BIP: Bi-level Information Transfer and Completion Prompting for Visual Recognition with Missing Modalities

Haoran Fan[1], Xu Han[2], Xianglong Bao[2], Zheng Gao[1], Qi Fan[3⋆], Yang Song[1], and Jiaojiao Jiang[1]

[1] University of New South Wales, Sydney, Australia
[2] Griffith University, Brisbane, Australia
[3] Nanjing University, Nanjing, China
haoran.fan@unsw.edu.au, fanqi@nju.edu.cn

Abstract. Existing multimodal visual recognition methods addressing missing modalities have achieved great progress, mainly based on the feature reconstruction framework. However, feature reconstruction often introduces distribution inconsistency between synthesized and real features. Considering each modality as a different view of the same instance, each view is noisy and incomplete, but important representations, such as semantic details, tend to be shared between all modalities. Thus, we propose that AlignPrompt alleviates the distribution inconsistency problem by completing and aligning the semantic representation between synthesized and non-missing prompts. AlignPrompt is an instance-level prompt, implemented via a modal-shared-specific structure including a modal-shared factor and a modal-specific factor. Specifically, for each modality pair, AlignPrompt aligns modality-specific factors of all modalities in a shared space by the modal-shared factor as an anchor, and completes the semantic representation by maximizing the mutual information between the synthesized factor and the non-missing factor. Besides, we also propose HyperPrompt to globally regulate the association of prompts of each layer based on different modalities missing scenarios: it transfers global prompts across layers by the inter-layer modules generated from an independent network. This collaborative prompts framework achieves state-of-the-art performance on multiple multimodal classification datasets while maintaining parameter efficiency.

Keywords: multimodal learning · prompt learning · missing modality

1 Introduction

Large-scale visual-language models (VLMs) [17, 21, 27] excel at multiple tasks, but their performance typically relies on complete, paired inputs. However, in real-world scenarios, data is often incomplete due to sensor failures, privacy concerns, or acquisition costs. The missing modality problem causes sharp

⋆ Corresponding author.

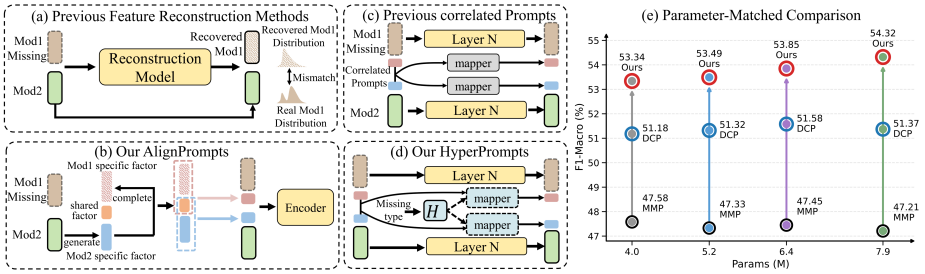

Fig. 1: Comparison of previous methods and our two prompts for missing modalities. (a) The feature reconstruction method [10, 30] predicts missing features but suffers from distribution inconsistency. (b) Our AlignPrompt aligns and completes instance-level semantics at the prompt level. (c) Previous correlated prompts [11] share mappers across missing types, causing interference. (d) Our HyperPrompt uses a missing-type-aware hypernetwork to generate per-layer mappers for prompt transfer. (e) Parameter-matched comparison of our method with MMP [20] and DCP [11] on MM-IMDb [1].

performance degradation and poses a critical bottleneck to practical deployment [11, 20, 24, 35].

The missing modality problem means that subsets of input modalities (image and text in this work) may be missing during model training and testing [20]. The common and intuitive approach is to reconstruct the missing modality using features from the available modalities [10, 30] (see Fig. 1 (a)). But these methods often suffer from distribution inconsistency between synthesized and real features, which stems from the following two core issues. When predicting missing modal features from available modalities, the inter-modality mapping is often non-deterministic (e.g., a single text can correspond to many plausible images) [3, 42], and an inherent modality distributions gap [22]. Consequently, the feature reconstruction method is unstable.

Existing methods attempt to solve this problem by leveraging the distribution relationships learned from complete modal pairs, such as applying distribution constraints to the reconstructed features [28, 29, 31, 34], or generating aggregations from the visible modality to align the modality distributions [9, 23, 25, 32]. Despite their success, these methods still struggle with the distribution inconsistency problem, bounded by inter-modality non-deterministic mapping and the modality distribution gap.

To avoid the distribution inconsistency caused by feature reconstruction, we treat the semantic representation as modality-invariant information and address the modality missing problem from the perspective of information consistency between modalities' prompts, thereby avoiding unstable feature reconstruction. We propose AlignPrompt, a novel strategy that directly completes and aligns the semantic representation of modality pairs with the prompt guidance. AlignPrompt is an instance-level prompt, consists of a modal-shared factor and a modal-specific factor. As shown in Fig. 1 (b), for each imodality of one instance, we generate an AlignPrompt and align them via the modal-shared factor in

a shared low-rank subspace, which alleviates the inter-modal distribution gap. Then, we use the semantic representation of the non-missing modality to synthesize the modal-specific factor of the missing modality by maximizing the mutual information between the synthesized factor and the non-missing factor.

In addition to the instance-level guidance, we also propose a global-level HyperPrompt, providing instance-agnostic, global guidance based on the modality missing scenario. We use this prompt to adapt the model to different modality missing scenarios and to control the associations among prompts of different layers. As shown in Fig. 1 (d), our HyperPrompt employs a shared hypernetwork [8] to generate dynamic, per-layer mappers. These mappers are conditioned on the missing-type scenario, enabling associative control of global prompt information across the layers.

