LEARNING FROM RELIABLE LATENT PROMPTS FOR VISUAL RECOGNITION WITH MISSING MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale multimodal models (LMMs) have achieved superior performance in visual recognition by synergizing information across diverse, massive-scale paired modalities. In real-world scenarios, however, missing-modality inputs are ubiquitous, causing models optimized for modality-complete data to exhibit precipitous performance degradation. Existing research has introduced prompt learning to mitigate this issue, typically by generating dynamic prompts from instance-level features, regardless of whether the input modalities are complete or partially absent. However, such input-conditioned strategies are hindered by the escalating unreliability of instance-level features; as higher missing rates increase the proportion of incomplete modalities, the resulting instability in prompt learning limits the model's performance. To address this limitation, we hypothesize that learnable latent prompts themselves encapsulate stable, modality-intrinsic priors that are decoupled from corrupted inputs. Consequently, we propose a novel paradigm: Learning from Reliable Latent Prompts. Unlike prior methods, we model input-agnostic learnable prompts as stable latent anchors that enable robust guidance and effective cross-modal knowledge compensation, even under extreme missing rates (e.g., 90%). Empirical results across three benchmark datasets demonstrate that our "learn-from-latent-prompts" approach achieves state-of-the-art performance across a wide range of missing-modality scenarios. Extensive experiments further confirm the effectiveness of this paradigm in providing a robust solution to the missing-modality problem. Code will be released after acceptance.

1 INTRODUCTION

The essence of human intelligence is rooted in the synergistic integration of cross-modal information to achieve robust perception, a capability that modern Large-scale Multimodal Models (LMMs) aspire to emulate. By leveraging advanced pre-trained architectures (Vaswani et al., 2017; Gu & Dao, 2023; Arnab et al., 2021) and massive-scale aligned datasets (Wang et al., 2015; Kiela et al., 2020), these models (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023) have demonstrated remarkable efficacy across a diverse spectrum of downstream tasks, ranging from cross-modal retrieval (Radford et al., 2021; Jiang & Ye, 2023; Kim et al., 2021) to image captioning (Sarto et al., 2023; Iashin & Rahtu, 2020) and visual question answering (Zhang et al., 2023; Sima et al., 2024). Despite these strides, the practical utility of LMMs is often bottlenecked by a rigid dependency on modality-complete inputs, an assumption that frequently falters in real-world deployments. Factors such as privacy constraints, sensor failures, and data collection environments often lead to partially missing modalities (Lee et al., 2023; Hu et al., 2024). Such modality-incomplete inputs can significantly degrade the cross-modal alignment, leading to a precipitous degradation in performance (Ma et al., 2022). Developing robust multimodal learning frameworks that can operate reliably under varying missing-modality scenarios has therefore become a critical challenge.

Existing approaches to missing-modality learning primarily rely on cross-modal generation or joint representation learning (Wu et al., 2024b; Ma et al., 2022; Wang et al., 2023a). While effective in controlled settings, these methods often introduce additional semantic noise through imperfect modality reconstruction and incur substantial computational overhead due to auxiliary model optimization (Wu et al., 2024a). Recently, prompt-based learning has emerged as an efficient alternative, adapting pre-trained LMMs by optimizing a small set of learnable tokens while keeping the backbone frozen. Early methods employ missing-aware prompts (Lee et al., 2023), whereas subsequent works adopt

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

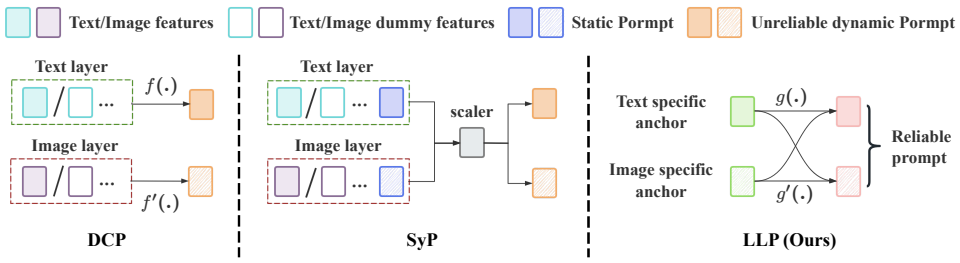


Figure 1: Comparison between input-conditioned prompting methods (e.g., DCP (Hu et al., 2024) and SyP (Zhang et al., 2025)) and our proposed LLP paradigm. Existing approaches generate prompts conditioned on instance-level features, which become unreliable under missing-modality settings due to their dependence on incomplete and modality-biased inputs. In contrast, LLP generates prompts from input-agnostic latent prompts (anchors), enabling more reliable prompt formation and robust cross-modal reasoning with missing modalities.

sample-aware dynamic prompting strategies that generate prompts conditioned on instance-level features (Hu et al., 2024).

Despite their empirical success, existing input-conditioned prompting methods (Hu et al., 2024; Zhang et al., 2025) fundamentally rely on incomplete observations, as illustrated in Figure 1. As the missing rate increases, the text and image features used to generate prompts become increasingly incomplete and modality-biased. Consequently, the generated dynamic prompts become more dependent on the available modality-specific signals, making prompt optimization sensitive to different missing conditions and potentially compromising the stability of learned prompts. Our theoretical analysis, representation-level visualizations, and gradient analysis provide complementary evidence for this limitation (see Appendix A). In particular, input-conditioned methods exhibit locally separable modality-specific structures but struggle to maintain a consistent class-level organization, while showing weaker gradient alignment across different missing conditions under high missing rates. These observations suggest that conditioning prompt generation on incomplete observations can cause dynamic prompts to collapse toward the available unimodal evidence, thereby disrupting cross-modal alignment and hindering robust optimization across missing conditions.

In this paper, we explore an alternative paradigm that decouples prompt learning from input-specific observations. We posit that learnable latent prompts, optimized independently of incomplete inputs, can capture stable modality-intrinsic priors and provide a more reliable foundation for multimodal reasoning under missingness. Based on this insight, we propose **Learning from Reliable Latent Prompts (LLP)**, a prompting framework that adopts a latent prior-driven paradigm, decoupling prompt learning from unreliable inputs. LLP introduces modality-specific latent anchors, implemented as input-agnostic learnable prompts. Unlike dynamic prompts that are directly conditioned on input features, these latent anchors are learned as input-agnostic parameters that encode modality-level priors. Crucially, these anchors serve as semantic foundations for learning refined latent prompts, rather than static offsets. To this end, we design a cross-modal co-adaptation (CMCA) mechanism that enables anchors to exchange complementary information across modalities while generating dual anchor-induced prompts through their interaction. Through this anchor-based interaction, LLP enables the model to approximate missing semantic information from a stable latent space, thereby mitigating the impact of input corruption even under extreme missing rates. Extensive evaluations across three widely-used datasets show that LLP consistently outperforms state-of-the-art methods, especially under high missing rates. Ablation studies further confirm the efficacy of our latent anchor design and the paradigm's robustness to diverse missing patterns. Our contributions are summarized:

- We propose LLP, a novel latent prior-driven prompting paradigm that decouples prompt learning from unreliable inputs and captures stable modality-level priors, leading to improved robustness under incomplete multimodal observations.
- Specifically, LLP introduces modality-specific latent anchors as input-agnostic learnable prompts, along with a cross-modal co-adaptation mechanism to generate dual anchor-induced prompts, enabling input-independent prompt learning and stable cross-modal interaction, thereby achieving robust semantic compensation under severe modality missing.

- We demonstrate state-of-the-art performance across multiple benchmarks and provide comprehensive analyses validating the robustness and generalization of the proposed approach, establishing a promising paradigm for robust multimodal learning under missing modalities.

2 Related Works

Missing-Modality in Multi-modal Learning. Despite the remarkable strides made in multimodal learning (Yuan et al., 2025; Li et al., 2022; Guo & Gu, 2025; Huang et al., 2026; Radford et al., 2021; Liu et al., 2025), ensuring the robustness of models under scenarios with missing modalities remains a formidable challenge (Yuan et al., 2025; Wu et al., 2024a; Lee et al., 2023). In real-world scenarios, multimodal data is often incomplete due to privacy constraints, sensor failures, and heterogeneous data collection environments (Hu et al., 2024; Wu et al., 2024a). Such missing modalities introduce significant distribution shifts and disrupt cross-modal alignment, making it difficult for models to learn consistent and informative multimodal representations (Wu et al., 2024a).To address these challenges, existing literature can be broadly categorized into three research directions: cross-modal generation (Wang et al., 2023b; Meng et al., 2024; Lang et al., 2025), joint representation learning (Ma et al., 2022; Wu et al., 2024b; Wang et al., 2023a), and prompt learning (Lee et al., 2023; Lu et al., 2025; Hu et al., 2024; Zhang et al., 2025). Cross-modal generation typically aims to reconstruct missing modalities from available ones (Meng et al., 2024; Wang et al., 2023b). Recent advances also explore leveraging auxiliary information from complete samples to guide the recovery process (Lang et al., 2025). However, such methods are often constrained by the inherent heterogeneity across modalities and the heavy computational overhead. Joint representation learning methods (Wu et al., 2024b; Wang et al., 2023a) have been proposed to align modalities into a unified space, but often heavily rely on data integrity. Prompt learning has recently surfaced as a promising alternative (Lee et al., 2023; Zhang et al., 2025), requiring only the optimization of compact learnable prompts while maintaining a frozen backbone. Existing prompting methods mostly rely on the instance-level dynamic prompts (Hu et al., 2024; Lu et al., 2025; Zhang et al., 2025). However, such dynamic prompts rely on instance-level features that may be incomplete or corrupted under missing-modality settings, which are unstable and susceptible to noise contamination.

