PROMISE: Prompt-Attentive Hierarchical Contrastive Learning for Robust Cross-Modal Representation with Missing Modalities

Jiajun Chen[1,*], **Sai Cheng**[1,*], **Yuan Yutao**[1], **YiRui Zhang**[1], **Haitao Yuan**[2], **Peng Peng**[3,†], **Yi Zhong**[4,†]
[1]Beijing University of Posts and Telecommunications
[2]Nanyang Technological University
[3]Tsinghua University
[4] Beijing Institute of Technology
{chenjiajun, chengsai, scsyuanyutao, zhangyirui650}@bupt.edu.cn, haitao.yuan@ntu.edu.sg, 1580259346@qq.com, yi.zhong@bit.edu.cn

Abstract

Multimodal models integrating natural language and visual information have substantially improved generalization of representation models. However, their effectiveness significantly declines in real-world situations where certain modalities are missing or unavailable. This degradation primarily stems from inconsistent representation learning between complete multimodal data and incomplete modality scenarios. Existing approaches typically address missing modalities through relatively simplistic generation methods, yet these approaches fail to adequately preserve cross-modal consistency, leading to suboptimal performance. To overcome this limitation, we propose a novel multimodal framework named **PROMISE**, a **PROM**pting-Attentive H**I**erarchical Contra**S**tive L**E**arning approach designed explicitly for robust cross-modal representation under conditions of missing modalities. Specifically, **PROMISE** innovatively incorporates multimodal prompt learning into a hierarchical contrastive learning framework, equipped with a specially designed prompt-attention mechanism. This mechanism dynamically generates robust and consistent representations for scenarios where particular modalities are absent, thereby effectively bridging the representational gap between complete and incomplete data. Extensive experiments conducted on benchmark datasets, along with comprehensive ablation studies, clearly demonstrate the superior performance of **PROMISE** compared to current state-of-the-art multimodal methods.

Introduction

Multimodal learning (**????**) is vital in intelligent healthcare (**?**), autonomous driving (**???**), and cross-modal retrieval (**??**) by leveraging complementary information from diverse modalities to achieve robust feature representations (**??**). However, multimodal datasets frequently suffer from incomplete information due to device failures, partial data collection, or transmission losses.

The challenge of missing modalities has garnered significant attention, yet existing solutions remain fundamentally limited. Joint representation-based approaches (**???**) attempt to learn shared semantic spaces but often fail to capture modality-specific nuances and struggle with effective cross-modal semantic alignment. Meanwhile, adaptive methods (**?**) dynamically adjust network architectures, but suffer from architectural complexity and training instability, particularly when missing rates become substantial.

In scenarios with high missing rates, multimodal learning faces three key challenges: (1) **information sparsity**, hindering accurate semantic representation inference for missing modalities; (2) ensuring **semantic consistency** between generated features and original complete modalities; and (3) **balancing discriminative power with semantic preservation** in the completed features.

To address these challenges, we propose **PROMISE**, a novel framework combining multimodal prompt learning with hierarchical contrastive learning. PROMISE employs a **Prompt Attention** mechanism to adaptively extract semantic information from available modalities, generating high-quality representations for missing ones. This is complemented by a dual-level contrastive learning strategy: **Fusion-driven Nexus Contrastive Learning (FNCL)** ensures semantic consistency between modalities, while **Cohesion-driven Core Contrastive Learning (CCCL)** enhances discriminative power. This integrated approach enables robust performance even with high missing rates.

Our main contributions are:

- Proposing a novel technical paradigm that combines prompt learning with hierarchical contrastive learning to address robust cross-modal representation learning under high missing rates.
- Designing an innovative Prompt Attention mechanism that effectively extracts semantic information from available modalities to generate high-quality semantically consistent representations for missing ones.
- Developing a dual-level contrastive learning strategy and incorporating it with prompt learning, which simultaneously ensures feature semantic consistency and discriminative power.

[*]Equal contribution.
[†]Corresponding author.

