforms state-of-the-art baselines in handling modality missing scenarios. The source codes to reproduce our results are available at https://github.com/Jian-Lang/REDEEM.

2 Related Work

2.1 Incomplete Multimodal Learning

Traditional multimodal methods often assume full modalities and struggle with incomplete data, leading to inaccurate or misleading decisions [1, 29, 30]. This limitation undermines the reliability and applicability of these methods in risk-sensitive scenarios, e.g., autonomous driving [22, 34], medical service [13, 14], and malicious content detection [4, 48]. To address this issue, researchers designed various methods that are broadly divided into three groups.

The first group of methods, referred to as *modality invariant learning*, primarily extracts inter-modal correlations to project multimodal features into a shared space and leverages the shared features for prediction. IF-MMIN [50], ShaSpec [36], DrFuse [43], CorrKD [21], and MoMKE [42] employed customized learning strategies (e.g., knowledge distillation, mixture of experts) to capture modality-invariant features across different modalities, thereby enhancing the robustness of missing multimodal learning. However, these methods simply fill incomplete inputs with dummy values, which introduces additional noise and causes unexpected behavior.

In contrast, *cross-modal imputation methods* aim to reconstruct missing modalities from the remaining available modalities with generative models [3, 26, 41, 45]. SMIL [26] and AcMAE [41] leveraged a general-purpose autoencoder to impute missing modalities

based on available modalities. TFR-Net [45] employed cross-modal attention mechanisms to generate representations for the missing modalities. However, these methods struggle with complex inter-modal relationships and modal distribution heterogeneity [37, 39], limiting the accuracy of recovered content.

With the rapid proliferation of MTs across various domains, a third category of approaches has emerged—*MT-based methods*-aimed at enhancing the robustness of pre-trained MTs in situations with missing modalities. Mainstream work MAPs [18] and MSPs [12] proposed to insert prompts at various layers within MTs to handle incomplete modalities without optimizing the entire framework. However, these MT-based methods are limited by the use of dummy values and static prompts. A very recent work, dubbed RAGPT [17], attempted to alleviate their limitations by simply averaging the retrieved modalities and utilizing the intra-modal prompting. Nevertheless, it struggles to accurately recover the missing modalities and overlooks what MTs actually "require" under severe modality-missing scenarios (i.e., the correct cross-modal patterns). In contrast, our Conditional Mixture of Experts Generator (CMoE Generator) explicitly conditions on the available modalities to guide the fine-grained aggregation of retrieved experts, enabling more accurate reconstruction of the missing modalities. Moreover, the Sample-Aware Inter-Modal Prompter (SAIM Prompter) effectively distills the cross-modal patterns from abundant retrieved modality-complete samples into dynamic prompts.

2.2 Mixture of Experts

The MoE was first introduced by Jacob *et al.*[11] as a method to integrate multiple experts, each trained on distinct subsets of data, into a unified, robust model. Eigen *et al.*[6] extended this concept to neural networks by designing a layer comprising expert networks and a trainable gating mechanism. This gating mechanism assigns weights to the experts on a per-sample basis, enabling MoE to generate weighted combinations of expert outputs dynamically. Recently, MoE has gained significant attention for its ability to expand model capacity without increasing computational costs [32], particularly in Natural Language Processing (NLP) [7, 19, 44] and Computer Vision (CV) [9, 27, 33, 42]. In multimodal learning, VL-MoE [33] explored the effectiveness of MoE in scaling vision-language models and investigated the trade-offs between model complexity and performance. In this work, we are the first to apply MoE for missing content recovery. By conditioning on the available modalities, our Conditional MoE Generator (CMoE Generator) dynamically prioritizes the experts—retrieved data of the same modality as the missing one—through a well-designed Conditional Soft Router. This router assigns higher importance to experts aligned with the reconstruction objectives while reducing the contributions of less relevant experts, enabling an effective within-modal reconstruction.

2.3 Prompt Tuning

Prompt tuning [23] utilizes a small number of learnable prompt parameters added to the pre-trained transformers, facilitating adjustments to the pre-trained models for alignment with downstream tasks. For incomplete modality learning, mainstream studies MAPs [18] and MSPs [12] introduced prompts strategically inserted at various layers within MTs to address incomplete modalities. However, their static, instance-agnostic prompts may not adapt effectively to different samples, particularly in datasets containing both modality-complete and modality-incomplete instances. The significant variability across samples under missing modality scenarios largely diminishes the effectiveness of these approaches. The just-released work RAGPT [17] leverage the retrieved instances to construct the intra-modal prompts. Nevertheless, it fails to guide the MTs in understanding the cross-modal relationships, which confines its effectiveness under severe missing modality conditions. In contrast, our Sample-Aware Inter-Modal Prompter (SAIM Prompter) generates inter-modal dynamic prompts from retrieved samples. These prompts act as few-shot exemplars, enabling the MTs to better capture the correspondence between text and image modalities within target-relevant modality-complete data, significantly improving MTs' robustness in handling missing modalities.

3 Methodology

3.1 Overview

In this section, we describe the proposed REDEEM in detail. The overall framework of our REDEEM is illustrated in Figure 2. We first provide the preliminaries for the incomplete modality learning problem. The subsequent subsections provide in-depth descriptions of the key components of REDEEM: the missing self-adaptive retriever, the conditional mixture of experts generator, and the sample-aware inter-modal prompter. The final subsection provides a complete workflow for classifying a modality-incomplete sample.

Problem Definition. Following the prior works [12, 18, 25], we consider a multimodal dataset with two modalities: text and image. Formally, we define the multimodal dataset as $\mathcal{D} = \{\mathcal{D}^c, \mathcal{D}^{m_1}, \mathcal{D}^{m_2}\}$. Then, $\mathcal{D}^c = \{(\mathcal{T}_j, \mathcal{I}_j, y_j)\}_{j=1}^{N^c}$ is the modality-complete subset, where y_j is the category of the j-th instance. $\mathcal{T}_j$ and $\mathcal{I}_j$ denote text and image modalities, respectively. N^c is the total number of instances in $\mathcal{D}^c$. In contrast, $\mathcal{D}^{m_1} = \{(\mathcal{T}_k, y_k)\}_{k=1}^{N^{m_1}}$ and $\mathcal{D}^{m_2} = \{(\mathcal{I}_n, y_n)\}_{n=1}^{N^{m_2}}$ represent the modality-incomplete subsets. The objective of this task is to improve the performance of models in the scenarios where modality-missing occurs in both training and inference phases.

3.2 Missing Self-Adaptive Retriever

In the context of incomplete modality learning, the target instance may be either modality-complete or missing certain modalities (e.g., missing text or image). To ensure effective retrieval in both cases, we propose a Missing Self-Adaptive Retriever (MSA Retriever), which dynamically switches its retrieval mechanism to fully utilize the currently available modalities to perform effective retrieval.