Prompt Learning. Prompt learning has emerged as an effective training paradigm across both Natural Language Processing (Brown et al., 2020; Liu et al., 2022; Li & Liang, 2021; Lester et al., 2021) and multimodal learning (Zhou et al., 2022b;a; Khattak et al., 2023; Zhang et al., 2024; Yang et al., 2026), demonstrating remarkable success in efficient model adaptation. CoOp (Zhou et al., 2022b) first introduced learnable soft prompts for vision-language models, replacing hand-crafted templates with optimized continuous vectors to enhance task-specific adaptation. MaPLe (Khattak et al., 2023) introduces a multi-modal prompting strategy to achieve synergistic learning across both vision and language branches. Meanwhile, DePT (Zhang et al., 2024) explores a decoupled feature space to enhance the generalization and efficiency of prompt tuning. To address modality incompleteness, MMP (Lee et al., 2023) pioneered the use of missing-aware prompts for efficient adaptation. Building upon this, subsequent works such as DCP (Hu et al., 2024) and SyP (Zhang et al., 2025) further develop different dynamic prompting paradigms, where prompts are adaptively generated conditioned on instance-level features to capture richer semantics. These methods have demonstrated strong effectiveness in adapting large-scale vision-language models to downstream tasks, and improving robustness under missing-modality scenarios. In this paper, we propose a *latent prior-driven prompting* framework that decouples prompt learning from unreliable inputs by leveraging latent anchors as learnable prompts, thereby mitigating the limitations of input-conditioned prompt generation and improving robustness under missing-modality scenarios.

3 Methods

3.1 Problem Formulation

We begin by providing a concise formulation of the missing-modality scenario considered in this paper. For simplicity and without loss of generality, this paper considers a multimodal dataset comprising M = 2 modalities (i.e., text and image) $m1$ and $m2$. Given a multimodal dataset $D = \{D^c, D^{m1}, D^{m2}\}$, $D^c = \{x_i^{m1}, x_i^{m2}, y_i\}$ is denoted as the modality complete subset, $D^{m1} = \{x_i^{m1}, y_i\}$ and $D^{m2} = \{x_i^{m2}, y_i\}$ are modality incomplete subsets. Similar to previous works (Lee et al., 2023; Hu et al.,

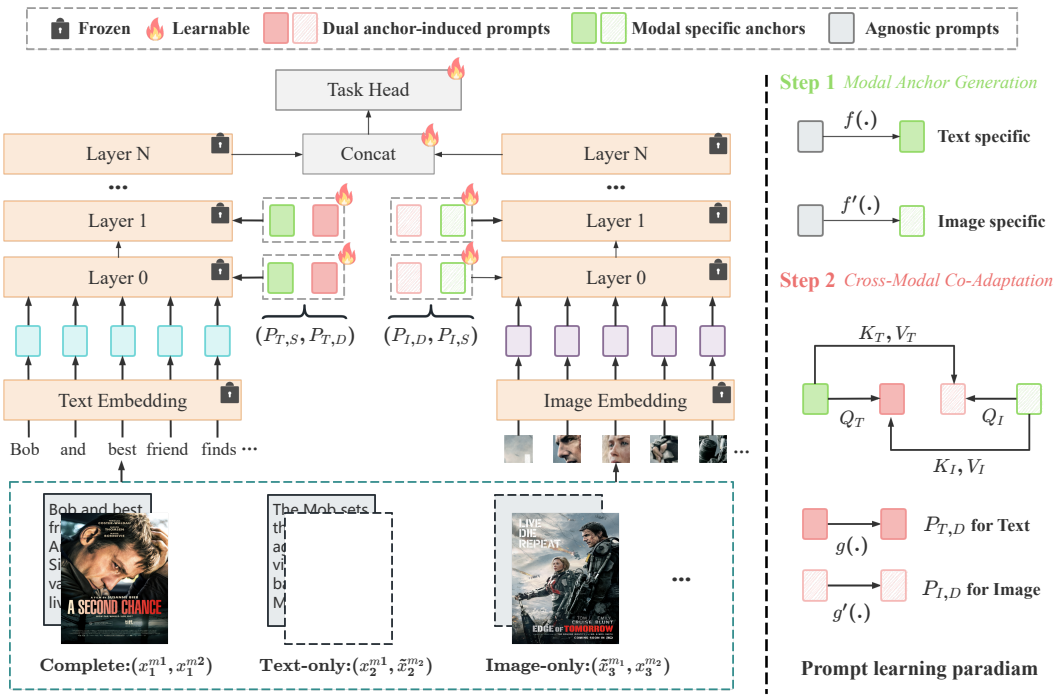


Figure 2: Overview of the proposed Learning from Reliable Latent Prompts (LLP) framework. Given multimodal inputs with potentially missing modalities, LLP introduces modality-specific latent anchors (i.e., learnable prompts) $P_{I,S}(P_{T,S})$ at each encoder layer. These anchors are input-agnostic and capture stable modality-level priors. A cross-modal co-adaptation (CMCA) mechanism lets anchors from different modalities exchange complementary information to jointly generate dual anchor-induced prompts $P_{I,D}(P_{T,D})$. The refined prompts are integrated into both text and image encoders, and the resulting representations are fused for downstream prediction.

2024), we also use dummy inputs $\tilde{x}^{m1}$ and $\tilde{x}^{m2}$ (i.e., empty string and pixel for missing texts and images) to preserve the format of multimodal inputs, so D^{m1} and D^{m2} can be rewritten as $\tilde{D}^{m1} = \{x_i^{m1}, \tilde{x}_i^{m2}, y_i\}$ and $\tilde{D}^{m2} = \{\tilde{x}_i^{m1}, x_i^{m2}, y_i\}$. Thus, a multimodal dataset with missing modalities can be formulated as $\tilde{D} = \{D^c, \tilde{D}^{m1}, \tilde{D}^{m2}\}$.

Figure 2 shows the overall framework of our proposed LLP. We adopt the widely-used two-stream multimodal architecture CLIP (Radford et al., 2021) as our model backbone. We first encode the input text and image into token sequences using the pre-trained embedding layers of pretrained CLIP (Radford et al., 2021). For each input sample, we select the corresponding prompts P_m^I and P_m^T for the image and text encoders, respectively, according to the missing-modality type $m \in \{c, m_1, m_2\}$. The prompt is composed of two types of prompts, including modality-specific latent prompts $P_{I,S}(P_{T,S})$, and dual anchor-induced prompts $P_{I,D}(P_{T,D})$. We prepend the prompts to the input tokens x_{m_1} and x_{m_2}, forming unified input sequences for subsequent processing. We combine the task-related tokens from both encoders to form the final representation, which is then fed into a fully connected layer for prediction. During training, only the fully connected layer and the proposed prompts are optimized, while the backbone networks (i.e., the text and image embedding layers and encoders) remain frozen. We next present the detailed design of our prompting mechanism.

3.2 Learning from Reliable Prompt Learning

3.2.1 Modality-Specific Latent Anchors

We first introduce modality-specific latent anchors, which are input-agnostic learnable prompt tokens that capture stable modality-level priors and serve as the foundation for subsequent prompt generation. Existing input-conditioned prompting methods (Hu et al., 2024; Zhang et al., 2025) rely on instance-

level features that become unreliable under missing-modality settings, leading to unstable prompt learning and error accumulation. Instead of building upon such corrupted signals, we construct prompts from clean and informative latent anchors. This design ensures that prompt generation is grounded in stable and reliable representations. These modality-specific prompts (anchors) are modeled as input-agnostic learnable prompts, which capture modality-level information for further robust cross-modal reasoning under different missing-modality scenarios.