Our approach is fundamentally different from conventional feature reconstruction [10, 30]. We are the first to perform mutual information completion on the model's prompt guidance. During training, we assume one modality is missing on the complete prompt pair, and supervise the completion using the real prompt, ensuring consistency with the distribution of the real prompt while preserving the semantic representation of the non-missing modality.

Recently, some methods [12, 14, 45, 46] also try prompt tuning to solve the modality missing problem, offering a lightweight alternative. They apply static, layer-independent prompts [20] or inter-layer correlations [11]. Despite their success, they typically lack a missing-type-aware and effective information transfer. For example, as shown in Fig. 1 (c), DCP [11] fails to prevent interference between prompts for different modality missing types, limiting their robustness on different missing types scenarios. Our HyperPrompt uses a hypernetwork based on modality missing type to control the information transfer across layers to improve robustness. Our main contributions are:

- We propose a collaborative prompts framework that combines instance-level semantic completion with global prompt information transfer across layers for missing modality robustness.
- We introduce a novel completion mechanism on prompts, completing the semantic representation of the missing modality to avoid the distribution inconsistency of feature reconstruction.
- We design a missing-type-aware transfer mechanism, using a hypernetwork, to adaptively control the associations among different layers.
- Our method achieves SOTA performance on multiple benchmarks while maintaining high parameter efficiency.

2 Related Work

Robustness to Missing Modalities. Prior work addressing this problem is divided into two main strategies. The first is feature-level completion or reconstruction. MMIN [44] uses cascade autoencoders to reconstruct latent features, DiCMoR [34] uses normalizing-flow-based to enforce distribution consistency between restored and real features, and MMPred [16] predicts the embeddings of

missing modality. As argued in the introduction, these methods, while effective, are suffer from distributional inconsistency [34]. This dilemma stems from non-unique mapping and the distribution gap between modalities.

The second strategy learns robust representations without explicit reconstruction. SMIL [25] applies Bayesian meta-learning. ShaSpec [32] learns shared versus specific by factorizing the model into shared and modality-specific parts. TATE [40] leverages missing type priors as input to guide the model. While innovative, these approaches often require specialized architectures or full model fine-tuning, which can be resource-intensive.

Our work aims to avoid both issues. Our AlignPrompt module recovers semantic information of missing modality (not features) in the prompt space, thereby avoiding the dilemma for feature-level reconstruction.

Parameter-Efficient Prompt Learning. Prompt tuning is a lightweight and effective alternative. This field has evolved from text-only prompts (e.g., CoOp [46], CoCoOp [45], and their variants [38,47]) to coupled vision-language prompts (MaPLe [14]) and other parameter-efficient decoupling designs (DePT [41], TCP [39], LPI [37]).

When applied to missing modalities, MMP [20] introduced the missing-aware prompts to handle diverse missing-modality cases. DCP [11] advances this concept by modeling inter-layer correlations. However, these methods address only part of the problem. Their mechanisms are typically static or locally adaptive, lacking the unified, missing-type-aware control signal identified in our introduction as a gap. While more recent works like SCP [26] (using semantic memory) or DisPro [36] (using distillation) are more adaptive, their focus is on sample-level adaptation. They do not provide an explicit mechanism to globally control the network's information flow based on the missing scenarios. Our HyperPrompt provides a global and hierarchical control mechanism to address this gap.

3 Method

Problem Definition. For simplicity and without loss of generality, the inputs of our task are considered as M = 2 modalities m_1 and m_2 (e.g., image and text). To be more specific, let $D = \{D^c, D^{m_1}, D^{m_2}\}$ be a multimodal dataset. We set the modality-complete subset as $D^c = \{x_{m_1}, x_{m_2}, y\}$, where x denotes the input and y the label. $D^{m_1} = \{x_{m_1}, y\}$ and $D^{m_2} = \{x_{m_2}, y\}$ are denoted as the modality-incomplete subsets, in which one modality is missing (e.g., text-only or image-only). Such two missing types can appear in both training and inference periods, and their frequencies may differ. To keep a unified input, dummy placeholders are used to represent the input of a missing modality.

3.1 Preliminaries

Prompt Learning for Missing Modalities. MMP [20] first addresses missing modalities via prompt tuning. For a task with multiple modalities, it enumerates the missing cases (e.g., with M=2: *complete*, *image-only*, *text-only*) and assigns