3.2.1 Missing Self-Adaptive Retrieval. For a modality-complete instance S, we first encode its text $\mathcal{T}$ and image $\mathcal{I}$ to query vectors through pre-trained text and vision encoders. Specifically, we employ the CLIP [31] text encoder to encode the $\mathcal{T}$, yielding the text query vector $\Psi_t(\mathcal{T}) \in \mathbb{R}^{d_t}$, where d_t is the dimension of text modality. Similarly, we leverage the CLIP vision encoder to process the image modality and generate the image query vector $\Psi_v(\mathcal{I}) \in \mathbb{R}^{d_v}$. Subsequently, we combine both modalities from the instance S to perform a joint within-modal retrieval through our MSA Retriever.

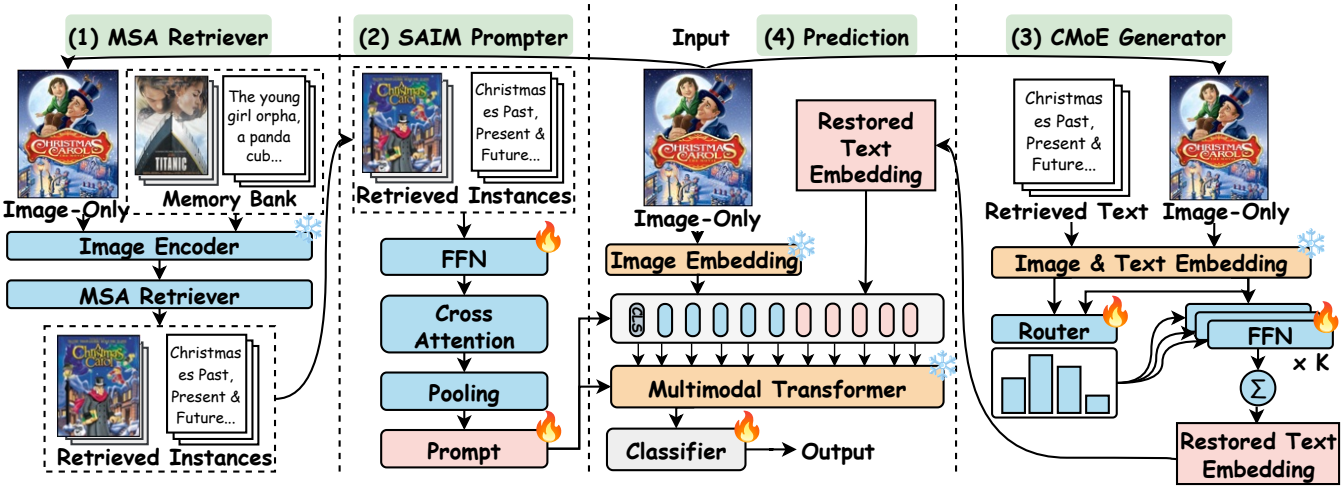


Figure 2: Overall framework of REDEEM. (1) The MSA Retriever identifies similar samples for the target image-only instance; (2) The CMoE Generator recovers the missing text through a within-modal conditional reconstruction; (3) The SAIM Prompter utilizes the retrieved instances to generate the sample-aware inter-modal prompts; (4) The recovered text and prompts are fed into the pre-trained MT with the remaining image modality to predict the result.

This process can be written as:

$$\mathcal{S}_r = \underset{b \in \mathcal{M}}{\text{Top-}K}\left(\frac{\Psi_t(\mathcal{T})^\top \Psi_t(\mathcal{T}_b)}{\|\Psi_t(\mathcal{T})\|\|\Psi_t(\mathcal{T}_b)\|} + \frac{\Psi_v(\mathcal{I})^\top \Psi_v(\mathcal{I}_b)}{\|\Psi_v(\mathcal{I})\|\|\Psi_v(\mathcal{I}_b)\|}\right), \quad (1)$$

where $\mathcal{S}_r = \{S_{r_1}, S_{r_2}, \cdots, S_{r_K}\}$ denotes the top-K retrieved modality-complete instances, $\mathcal{M}$ is the memory bank which stores (image, text) tuples. For modality-incomplete target instances, MSA Retriever can flexibly switch the retrieval mechanism to adapt any missing case by only employing the remaining available modality as query to perform the within-modal retrieval.

The features from the top-K retrieved instances $\mathcal{S}_r$ provide additional contextual information, guiding the recovery of the missing content for the target modality-incomplete instance and the generation of the sample-aware dynamic prompts.

3.3 Conditional Mixture of Experts Generator

Prior MT-based methods and modality invariant learning approaches simply leverage dummy values (e.g., zeros) to pad the missing content, which overlooks recovering the missing content and thus introduces unpredictable noise. In contrast, we introduce a Conditional Mixture of Experts Generator (CMoE Generator), which first conditions on the remaining modalities and weightily aggregates the within-modal experts (i.e., retrieved instances with the same modality as the missing one) to reconstruct the missing content. This generator realizes a within-modal reconstruction that effectively resolves the heterogeneity issues in cross-modal imputation.

3.3.1 Conditional Soft Router Network. In the CMoE Generator, we first define within-modal experts as data from the retrieved instances that match the same modality as the missing one. Conditioning on the remaining modality in the target instance, we design a Conditional Soft Router (CSR) network to facilitate interactions between the remaining modality and these experts, assigning each expert a role in the reconstruction process by generating routing scores based on their assessed contribution. This conditional routing enables dynamic prioritization of experts, ensuring that those most aligned with the reconstruction objectives are emphasized, while the weights of less relevant experts are accordingly reduced.

Specifically, for an image-only target instance S, we aim to recover its text modality using the remaining image modality $\mathcal{I}$ and a set of retrieved text modality data (i.e., within-modal experts) $\mathcal{T}_r = \{\mathcal{T}_{r_1}, \mathcal{T}_{r_2}, \ldots, \mathcal{T}_{r_K}\}$. The CMoE Generator first applies the frozen embedding layers from the MT to convert both the image and retrieved text data into embeddings, resulting in the image embedding $\mathbf{E}^i \in \mathbb{R}^{n \times d}$ and retrieved text embeddings $\mathbf{E}_r^t = \{\mathbf{E}_{r_i}^t\}_{i=1}^K \in \mathbb{R}^{K \times m \times d}$, where n and m are the number of image patches and word tokens, respectively. Subsequently, a pooling strategy is applied to these embeddings to reduce the sequence length, yielding $\tilde{\mathbf{E}}^i \in \mathbb{R}^d$ and $\tilde{\mathbf{E}}_r^t = \{\tilde{\mathbf{E}}_{r_i}^t\}_{i=1}^K \in \mathbb{R}^{K \times d}$.