Specifically, we derive modality-specific latent anchors from input-agnostic prompts for each encoder. The latent anchors for the image and text encoders are defined as:

$$P_{I,S} = f_I(P_{I,A}), \quad P_{T,S} = f_T(P_{T,A}), \tag{1}$$

where $P_{I,A}$ and $P_{T,A}$ denote the input-agnostic prompts for the image and text encoders, respectively.

The transformation functions $f_I(\cdot)$ and $f_T(\cdot)$ share the same architecture, implemented as a fully connected layer (FC) with GELU activation (Hendrycks & Gimpel, 2016) followed by Layer Normalization (LN) (Ba et al., 2016). Specifically, the function $f(\cdot)$ is formulated as:

$$f(\cdot) = \mathrm{LN}\big(\mathrm{FC}(\mathrm{GELU}(\mathrm{FC}(\cdot)))\big). \tag{2}$$

The latent anchor $P_{I,S}(P_{T,S})$ stores the stable latent information for each modality under different missing cases.

3.2.2 Dual Anchor-induced Prompts

Building upon the modality-specific latent anchors introduced above, we next refine these priors through cross-modal interaction. Since the anchors capture modality-specific information, we introduce a cross-modal co-adaptation (CMCA) mechanism to enable interaction between modality-specific priors, facilitate cross-modal semantic compensation, and adaptively adjust the model's focus across modalities. This interaction refines the modality-specific priors into dual anchor-induced prompts, enhancing the modeling of cross-modal relationships while preserving their input-independent nature for robust semantic guidance under severe modality missing.

Specifically, given the latent anchors for the text and image modalities, denoted as $P_{T,S}$ and $P_{I,S}$, respectively, we perform bidirectional cross-attention to enable complementary information exchange between the two modalities. For the text branch, we use $P_{T,S}$ as queries and $P_{I,S}$ as keys and values:

$$\tilde{P}_T = \mathrm{MHCA}(Q = P_{T,S},\ K = P_{I,S},\ V = P_{I,S}), \tag{3}$$

Similarly, for the image branch, we use $P_{I,S}$ as queries and $P_{T,S}$ as keys and values:

$$\tilde{P}_I = \mathrm{MHCA}(Q = P_{I,S},\ K = P_{T,S},\ V = P_{T,S}), \tag{4}$$

where $\mathrm{MHCA}(\cdot)$ denotes multi-head cross-attention (Vaswani et al., 2017). Through this bidirectional interaction, the anchors exchange complementary information across modalities. The final modality-specific prompts are obtained by applying a lightweight transformation with a bottleneck multi-layer perceptron (MLP) and Layer normalization (Ba et al., 2016):

$$P_{I,D} = \mathrm{LN}(\mathrm{MLP}(\mathrm{LN}(\tilde{P}_I))), \tag{5}$$

$$P_{T,D} = \mathrm{LN}(\mathrm{MLP}(\mathrm{LN}(\tilde{P}_T))). \tag{6}$$

By grounding prompt generation in input-agnostic latent anchors and modeling cross-modal interactions via dual-path attention, our approach mitigates noise introduced by incomplete inputs, facilitates effective information exchange, and yields robust and reliable semantic compensation even under extreme missing-modality scenarios. The final prompts P_m^I and P_m^T are constructed by concatenating the modality-specific latent anchors ($P_{I,S}, P_{T,S}$) and the dual anchor-induced prompts ($P_{I,D}, P_{T,D}$), respectively:

$$P_m^I = [P_{I,S}, P_{I,D}], \quad P_m^T = [P_{T,S}, P_{T,D}], \tag{7}$$

where $[\cdot, \cdot]$ denotes the concatenation operation.

3.2.3 LAYER-WISE PROMPT GENERATION

Following prior works (Hu et al., 2024; Zhang et al., 2025), we adopt a layer-wise prompt generation strategy to maintain coherent information flow across encoder layers. Specifically, the generated prompts are injected into each layer and progressively refined along the network depth. For the image encoder, we denote the prompt at the i-th layer as P_m^{I,R_i}, where $i \in [1, N]$. The prompt is updated based on the previous layer via:

$$P_m^{I,R_i} = \mathrm{LN}\big(\mathrm{FC}(\mathrm{GELU}(\mathrm{FC}(P_m^{I,R_{i-1}})))\big), \tag{8}$$

where FC denotes a bottleneck MLP, GELU (Hendrycks & Gimpel, 2016) is the activation function, and LN (Ba et al., 2016) denotes Layer Normalization. This design enables the prompts to capture hierarchical representations across layers, enhancing their expressiveness for multimodal reasoning.

Comparisons with MMP, DCP, and SyP. LLP differs from MMP (Lee et al., 2023) by moving from modality-shared static prompts to modality-specific latent prompts, which serve as modality anchors for further learning and cross-modal interaction. It differs from DCP (Hu et al., 2024) and SyP (Zhang et al., 2025) by learning prompts from stable latent anchors rather than potentially incomplete input features, thereby establishing a new "learning from reliable latent prompts" paradigm for missing-modality scenarios.

4 EXPERIMENTS

4.1 EXPERIMENT SETTINGS

Datasets. We follow prior works (Lee et al., 2023; Hu et al., 2024; Zhang et al., 2025) and evaluate our method on three widely used benchmark datasets: 1) MM-IMDb (Arevalo et al., 2017) is a large-scale multimodal dataset for movie genre prediction, containing 25,959 movies annotated with both images and textual plot descriptions. Since each movie may belong to multiple genres, the task is formulated as a multi-label classification problem. This dataset evaluates the model's ability to jointly reason over visual and textual modalities for semantic understanding. 2) UPMC Food-101 (Wang et al., 2015) is a large multimedia dataset consisting of noisy image-text pairs collected from the web, covering 101 food categories. The textual descriptions are often unstructured and noisy, making the dataset particularly challenging for multimodal learning. It is widely used for food classification tasks requiring robust cross-modal alignment. 3) Hateful Memes (Kiela et al., 2020) is a multimodal benchmark for hateful meme detection, containing over 10,000 image-text pairs. The dataset is carefully constructed such that neither modality alone is sufficient for accurate prediction, explicitly requiring joint multimodal reasoning. This makes it a challenging testbed for evaluating robustness under incomplete or misleading modality inputs.

Implementation details. We leverage CLIP (Radford et al., 2021) as our multimodal backbone with a ViT-B/16 (Dosovitskiy et al., 2020) image encoder. We resize input images to 224×224 and split them into patches of size 16×16. Text inputs are processed using the CLIP tokenizer with a maximum sequence length of 77. Both the image and text encoders are kept frozen during training, and only the prompt parameters, together with the task-specific fully connected (FC) layer are optimized. The feature dimensions for both modalities are set to 512. We use learnable prompts of length $L_p = 16$, which are injected into features from $M = 6$ layers. For missing modalities, we replace the corresponding inputs with zero-filled tensors to maintain a consistent input structure. We train the model using the AdamW optimizer Loshchilov & Hutter (2017) with a learning rate of 1×10^{-2} and a weight decay of 2×10^{-2}. The learning rate is linearly warmed up for 10% of the total training steps and then decayed to zero. All experiments are conducted on an NVIDIA RTX A6000 GPU with a batch size of 32.

Missing Modality Setting. In this paper, we consider a realistic multimodal setting where modalities may be partially missing during both training and evaluation. Following prior works, we characterize the level of missing data by a missing rate η, defined as the fraction of samples with incomplete modalities. For vision-language tasks, we consider three missing scenarios: missing-text, missing-image, and missing-both. In the missing-text or missing-image setting, a proportion η of the samples contain only a single modality, while the remaining $(1 - \eta)$ samples retain both modalities. In the missing-both setting, the dataset is constructed such that $\frac{\eta}{2}$ of the samples are text-only, $\frac{\eta}{2}$ are image-only, and the remaining $(1 - \eta)$ are modality-complete. This formulation can be extended to

Table 1: Comparison with CoOp (Zhou et al., 2022b), MMP (Lee et al., 2023), MaPLe (Khattak et al., 2023), DePT (Zhang et al., 2024), DCP (Hu et al., 2024), and SyP (Zhang et al., 2025) on the MM-IMDb (Arevalo et al., 2017), UPMC Food-101 (Wang et al., 2015), and Hateful Memes (Kiela et al., 2020) datasets under various missing-modality cases with different missing rates. The bold number indicates the best performance.