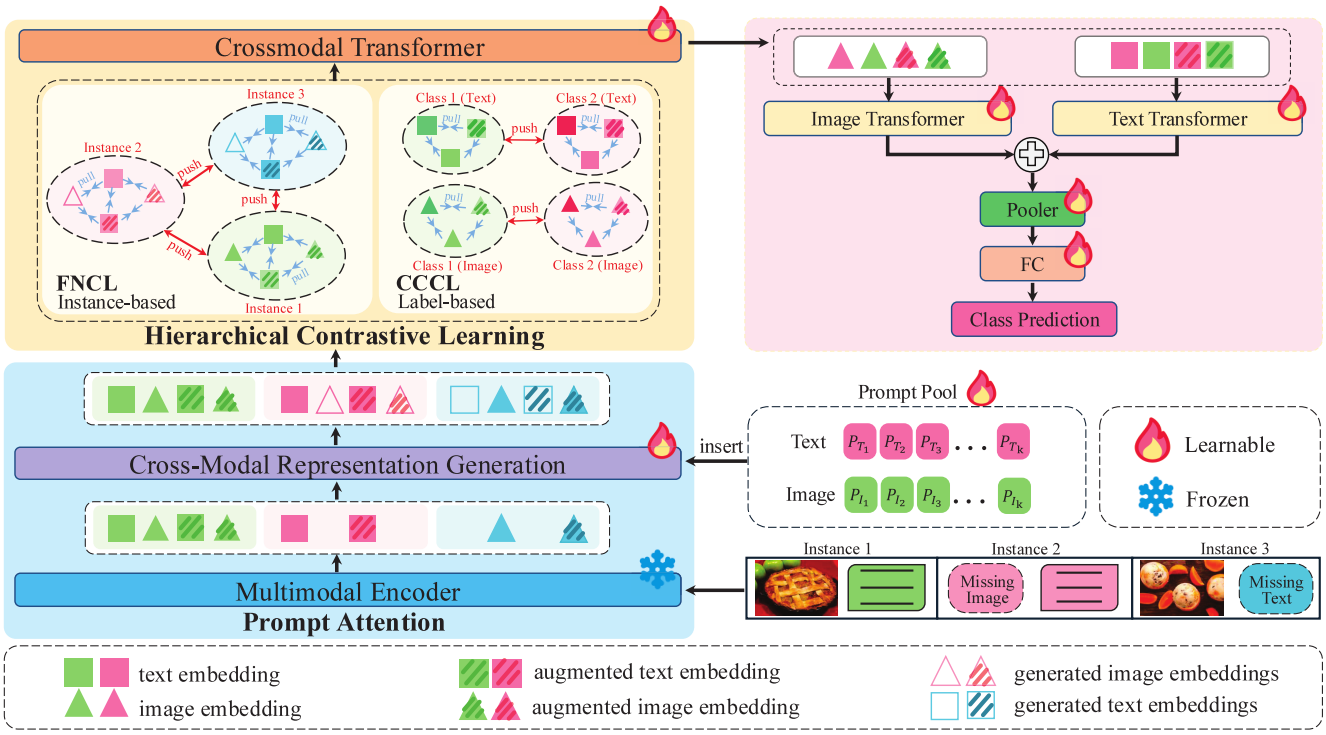


Figure 1: Architectural overview of the PROMISE framework for robust multimodal representation learning with missing modalities. The framework integrates: (1) a frozen multimodal encoder backbone, (2) a prompt-attention mechanism with modality-specific prompt pools for generating missing representations, (3) fusion-driven nexus contrastive learning (FNCL) for cross-modal alignment, and (4) cohesion-driven core contrastive learning (CCCL) for enhancing discriminative power. This hierarchical design effectively bridges the representation gap between complete and incomplete modality scenarios through dynamic prompt-based generation and multi-level contrastive optimization.

Related Work

Multimodal Prompt Learning

Multimodal prompt learning has garnered increasing attention for its ability to efficiently adapt large models to new tasks and domains. Existing approaches can be categorized along two key dimensions.

First, methods differ in prompt modality: early studies in NLP prompt tuning (**??**) focused on single modalities such as natural language, while recent approaches (**?**) jointly exploit image and text inputs.

Second, prompts can be either *hard* or *soft*: *hard* prompt (**?**) directly modifies input features, whereas *soft* prompt (**??**) manipulates internal representations through prefix tokens in transformer architectures.

Prompt-based methods excel in parameter-efficient adaptation and cross-context generalization. Soft multimodal prompt particularly offers flexibility in handling varied input conditions by modulating internal representations without architectural modifications. This motivates our integration of prompt learning into multimodal frameworks, treating distinct missing modality scenarios as separate learning tasks within the broader context of modality incompleteness.

Contrastive Learning

Contrastive learning (CL) is a powerful self-supervised paradigm that aligns similar instances while separating dissimilar ones. Seminal works like SimCLR (**?**) and MoCo (**?**) established foundational frameworks for invariant feature learning. CL was extended to multimodal settings, with CLIP (**?**) achieving remarkable cross-modal alignment by contrasting image-text pairs. However, these methods typically assume full modality availability during both training and inference, limiting their applicability in real-world scenarios with missing modalities.