To compute the routing scores, we design the CSR network to utilize the remaining modality to interact with each within-modal expert, generating the weights that determine each expert's contribution in completing the missing modality:

$$\text{CSR}\left(\tilde{\mathbf{E}}^i, \tilde{\mathbf{E}}_{r_k}^t\right) = \frac{1}{\sqrt{d}}\left(\tilde{\mathbf{E}}^i \mathbf{W}_1\right)\left(\tilde{\mathbf{E}}_{r_k}^t \mathbf{W}_2\right)^\top, \quad (2)$$

where $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable projection matrices, and d is the embedding dimensionality. Finally, the softmax operation is adopted to generate the normalized routing scores:

$$w_k = \frac{\exp\left(\text{CSR}\left(\tilde{\mathbf{E}}^i, \tilde{\mathbf{E}}_{r_k}^t\right)\right)}{\sum_{j=1}^K \exp\left(\text{CSR}\left(\tilde{\mathbf{E}}^i, \tilde{\mathbf{E}}_{r_j}^t\right)\right)}, \quad (3)$$

where w_k is the routing score assigned to the k-th expert, reflecting its relative importance in recovering the missing text modality for instance S.

3.3.2 Within-Modal Conditional Reconstruction. With the routing scores $\boldsymbol{w} = [w_1, w_2, \ldots, w_K]$, the CMoE Generator first exerts a

set of Feed-Forward Networks (FFNs) to each expert's embedding, and aggregates the expert-specific outputs by the routing scores to generate the missing text modality:

$$\hat{\mathbf{E}}^t = \sum_{k=1}^{K} w_k \cdot \mathrm{FFN}_k(\mathbf{E}^t_{r_k}), \tag{4}$$

where $\hat{\mathbf{E}}^t \in \mathbb{R}^{m\times d}$ denotes the reconstructed embedding of the missing text modality for target instance S. The way to recover the image modality follows the same manner. This within-modal conditional reconstruction strategy offers a semantics- and distribution-aligned recovery of the missing content, providing more robust multimodal inputs for pre-trained MTs.

3.4 Sample-Aware Inter-Modal Prompter

Prior MT-based methods [12, 18] often rely on static prompts to guide the MT in addressing modality-missing challenges. However, these prompts are sample-agnostic and not optimal for every sample, especially when handling datasets with the mixture of modality-complete and -incomplete instances. To address this, we propose a Sample-Aware Inter-Modal Prompter (SAIM Prompter). By extracting inter-modal informative patterns from modality-complete instances similar to the target, our prompter generates sample-specific dynamic prompts that aid the MT in understanding text-image correspondences present in these samples.

Given an instance S, the SAIM Prompter generates both text and image prompts by extracting cross-modal relationships from retrieved modality-complete instances similar to S. Specifically, for text prompts, the SAIM Prompter applies cross-modal attention [40] by utilizing the retrieved image embeddings $\mathbf{E}^i_r$ as query and the retrieved text embeddings $\mathbf{E}^t_r$ as key and value:

$$\mathbf{P}^t = \mathrm{Pool}(\mathrm{CrossAttn}(f_Q(\mathbf{E}^i_r), f_K(\mathbf{E}^t_r), f_V(\mathbf{E}^t_r))), \tag{5}$$

$$\mathrm{CrossAttn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}, \tag{6}$$

where Pool is the adaptive pooling operation, f_Q, f_K, and f_V denote the linear mapping functions, $\mathbf{P}^t \in \mathbb{R}^{b\times d}$ represents the text prompt, and b is the prompt length. Similarly, the image prompt $\mathbf{P}^i \in \mathbb{R}^{b\times d}$ can be obtained in the same manner. These prompts act as few-shot examples, guiding the MT to understand the cross-modal relationship from the target-relevant complete data, enhancing the robustness of MT in tackling incomplete modality problems.

3.5 Prediction

The model prediction process unfolds as follows. For an image-only instance S, we first retrieve its top-K most relevant modality-complete instances $\mathcal{S}_r$ by the MSA Retriever, recovering its text embedding $\hat{\mathbf{E}}^t$ via the CMoE Generator and feeding both image and reconstructed text embeddings into MT. Subsequently, the prompts $\mathbf{P}^i$ and $\mathbf{P}^t$ are generated through the SAIM Prompter and inserted into the l-th layer of MT for prompt-tuning:

$$\mathbf{H}_l = [\mathbf{P}^i, \mathbf{P}^t, \mathbf{H}^i_l, \mathbf{H}^t_l], \tag{7}$$

where $l \in [0, L]$, $\mathbf{H}_l$ represents the input features to the l-th layer of the MT, $\mathbf{H}^i_l$ and $\mathbf{H}^t_l$ denote the hidden representations of image and text modalities input to l-th layer, with $\mathbf{H}^i_0$ and $\mathbf{H}^t_0$ initialized as $\mathbf{E}^i$ and $\tilde{\mathbf{E}}^t$, respectively.

The features from the last layer of the MT are fed into a pooler layer, followed by a classification layer, yielding the final output $\hat{y}$. For text-only or modality-complete instances, the generation of classification results follows a similar process. During training, all MT parameters remain frozen, and only the CMoE Generator and SAIM Prompter, which comprise a relatively small number of parameters, are updated. The framework is optimized using a combination of Cross-Entropy loss and reconstruction loss (i.e., mean square error loss).

4 Experiments

4.1 Experimental Setup

A concise summary of the experimental settings is outlined below, with detailed descriptions regarding datasets, baselines, and implementation available in the Appendix A.

Datasets. Following prior works [12, 18], we evaluate our REDEEM on three diverse multimodal downstream datasets: (1) MM-IMDb [2], a movie genre classification dataset involving both image and text modalities. Given that each movie may belong to multiple genres, the task is a multi-label classification. (2) HateMemes [15], a challenging multimodal dataset designed to identify hate within memes using image and text modalities. (3) Food101 [38], a food classification dataset containing noisy image-text pairs with 101 categories, sourced from Google Image Search. Detailed dataset statistics are provided in Table 2. The splits for each dataset are consistent with the original paper.

Baselines. We compare our REDEEM with 11 competitive baselines, which are roughly grouped into three categories: (1) *Modality invariant learning methods*: IF-MMIN [50], ShaSpec [36], DrFuse [43], CorrKD [21], and MoMKE [42]. (2) *Cross-modal imputation methods*: SMIL [26], TFR-Net [45], and AcMAE [41].(3) *MT-based methods*: MAPs [18], MSPs [12], RAGPT [17].

Metrics. Following prior works [12, 18], we adopt appropriate metrics for each dataset. For MM-IMDb, we use F1-Micro (F1-M) and F1-Samples (F1-S) to assess multi-label classification performance. For Hateful Memes, we utilize the Area Under the Receiver Operating Characteristic Curve (AUROC) as the metric. For Food101, we employ classification accuracy (ACC).

Setting of Missing Modality. Following prior work [17], we assume training set is fully available and define the missing rate η% as the rate of modality-incomplete samples in the test set: (1) text/image missing with η% indicates that there are η% image-only/text-only instances and (1-η%) modality-complete instances. (2) both modalities missing with η% indicates that there are $\frac{\eta}{2}$% text-only instances, $\frac{\eta}{2}$% image-only instances and (1-η%) modality-complete instances. We set missing rate η% = 70% by default. For training of each model, we simulate the same 70% missing rate to align model optimization well with test conditions, but allow each model to access the full modality information in the training set.