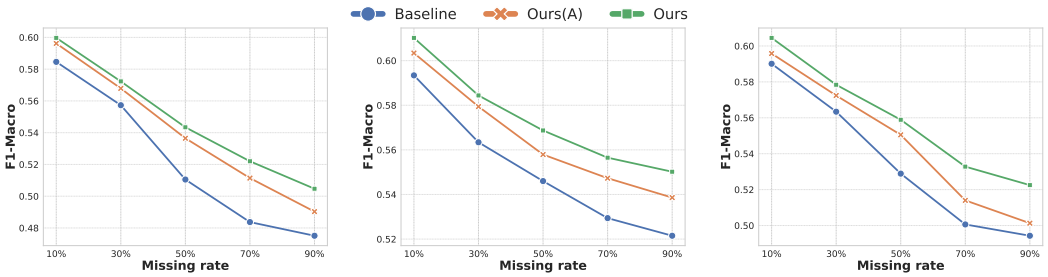
Datasets	Missing rate η	Train/Test Image	Train/Test Text	CoOp	MMP	MaPLe	DePT	DCP	SyP	**LLP**
MM-IMDb (F1-Macro)	50%	100%	50%	48.06	48.88	49.58	50.64	52.13	53.90	**54.34**
		50%	100%	49.89	51.46	52.32	52.78	54.32	56.27	**56.87**
		75%	75%	48.37	49.32	49.56	50.87	52.32	55.02	**55.89**
	70%	100%	30%	44.13	45.64	45.52	46.38	48.52	51.37	**52.20**
		30%	100%	48.82	50.52	50.64	52.13	53.14	54.20	**55.65**
		65%	65%	46.84	48.12	49.16	50.32	51.42	52.90	**53.28**
	90%	100%	10%	44.76	45.32	46.84	47.56	49.26	50.21	**50.47**
		10%	100%	48.32	49.12	50.13	50.88	52.22	53.72	**55.02**
		55%	55%	44.12	44.87	45.12	46.54	48.04	49.63	**51.25**
Food101 (Accuracy)	50%	100%	50%	77.45	77.89	79.64	80.16	82.11	83.20	**84.28**
		50%	100%	87.02	87.16	87.35	82.14	89.12	89.64	**89.85**
		75%	75%	81.24	81.72	82.34	83.12	85.24	86.17	**86.74**
	70%	100%	30%	76.34	76.52	77.02	77.34	78.87	79.56	**80.68**
		30%	100%	84.78	85.64	85.89	86.12	87.32	88.67	**88.94**
		65%	65%	78.87	79.12	79.84	81.46	81.87	82.45	**83.22**
	90%	100%	10%	71.87	73.14	73.46	74.12	75.26	76.33	**78.10**
		10%	100%	81.67	82.14	83.12	83.56	85.78	86.41	**86.83**
		55%	55%	76.46	76.58	77.85	78.12	79.87	81.03	**81.69**
Hateful Memes (AUROC)	50%	100%	50%	60.56	60.31	60.87	61.87	62.32	68.25	**70.84**
		50%	100%	62.41	62.35	63.13	63.88	64.46	66.80	**68.07**
		75%	75%	64.87	65.84	65.46	65.86	66.02	68.16	**69.38**
	70%	100%	30%	60.74	61.12	61.26	61.56	62.82	68.94	**70.28**
		30%	100%	62.74	63.24	63.14	63.48	64.12	66.98	**67.25**
		65%	65%	64.82	65.04	65.23	65.48	66.08	68.42	**69.62**
	90%	100%	10%	60.03	57.21	60.74	61.14	62.08	69.70	**70.90**
		10%	100%	61.46	61.52	61.87	62.42	63.87	64.54	**65.19**
		55%	55%	64.32	63.34	64.85	65.37	66.78	68.93	**69.17**

multimodal settings with M modalities, where each incomplete case is sampled proportionally based on $\frac{\eta}{M^2-2}$, while $(1-\eta)$ of the data remains fully observed.

Baselines and Metrics. We compare our proposed LLP with six competitive baselines: (1) CoOp (Zhou et al., 2022b), which learns input-level prompts; (2) MMP (Lee et al., 2023), which introduces missing-aware prompts for both input and intermediate features; (3) MaPLe (Khattak et al., 2023), which performs cross-modal prompt learning between image and text encoders; (4) DePT (Zhang et al., 2024), which disentangles modality-specific knowledge into separate feature spaces; (5) DCP (Hu et al., 2024), which models prompt correlations and inter-layer features interactions for enhanced representation learning; and (6) SyP (Zhang et al., 2025), which employs an adapter to jointly learn from instance features and static prompts. Following prior works, we leverage F1-Macro to measure the multi-label classification performance on the MM-IMDb dataset (Arevalo et al., 2017), top-1 classification accuracy to evaluate the recognition performance on the UPMC Food-101 dataset (Wang et al., 2015), and Area Under the Receiver Operating Characteristic Curve (AUROC) for Hateful Memes (Kiela et al., 2020).

4.2 Experiment results

Comparison with the State-of-the-arts. We first conduct experiments on three widely-used datasets: MM-IMDb (Arevalo et al., 2017), UPMC Food-101 (Wang et al., 2015), and Hateful Memes (Kiela et al., 2020). we compare our proposed LLP method with six state-of-the-art methods, including: (1) CoOp (Zhou et al., 2022b), (2) MMP (Lee et al., 2023), (3) MaPLe (Khattak et al., 2023), (4) DePT (Zhang et al., 2024), (5) DCP (Hu et al., 2024), and (6) SyP (Zhang et al., 2025). We evaluate these methods across three different missing rates, including $\eta = 50\%$, $\eta = 70\%$, and $\eta = 90\%$, across three various missing-modality cases, including missing-image, missing-text



(a) Missing Text (b) Missing Image (c) Missing Both

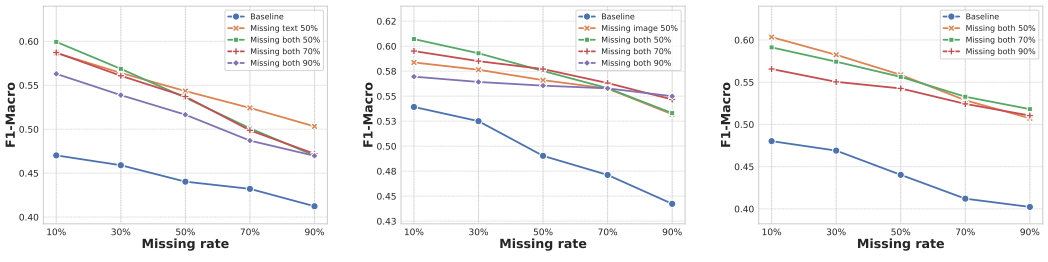
Figure 3: Comparison of our final model (Ours) with (1) baseline, which directly drops the features when a modality is missing; (2) Ours (A), which only equips the Modality-Specific Latent Anchors. The experiments are conducted on the MM-IMDb dataset (Arevalo et al., 2017) across different missing rates under three missing-modality scenarios, with all models trained and evaluated at the same missing rate. Additionally, Ours (A) and Ours use the same prompt length and depth.

and missing-both, in Table 1. As shown in Table 1, LLP consistently outperforms state-of-the-art methods, including SyP (Zhang et al., 2025) and DCP (Hu et al., 2024), across all missing-modality settings. The performance gains are particularly pronounced under high missing rates ($\eta = 90\%$), demonstrating the effectiveness of the latent prior-driven prompting paradigm in improving the robustness of multimodal models. The gains are particularly pronounced on more challenging benchmarks. On Hateful Memes (Kiela et al., 2020), which requires strong cross-modal reasoning, LLP achieves significant improvements, highlighting its ability to capture complementary multimodal information. Similarly, on MM-IMDb (Arevalo et al., 2017), a difficult multi-label classification task with relatively low overall performance, LLP consistently outperforms all baselines across different missing settings. Despite the performance being in the 40-50 F1 range, LLP still delivers stable and noticeable gains, demonstrating its effectiveness in handling complex and noisy multimodal scenarios. These results suggest that the advantage of LLP is not dataset-specific, but generalizes well to challenging multimodal learning problems.

An interesting observation is that different datasets exhibit asymmetric sensitivity to modality missing. Specifically, MM-IMDb (Arevalo et al., 2017) and Food101 (Wang et al., 2015) suffer more performance degradation when the text modality is missing, while Hateful Memes (Kiela et al., 2020) is more sensitive to missing visual information. Despite these varying modality dependencies, LLP consistently achieves larger performance gains in the more challenging missing cases for each dataset (i.e., missing text for MM-IMDb (Arevalo et al., 2017) and Food101 (Wang et al., 2015), and missing images for Hateful Memes (Kiela et al., 2020)). This suggests that, unlike prior input-conditioned methods that rely on potentially incomplete or corrupted inputs, LLP effectively leverages modality-specific latent anchors as stable sources of information. By grounding prompt generation in these input-independent representations, LLP enables more reliable cross-modal compensation, leading to improved robustness across diverse missing-modality scenarios.