Recent research has explored prompt-based learning for adapting pretrained models. Multimodal prompt tuning methods, such as MaPLe (**?**) and MPLMM (**?**), enhance task-specific adaptation by refining cross-modal interactions via prompts. While some efforts combine contrastive objectives with prompt techniques (e.g., CP-Tuning (**?**)), they exhibit significant limitations, primarily restricting to single-modal missing scenarios and employing rigid alignment. This creates a critical gap in dynamically aligning partial modality inputs with learnable prompts while ensuring robust feature discrimination under varying missing modality conditions. Our work, PROMISE, bridges this gap by jointly optimizing contrastive objectives and multimodal prompts to address both modality-incomplete and cross-modal align-

ment challenges.

Methodology

In this section, we detail our methodology by presenting a clear problem definition and introducing our proposed PROMISE.

Problem Formulation

We consider a multimodal dataset comprising two modalities, denoted as m_1 and m_2 (e.g., image and text). Let $\mathcal{D} = \{\mathcal{D}^c, \mathcal{D}^{m_1}, \mathcal{D}^{m_2}\}$ represent the complete dataset, where:

- $\mathcal{D}^c = \{(x_i^{m_1}, x_i^{m_2}, L_i)\}_{i=1}^{N_c}$ contains samples with both modalities
- $\mathcal{D}^{m_1} = \{(x_i^{m_1}, \emptyset, L_i)\}_{i=1}^{N_{m_1}}$ comprises samples with only modality m_1
- $\mathcal{D}^{m_2} = \{(\emptyset, x_i^{m_2}, L_i)\}_{i=1}^{N_{m_2}}$ contains samples with only modality m_2

Here, $x_i^{m_1}$ and $x_i^{m_2}$ represent features from respective modalities, $\emptyset$ denotes missing modality, and L_i represents the corresponding label. We define missing rate as $\eta = 1 - \frac{N_c}{N}$ where $N = N_c + N_{m_1} + N_{m_2}$.

For each available modality, we apply data augmentation to obtain augmented representations $x_{a,i}^{m_1}$ and $x_{a,i}^{m_2}$:

$$x_{a,i}^{m_j} = \mathcal{T}_{m_j}(x_i^{m_j}), j \in \{1, 2\} \tag{1}$$

where $\mathcal{T}_{m_j}$ represents modality-specific transformation that preserves semantic content.We will discuss the details in the Implementation Details section.

Overall Framework

PROMISE (Figure 1) employs a hierarchical design with frozen encoders and three learnable components: modality-specific prompt pools with attention mechanisms for missing representation generation, and dual-level contrastive learning (FNCL and CCCL) to maintain both cross-modal alignment and within-modality discriminability. This architecture effectively handles high missing rates by preserving semantic consistency between available and reconstructed modalities.

Prompt Attention

Prior approaches (**??**) to missing modality problems predominantly rely on direct generation techniques without adequately addressing semantic consistency between original and generated representations. These methods often fail to preserve the intrinsic semantic properties of the missing modality, resulting in representations that do not align well with the original feature space.

To address these limitations, we propose the Prompt Attention mechanism, which enables precise cross-modal representation generation by leveraging modality-specific prompt pools in conjunction with multi-head attention. This approach allows for fine-grained semantic alignment between available and missing modalities through learnable prompts that capture modality-specific characteristics.

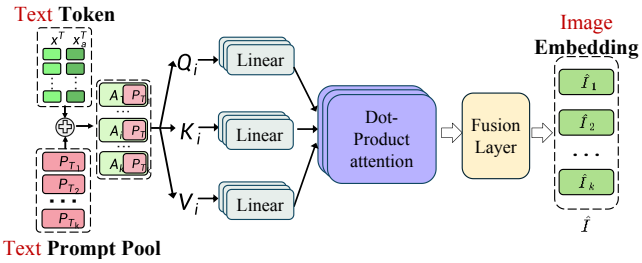


Figure 2: The proposed Prompt Attention mechanism. It generates representations for a missing modality by feeding a multi-head attention module with a concatenation of available modality features and learnable, modality-specific prompts.

Modality-Specific Prompt Pools: We initialize separate prompt pools for each modality to account for their distinct feature distributions. Formally, we define these prompt pools as $\mathcal{P} = \{\mathbf{P}^{m_1}, \mathbf{P}^{m_2}\}$, where $\mathbf{P}^{m_1} = \{p_1^{m_1}, p_2^{m_1}, \ldots, p_N^{m_1}\}$ and $\mathbf{P}^{m_2} = \{p_1^{m_2}, p_2^{m_2}, \ldots, p_N^{m_2}\}$ represent the trainable prompt pools for text and image modalities, respectively. Each prompt pool contains N prompts corresponding to the N attention heads in our multi-head attention mechanism. During training, these prompt pools are optimized end-to-end alongside other parameters through backpropagation, enabling them to adapt to modality-specific characteristics and cross-modal relationships in the dataset.