Implementation Details. In this study, following prior MT-based works [12, 17, 18, 25], we employ the pre-trained ViLT [16] as our MT backbone. However, REDEEM is a model-agnostic framework that can be seamlessly adapted to a wide range of MT architectures, and we also evaluate the effectiveness of REDEEM on additional MT

Table 1: Performance comparison on three datasets with a 70% missing rate across various missing-modality scenarios. The best results are in red bold and the second black bold. Higher values of F1-M, F1-S, AUROC, and ACC indicate better performance.

	MM-IMDb						HateMemes			Food101		
Missing Type	**Text**		**Image**		**Both**		**Text**	**Image**	**Both**	**Text**	**Image**	**Both**
Methods	F1-M	F1-S	F1-M	F1-S	F1-M	F1-S	AUROC	AUROC	AUROC	ACC	ACC	ACC
IF-MMIN	39.63	38.10	31.95	26.89	31.98	29.33	57.62	53.44	55.19	66.76	64.36	68.53
ShaSpec	44.04	42.05	44.23	42.53	44.06	42.13	58.75	60.30	60.96	60.99	74.87	70.02
DrFuse	47.05	45.22	43.58	42.19	48.83	47.15	57.60	60.66	55.84	66.30	75.09	68.23
CorrKD	44.82	45.27	39.48	39.11	41.20	40.51	58.74	55.59	57.91	61.37	66.83	62.87
MoMKE	50.98	50.06	45.67	44.28	46.99	45.30	63.08	61.35	62.53	66.85	68.40	67.38
SMIL	38.32	38.55	27.57	35.27	35.12	31.87	50.32	58.50	54.63	61.83	58.86	60.77
TFR-Net	37.70	38.82	38.14	39.45	37.24	38.11	51.18	55.57	52.12	65.91	67.58	63.41
AcMAE	47.47	46.73	43.82	42.20	44.05	43.75	55.74	59.66	57.25	69.28	73.75	71.15
MAPs	46.12	45.47	44.86	43.19	45.48	44.30	58.62	60.16	58.89	67.02	75.62	72.52
MSPs	49.16	48.81	44.62	43.06	48.28	46.71	59.60	60.05	59.08	71.74	79.09	74.46
RAGPT	**55.16**	**55.00**	**46.44**	**45.12**	**50.89**	**50.22**	**64.10**	**62.57**	**63.47**	**75.53**	**81.98**	**76.94**
REDEEM	**59.94**	**58.50**	**49.54**	**49.51**	**51.49**	**50.93**	**71.45**	**64.30**	**67.02**	**79.81**	**83.71**	**78.65**
Improv. (%)	8.67↑	6.36↑	6.68↑	9.73↑	1.18↑	1.41↑	11.47↑	2.76↑	5.59↑	5.67↑	2.11↑	2.22↑
p-val.	$3.86e^{-6}$	$2.02e^{-6}$	$2.29e^{-5}$	$4.81e^{-5}$	$4.49e^{-4}$	$1.82e^{-5}$	$4.98e^{-7}$	$1.89e^{-5}$	$9.17e^{-6}$	$1.05e^{-7}$	$1.09e^{-8}$	$6.81e^{-6}$

Table 2: Statistics of three multimodal downstream datasets.

Dataset	# Image	# Text	# Train	# Val	# Test
MM-IMDb	25,959	25,959	15,552	2,608	7,799
HateMemes	10,000	10,000	8,500	500	1,500
Food101	90,688	90,688	67,972	-	22,716

backbones in Section 4.9. To avoid heavy overhead, all parameters in the ViLT remain frozen, and only the proposed CMoE Generator, SAIM Prompter, and downstream task-specific parameters (e.g., pooler and classifier) are trained. For pair comparison, the memory bank $\mathcal{M}$ for each task is constructed with the corresponding training and validation set without test data leakage. The number of retrieved instances K is set to 5, the sample-aware inter-modal prompt length b is set to 2, and the insertion position l is set to 1. We use the AdamW [24] optimizer with a learning rate of 1×10^{-3} and weight decay of 5×10^{-5}. All experiments are conducted on a system equipped with an Intel(R) Core(TM) i7-13700KF processor, a RTX 3090 GPU, and 128 GB of system RAM.

4.2 Overall Performance

To verify the superiority of our proposed framework REDEEM in tackling modality-missing problems, we compare it with 11 competitive baseline models across three datasets under a 70% missing rate and the results are summarized in Table 1. From the results, we have the following observations.

First, our proposed **REDEEM** consistently outperforms all the competitive baseline models across three datasets under various modality-missing scenarios. Notably, REDEEM achieves an average improvement of 9.68% on all three datasets across all metrics and missing cases. To further validate REDEEM's superiority, we compute the statistical differences between REDEEM and the best-performing baseline by retraining both models five times. The resulting p-values, all far below 0.05, confirm that REDEEM's improvement over the baselines is statistically significant. These performance gains are attributed to the effectiveness of REDEEM's novel framework, which pioneers a retrieval-guided paradigm for both missing modality recovery and prompt generation. The MSA Retriever flexibly utilizes the current available modalities to search for the most relevant instances for the target sample. The CMoE Generator and the SAIM Prompter carefully leverage the retrieved instances to recover the missing modality and generate the sample-aware inter-modal prompts, significantly enhancing the MT's robustness in tackling various modality-missing challenges.

Second, both **cross-modal imputation** and **modality invariant learning baselines** demonstrate a certain degree of capacity in handling missing modalities. However, they consistently fall short of the performance achieved by our proposed REDEEM framework due to their inherent limitations. Specifically, modality invariant learning methods employ dummy values to substitute missing data, which introduces noise and results in unstable model performance. Cross-modal imputation methods, on the other hand, directly generate the missing modality from available ones but fail to address the fundamental modality gap in the reconstruction. In contrast, REDEEM effectively addresses these limitations by introducing the CMoE Generator, which recovers the missing modality via a within-modal reconstruction, achieving superior performance across diverse modality-missing scenarios.

Third, within **MT-based baselines**, each approach demonstrates advanced performance over the vanilla pre-trained MT, i.e., ViLT, underscoring the effectiveness of designs in strengthening pre-trained MTs against modality-missing challenges. For instance, MAPs and MSPs adopt prompt-tuning to improve the ViLT's robustness under missing situations. However, these methods remain less effective than our REDEEM due to inherent drawbacks: (1) They overlook the missing content recovery, similar to joint methods that rely on masking values padding strategy. (2) Their sample-agnostic, static prompts lack fine-grained guidance for individual instances, especially when handling datasets that contain a mix of modality-complete and -incomplete samples. Although RAGPT

Table 3: Ablation study of the core components within REDEEM under 70% text missing.