Ablation Studies. We also conduct numerous ablation studies to evaluate the effectiveness of the proposed components of LLP. As shown in Figure 3, we first verify the effectiveness of our proposed components across different missing cases on the MM-IMDb dataset (Arevalo et al., 2017), including missing-image, missing-text, and missing-both. Each model variant was trained and evaluated with a consistent missing rate. We first observe that all our variants consistently outperform the baseline across different missing-modality settings, demonstrating improved robustness under incomplete multimodal inputs. Modality-specific latent anchors yield substantial gains over the baseline, demonstrating the value of stable modality-level representations. Notably, as the missing rate increases, further incorporating dual anchor-induced prompts leads to increasingly larger improvements, highlighting the growing effectiveness of the proposed **learning from reliable latent prompts** paradigm under severe modality missingness.

The results in the Appendix B.2 further support this observation, providing additional evidence for the advantage of learning from reliable latent anchors over directly learning static prompts and other component analyses. This confirms the effectiveness of our design, where latent anchors serve



(a) Missing Text (b) Missing Image (c) Missing Both

Figure 4: Generalizability Analysis of Our Method to Different Missing Rates on MM-IMDb dataset (Arevalo et al., 2017). (a) All Models are trained on missing-both or missing-text cases, and evaluated on missing-text cases with different missing rates. (b) All models are trained on missing-both or missing-image cases, and evaluated on missing-image cases with different missing rates. (c) All models are trained on missing-both cases, and evaluated on missing-both cases with different missing rates.

as robust information sources that facilitate more reliable cross-modal interaction and semantic compensation. We also conduct ablation studies on prompt length and depth (see Appendix B.1). Overall, these results demonstrate the efficacy of the proposed LLP.

Generalization Capability. To assess the generalization ability of LLP under varying missing-modality conditions, we train models with different missing rates and evaluate them across three scenarios: missing text, missing image, and missing both modalities on the MM-IMDb dataset. The results are presented in Figure 4. We observe that all variants of LLP consistently outperform the baselines across a wide range of missing rates (10%–90%), demonstrating strong robustness to incomplete multimodal inputs. Moreover, models trained under single-modality missing settings exhibit competitive performance when evaluated on the corresponding missing scenario, indicating effective adaptation to modality-specific absence. Notably, models trained under the missing-both setting achieve strong and stable performance across all evaluation scenarios. This suggests that jointly modeling multiple missing patterns enables more flexible and robust behavior. In addition, more details on the generalization ability of our LLP framework across different backbones are provided in the Appendix C.2.

Robustness. We further evaluate the robustness of LLP under varying missing rates (see Appendix C.1 for details). We observe that all baseline methods suffer from significant performance degradation as the missing rate increases, highlighting their sensitivity to incomplete multimodal inputs. In contrast, LLP consistently exhibits the smallest performance drop across different missing scenarios, and even shows performance improvements under high missing rates. These results demonstrate the effectiveness of our latent anchor-based design, which provides stable modality-level priors and enables reliable cross-modal interaction through dual anchor-induced prompts.

5 Conclusion

In this paper, we address the missing-modality problem in multimodal learning by proposing Learning from Reliable Latent Prompts (LLP), a latent prior-driven prompting paradigm that decouples prompt learning from unreliable input observations. By introducing modality-specific latent anchors and dual anchor-induced prompts, LLP enables stable prompt generation and effective cross-modal interaction, providing reliable semantic compensation even under severe modality missing. Extensive experiments across multiple benchmarks demonstrate that LLP consistently outperforms state-of-the-art methods under diverse missing-modality settings. Notably, the performance gains become more pronounced as the missing rate increases, and LLP shows strong robustness across both modality-specific and balanced missing scenarios. Further analyses, including ablation and generalization studies, validate the effectiveness of each component and highlight the importance of stable latent priors for robust multimodal reasoning. Overall, LLP offers a simple yet effective solution for handling missing modalities, providing a promising direction for robust and generalizable multimodal learning.

6 AI USE STATEMENT

In this work, we only used generative AI tools for partial polishing based on our original statements for better presentation, and we have reviewed all related content. Other parts do not use LLMs.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

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

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6836–6846, 2021.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Yuncheng Guo and Xiaodong Gu. Mmrl: Multi-modal representation learning for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25015–25025, 2025.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Lianyu Hu, Tongkai Shi, Wei Feng, Fanhua Shang, and Liang Wan. Deep correlated prompting for visual recognition with missing modalities. *Advances in Neural Information Processing Systems*, 37:67446–67466, 2024.

Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Usman Naseem, Chunyu Wang, Qi Dai, Xiyang Dai, Dongdong Chen, et al. Llm2clip: Powerful language model unlocks richer cross-modality representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5131–5139, 2026.

Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 958–959, 2020.

Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2787–2797, 2023.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19113–19122, 2023.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pp. 5583–5594. PMLR, 2021.

Jian Lang, Rongpei Hong, Zhangtao Cheng, Ting Zhong, Yong Wang, and Fan Zhou. 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*, pp. 1241–1252, 2025.

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

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 3045–3059, 2021.

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

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10965–10975, 2022.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 61–68, 2022.

Xiaohao Liu, Xiaobo Xia, See-Kiong Ng, and Tat-Seng Chua. Continual multimodal contrastive learning. *arXiv preprint arXiv:2503.14963*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Andong Lu, Chenglong Li, Jiacong Zhao, Jin Tang, and Bin Luo. Modality-missing rgbt tracking: Invertible prompt learning and high-quality benchmarks. *International Journal of Computer Vision*, 133(5):2599–2619, 2025.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18177–18186, 2022.

Xiangxi Meng, Kaicong Sun, Jun Xu, Xuming He, and Dinggang Shen. Multi-modal modality-masked diffusion network for brain mri synthesis with random modality missing. *IEEE Transactions on Medical Imaging*, 43(7):2587–2598, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-augmented contrastive learning for image and video captioning evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6914–6924, 2023.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pp. 256–274. Springer, 2024.

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

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15878–15887, 2023a.

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

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

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024a.

Zongwei Wu, Jilai Zheng, Xiangxuan Ren, Florin-Alexandru Vasluianu, Chao Ma, Danda Pani Paudel, Luc Van Gool, and Radu Timofte. Single-model and any-modality for video object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19156–19166, 2024b.

Xi Yang, Yuanrong Xu, Weigang Zhang, Guangming Lu, David Zhang, and Jie Wen. Prompt tuning for clip on the pretrained manifold. *arXiv preprint arXiv:2602.19198*, 2026.

Yuan Yuan, Zhaojian Li, and Bin Zhao. A survey of multimodal learning: Methods, applications, and future. *ACM Computing Surveys*, 57(7):1–34, 2025.

Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12924–12933, 2024.

Xi Zhang, Feifei Zhang, and Changsheng Xu. Vqacl: A novel visual question answering continual learning setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19102–19112, 2023.

Zhihui Zhang, Luanyuan Dai, Qika Lin, Yunfeng Diao, Guangyin Jin, Yufei Guo, Jing Zhang, and Xiaoshuai Hao. Synergistic prompting for robust visual recognition with missing modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1881–1890, 2025.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022b.

APPENDIX

A Additional Analysis of Proposed Mechanism **14**

A.1 Visualization Study 14

A.2 Theoretical Analysis 14

A.3 Empirical Validation of the Gradient-Bias Mechanism. 15

A.4 Positioning of LLP with Existing Prompting Methods 16

B Additional Ablation Study **16**

B.1 Effect of Prompt Length and Depth. 16

B.2 More Component-wise Analysis 17

C Additional Results **18**

C.1 Robustness Analysis 18

C.2 Generalization Across Backbones 18

C.3 Multi-seed Experiments. 19

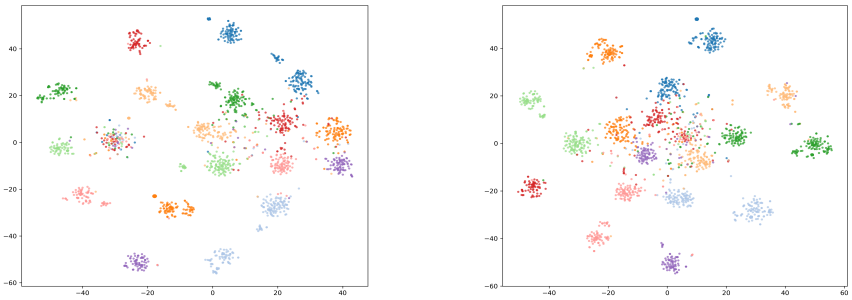
C.4 Complete-Data Performance. 19

D More Implementation Details **20**

D.1 Training Objective. 20

E Limitations and Discussion. **21**

A Additional Analysis of Proposed Mechanism



(a) SyP (b) Ours

Figure 5: Visualization of concatenated prompt-enhanced features from the two encoders on the Food101 dataset (Wang et al., 2015) under the missing-both setting with a high missing rate ($\eta = 90\%$). Compared to SyP (Zhang et al., 2025), our method produces more compact intra-class distributions and clearer inter-class separation, indicating more consistent and robust representations under severe modality missing.