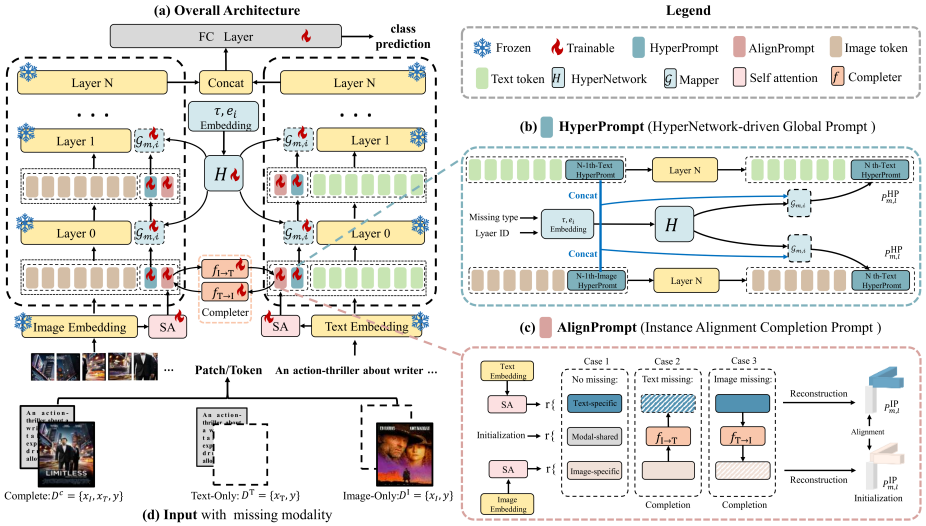

Fig. 2: Overview of our framework. (a) *Overall architecture.* We freeze the image/text encoders and train only light modules. Each layer consumes two prompts: a blue **HyperPrompt** and a red **AlignPrompt**. (b) *HyperPrompt.* A shared hypernetwork H takes the missing type τ and the layer ID e_l as input and outputs a layer-wise mapper $g_{m,l}$ to generate $P^{\mathrm{HP}}_{m,l}$. This provides layer-correlated guidance that selects and passes forward information relevant to the current missing pattern. (c) *AlignPrompt.* We factorize the instance prompt into modality-shared and modality-specific factors. When a modality is missing, a lightweight completer $f_{I\to T}$ or $f_{T\to I}$ predicts the missing-side factors from the visible modality and reconstructs $P^{\mathrm{IP}}_{m,l}$, which are aligned by the modality-shared factor. (d) *Input forms.* The inputs include complete samples $D^c = \{x_I, x_T, y\}$, text-only D^T, and image-only D^I at training and testing.

a learnable prompt set to each case. For each sample, it selects the matched prompts for its missing cases. Then, it places them before the input tokens, which can help MMP to make predictions under corresponding missing settings.

3.2 Overall Framework

We consider two modalities $m \in \{\mathrm{I}, \mathrm{T}\}$ with L-layer encoders and dataset $D = \{D^c, D^{\mathrm{I}}, D^{\mathrm{T}}\}$, where $D^c = \{x_{\mathrm{I}}, x_{\mathrm{T}}, y\}$ and $D_{\mathrm{I}}, D_{\mathrm{T}}$ are incomplete cases.

As shown in Fig. 2, our framework builds upon the MMP [20] model by introducing two complementary prompts. (1) **HyperPrompt** (HyperNetwork-driven Global Prompt) employs a shared hypernetwork H conditioned on the missing-type/task embedding τ to generate the parameters of a layer-wise prompt-transfer module, which then produces prompts $P^{\mathrm{HP}}_{m,l}$ for each layer. This preserves cross-sample adaptability yet confines missing-type perturbations to mapping parameters. (2) **AlignPrompt** (Instance alignment completion prompt) assigns each instance a low-rank structure comprising modality-shared (cross-modal) and modality-specific factors. These factors reconstruct to produce our

AlignPrompt $P^{\mathrm{IP}}_{m,l}$ and align semantics by the anchors - modality-shared factors; when a modality is missing, a Cross-Modal Completion head C predicts absent coefficients from the visible side, so the available modality continues to guide semantics. At each layer, we merge the two into $\tilde{P}_{m,l}$ and prepend $\tilde{P}_{m,l}$ to the inputs and intermediate features of both encoders for comparability with prior placements. We keep the backbone frozen, and only the H, C, and prompts.

3.3 HyperNetwork-driven Global Prompt

To enable the mappers at each layer to distinguish the missing type of the input and filter the useful information to transfer. We use a shared, missing-type-conditioned hypernetwork that generates the weights of lightweight, modality-specific transfer mappers at each layer [11, 20]. Our structure follows the hypernetwork recipe [8].

For modality $m \in \{\mathrm{I}, \mathrm{T}\}$ and layer i ($i = 0, \ldots, L-1$), the following equation demonstrates the process of the current prompt–feature pair update.

$$\big(P^{\mathrm{HP}}_{m,i-1},\, X_{m,i-1}\big) \mapsto \big(P^{\mathrm{HP}}_{m,i},\, X_{m,i}\big) \tag{3}$$

here $P^{\mathrm{HP}}_{m,i}$ is the prompt for modality m at layer i (generated by the hypernetwork) and $X_{m,i}$ is the hidden feature of modality m after layer i. We use a hypernetwork-driven, modality-separate transfer to produce these prompts at each layer. Let τ be the missing-type/task embedding and e_i the layer-id embedding (both learnable). We first fuse the two prompts from the previous layer via a cross-modal fusion:

$$\tilde{p}^{(i-1)} \;=\; \mathrm{concat}\big(P^{\mathrm{HP}}_{\mathrm{I},\,i-1},\, P^{\mathrm{HP}}_{\mathrm{T},\,i-1}\big) \tag{4}$$

here $P^{\mathrm{HP}}_{\mathrm{m},\,i-1}$ is the per-modality HyperPrompt used for fusion at layer $i{-}1$; $\tilde{p}^{(i-1)}$ matches the mapper input width.

We then obtain the layer-i HyperPrompt for modality m using a modality-specific mapper $\mathcal{G}_{m,i}$. Its parameters come from the shared hypernetwork H, conditioned on τ, and e_i; $[\cdot \,\|\, \cdot]$ denotes concatenation:

$$P^{\mathrm{HP}}_{m,i} = \mathcal{G}_{m,i}\Big(\tilde{p}^{(i-1)};\; H\big([\tau \,\|\, e_i\,]\big)\Big), \quad m \in \{\mathrm{I}, \mathrm{T}\}. \tag{5}$$

Both modalities share H but use their own $\mathcal{G}_{m,i}$; each $\mathcal{G}_{m,i}$ is a two-layer bottleneck MLP with LayerNorm (choose a hidden-width reduction ratio r, e.g., r=1/16).

3.4 Instance Alignment Completion Prompt

Our Instance Alignment Completion Prompt (AlignPrompt) aims to avoid the distributional inconsistency of feature reconstruction. It employs a low-rank modal-shared-specific structure [18, 37] to align modal-specific factors and complete missing modal prompts by maximizing the mutual information between the predicted factor and the available one.