Cross-Modal Representation Generation: Given an input tuple $\mathcal{D}^{m_2} = \{x^{m_2}, x_a^{m_2}, L\}$ where x^{m_2} and $x_a^{m_2}$ denote the original and augmented representations of modality m_2, and L represents the corresponding labels, our objective is to generate representations $\hat{x}^{m_1}$ and $\hat{x}_a^{m_1}$ for the missing modality m_1. As illustrated in Figure 2, the Prompt Attention mechanism operates as follows:

Given that $[\cdots;\cdots]$ represents the concatenation operation, we concatenate the available modality with its augmented data. Then, we map $[x^{m_2}; x_a^{m_2}]$ to a higher dimension through a function f and concatenate $f([x^{m_2}; x_a^{m_2}])$ with prompt $p_i^{m_1}$.

$$A_i^{m_1} = [f([x^{m_2}; x_a^{m_2}]); p_i^{m_1}], \quad i \in 1, 2, \ldots, N \tag{2}$$

where $f(\cdot)$ indicates the linear projection to mitigate the modality gap between modality m_1 and m_2, and i ranges from 1 to N corresponding to each attention head in our multi-head attention mechanism.

We then derive query, key, and value matrices for the multi-head self-attention mechanism (**?**) as:

$$Q_i = A_i^{m_1} W_i^Q, \quad K_i = A_i^{m_1} W_i^K, \quad V_i = A_i^{m_1} W_i^V \tag{3}$$

where W_i^Q, W_i^K, and W_i^V are learnable parameter matrices for the i-th attention head.

The attention mechanism computes attention scores to capture relevant information for reconstructing the missing modality representations:

$$Attention^i = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i \tag{4}$$

Finally, we generate the representations for the missing modality through separate linear fusion functions for original and augmented representations:

$$\hat{x}^{m_1} = f_{ori}([Attn^1; ...; Attn^N]) \tag{5}$$

$$\hat{x}_a^{m_1} = f_{aug}([Attn^1; ...; Attn^N]) \tag{6}$$

After applying the Prompt Attention mechanism, we construct a comprehensive multimodal dataset $\hat{\mathcal{D}} = \{x^{m_2}, x_a^{m_2}, \hat{x}^{m_1}, \hat{x}_a^{m_1}, L\}$ that integrates both the available modality and the synthetically generated representations for the absent modality.

Despite this generation capability, representation synthesis alone may introduce potential information asymmetry between modalities. To address this challenge, we incorporate *hierarchical contrastive learning* into our framework, which establishes explicit cross-modal semantic alignment at multiple abstraction levels, thereby ensuring consistency between the original and generated representations.

Hierarchical Contrastive Learning

Having addressed missing modalities via the **Prompt Attention** mechanism, we further ensure semantic and modality consistency through **hierarchical contrastive learning**. This strategy consists of two main components: **Fusion-driven Nexus Contrastive Learning (FNCL)** and **Cohesion-driven Core Contrastive Learning (CCCL)**.

Fusion-driven Nexus Contrastive Learning (FNCL): FNCL aims to create a unified representation space, bridging the heterogeneity between modalities. It posits that cross-modal embeddings from the same instance should be semantically consistent, while those from different instances should be distinguishable. We leverage the normalized temperature-scaled cross-entropy (NT-Xent) (**?**) loss for this purpose, where $\text{sim}(u, v) = u^T v/(\|u\|_2\|v\|_2)$ denotes cosine similarity and τ is the temperature parameter.

For each instance i, we define positive pairs as all cross-modal combinations of its original and augmented representations. Given a batch of B instances, the FNCL loss aggregates four alignment objectives:

$$\begin{aligned}\mathcal{L}_{FNCL} =& \mathcal{L}(x^{m_1}, \hat{x}^{m_2}) + \mathcal{L}(x^{m_1}, \hat{x}_a^{m_2}) \\ &+ \mathcal{L}(x_a^{m_1}, \hat{x}^{m_2}) + \mathcal{L}(x_a^{m_1}, \hat{x}_a^{m_2})\end{aligned} \tag{7}$$

where $\mathcal{L}(X, Y)$ is the symmetric NT-Xent loss between sets of features X and Y:

$$\begin{aligned}\mathcal{L}(X, Y) =& \frac{1}{2B}\sum_{i=1}^{B}\Bigg[-\log\frac{\exp(\text{sim}(x_i, y_i)/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(x_i, y_j)/\tau)} \\ &- \log\frac{\exp(\text{sim}(y_i, x_i)/\tau)}{\sum_{j=1}^{B}\exp(\text{sim}(y_i, x_j)/\tau)}\Bigg]\end{aligned} \tag{8}$$

Here, $x_i \in X$ and $y_i \in Y$ are features for instance i. This formulation ensures comprehensive alignment across original and augmented views for both modalities.