A.1 Visualization Study

To better understand the effectiveness of our method, we visualize the concatenated prompt-enhanced features from the two encoders with a high missing rate ($\eta = 90\%$), as shown in Figure 5. We observe that SyP (Zhang et al., 2025), which relies on input-conditioned dynamic prompts, tends to produce fragmented representations for each class. Specifically, samples from the same class are split into two distinct clusters. This phenomenon arises because, under the missing-both setting, representations are constructed from different incomplete sources, namely prompts combined with image features or prompts combined with text features. Due to the unreliability of input-conditioned features at high missing rates, these clusters are often significantly separated and may even overlap with clusters from other classes. As a result, while individual clusters remain locally separable, the overall class-level separability is degraded.

In contrast, our method produces more coherent and compact class distributions. The two clusters corresponding to the same class are much closer in the feature space, and the boundaries between different classes are better preserved. This indicates that LLP effectively aligns representations across different missing-modality conditions.

We attribute this improvement to the use of modality-specific latent anchors, which provide stable and input-independent representations for prompt generation. By avoiding direct reliance on incomplete input features, LLP enables more reliable cross-modal compensation and maintains consistent semantic structures even under extreme missing conditions.

A.2 Theoretical Analysis

We provide a theoretical perspective to understand the optimization bias of input-conditioned prompting under missing-modality settings. Consider a multimodal input $x = (x_I, x_T)$ and a missing mask $m \in \{0,1\}^2$, where $m_I = 0$ (or $m_T = 0$) indicates the absence of the image (or text) modality. Input-conditioned prompts are generated as

$$P = f(x \odot m; \theta), \tag{9}$$

where $\odot$ denotes element-wise masking and θ denotes the prompt-generation parameters.

Let $\mathcal{L}$ denote the training objective. By the chain rule, the gradient with respect to the prompt-generation parameters can be written as

$$\nabla_\theta \mathcal{L} = \mathbb{E}_m \left[\frac{\partial \mathcal{L}}{\partial P} \frac{\partial P}{\partial \theta} \right], \qquad P = f(x \odot m; \theta). \tag{10}$$

Table 2: Gradient alignment analysis under the extreme missing setting ($\eta = 1.0$) across three datasets. $\cos(\mathcal{B}, \mathcal{T})$ and $\cos(\mathcal{B}, \mathcal{I})$ measure gradient alignment between the missing-both condition and the missing-text and missing-image conditions, respectively. A smaller Alignment Gap indicates more balanced optimization across missing patterns.

Dataset	Method	$\cos(\mathcal{B}, \mathcal{T}) \uparrow$	$\cos(\mathcal{B}, \mathcal{I}) \uparrow$	Alignment Gap $\downarrow$
Food101	DCP	0.0486	0.7549	0.7063
	SyP	0.0267	0.6648	0.6381
	LLP	**0.4511**	**0.7826**	**0.3315**
Hateful Memes	DCP	0.0316	0.7174	0.6857
	SyP	0.4665	0.2167	0.2498
	LLP	**0.7953**	**0.7814**	**0.0139**
MM-IMDb	DCP	0.2539	0.5432	0.2893
	SyP	0.4704	0.1903	0.2802
	LLP	**0.4938**	**0.5693**	**0.0756**

Here, the task-loss term $\partial\mathcal{L}/\partial P$ generally depends on the observed input and therefore remains input-dependent. More importantly, for input-conditioned prompting, the prompt-generation Jacobian $\partial P/\partial\theta$ also depends on the incomplete observation $x \odot m$.

Under standard missing-modality sampling with missing rate η, the expected optimization signal can be decomposed as

$$\mathbb{E}[\nabla_\theta \mathcal{L}] = (1-\eta)\nabla_{\text{multi}} + \eta\nabla_{\text{uni}}, \tag{11}$$

where ∇_{multi} and ∇_{uni} denote the optimization signals induced by multimodal and unimodal observations, respectively. As η increases, the contribution from multimodal observations decreases:

$$\lim_{\eta \to 1} \mathbb{E}[\nabla_\theta \mathcal{L}] \approx \nabla_{\text{uni}}. \tag{12}$$

Consequently, under a high missing rate, input-conditioned prompt generation is increasingly optimized through incomplete unimodal observations. This creates a missing-pattern-dependent optimization bias, since the generated prompts and their parameterization vary according to which modality is observed. Such dependence can make the learned prompts more sensitive to specific missing patterns and hinder consistent cross-modal alignment, motivating the need for an input-agnostic prompt formulation.

In contrast, LLP generates prompts from modality-specific latent anchors P_S, which are independent of the input and missing mask:

$$P = g(P_S; \theta), \tag{13}$$

where P_S does not depend on x or m. The corresponding gradient is

$$\nabla_\theta \mathcal{L} = \mathbb{E}_m \left[\frac{\partial \mathcal{L}}{\partial P} \frac{\partial P}{\partial \theta} \right]. \tag{14}$$

Although the task-loss term $\partial\mathcal{L}/\partial P$ may still depend on the observed input, the prompt-generation Jacobian $\partial P/\partial\theta$ is independent of the corrupted observation $x \odot m$. Therefore, LLP decouples prompt generation from incomplete observations and removes this source of missing-pattern-dependent optimization bias.

This input-agnostic formulation provides a more stable basis for learning modality-level prompt priors under different missing conditions. As a result, LLP can maintain more consistent optimization across missing patterns, providing a theoretical motivation for its improved robustness under missing modalities.

A.3 Empirical Validation of the Gradient-Bias Mechanism.

To empirically validate the theoretical motivation of LLP, we conduct a gradient alignment analysis under the extreme missing setting ($\eta = 1.0$) on Food101, where the optimization bias predicted by our theoretical analysis is expected to be most pronounced. We compare the gradient directions induced

by different missing-modality conditions for LLP and representative input-conditioned prompting methods, DCP (Hu et al., 2024) and SyP (Zhang et al., 2025).

For each method, we compute the cosine similarity between the gradient induced by the missing-both condition and those induced by the missing-text and missing-image conditions, denoted as $\cos(\mathcal{B}, \mathcal{T})$ and $\cos(\mathcal{B}, \mathcal{I})$, respectively. To ensure a fair comparison, cosine similarities are computed only over the intersection of parameter coordinates that receive gradients under all missing conditions, such that all gradient directions are compared within a common parameter space. We further define the *Alignment Gap* as the difference between the two alignment scores, with a smaller gap indicating more balanced optimization across missing patterns.

As shown in Table 2, LLP exhibits substantially higher and more balanced gradient alignment across missing conditions, resulting in the smallest Alignment Gap among all compared methods. In contrast, the substantially larger gaps observed for DCP and SyP indicate that their optimization directions vary more strongly across missing patterns. This behavior is consistent with our theoretical analysis: input-conditioned prompt generation depends on incomplete input observations, making the resulting optimization directions more sensitive to the missing modality. In contrast, LLP learns input-agnostic latent anchors, which leads to more stable optimization across different missing conditions. These results provide empirical evidence supporting the proposed gradient-bias mechanism and establish a direct connection between our theoretical motivation and the observed optimization behavior.

A.4 Positioning of LLP with Existing Prompting Methods

To further clarify the distinction between LLP and existing prompt learning approaches, we compare our design with representative missing-modality methods, including MMP (Lee et al., 2023), DCP (Hu et al., 2024), and SyP (Zhang et al., 2025). MMP (Lee et al., 2023) introduces missing-aware prompts to provide stable information when modalities are unavailable, using static prompts shared across modality encoders. In contrast, DCP (Hu et al., 2024) and SyP (Zhang et al., 2025) introduce input-conditioned or dynamically adapted prompts to improve flexibility across different missing-modality conditions. While such input-dependent designs provide greater adaptability, the generated prompts depend on available input features, which can become incomplete or unreliable as the missing rate increases. LLP instead adopts a complementary design: it first learns input-independent, modality-specific latent anchors as stable modality-level priors, and then further learns from these latent prompts through cross-modal interaction to obtain refined prompts. This moves beyond MMP's modality-shared static prompts while avoiding the input dependence of DCP and SyP, establishing a distinct prompt-learning paradigm for missing-modality scenarios.

Importantly, the proposed framework is not tied to cross-attention as the refinement operator. As shown in Table 3, replacing cross-attention with a lightweight MLP still yields substantial improvements over static prompts, demonstrating the effectiveness and flexibility of the **learning from reliable latent prompts** paradigm. Cross-attention is used as one explicit mechanism to model bidirectional complementary information between modality-specific priors.