Low-rank Instance-Alignment Structure. Let K denote the prompt length and let d_m be the embedding size of the Modal m branch. We parameterize the layer-0 prompt with a rank-R form consisting of a shared factor $B \in \mathbb{R}^{K\times R}$ (a learnable parameter shared across modalities/instances) and modality-specific factors $C^{(m)}(x) \in \mathbb{R}^{d_m\times R}$ that depend on the input x.

The sharing factors are initialized randomly, and the modality-specific factors are produced from self-attention over shallow features $X_{m,0}(x)$:

$$C^{(m)}(x) = \mathrm{reshape}\Big(\mathrm{SelfAtt}\big(X_{m,0}(x)\big)\Big) \in \mathbb{R}^{d_m\times R} \tag{6}$$

The layer-0 instance prompt is reconstructed in this low-rank form and injected by prepending a prompt of length K to the modality-m sequence (same placement as Sec. 3.3):

$$P^{\mathrm{IP}}_{m,0}(x) = \sum_{r=1}^{R} B_{:,r}\, C^{(m)\top}_{:,r}(x) = B\, C^{(m)}(x)^{\top} \in \mathbb{R}^{K\times d_m} \tag{7}$$

We apply this construction only at the input layer; for the deeper layers, the prompts are inherited from the previous layers and updated together with intermediate features.

Completion via Maximum Mutual Information. Following Information Maximization [2], we start from the Barber–Agakov (BA) lower bound on mutual information:

$$\mathcal{I}(Z;Y) \;\geq\; \mathbb{E}_{p(z,y)}\big[\log q_\psi(y \mid z)\big] + H(Y) \;\triangleq\; \mathcal{L}_{\mathrm{BA}}. \tag{8}$$

here $\mathcal{I}(Z;Y)$ is the mutual information between the Z and Y. $H(Y)$ quantifies the uncertainty of the prediction target Y before observing Z and it is determined by the data distribution $p(y)$. The term $\mathbb{E}_{p(z,y)}[\log q_\psi(y \mid z)]$ is the expected conditional log-likelihood of the observed Y given Z; maximizing it improves the accuracy of predicting Y from Z.

For a modal-complete pair $(x_{\mathrm{I}}, x_{\mathrm{T}})$ at input layer, we assume one of them is missing. For example, we set $Z = C^{(m)}(x) \in \mathbb{R}^{d_m\times R}$ and $Y = C^{(\bar{m})}(x) \in \mathbb{R}^{d_{\bar{m}}\times R}$, where $m \in \{\mathrm{I},\mathrm{T}\}$ indexes the complete side and $\bar{m}$ the missing side. We directly predict the missing-side factor from the complete-side factor using two lightweight completion heads. The two-layer MLP is applied along the rank axis so that each column $C^{(m)}(x)_{:,r} \in \mathbb{R}^{d_m}$ maps to $\widehat{C}^{(\bar{m})}(x)_{:,r} \in \mathbb{R}^{d_{\bar{m}}}$ for $r = 1,\ldots,R$:

$$\widehat{C}^{(\mathrm{T})}(x) = f_{\mathrm{I}\to\mathrm{T}}\big(C^{(\mathrm{I})}(x)\big), \qquad \widehat{C}^{(\mathrm{I})}(x) = f_{\mathrm{T}\to\mathrm{I}}\big(C^{(\mathrm{T})}(x)\big) \tag{9}$$

We then parameterize the BA term with a diagonal-Gaussian conditional on Y. Let $\mathbf{M} \triangleq \widehat{C}^{(\bar{m})}(x) \in \mathbb{R}^{d_{\bar{m}}\times R}$ denote the completion produced by the MLP head, and let $\mathbf{V} \triangleq \sigma^2_\psi\big(C^{(m)}(x)\big) \in \mathbb{R}^{d_{\bar{m}}\times R}_{+}$ denote the entry-wise variance predicted by a small variance head with softplus to ensure nonnegativity. The factorized Gaussian can be expressed as:

$$q_\psi\big(C^{(\bar{m})}(x) \,\big|\, C^{(m)}(x)\big) = \prod_{i=1}^{d_{\bar{m}}}\prod_{r=1}^{R} \mathcal{N}\Big(C^{(\bar{m})}_{i,r}(x)\;;\; M_{i,r},\; V_{i,r}\Big), \tag{10}$$

At inference, we use the completion head's output, as the completed factor. Next, we reconstruct the full K-length prompt by composing the shared factor $B \in \mathbb{R}^{K\times R}$ with the modality-specific factors $C^{(\bar{m})}(x)$ and $\widehat{C}^{(\bar{m})}(x)$.

$$\widehat{P}_{\bar{m},0}(x) = B\,\widehat{C}^{(\bar{m})}(x)^{\top}, \qquad P_{\bar{m},0}(x) = B\,C^{(\bar{m})}(x)^{\top}, \tag{11}$$

both in $\mathbb{R}^{K\times d_{\bar{m}}}$.

$$\mathcal{L}_{\mathrm{recP}} = \big\|\widehat{P}_{\bar{m},0}(x) - P_{\bar{m},0}(x)\big\|_1. \tag{12}$$

$$\mathcal{L}_{\mathrm{mom}} = \big\|\mathrm{mean}_K(\widehat{P}_{\bar{m},0}) - \mathrm{mean}_K(P_{\bar{m},0})\big\|_2^2 + \big\|\mathrm{var}_K(\widehat{P}_{\bar{m},0}) - \mathrm{var}_K(P_{\bar{m},0})\big\|_1. \tag{13}$$

As shown in Algorithm 1, when training, for complete samples, we simulate one side as missing and predict the missing-side factor by the complete-side factor(detached). We evaluate (8) with $Z = C^{(m)}(x)$ and $Y = C^{(\bar{m})}(x)$, and apply (12)–(13). On truly missing-modality samples, we inject $\widehat{P}_{\bar{m},0}(x)$ downstream and train only with the CLIP-style contrastive objective on available pairings. The overall objective is

$$\mathcal{L} = \lambda_{\mathrm{BA}} \cdot \big(-\mathcal{L}_{\mathrm{BA}}\big) + \lambda_{\mathrm{recP}}\mathcal{L}_{\mathrm{recP}} + \lambda_{\mathrm{mom}}\mathcal{L}_{\mathrm{mom}} + \mathcal{L}_{\mathrm{CLIP}}, \tag{14}$$

where $\lambda_{\{\cdot\}}$ balances the terms; BA losses are enabled only when both modalities are present.