Cohesion-driven Core Contrastive Learning (CCCL): While FNCL focuses on inter-modal alignment, CCCL enhances discriminative power within each modality by promoting feature compactness for same-class instances and separation for different-class instances. For each modality $m \in \{m_1, m_2\}$, we define positive pairs based on both augmentation invariance and semantic label information. For an instance i with representation x_i^m and augmented view $x_{a,i}^m$, and label L_i, the positive set for x_i^m includes $x_{a,i}^m$ and all x_j^m where $L_j = L_i$ and $j \neq i$.

The CCCL loss for modality m is defined as:

$$\mathcal{L}_{CCCL}^m = -\frac{1}{2B}\sum_{i=1}^{B}\log\frac{\sum_{j\in P(i)}\exp(\text{sim}(x_i^m, x_j^m)/\tau)}{\sum_{k\in K(i)}\exp(\text{sim}(x_i^m, x_k^m)/\tau)} \tag{9}$$

where $P(i)$ denotes the set of indices of positive samples for x_i^m (including its augmented view and same-label samples in the batch), and $K(i)$ denotes the set of all other samples in the batch (excluding x_i^m itself).

The total CCCL loss combines losses from both modalities:

$$\mathcal{L}_{CCCL} = \mathcal{L}_{CCCL}^{m_1} + \mathcal{L}_{CCCL}^{m_2} \tag{10}$$

Integrated Contrastive Framework:

To leverage the complementary strengths of both contrastive learning components, we combine them into an integrated hierarchical framework with weight α :

$$\mathcal{L}contrast = \alpha\mathcal{L}_{FNCL} + (1-\alpha)\mathcal{L}_{CCCL} \tag{11}$$

Here, α is hyperparameter that determines the relative importance of the Fusion-driven Nexus Contrastive Learning (FNCL) and Cohesion-driven Core Contrastive Learning (CCCL) losses. This allows for flexible optimization and fine-tuning to achieve the best balance between inter-modal semantic consistency and intra-modal discriminative power, resulting in more robust and generalizable representations under missing modality scenarios.

Experiments

Datasets and Evaluation Metrics

To evaluate our methods, we follow the work (**?**) by conducting experiments on the same three multimodal datasets, while introducing a fourth dataset to provide a more comprehensive assessment of our approaches' adaptability and robustness.

- *UPMC Food-101* (**?**) comprises recipe descriptions and images across 101 food categories;
- *MM-IMDb* (**?**) is a multi-label dataset for movie genre classification that integrates visual and textual features;
- *Hateful Memes* (**?**) is a challenging benchmark for detecting hate speech in memes through multimodal analysis;
- *N24News* (**?**) contains news images paired with four text types (Heading, Caption, Abstract, Body).

For consistent evaluation against baselines, we employ dataset-specific metrics: classification accuracy (Acc) for *UPMC Food101* and *N24News*, Area Under the ROC Curve (AUROC) for *Hateful Memes*, and macro-averaged F1 score (F1-Macro) for *MM-IMDb*.

Datasets	Missing rate η	Training		Testing		Ma Model	MPVR	MSPs	MMP	DCP	PROMISE
		Image	Text	Image	Text						
Hateful Memes (AUROC)	70%	100%	30%	100%	30%	60.96	61.01	-	61.39	62.82	**63.63**
		30%	100%	30%	100%	63.56	62.34	-	**68.97**	64.12	64.40
		65%	65%	65%	65%	64.41	63.53	-	65.17	66.08	**67.16**
UPMC Food101 (ACC)	70%	100%	30%	100%	30%	73.41	73.85	71.58	78.81	78.87	**79.97**
		30%	100%	30%	100%	86.38	86.09	85.91	87.11	87.32	**88.10**
		65%	65%	65%	65%	78.58	77.49	78.89	82.67	81.87	**83.22**
MM-IMDb (F1-Macro)	70%	100%	30%	100%	30%	38.65	39.19	38.34	43.24	48.52	**49.62**
		30%	100%	30%	100%	46.63	46.30	47.45	**56.03**	53.14	54.29
		65%	65%	65%	65%	41.28	42.41	42.03	49.24	51.42	**52.19**

Table 1: Quantitative results on the *MM-IMDb*, *UPMC Food-101*, and *Hateful Memes* datasets with missing rate $\eta = 70\%$ under various modality-missing scenarios. **Bold** numbers indicate the best performance.