B Additional Ablation Study

B.1 Effect of Prompt Length and Depth.

We conduct an ablation study on the prompt length and depth on the Hateful Memes dataset (Kiela et al., 2020) under the missing-both setting with a missing rate of $\eta = 70\%$. As shown in Figure 6, both factors have a noticeable impact on model performance. Increasing the prompt depth improves performance at shallow levels, while excessive depth yields diminishing returns. Similarly, longer prompts provide richer representations but may introduce redundancy when overly large. Overall, the best performance is achieved with a prompt depth of 6 and a prompt length of 16, suggesting that a moderate configuration strikes an effective balance between representation capacity and optimization efficiency.

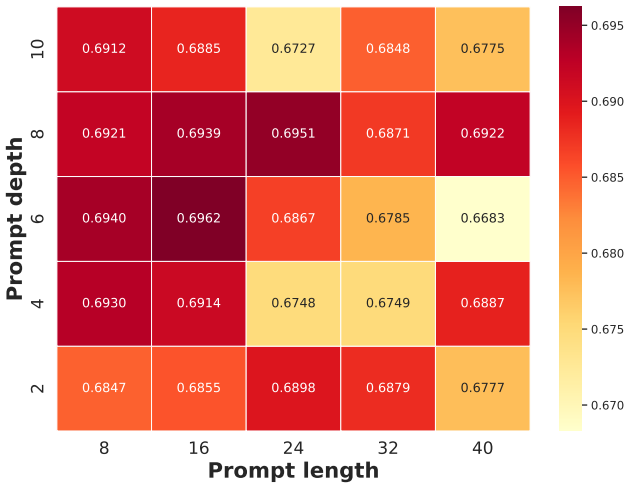


Figure 6: Ablation study on prompt length and depth on the Hateful Memes dataset (Kiela et al., 2020) under the missing-both setting with a missing rate of $\eta = 70\%$. The best performance is achieved with a prompt depth of 6 and a prompt length of 16, indicating a balanced trade-off between representation capacity and optimization stability.

Table 3: Ablation study of different components under the missing-both setting ($\eta = 0.9$) on three datasets. All variants are parameter-controlled, using the same prompt length and depth for a fair comparison. The *+CMCA (MLP)* variant replaces the cross-attention operation in the cross-modal co-adaptation (CMCA) mechanism with a parameter-controlled MLP while retaining the proposed LLP framework and latent-anchor design, demonstrating the general applicability of LLP across different prompt refinement strategies.

Methods	MM-IMDb	Food101	Hateful Memes
Static	48.28%	77.87%	63.22%
Anchors	50.12%	79.90%	64.39%
+ CMCA (MLP)	50.83%	80.61%	64.70%
+ CMCA	**51.25%**	**81.69%**	**69.17%**

B.2 More Component-wise Analysis

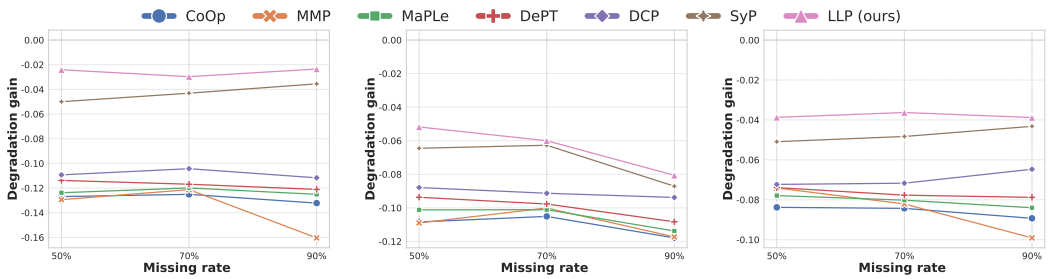
To investigate the contribution of different components in LLP, we conduct a parameter-controlled ablation study under the missing-both setting ($\eta = 0.9$) on MM-IMDb, Food101, and Hateful Memes. All variants use the **same prompt length and depth** to ensure a fair comparison. We progressively introduce the latent-anchor formulation and the proposed dual-path cross-attention, while using an MLP variant as a **parameter-controlled** alternative to cross-attention. Specifically, **Static** directly learns prompt tokens without latent anchors or transformation modules; **Anchors** directly optimizes modality-specific latent anchors ($P_{I,S}$ and $P_{T,S}$); **+CMCA (MLP)** replaces the proposed cross-attention operation in the cross-modal co-adaptation (CMCA) mechanism with an MLP; and **+CMCA** corresponds to the complete LLP. The results show that introducing latent anchors consistently improves over static prompts across all three datasets, demonstrating the benefit of learning modality-specific prompts from latent anchors.

Moreover, adding the proposed cross-modal co-adaptation mechanism-induced prompts further improves performance by 1.13, 1.79, and 4.78 percentage points on MM-IMDb, Food101, and Hateful Memes, respectively, compared with the parameter-controlled **Anchors** variant. Notably, the **+CMCA (MLP)** variant also consistently improves over **Anchors** across all three datasets, despite using a simpler refinement operator. This indicates that the effectiveness of LLP is not restricted to a specific prompt refinement mechanism, but generalizes across different refinement strategies within the latent-anchor-based framework. While the magnitude of the improvement varies across datasets, this is expected because different tasks rely on cross-modal interaction and modality compensation

to different extents. Importantly, the improvement is consistently positive across all three datasets, supporting the effectiveness of the proposed cross-modal anchor interaction beyond generic nonlinear transformation.

C Additional Results

C.1 Robustness Analysis



(a) Missing Text (b) Missing Image (c) Missing Both

Figure 7: Analysis of Robustness to Different Missing Rates on Hateful Memes dataset (Kiela et al., 2020). Notably, LLP maintains stable performance and even exhibits slight improvements under extreme missing conditions ($\eta = 90\%$), highlighting its effectiveness in mitigating missing-modality effects through the proposed latent anchor-based design.

We provide a detailed analysis of model robustness under varying missing rates on the Hateful Memes dataset (Kiela et al., 2020), as illustrated in Figure 7. Specifically, we evaluate different methods under three missing-modality scenarios: missing text, missing image, and missing both modalities, with missing rates ranging from 50% to 90%. Overall, all baseline methods exhibit a clear trend of performance degradation as the missing rate increases across all scenarios. This behavior highlights their strong dependence on input-conditioned features, which become increasingly unreliable when a larger proportion of modalities is absent. In particular, methods such as MMP and MaPLe show sharper performance drops at higher missing rates, indicating limited robustness under severe modality incompleteness.

In contrast, LLP consistently demonstrates superior robustness across all settings. As shown in Figure 7, LLP achieves the smallest performance degradation in all three scenarios. Notably, under extreme missing conditions (e.g., $\eta = 90\%$), LLP not only maintains stable performance but in some cases even shows slight improvements. This phenomenon suggests that LLP is less reliant on potentially corrupted input features and can instead leverage more stable internal representations. Furthermore, LLP exhibits consistent behavior across different types of missing modalities. Whether the missing information lies in the text, image, or both modalities, LLP maintains a relatively flat degradation curve compared to other methods. This indicates that LLP is capable of adapting to diverse missing-modality patterns without overfitting to any specific condition. The observed robustness can be attributed to the proposed latent anchor-based design. Unlike input-conditioned prompting methods, LLP introduces modality-specific latent anchors that serve as input-independent and stable representations of modality-level priors. These anchors provide a reliable foundation for prompt generation, especially when input features are incomplete or noisy. In addition, the dual anchor-induced prompting mechanism enables effective cross-modal interaction, allowing the model to compensate for missing information by leveraging complementary signals from other modalities. Overall, these results demonstrate that LLP significantly mitigates the impact of missing modalities and achieves strong robustness under varying missing rates. This further validates the effectiveness of learning from reliable latent prompts for handling incomplete multimodal data.

C.2 Generalization Across Backbones

To further evaluate the architectural generalizability of LLP, we extend our experiments beyond the CLIP-based architecture to ViLT (Kim et al., 2021). Following the same evaluation protocol, we compare LLP with the state-of-the-art input-conditioned prompting method SyP (Zhang et al.,

Table 4: Generalization of LLP to the ViLT backbone under different missing-modality settings at $\eta = 0.9$. We compare LLP with the state-of-the-art prompting method SyP (Zhang et al., 2025) across three datasets and three missing-modality patterns, including missing image, missing text, and missing both. The consistent improvements across all nine evaluation settings provide additional evidence for the architectural generalizability of LLP.