4 Experiments

Datasets. Following prior works [11,20] on missing-modality multimodal recognition, we evaluate on three classification datasets using their official splits and metrics.

MM-IMDb [1] pairs each movie poster with its plot synopsis and targets multi-label genre prediction (23 genres). The corpus contains 25,959 movies with an official stratified split of 15,552 train, 2,608 validation, and 7,799 test examples. We follow the common protocol and report Macro F1 on the multi-label classification task.

UPMC Food-101 [6] is a large-scale image–text collection for food recognition (classification), aligned to the same 101 categories as ETHZ Food-101 [4]. Each sample consists of a food image and the HTML page it was crawled from (recipe/title text). We use the official split with 67,988 training and 22,716 test pairs and report top-1 accuracy.

Hateful Memes [15] is a binary meme classification benchmark designed to require joint image–text reasoning. The challenge set contains over 10K newly created memes and includes "benign confounders" that flip the label when only one modality is considered. We use the official splits (8,500 train, 500 dev, 1,000 test); the dev and test sets are balanced. Following prior work, we use AUROC as the primary metric.

Table 1: Comparison with CoOp [46], MMP [20], MaPLe [14], DePT [41], and DCP [11] on MM-IMDb [1], Food101 [6], and Hateful Memes [15] under various missing-modality cases with different missing rates. $^{\uparrow\Delta}$ denotes the gain of **Ours** over the baseline **MMP** [20].

Datasets	Missing rate η	Train/Test		Validation set						Testing set					
		Image	Text	CoOp	MMP	MaPLe	DePT	DCP	**Ours**	CoOp	MMP	MaPLe	DePT	DCP	**Ours**
MM-IMDb (F1-Macro)	50%	100%	50%	51.23	52.07	52.76	53.87	55.23	**56.62**$^{\uparrow 4.55}$	48.06	48.88	49.58	50.64	52.13	**55.77**$^{\uparrow 6.89}$
		50%	100%	53.04	54.52	55.26	56.04	57.32	**57.61**$^{\uparrow 3.09}$	49.89	51.46	52.32	52.78	54.32	**57.39**$^{\uparrow 5.93}$
		75%	75%	51.46	52.12	52.87	54.02	55.45	**57.65**$^{\uparrow 5.53}$	48.37	49.32	49.56	50.87	52.32	**55.85**$^{\uparrow 6.53}$
	70%	100%	30%	47.26	48.23	48.75	49.87	51.35	**53.31**$^{\uparrow 5.08}$	44.13	45.64	45.52	46.38	48.52	**52.28**$^{\uparrow 6.64}$
		30%	100%	52.32	53.21	53.98	55.04	56.21	**56.28**$^{\uparrow 3.07}$	48.82	50.52	50.64	52.13	53.14	**56.92**$^{\uparrow 6.40}$
		65%	65%	50.22	51.34	52.31	53.17	54.24	**54.97**$^{\uparrow 3.63}$	46.84	48.12	49.16	50.32	51.42	**53.89**$^{\uparrow 5.77}$
	90%	100%	10%	47.86	48.84	50.12	50.98	52.36	51.79$^{\uparrow 2.95}$	44.76	45.32	46.84	47.56	49.26	**49.55**$^{\uparrow 4.23}$
		10%	100%	51.65	52.36	53.14	54.12	55.42	54.55$^{\uparrow 2.19}$	48.32	49.12	50.13	50.88	52.22	**54.74**$^{\uparrow 5.62}$
		55%	55%	47.44	48.04	48.82	49.98	51.26	**52.43**$^{\uparrow 4.39}$	44.12	44.87	45.12	46.54	48.04	**52.67**$^{\uparrow 7.80}$
Food101 (Accuracy)	50%	100%	50%	77.36	78.24	79.87	80.24	82.33	**84.07**$^{\uparrow 5.83}$	77.45	77.89	79.64	80.16	82.11	**83.59**$^{\uparrow 5.70}$
		50%	100%	86.98	87.12	87.48	87.85	89.23	**89.40**$^{\uparrow 2.28}$	87.02	87.16	87.35	82.14	89.12	**89.28**$^{\uparrow 2.12}$
		75%	75%	81.76	81.98	82.58	83.26	85.25	**85.79**$^{\uparrow 3.81}$	81.24	81.72	82.34	83.12	85.24	**85.67**$^{\uparrow 3.95}$
	70%	100%	30%	76.65	76.74	76.87	76.87	79.18	**80.96**$^{\uparrow 4.22}$	76.34	76.52	77.02	77.34	78.87	**84.05**$^{\uparrow 7.53}$
		30%	100%	85.21	86.12	86.36	86.52	87.53	**87.73**$^{\uparrow 1.61}$	84.78	85.64	85.89	86.12	87.32	**87.46**$^{\uparrow 1.82}$
		65%	65%	79.14	79.56	80.06	81.85	82.38	**83.56**$^{\uparrow 4.00}$	78.87	79.12	79.84	81.46	81.87	**83.35**$^{\uparrow 4.23}$
	90%	100%	10%	72.65	73.74	73.25	74.22	75.54	**77.97**$^{\uparrow 4.23}$	71.87	73.14	73.46	74.12	75.26	**77.87**$^{\uparrow 4.73}$
		10%	100%	82.16	82.78	83.42	84.02	86.26	**86.82**$^{\uparrow 4.04}$	81.67	82.14	83.12	83.56	85.78	**85.97**$^{\uparrow 3.83}$
		55%	55%	77.36	77.78	78.26	78.66	80.39	**81.70**$^{\uparrow 3.92}$	76.46	76.58	77.85	78.12	79.87	**80.52**$^{\uparrow 3.94}$
Hateful Memes (AUROC)	50%	100%	50%	58.32	58.56	58.78	59.31	60.24	**65.81**$^{\uparrow 7.25}$	60.56	60.31	60.87	61.87	62.32	**70.71**$^{\uparrow 10.40}$
		50%	100%	60.34	61.12	61.34	61.78	62.34	**63.35**$^{\uparrow 2.23}$	62.41	62.35	63.13	63.88	64.46	**66.51**$^{\uparrow 4.16}$
		75%	75%	62.34	62.87	63.14	63.24	63.78	**65.65**$^{\uparrow 2.78}$	64.87	65.84	65.46	65.86	66.02	**67.81**$^{\uparrow 1.97}$
	70%	100%	30%	58.54	59.02	59.36	60.02	60.56	**65.57**$^{\uparrow 6.55}$	60.74	61.12	61.26	61.56	62.82	**69.56**$^{\uparrow 8.44}$
		30%	100%	60.12	60.78	61.32	61.54	62.32	**63.24**$^{\uparrow 2.46}$	62.74	63.24	63.14	63.48	64.12	**67.80**$^{\uparrow 4.56}$
		65%	65%	62.34	62.56	63.12	63.32	63.78	**65.17**$^{\uparrow 2.61}$	64.82	65.04	65.23	65.48	66.08	**66.62**$^{\uparrow 1.58}$
	90%	100%	10%	58.02	57.34	58.32	59.02	60.34	**63.34**$^{\uparrow 6.00}$	60.03	57.21	60.74	61.14	62.08	**67.86**$^{\uparrow 10.65}$
		10%	100%	59.02	59.32	60.21	60.56	61.34	**62.52**$^{\uparrow 3.20}$	61.46	61.52	61.87	62.42	63.87	**64.52**$^{\uparrow 3.00}$
		55%	55%	62.32	62.56	63.24	63.78	64.34	**64.73**$^{\uparrow 2.17}$	64.32	63.13	64.85	65.37	66.78	**67.32**$^{\uparrow 4.19}$