Module	Variant	MM-IMDb F1-M	HateMemes AUROC	Food101 ACC
REDEEM	**All**	**59.94**	**71.45**	**79.81**
MSA Retriever	Random	53.31	61.37	67.31
	w/o Retriever	47.56	58.87	63.78
CMoE Generator	Cross-modal	54.40	69.19	77.92
	w/o Router	54.01	68.87	76.97
	w/o Generator	52.64	61.37	72.87
SAIM Prompter	Static Prompt	55.26	70.15	78.15
	Intra-modal	54.62	69.49	77.55
	w/o Prompter	54.18	69.02	76.62

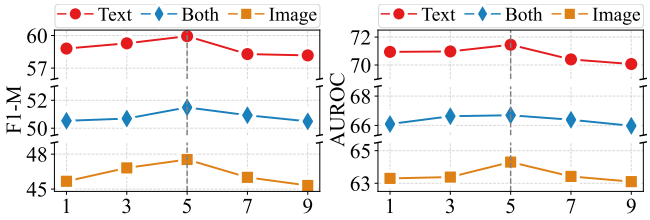
makes an initial attempt to address the above limitations, its design remains simplistic: it reconstructs missing modalities by merely averaging the retrieved ones, which limits its ability to capture modality-specific nuances. Furthermore, its reliance on intra-modal prompts to guide the MTs, neglecting the crucial role of cross-modal relationships under severe modality-missing conditions. Instead, REDEEM proposes a MoE-based conditional generation paradigm, which explicitly leverages the remaining modalities as conditions to guide the fine-grained aggregation of the retrieved modality representations. Furthermore, REDEEM designs the SAIM Prompter that distills informative cross-modal patterns from the retrieved instances, generating more informative dynamic prompts. These prompts serve as few shots, which guide the MTs in understanding the correspondence between text and image modalities, largely enhancing the missing robustness of MTs.

4.3 Ablation Study

To investigate the contributions of REDEEM's core components, we conduct a comprehensive ablation study across all datasets. The results under text missing case are reported at Table 3.

4.3.1 Effect of MSA Retriever. To validate the efficacy of the MSA Retriever, we design two variant models: (1) **Random**: where instances are randomly selected to replace the retrieved samples, and (2) **w/o Retriever**: where the retriever is removed, and features of retrieved instances are replaced with random values. In both variants, the retrieved samples lack relevance to the target sample, leading to a significant drop in performance. This decline highlights the critical role of retrieval quality in enhancing the REDEEM 's overall effectiveness within this retrieval-guided framework.

4.3.2 Effect of CMoE Generator. To evaluate the efficacy of the CMoE Generator, we design three variant models: (1) **Cross-modal**: where the CMoE Generator is replaced with an encoder-decoder based cross-modal generator, (2) **w/o Router**: where the Conditional Soft Router (CSR) network is removed and the retrieved data are simply averaged to impute the missing content, and (3) **w/o Generator**: where the missing modalities are padded with masking values. For the Cross-modal variant, we observe a decline in performance, attributed to the modal heterogeneity issues during reconstruction. Additionally, without the router's weighting, irrelevant or low-quality retrieved data may contribute noise, leading to inaccuracies in the reconstructed modality and inferior performance. Finally, in the w/o Generator variant, padding missing content with masking values results in a substantial performance drop, reflecting the detrimental effect of masking noise on pre-trained MTs.



(a) K on MM-IMDb **(b) K on HateMemes**

Figure 3: Sensitivity analysis of hyper-parameters K under text missing, image missing and both modalities missing scenarios on the MM-IMDb and HateMemes datasets.

4.3.3 Effect of SAIM Prompter. To assess the efficacy of the SAIM Prompter, we design three variant models: (1) **Static Prompt**: where the static prompts are inserted into the MT, (2) **Intra-modal**: where the prompts are generated by intra-modal attention between the retrieved modalities and the same modalities from the target instances, and (3) **w/o Prompter**: where the SAIM Prompter is removed. Entirely removing the SAIM Prompter yields a substantial performance drop, as MT lacks the necessary guidance to effectively manage modality-missing scenarios without any prompt cues. By contrast, the static and intra-modal prompt variants demonstrate improved performance but still encounters a noticeable decline. These two type of prompts provide suboptimal guidance, constraining MT's capacity to handle severe modality-missing cases.

4.4 Hyper-Parameter Analysis

In this section, we perform a sensitivity analysis on the number of retrieved instances K on the MM-IMDb and HateMemes datasets, and the results are shown in Figure 3. From the results, we obtain the conclusion: Adding retrieved instances improves REDEEM 's performance. However, a large number of instances results in a decline in performance due to the noise (i.e., the irrelevant samples). The optimal performance is generally achieved with $K = 5$.

4.5 Robustness to Varying Missing Rate

To evaluate the robustness of our proposed REDEEM framework against varying degrees of data loss, we conduct experiments on the HateMemes dataset under conditions of text missing and both modalities missing at different missing rates. We compare REDEEM with another two competitive MT-based baselines: MAPs and MSPs. Figure 4(a) and (b) illustrate the performance of each model across different missing rates, while Figure 4(c) and (d) present the performance degradation relative to the complete modality for each model at each missing rate. From Figure 4(a) and (b), we observe that our proposed REDEEM consistently outperforms the baselines, MAPs, and MSPs, achieving the highest AUROC scores under all missing rates in the text missing and both modalities missing scenarios. Additionally, as shown in Figure 4(c) and (d), REDEEM exhibits the smallest relative performance degradation compared to the complete modality condition at each evaluated missing rate, underscoring its superior modality-missing robustness. These results confirm that REDEEM effectively utilizes contextual knowledge

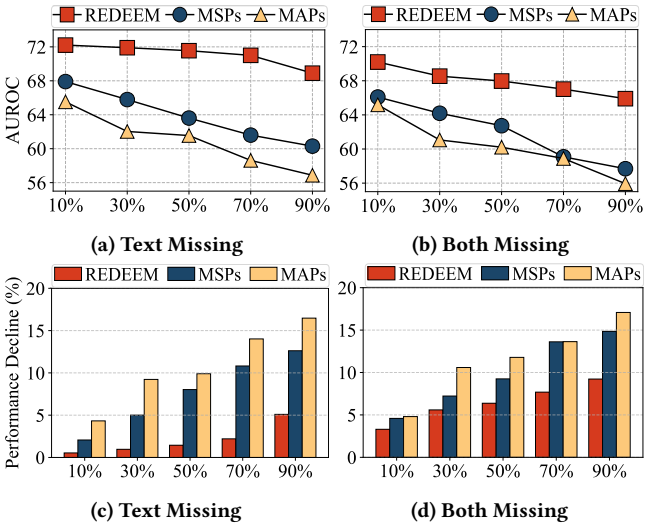


Figure 4: Robustness of REDEEM and baselines MAPs, MSPs on the HateMemes dataset across various missing rate.

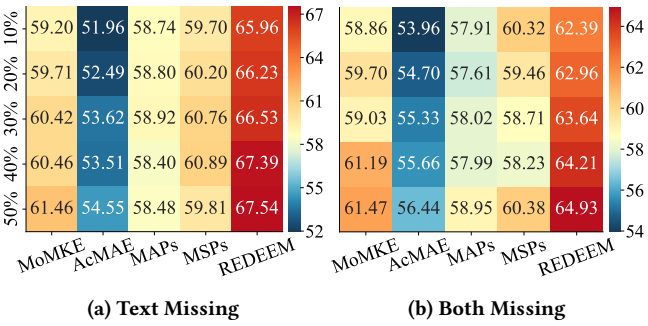


Figure 5: Generalization analysis of REDEEM and baselines on the HateMemes dataset under various training missing rates and a 90% inference missing rate, in terms of AUROC.

from retrieved instances to mitigate the impact of missing content, ensuring robust performance across varying degrees of missing.