Baselines

We benchmark against five state-of-the-art multimodal models:

- *Ma Model* (**?**) combines VILT pre-training with multi-task learning and algorithmic modality fusion.
- *MPVR* (**?**) enhances pre-trained VILT with missing-aware prompts in its transformer architecture.
- *MSPs* (**?**) introduces modality-specific prompts and employs an orthogonal loss term.
- *MMP* (**?**) integrates parameter-efficient fine-tuning of unimodal models with self-supervised joint-embedding learning.
- *DCP* (**?**) adapts large pretrained multimodal models for missing modality scenarios by leveraging correlations between prompts and input features, inter-layer prompt relationships, and complementary modality semantics.

Implementation Details

Input & Feature Extraction: We apply standard data augmentation techniques to images, including horizontal flipping, rotation, color adjustment, random cropping, and Gaussian blur. For text augmentation, we utilize GPT-4o (**?**) to generate synonyms. Feature extraction is performed using a frozen CLIP-ViT-Large-Patch14 (**?**) backbone. The Prompt Attention mechanism employs 4 attention heads. Detailed hyperparameters are provided in the Appendix.

Missing Modality Settings: By following the work in MPVR, we define the missing rate $\eta = 1 - \frac{N_c}{N}$ where N_c is the number of samples with both modalities and N is the total sample size. For *balanced scenarios*, we retain $\left(100 - \frac{\eta}{2}\right)\%$ of each modality; for *single-modality* scenarios, we keep 100% of one modality and $(100 - \eta)\%$ of the other, with at most one modality missing per sample.

Training Configuration: Our model trains directly under missing data conditions without complete-data pre-training. Each prompt pool contains 4 prompts (length $\ell_p = 16$, initialized with $\mathcal{N}(0, 0.02)$). Using Adam optimizer (**?**) (batch size 64, temperature parameter $\tau = 0.07$, learning rate 1×10^{-4}). We simulate missing scenarios by randomly discarding modalities at a rate of $\eta = 70\%$ **during training, validation and testing**

Main Results

We focus on studying the robustness and generalization ability of our **PROMISE** against partial incompleteness in multimodal data.

As shown in Table 1, our proposed PROMISE consistently outperforms all baselines across all datasets and missing modality configurations. Notably, PROMISE demonstrates robust performance in the balanced-missing scenario where each modality is missing 35% of its data, achieving the highest gains compared to state-of-the-art methods. The performance patterns also reveal distinct modality importance across datasets. For instance, PROMISE on the *Hateful Memes* dataset benefits significantly from balanced modality information, while it achieves superior performance on *UPMC Food101* and *MM-IMDb* when text modality is complete, highlighting its crucial role in these specific tasks.

Different Missing Patterns: While our prior experiment at a fixed 70% missing rate demonstrates the model's potential, it overlooks diverse real-world patterns, such as complete image absence due to sensor damage. For a comprehensive evaluation, we assess PROMISE across three scenarios—missing-image, missing-text, and missing-both—at missing rates from 10% to 100%. As shown in Figure 3, this analysis examines the individual and synergistic contributions of our hierarchical contrastive learning losses: PROMISE-Full (both losses), PROMISE-FNCL (FNCL only), and PROMISE-CCCL (CCCL only). Detailed ablation results are provided in the Ablation Study section.

Missing-image & Missing-text Cases As Figure 3 illustrates, the CCCL-only model significantly outperforms the baseline MPVR, indicating that it enables the model to learn a more robust representation with missing data through focusing on intra-modal discriminability. While the FNCL-only model performs slightly below the full model, its strong performance underscores the critical role of inter-modal consistency (FNCL), especially when an entire modality is completely absent. This suggests that bridging the gap between modalities becomes even more crucial in such extreme missing data scenarios.

Missing-both Case As shown in Figure 3, the full model consistently outperforms the FNCL-only model, which in turn surpasses the CCCL-only model, all exceeding the