Implementation Details. Following prior prompt-based approaches for missing-modality recognition [11,20,43], we use a frozen CLIP backbone with a ViT-B/16 image encoder and a Transformer text encoder [5, 27]. Images are resized to 224×224 and tokenized into 16×16 patches; texts are processed by the CLIP tokenizer with a maximum length of 77 tokens. Unless otherwise specified, backbone parameters are kept frozen, and we only optimize the task head and learnable prompts. In line with [11], we prepend prompts of length L_p=36 to features at M=6 selected transformer layers. For missing inputs, we skip the corresponding encoder and substitute a zero-filled feature tensor as in [11]; for reference, [20] uses blank-text / uniform-image dummies with similar effect at the feature level. Optimization follows prompt-learning baselines [11, 20]: Adam-type optimizer with a short warm-up (10% of steps) and linear decay (default learning rate 1×10^{-2} and weight decay 2×10^{-2} unless noted).

Metrics. We follow the metrics used in prior studies [11, 20]. On MM-IMDb, we report macro-F1 for multi-label genre prediction [1]. On UPMC Food-101 we report top-1 accuracy for 101-way classification [4, 33]. On Hateful Memes, we use AUROC as the primary score following the challenge protocol [15].

Algorithm 1: Our mask training algorithm with MI.

Input : Batch $\{(x_{\mathrm{I}}, x_{\mathrm{T}})\}$; mask $M \in \{\text{Complete}, \text{I-only}, \text{T-only}\}$.
Output : Updated completion heads and q_ψ.

Factorize: $C^{(\mathrm{I})}, C^{(\mathrm{T})} \leftarrow \textsc{SelfAttn}(X_{\mathrm{I},0}, X_{\mathrm{T},0})$ `// specific factors`
if $M =$ *Complete* **then**
  Sample a missing direction $m \in \{\text{I-only}, \text{T-only}\}$ `// mask training`
  if *assume* $m =$ *I-only* **then**
    $\widehat{C}^{(\mathrm{T})} \leftarrow f_{\mathrm{I}\to\mathrm{T}}\big(C^{(\mathrm{I})}\big)$;
    compute $\mathcal{L}_{\mathrm{BA}}$ (Eq. (8)) `// complete` T
    $\widehat{P}_{\mathrm{T},0} \leftarrow B\widehat{C}^{(\mathrm{T})\top},\ P_{\mathrm{T},0} \leftarrow BC^{(\mathrm{T})\top}$
  else if *assume* $m =$ *T-only* **then**
    $\widehat{C}^{(\mathrm{I})} \leftarrow f_{\mathrm{T}\to\mathrm{I}}\big(C^{(\mathrm{T})}\big)$;
    compute $\mathcal{L}_{\mathrm{BA}}$ (Eq. (8)) `// complete` I
    $\widehat{P}_{\mathrm{I},0} \leftarrow B\widehat{C}^{(\mathrm{I})\top},\ P_{\mathrm{I},0} \leftarrow BC^{(\mathrm{I})\top}$
  Compute $\mathcal{L}_{\mathrm{recP}}, \mathcal{L}_{\mathrm{mom}}$ and $\mathcal{L}_{\mathrm{CLIP}}$
else
  $\mathcal{L}_{\mathrm{task}} \leftarrow \mathcal{L}_{\mathrm{CLIP}}$
Update completion heads and q_ψ by minimizing the total loss $\mathcal{L}$ (Eq. (14))