4.6 Generalizability to Severe Missing Challenge

To analyze the generalizability of REDEEM, we conduct experiments with varying missing rates in the training set (i.e., 10%, 20%, 30%, 40% and 50%) and evaluate their performance under a 90% missing rate test set. These experiments are performed on the HateMemes dataset, comparing REDEEM against four competitive baselines: MoMKE, AcMAE, MAPs and MSPs. The results for the text missing are presented in Figure 5(a), while results for the both modalities missing are shown in Figure 5(b).

From the results, we observe that as the missing rate in the training set increases, all models exhibit improved performance in the test set. This trend indicates that a high rate of missing data during training enhances the models' ability to handle severe modality-missing cases at inference phase. Notably, the MT-based models (MAPs, MSPs, and REDEEM) demonstrate minimal performance variation due to their reliance on a pre-trained, frozen MT backbone with limited additional trainable parameters. This design reduces

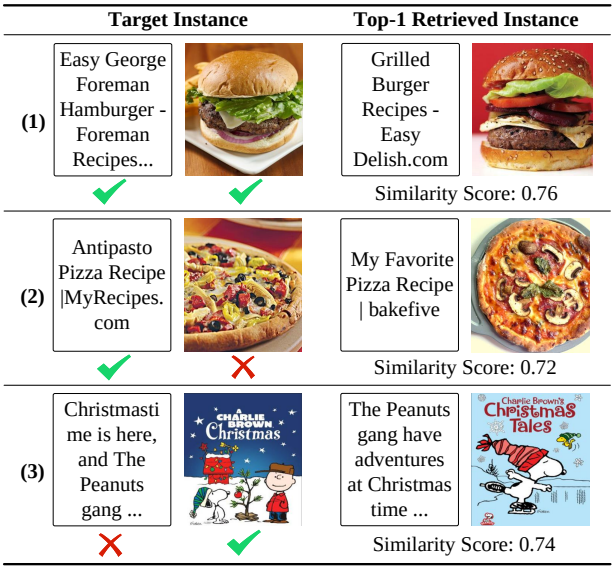


Figure 6: Presentation of retrieval results. The first target instance is modality-complete, while the second and third are text-only and image-only instances, respectively.

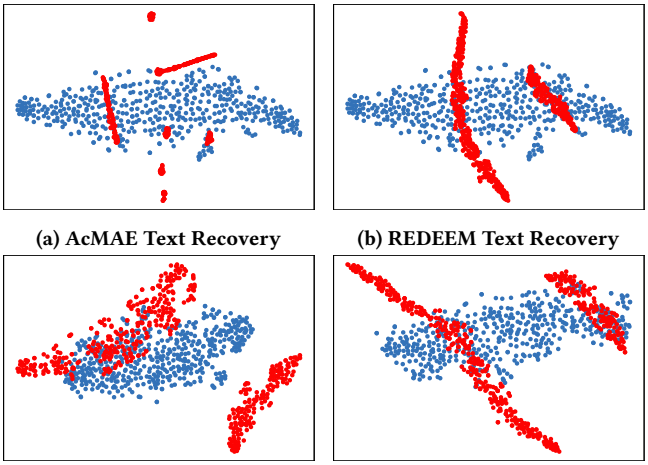
their sensitivity to training data variability, ensuring consistent performance. Moreover, our REDEEM achieves the best missing generalizability, attributed to its retrieval-guided paradigm which incorporates expressive information from retrieved instances, significantly enhancing the REDEEM's ability to effectively generalize on severe modality-missing scenarios.

4.7 Retrieval Quality Presentation

To evaluate the effectiveness of our proposed MSA Retriever, we randomly select three target instances from the MM-IMDb and Food101 datasets, along with their Top-1 retrieved samples. The selected targets include one modality-complete instance and two modality-incomplete instances. As visualized in Figure 6, the retrieved instances exhibit a strong semantic correlation with their respective targets. These results highlight the high relevance of the retrieved content, underscoring the MSA Retriever's capability to flexibly adapt its retrieval mechanism to the specific modality-missing scenarios and identify the most relevant information.

4.8 T-SNE Visualization

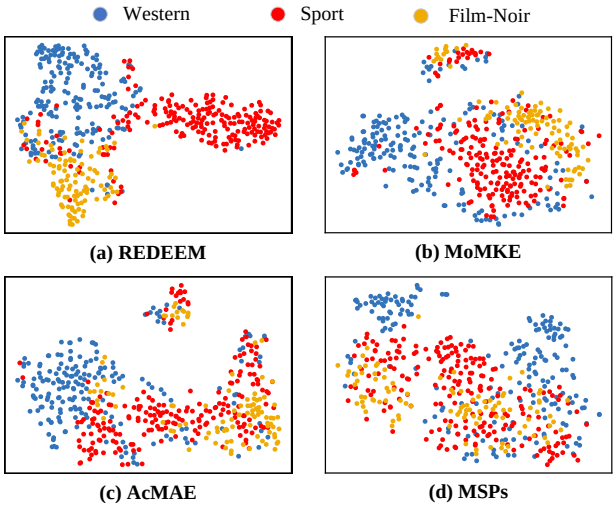
4.8.1 Modality Recovery Visualization. Figure 7 illustrates the distribution of recovered text and image data and ground truth for both cross-modal imputation baseline AcMAE and our proposed REDEEM. To generate this visualization, we randomly select 500 samples from the test set of the HateMemes dataset and project the features of these samples into a 2D space using t-SNE [35]. As shown in the Figure 7, the distribution of original and recovered modalities imputed by REDEEM aligns closer compared to AcMAE. This demonstrates that the Conditional MoE Generator (CMoE Generator) within our proposed REDEEM effectively reconstructs missing content while resolving the limitations (i.e., modal heterogeneity [37]) inherent in the cross-modal generation process.



(a) AcMAE Text Recovery (b) REDEEM Text Recovery

(c) AcMAE Image Recovery (d) REDEEM Image Recovery

Figure 7: Text and image data recovery visualization of AcMAE and our REDEEM. Blue points indicate the ground truth while red points represent the recovered data.

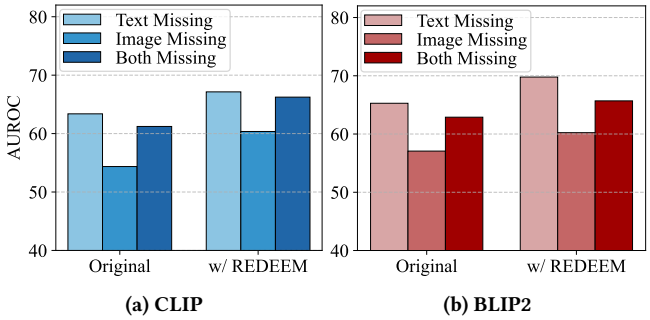


(a) REDEEM (b) MoMKE

(c) AcMAE (d) MSPs

Figure 8: T-SNE visualization of classification for REDEEM and baselines MoMKE, AcMAE, and MSPs on the MM-IMDb dataset under a 90% text missing.