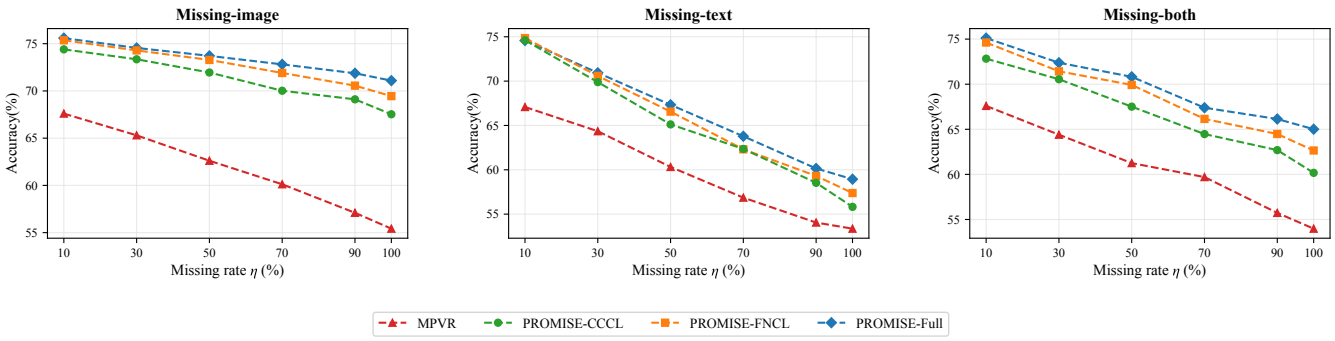


Figure 3: Quantitative results and the chart on the *N24News* dataset with different missing rates under different missing-modality scenarios. Each data point on the figure represents that training with are the same 70% missing rate and testing are with η% missing rate.

baseline across scenarios. The missing-both case introduces greater complexity, as both modalities may be absent, exacerbating the modality gap. Compared to single-modality missing cases, the performance gap between the CCCL-only model and the FNCL-only model widens, highlighting FNCL's crucial role in aligning semantics across modalities. Nonetheless, the CCCL-only model, leveraging intra-modal discriminability, outperforms the baseline, underscoring our method's robustness.

Ablation Study

To evaluate the effectiveness of our proposed PROMISE framework, we conduct two comprehensive ablation studies: (1) component-wise analysis and (2) missing rate analysis.

Prompt	FNCL	CCCL	HatefuleMeMes		N24News	
			F1 (%)	ACC (%)	F1 (%)	ACC (%)
✓	✓	✓	**66.56**	**65.47**	**68.53**	**68.89**
✓	✓		60.76	60.41	67.23	67.82
✓		✓	58.13	58.01	67.03	67.34
✓			57.57	57.11	66.01	66.31
			56.50	56.89	63.69	64.01

Table 2: Ablation study showing the impact of each PROMISE component across three datasets, with checkmarks indicating active components in each configuration.

Component-wise Analysis: To comprehensively understand the contribution of each component within PROMISE, we conducted a detailed ablation study. The results in Table 2 show that the full PROMISE framework consistently achieves the best performance, confirming the benefits of its integrated design.

The Prompt Attention mechanism provides a solid foundation, enabling the model to reconstruct missing information. However, our experiment shows that its full potential is unlocked only when combined with hierarchical contrastive learning.

The table reveals that adding either FNCL or CCCL alone results in only a modest performance increase. The true synergy emerges when both FNCL and CCCL are employed together. The combined approach is crucial because it allows the model to simultaneously address inter-modal consistency (via FNCL) and intra-modal discriminability (via CCCL). Such a dual-level strategy is key to achieving robust and generalizable representations, especially under high missing rates, and ultimately leads to the superior performance of the full PROMISE framework.

Different rate of missing modality: To further validate the generalization of our Component-wise Analysis across varying missing rates, we conduct experiments with missing rates ranging from 10% to 100% on three different datasets.

Figure 4 illustrates the performance of our PROMISE method compared to MPVR across various missing rate scenarios on the *N24News*, *Hateful Memes*, and *UPMC Food101* datasets, respectively. Similar trends can be concluded. Across all three datasets and varying missing rates (from 10% to 100%), PROMISE consistently demonstrates superior performance. This indicates that our approach maintains robustness and effectiveness even as the proportion of missing modalities increases, significantly outperforming existing baselines in handling data incompleteness.

Parameter Sensitivity

To assess parameter sensitivity in the Prompt Attention module of PROMISE, we evaluated the impacts of layer count and prompt dimension on performance across the *Hateful Memes* dataset. As depicted in Figure 5, the analysis explores configurations spanning 1 to 8 layers and prompt dimensions from 16 to 128.

The Effect of Prompt Dimension: Performance exhibits a non-linear relationship with prompt dimension. Moderate dimensions, such as 32 or 48, consistently yield strong results. Performance typically declines beyond 48, potentially due to increased noise or redundancy. In contrast, smaller dimensions like 16 can deliver robust outcomes, especially with 6 layers, highlighting the efficacy of compact representations.

The Effect of Layer Count: Performance does not increase monotonically with layer count. Optimal results are achieved at 6 layers with a dimension of 16 (67.17%). Although 1- and 2-layer models generally underperform, a 3-layer model with a dimension of 16 attains a near-optimal