Setting of Missing Modality. We evaluate models when any modality may be absent during training or testing [11,20]. Let η be the proportion of modality-incomplete samples. For vision–language tasks, we consider three cases: *missing-text*, *missing-image*, and *missing-both*. With rate η, the first two correspond to η portion text-only (or image-only) plus $(1-\eta)$ portion complete pairs; for *missing-both*, the mix is $\eta/2$ text-only, $\eta/2$ image-only, and $(1-\eta)$ complete pairs. Unless stated otherwise, we set η=70% following prior work [11,20]. This protocol naturally extends to M modalities by distributing the incomplete portion across single-modality-missing patterns and leaving $(1-\eta)$ for complete samples [11,20].

4.1 Comparison with other methods

In Tab. 1, we conduct experiments on three widely used datasets to validate the effectiveness of our method in various missing-modality scenarios. To comprehensively evaluate our method, we conduct extensive experiments across various missing-modality scenarios on three standard benchmarks: MM-IMDb [1], UPMC Food-101 [6], and Hateful Memes [15]. We compare our method against a suite of state-of-the-art baselines, namely CoOp [46], MMP [20], MaPLe [14], DePT [41], and DCP [11]. The evaluation protocol assesses performance under three missing data rates ($\eta = 50\%$, 70%, and 90%) for scenarios involving missing text, missing images, or missing both modalities.

As evidenced by the results in Tab. 1, our method establishes a new state-of-the-art, consistently outperforming all baselines across a diverse range of missing-modality scenarios. On the MM-IMDb [1], Food101 [6], and Hateful Memes [15] datasets, our average performance improved over DCP [11] by 3.08, 1.37, and 3.35, respectively. Notably, our method achieves significant performance gains

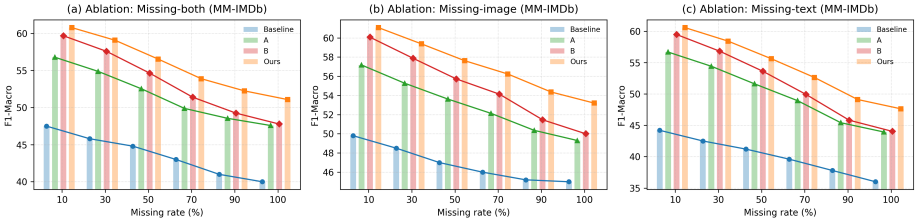

Fig. 3: An ablation study assessing the robustness of our proposed model to missing modalities on the MM-IMDb [1] Test set. We compare our full model ('Ours') with (1) a naive baseline that discards features from absent modalities; (2) an intermediate version equipped only with HyperNetwork-driven Global Prompts ('A'); and (3) a version only with Instance Alignment Completion Prompt ('B'). All models are evaluated under three distinct scenarios—missing image, missing text, and missing both—across a full spectrum of data missingness rates (10% to 100%).

over the previous leading method, DCP [11]. A deeper dive into the results reveals distinct modality impacts contingent on the dataset. For instance, performance on MM-IMDB [1] and Food101 [6] is more severely affected by text removal, likely due to their dependence on semantic cues. In contrast, Hateful Memes [15] experiences a greater performance drop from image removal, showing the differential importance of missing type. This observation validates the importance of the core design of our method: the adaptive modal fusion mechanism based on missing type, and the Prompt-level modal information completion mechanism. We also compared our method with the feature-based reconstruction method in Tab. 4 when the image missing rate on MM-IMDb was 0.5, which shows our method significantly outperformed the feature reconstruction method.

4.2 Ablation Study

Effectiveness of Proposed Prompts. As shown in Fig. 3, we train a single model under the 70% missing-both setting and evaluate on three test scenarios (both/image/text missing) over 10–100% missing rates. The baseline in Fig. 3 fine-tunes only the classifier using frozen encoders and achieves much lower performance. This result shows the inherent limitations of a static feature extraction model that cannot adapt to input with missing modality. Our proposed models consistently outperform the baseline across all rates and scenarios, with a smoother degradation at high missing rates. A (HyperNetwork-driven Global Prompt) provides structural guidance. It generates per-layer mappers based on the missing type, which helps stabilize features at high missing rates. B (Instance Alignment Completion Prompt) recovers instance-level semantics for the missing modality using the visible one. This method performs best at mild-to-moderate missing rates. As the missing rate increases, the performance gap between B and A narrows. This suggests a shift in reliance from semantic completion (B) to structural robustness (A). Furthermore, combining both components achieves the best performance in all scenarios.

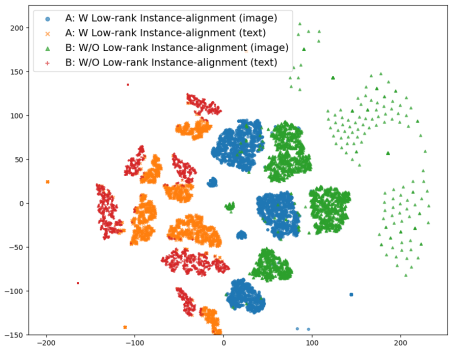

Fig. 4: t-SNE on MM-IMDb [1]. Our AlignPrompt (A: blue circles for image, orange crosses for text) forms tighter and more co-located image/text clusters than the common prompt (B: green triangles for image, red pluses for text), which verifies the alignment effect of the shared factors.