4.8.2 Classification Visualization. Figure 8 shows the t-SNE visualization of the embedding distributions of REDEEM and several competitive baselines for three movie genres (Sport, Film-Noir, and Western) from the MM-IMDb test set under a severe 90% text missing situation. These embeddings correspond to the output of the final layer before classification. The baselines MoMKE, AcMAE, and MSPs exhibit limited separation with features intertwined across different labels. In contrast, the embeddings learned by our REDEEM form more distinct clusters, demonstrating clearer and more discriminative boundaries. This result highlights the effectiveness of REDEEM in tackling severe modality-missing problems.



(a) CLIP (b) BLIP2

Figure 9: Model scalability analysis on the HateMemes dataset under various missing scenarios.

4.9 Model Scalability

To evaluate the scalability ability of our proposed model-agnostic framework REDEEM, we also conducted experiments with additional popular multimodal transformer backbones. Specifically, we utilized CLIP [31] and BLIP2 [20] as backbones on the HateMemes dataset, considering three scenarios where 90% of the text modality, image modality and both modalities are missing. In these experiments, we firstly leverage retrieved instances to recover missing content via the CMoE Generator and generates contextual prompts through the SAIM Prompter to guide the backbones in addressing challenges associated with missing modalities. Finally, we utilize the image and text features from the last hidden layer of each backbone and feed these features into a MLP-based predictor to generate the classification result. The performance of the original backbones and the improvements achieved by integrating the core components of REDEEM are presented in Figure 9. The results demonstrate that all backbones benefit significantly from our framework, with substantial performance improvements observed in both missing modality scenarios. These findings highlight the scalability of REDEEM across different MT backbones.

5 Conclusion

In this study, we introduced REDEEM, a novel retrieval-guided conditional generative framework to address the challenge of modality missing. REDEEM dynamically selects the most relevant instances for both modality-complete and -incomplete targets to ensure robust retrieval regardless of the modality completeness of the target instance. Leveraging the available modalities alongside the retrieved instances, REDEEM effectively reconstructs the missing content through within-modal reconstruction techniques. By extracting informative cross-modal patterns from the retrieved modality-complete instances, REDEEM generates sample-specific, dynamic prompts to guide pre-trained MTs in handling severe missing modality scenarios with greater precision. Extensive experiments on three benchmarks demonstrated the superior performance of our framework compared to existing methods. In the future, we aim to apply REDEEM in important practical scenarios (e.g., autonomous driving) to further validate its applicability.

6 Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No.62176043, No.62072077, and No.U22A2097).

References

[1] Gustavo Aguilar, Viktor Rozgic, Weiran Wang, and Chao Wang. 2019. Multimodal and Multi-view Models for Emotion Recognition. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 991–1002.

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

[3] Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. 2018. Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 1158–1166.

[4] Tong Chen, Danny Wang, Xurong Liang, Marten Risius, Gianluca Demartini, and Hongzhi Yin. 2024. Hate speech detection with generalizable target-aware fairness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 365–375.

[5] Shizhe Diao, Pengcheng Wang, Yong Lin, Rui Pan, Xiang Liu, and Tong Zhang. 2024. Active prompting with chain-of-thought for large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 1330–1350.

[6] David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. 2013. Learning factored representations in a deep mixture of experts. In *International Conference on Learning Representations (ICLR)*.

[7] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39.

[8] Yan Gao, Tong Xu, and Enhong Chen. 2024. Are Mixture-of-Modality-Experts Transformers Robust to Missing Modality During Training and Inferring?. In *International Conference on Intelligent Information Processing (ICIIP)*. Springer, 157–172.

[9] Zixian Gao, Disen Hu, Xun Jiang, Huimin Lu, Heng Tao Shen, and Xing Xu. [n. d.]. Enhanced Experts with Uncertainty-Aware Routing for Multimodal Sentiment Analysis. In *Proceedings of the ACM International Conference on Multimedia (MM)*.

[10] Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. [n. d.]. Retrieval-Enhanced Contrastive Vision-Text Models. In *The Twelfth International Conference on Learning Representations*.

[11] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation* 3, 1 (1991), 79–87.

[12] Jaehyuk Jang, Yooseung Wang, and Changick Kim. 2024. Towards Robust Multimodal Prompting with Missing Modalities. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 8070–8074.

[13] Congyun Jin, Ming Zhang, Weixiao Ma, Yujiao Li, Yingbo Wang, Yabo Jia, Yuliang Du, Tao Sun, Haowen Wang, Cong Fan, et al. 2024. RJUA-MedDQA: A Multimodal Benchmark for Medical Document Question Answering and Clinical Reasoning. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 5218–5229.

[14] Amin Karimi Monsefi, Payam Karisani, Mengxi Zhou, Stacey Choi, Nathan Doble, Heng Ji, Srinivasan Parthasarathy, and Rajiv Ramnath. 2024. Masked LoGoNet: Fast and Accurate 3D Image Analysis for Medical Domain. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 1348–1359.

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

[16] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning (ICML)*. PMLR, 5583–5594.

[17] Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. 2025. Retrieval-Augmented Dynamic Prompt Tuning for Incomplete Multimodal Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

[18] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. 2023. Multimodal prompting with missing modalities for visual recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14943–14952.

[19] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In *International Conference on Learning Representations (ICLR)*.

[20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*. PMLR, 19730–19742.

[21] Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. 2024. Correlation-Decoupled Knowledge Distillation for Multimodal Sentiment Analysis with Incomplete Modalities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12458–12468.

[22] Rongqing Li, Changsheng Li, Yuhang Li, Hanjie Li, Yi Chen, Ye Yuan, and Guoren Wang. 2024. Itpnet: Towards instantaneous trajectory prediction for autonomous driving. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 1643–1654.

[23] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

[24] Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

[25] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality?. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18177–18186.

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

[27] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. 2022. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems (Neurips)* 35 (2022), 9564–9576.

[28] Keon-Hee Park, Kyungwoo Song, and Gyeong-Moon Park. 2024. Pre-trained Vision and Language Transformers Are Few-Shot Incremental Learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 23881–23890.

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

[30] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 33. 6892–6899.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*. PMLR, 8748–8763.

[32] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2016. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations (ICLR)*.

[33] Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. 2023. Scaling Vision-Language Models with Sparse Mixture of Experts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 11329–11344.

[34] Qian Sun, Le Zhang, Huan Yu, Weijia Zhang, Yu Mei, and Hui Xiong. 2023. Hierarchical reinforcement learning for dynamic autonomous vehicle navigation at intelligent intersections. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 4852–4861.

[35] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 11 (2008).

[36] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. 2023. Multi-modal learning with missing modality via shared-specific feature modelling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15878–15887.