Dataset	Image	Text	SyP	LLP
MM-IMDb	100%	10%	37.71%	**38.48%**
MM-IMDb	10%	100%	43.76%	**44.10%**
MM-IMDb	55%	55%	41.12%	**43.54%**
Food101	100%	10%	72.11%	**72.53%**
Food101	10%	100%	81.57%	**83.35%**
Food101	55%	55%	78.39%	**79.49%**
Hateful Memes	100%	10%	58.58%	**60.94%**
Hateful Memes	10%	100%	63.32%	**65.83%**
Hateful Memes	55%	55%	63.88%	**64.60%**

2025) under three missing-modality patterns: missing image, missing text, and missing both. All experiments are conducted at $\eta = 0.9$ across MM-IMDb, Food101, and Hateful Memes, resulting in nine evaluation settings in total.

As shown in Table 4, LLP consistently outperforms SyP across all nine evaluation settings, covering all three datasets and all three missing-modality patterns. In particular, LLP achieves improvements of 0.77, 0.64, and 2.42 percentage points over SyP on MM-IMDb under missing image, missing text, and missing both, respectively. On Food101, LLP improves over SyP by 0.42, 1.78, and 1.10 points, while on Hateful Memes, the corresponding improvements are 2.36, 2.51, and 0.72 points. These results demonstrate that the effectiveness of LLP is not restricted to the CLIP-based two-stream architecture or to a particular missing-modality pattern. Instead, the consistent improvements across different datasets, architectures, and missing conditions provide broader evidence for the architectural generalizability and robustness of the proposed input-agnostic latent-anchor prompting strategy.

C.3 Multi-seed Experiments.

To assess the robustness of LLP across different missing-modality patterns, we repeat the experiments under all three missing settings, including missing image, missing text, and missing both, using three independent random seeds at the standard evaluation setting ($\eta = 0.9$). We report the mean performance and standard deviation across the three runs and perform Welch's t-tests to assess the statistical significance of the improvements over the strongest competing methods.

As shown in Table 5, LLP consistently achieves the highest mean performance across all nine evaluation settings, covering three datasets and three missing-modality patterns. Moreover, the standard deviations remain small across different seeds, indicating that the performance of LLP is relatively insensitive to random initialization and experimental variation. The statistical tests further support the consistency of these improvements: all pairwise comparisons between LLP and the competing methods yield $p \leq 0.05$. In several settings, the improvements are highly significant, with p-values as low as 6×10^{-5} on Food101 and 4×10^{-4} on Hateful Memes. Together, these results demonstrate that the performance gains of LLP are not attributable to a particular random seed, but remain robust across repeated runs and diverse missing-modality conditions.

C.4 Complete-Data Performance.

To examine whether LLP preserves performance when all modalities are available, we additionally evaluate the model under the complete-data setting ($\eta = 0$) on MM-IMDb, Food101, and Hateful Memes. We compare a baseline that directly uses the complete multimodal features with the corresponding model equipped with LLP.

Table 5: Multi-seed results under three missing-modality settings ($\eta = 0.9$). We report the mean performance $\pm$ standard deviation over three independent random seeds. The p-values for pairwise comparisons between each baseline and LLP are reported in parentheses.

Dataset	Image	Text	DCP	SyP	LLP
MM-IMDb	100%	10%	49.63% $\pm$ 0.004 (0.02)	50.36% $\pm$ 0.002 (0.05)	**50.77%** $\pm$ 0.002
MM-IMDb	10%	100%	51.75% $\pm$ 0.004 (0.01)	52.99% $\pm$ 0.015 (0.03)	**55.21%** $\pm$ 0.005
MM-IMDb	55%	55%	48.74% $\pm$ 0.005 (0.01)	50.32% $\pm$ 0.006 (0.03)	**51.81%** $\pm$ 0.002
Food101	100%	10%	75.64% $\pm$ 0.004 (0.01)	76.47% $\pm$ 0.005 (0.02)	**78.20%** $\pm$ 0.003
Food101	10%	100%	85.32% $\pm$ 0.001 (0.03)	86.21% $\pm$ 0.002 (0.03)	**86.72%** $\pm$ 0.002
Food101	55%	55%	79.95% $\pm$ 0.002 (6e-5)	80.53% $\pm$ 0.003 (0.03)	**81.68%** $\pm$ 0.001
Hateful Memes	100%	10%	62.49% $\pm$ 0.002 (6e-4)	68.26% $\pm$ 0.023 (0.02)	**70.69%** $\pm$ 0.002
Hateful Memes	10%	100%	63.41% $\pm$ 0.004 (0.02)	64.52% $\pm$ 0.003 (0.04)	**66.06%** $\pm$ 0.007
Hateful Memes	55%	55%	66.92% $\pm$ 0.002 (4e-4)	68.39% $\pm$ 0.004 (0.03)	**69.02%** $\pm$ 0.002

Table 6: Performance under the complete-data setting ($\eta = 0$). LLP consistently improves over the baseline across all three datasets, demonstrating that the proposed prompt learning mechanism does not sacrifice performance when no modality is missing.

Methods	MM-IMDb	Food101	Hateful Memes
Baseline	59.17	90.05	71.18%
+ Ours	**61.80**	**93.36**	**74.39%**

As shown in Table 6, LLP consistently improves performance over the baseline across all three datasets, with gains of 2.63, 3.31, and 3.21 percentage points on MM-IMDb, Food101, and Hateful Memes, respectively. These results demonstrate that LLP does not sacrifice complete-data performance and that the proposed latent-anchor prompt learning mechanism remains beneficial even when no modality is missing.

D More Implementation Details

D.1 Training Objective.

Following prior work (Hu et al., 2024; Zhang et al., 2025), we adopt the standard supervised classification objective to train LLP. Specifically, for Food101 and Hateful Memes, which are single-label classification tasks, we use the standard cross-entropy (CE) loss:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log p_{i,c}, \tag{15}$$

where N is the number of training samples, C is the number of classes, $y_{i,c}$ is the ground-truth label for class c, and $p_{i,c}$ denotes the predicted probability for class c.

For MM-IMDb, which is a multi-label classification task, we instead adopt the binary cross-entropy (BCE) loss:

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{NC}\sum_{i=1}^{N}\sum_{c=1}^{C}\left[y_{i,c}\log p_{i,c} + (1-y_{i,c})\log(1-p_{i,c})\right]. \tag{16}$$

During training, only the proposed prompt parameters and the task-specific fully connected (FC) classification layer are optimized, while the pretrained multimodal backbone remains frozen. The overall training objective is therefore given by

$$\min_{\theta_P,\theta_{\mathrm{FC}}} \mathcal{L}, \tag{17}$$

where θ_P denotes the learnable parameters of LLP and θ_{FC} denotes the parameters of the task-specific FC layer. For Food101 and Hateful Memes, $\mathcal{L} = \mathcal{L}_{\mathrm{CE}}$, while for MM-IMDb, $\mathcal{L} = \mathcal{L}_{\mathrm{BCE}}$.

E LIMITATIONS AND DISCUSSION.

Despite the promising results, our work has several limitations. First, we focus on bimodal image-text settings; extending LLP to additional modalities (e.g., audio or video) remains an important direction for future work. Second, following prior missing-modality prompting methods (Lee et al., 2023; Hu et al., 2024; Zhang et al., 2025), our evaluation adopts a synthetic and approximately missing-at-random protocol, where the missing-modality type is assumed to be known at inference time. This setting enables fair and controlled comparisons with existing methods, but may not fully capture the more complex or structured missingness patterns encountered in real-world applications. Extending LLP to unknown or structured missingness is therefore an important direction for future research. Third, while the input-agnostic latent prompts provide robust modality-level priors under missing modalities, they are not intended to reconstruct instance-specific information that is absent from the input. Incorporating uncertainty estimation and recovering instance-specific information under missing modalities are also promising directions for future work. Fourth, although we demonstrate generalization across multiple two-stream architectures, including CLIP and ViLT, the applicability of LLP to other types of multimodal models, such as unified architectures or large-scale multimodal foundation models, remains unexplored.

Beyond these limitations, LLP offers several potential broader impacts. The missing-modality problem is prevalent in real-world multimodal applications, and our framework provides a new prompting paradigm for improving robustness under such scenarios. Moreover, LLP achieves strong performance with high parameter efficiency by optimizing only a small set of learnable parameters, which may facilitate practical deployment. Our findings also highlight the potential of learning from input-independent latent prompts and may inspire further research on robust prompt-based multimodal learning.

Nevertheless, LLP does not eliminate the risks associated with missing or unreliable modalities. In sensitive applications, such as hateful-content moderation, severe missingness may lead to unreliable predictions or amplify existing dataset and model biases. Therefore, LLP should not be interpreted as eliminating such risks, and careful uncertainty assessment, validation, and human oversight remain important when deploying multimodal systems in sensitive settings.