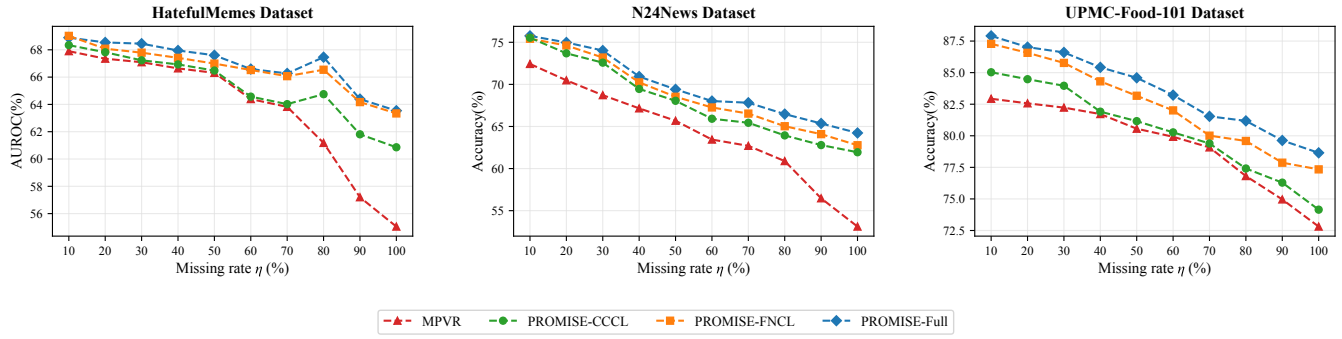


Figure 4: Quantitative results and the chart on the *N24News*, *Hateful Memes*, *UPMC Food101* datasets with different missing rates. Each data point represents training and testing with the same missing rate η.

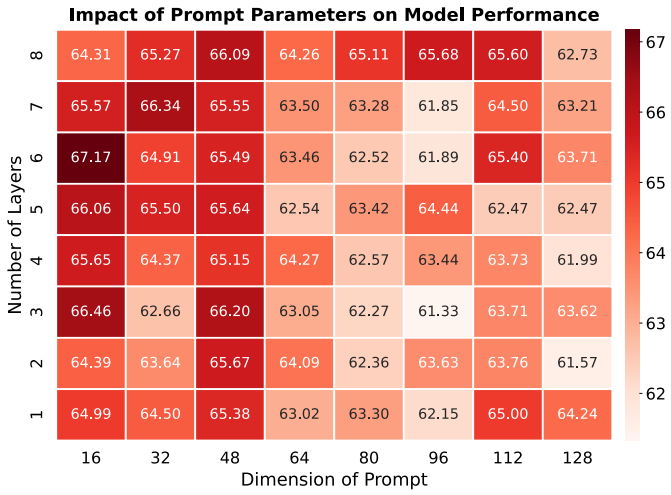


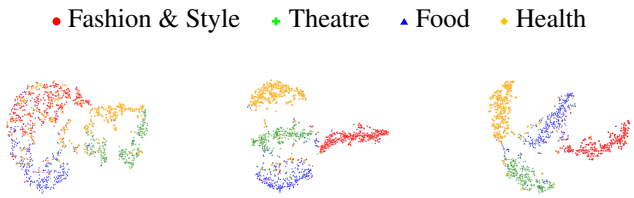
Figure 5: Parameter sensitivity study on the number and length of prompts, using AUROC as evaluation metric on the *Hateful Memes* dataset with 70% missing rate in training and testing.

value of 66.46%. These results indicate that fewer layers, when combined with suitable prompt dimensions, can produce competitive performance, thereby reducing computational overhead without substantial degradation. Models with 7-8 layers exhibit variable outcomes, suggesting diminishing returns from additional depth.

Representation Discriminability and Semantic Consistency

To demonstrate the robustness of the learned representations, we conducted a t-SNE visualization analysis on four *N24News* categories, as shown in Figure 6.

The t-SNE visualization in Figure 6 reveals that our PROMISE approach successfully disentangles the latent embeddings across different categories, achieving superior performance compared to the image-only baseline (LowerBound). Remarkably, PROMISE attains comparable embedding separation using only 30% text data combined with



(a) LB (b) 30%text + 100%image (c) UB

Figure 6: t-SNE comparison on the *N24News* dataset: (a) Lower bound (image-only), (b) PROMISE (100% image + 30% text), (c) Upper bound (full data).

100% image data, matching the performance of our UpperBound model that utilizes complete modalities (100% image and 100% text). These results validate PROMISE's effectiveness in leveraging limited cross-modal information and its robustness under severe modality missingness scenarios.

Conclusion

We proposed PROMISE, a novel framework that combines Prompt Attention mechanism with hierarchical contrastive learning to address missing modalities in multimodal tasks. By generating trainable modality-specific prompts and enforcing cross-modal consistency through dual-level contrastive learning, our approach significantly outperforms state-of-the-art methods on benchmark datasets. Experiments demonstrate PROMISE's robustness even at high missing rates, maintaining semantic consistency between available and missing modalities.

Acknowledgments

This work was supported by the National Science Foundation of China (NSFC) under Grant 62201061.