[37] Hao Wang, Shengda Luo, Guosheng Hu, and Jianguo Zhang. 2024. Gradient-Guided Modality Decoupling for Missing-Modality Robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 38. 15483–15491.

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

[39] Yuanzhi Wang, Yong Li, and Zhen Cui. 2023. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems* 36 (2023), 17117–17128.

[40] A Waswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, L Kaiser, and I Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (Neurips)*.

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

[42] Wenxin Xu, Hexin Jiang, et al. 2024. Leveraging Knowledge of Modality Experts for Incomplete Multimodal Learning. In *Proceedings of the ACM International Conference on Multimedia (MM)*.

[43] Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. 2024. DrFuse: Learning Disentangled Representation for Clinical Multi-Modal Fusion with Missing Modality and Modal Inconsistency. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 38. 16416–16424.

[44] Haofei Yu, Zhengyang Qi, Lawrence Jang, Russ Salakhutdinov, Louis-Philippe Morency, and Paul Pu Liang. 2024. MMoE: Enhancing Multimodal Models with Mixtures of Multimodal Interaction Experts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 10006–10030.

[45] Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the ACM International Conference on Multimedia (MM)*. 4400–4407.

[46] Yaohua Zha, Jinpeng Wang, Tao Dai, Bin Chen, Zhi Wang, and Shu-Tao Xia. 2023. Instance-aware dynamic prompt tuning for pre-trained point cloud models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 14161–14170.

[47] Ziqin Zhou, Hai-Ming Xu, Yangyang Shu, and Lingqiao Liu. 2024. Unlocking the Potential of Pre-trained Vision Transformers for Few-Shot Semantic Segmentation through Relationship Descriptors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3817–3827.

[48] Junyou Zhu, Chao Gao, Ze Yin, Xianghua Li, and Jürgen Kurths. 2024. Propagation Structure-Aware Graph Transformer for Robust and Interpretable Fake News Detection. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. 4652–4663.

[49] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, Yang You, Zhaoxiang Zhang, Dawei Zhao, Liang Xiao, Jian Zhao, Jiwen Lu, and Guan Huang. 2024. Is Sora a World Simulator? A Comprehensive Survey on General World Models and Beyond. *arXiv* abs/2405.03520 (2024).

[50] Haolin Zuo, Rui Liu, Jinming Zhao, Guanglai Gao, and Haizhou Li. 2023. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

A Detailed Experimental Settings

In this section, we provide detailed descriptions about the datasets, baselines, and implementation.

A.1 Datasets

Following prior works [12, 18], we evaluate our REDEEM across three diverse multimodal downstream datasets: MM-IMDb [2], HateMemes [15], and Food101 [38]. Below, we provide detailed descriptions of each dataset.

- **MM-IMDb** is a multimodal dataset developed for movie genre classification, incorporating two modalities: images (movie posters) and text (plot summaries). Each movie may belong to multiple genres, making it a multi-label binary classification task.
- **HateMemes** is a benchmark dataset for identifying hate in memes by leveraging both image and text modalities. By introducing challenging samples, termed "benign confounders", it makes unimodal models more likely to fail, while multimodal approaches are more likely to make correct predictions.
- **Food101** is a large-scale multimodal dataset curated for the multi-class classification task of 101 food categories. This dataset uniquely pairs noisy image and text data across a diverse range of food categories. Collected via Google Image Search, the dataset introduces real-world noise and variability, offering both challenges and opportunities for developing robust multimodal models in food recognition.

A.2 Baselines

To evaluate the efficacy of REDEEM, we compare it with 11 competitive baseline models, which can be classified into three distinct groups: (1) Modality invariant learning methods, (2) Cross-modal imputation methods, and (3) MT-based methods. Below, we provide detailed descriptions of each baseline.

(1) Modality invariant learning methods:

- **IF-MMIN** [50] incorporates invariant features into cross-modality imagination, reducing modality gaps and improving the robustness of joint multimodal representations.
- **ShaSpec** [36] tackles the missing modality problem by learning shared and specific features from the available inputs. It further incorporates auxiliary tasks, including distribution alignment and domain classification, to strengthen feature representations.
- **DrFuse** [43] addresses missing modality challenges by disentangling shared and modality-specific features. It then effectively preserves crucial shared information from the available modalities, enhancing robustness in scenarios with incomplete modalities.
- **CorrKD** [21] is a correlation-decoupled knowledge distillation framework designed to enhance the learning of joint representations under incomplete modality situations by refining and transferring cross-sample, cross-category, and cross-target correlations.
- **MoMKE** [42] utilizes a MoE based framework to dynamically integrate unimodal and joint representations via a Soft Router, enabling robust modality representation under modality missing conditions.

(2) Cross-modal imputation methods:

- **SMIL** [26] develops a Bayesian meta-learning based solution to tackle the modality missing problems and utilizes the remaining modalities to reconstruct the missing modalities.
- **TFR-Net** [45] leverages attention-based extractors to capture intra-modal and inter-modal robust representations for generating missing modality features.
- **AcMAE** [41] adopts missing modality predictive coding by randomly dropping modality features and reconstructing them using the remaining features via an autoencoder network.

(3) MT-based methods:

- **MAPs** [18] introduces missing-aware prompts, which are inserted into different layers of multimodal transformers to effectively address missing modality challenges.
- **MSPs** [12] constructs modality-specific prompts to enhance the robustness of pre-trained multimodal transformers under various modality missing scenarios.
- **RAGPT** [17] leverage a multimodal retrieval to simply recover the missing content via average the retrieved modalities. It also constructs intra-modal prompts with retrieved instances to guide the pre-trained multimodal transformers in tackling the missing modalities.

A.3 Implementation Details

A.3.1 Multimodal Transformer Backbone. In our experiments, we follow the prior works to adopt ViLT [16] as our multimodal transformer backbone.

A.3.2 Details of Memory Bank Construction. In this work, the memory bank $\mathcal{M}$ is composed of only the instances from the training sets, thereby preventing data leakage during inference. Specifically, we utilize the pre-trained ViLT to extract embeddings for both text and image modalities from the training and validation instances. These embeddings are then employed to build the memory bank.

A.3.3 Training Configuration. During the training, to avoid heavy overhead, all parameters in the ViLT remain frozen, and only the proposed CMoE Generator, SAIM Prompter, and downstream task-specific parameters (e.g., pooler and classifier) are optimized. The number of retrieved instances K is set to 5, the context-aware dynamic prompt length b is set to 2, and the insertion position l is set to 1. We use the AdamW [24] optimizer with a learning rate of 1×10^{-3} and weight decay of 5×10^{-5}. Both the baseline models and our REDEEM are trained and tested five times, with the average values reported as the final results. For fair comparisons, we strictly follow the parameter configurations specified in the original papers of the baseline methods.

A.3.4 Implementation Environment. All experiments are conducted on a system equipped with an Intel(R) Core(TM) i7-13700KF processor, an NVIDIA GeForce RTX 3090 GPU with 24 GB of VRAM, and 128 GB of system RAM.