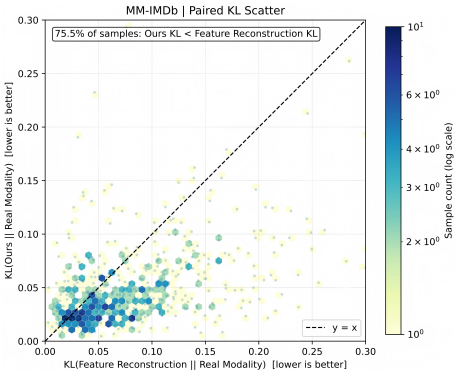

Fig. 5: KL on MM-IMDb [1]. For image-missing samples in the Validation set, we compare the KL divergence between predicted logits distributions and the real modality. Our prompt completion method (y-axis) yields lower KL than feature reconstruction (x-axis) on 75.5% of samples.

Analysis for HyperPrompt. As shown in Tab. 2, the upper block compares three prompting modules' results on MM-IMDB [1]: DCP's Correlated Prompt, our HyperPrompt (w/o Missing type), and our HyperPrompt (w Missing type). Adding the missing-type condition yields the best F1-Macro, indicating that conditioning on the missing pattern consistently benefits multimodal classification. The lower block analyzes the TXT-missing case: for the image side, it reports CKA(%) [19] between the input-layer prompt and the mapper outputs at Layers 4/5/6. Across layers, the w (with missing type) column typically shows higher CKA (Centered Kernel Alignment) than w/o, implying that the conditional hypernetwork better preserves the useful structure of the input prompt and suppresses cross-prompt interference during mapping. CKA is a representation-similarity metric that compares the sample-similarity structures of two representations.

Table 2: Analysis of HyperPrompt and its effectiveness in preventing crosstalk (via CKA) on MM-IMDb [1].

Modules	**F1-Macro (%)**	
Correlated Prompt [11]	53.41	
HyperPrompt (w/o Missing type)	53.79	
HyperPrompt (w Missing type)	**54.32**	
CKA Similarity (%)	**Missing type**	
Layer	**w/o**	**w**
Layer4	68.34	**77.99**
Layer5	78.27	**83.14**
Layer6	31.13	**52.01**

Analysis for AlignPrompt. In Fig. 4, the t-SNE on MM-IMDb [1] shows that our AlignPrompt forms tighter, co-located cross-modal clusters than the common prompt, demonstrating stronger image–text alignment of AlignPrompt on semantic representation. As shown in Tab. 3 compared with Dynamic Prompt [11], our two single-module variants (w/o Low-rank Instance Alignment and w/o MI Completion) improve F1-Macro by about 0.5 and 0.9. Removing either

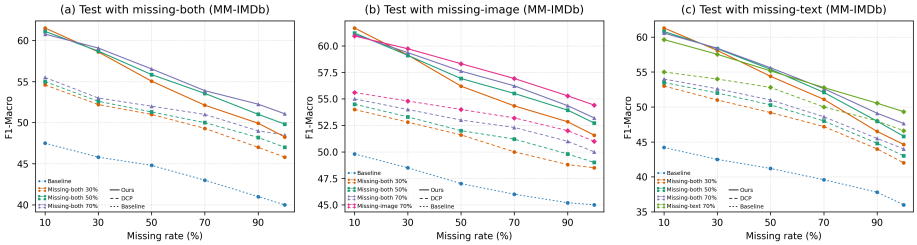

Fig. 6: We performed ablation studies of generalizability on the test set of the MM-IMDb [1] dataset to assess model generalizability across various testing scenarios with differing missing modality rates. (a) Models were trained on data where either modality could be absent and were then evaluated under the same conditions across a spectrum of missing rates. (b) Models trained under either the 'missing-both' or 'missing-image' paradigms were evaluated on text-only test cases at various missing rates. (c) Models trained on 'missing-both' or 'missing-text' data were assessed on image-only test cases across different missing rates.

Table 3: Analysis of AlignPrompt: Ablation of Low-rank alignment and MI completion in AlignPrompt on the MM-IMDb [1].

Modules	F1-Macro
AlignPrompt (w/o Alignment)	53.89
AlignPrompt (w/o MI Completion)	53.47
AlignPrompt (w/o Both)	53.01
AlignPrompt	**54.32**

Table 4: Compare our method with **feature reconstruction** methods on MM-IMDb [1] of 0.5 images missing rate.

Methods	F1-Macro
MMIN [44]	49.56
DiCMoR [34]	43.71
Knowledge Bridger [13]	51.95
Ours (only AlignPrompt)	55.89
Ours	**57.39**

module from the full AlignPrompt leads to about a 0.4 and 0.9 drop, and the full model achieves the best score.

Fig. 5 shows a paired Logit-KL scatter on MM-IMDb [1] image-missing validation samples. For each sample, we compute the KL divergence to the real-modality (full-modality) prediction using the classifier output logits z, and compare Feature Reconstruction (x-axis) with our prompt completion (y-axis) (lower is better). Points below the diagonal $y = x$ indicate that our prompt completion method is closer to the real-modality prediction in the decision space. 75.5% of samples lie below the diagonal, indicating that our prompt completion method aligns more closely with the full-modality predictions in the decision space.

Generalizability. In Fig. 6 (a)-(c), we train models under various missing cases and evaluate them on three test scenarios: missing-both, missing-image, and missing-text. In all scenarios, our method consistently outperforms the MMP [20] baseline and is better than DCP [11] over a broad range of missing rates. When the test case matches the training case, that case-specific model is slightly stronger than the counterpart trained on missing-both, indicating a targeted benefit. Meanwhile, models trained on missing-both (30/50/70%) transfer well across all three test cases, showing stable behavior even when one modal-

ity is always absent. We also observe a rate-robustness pattern: using a higher training rate (e.g., 70%) tends to slow the performance drop at high test rates, whereas 30%/50% can be marginally better at lower rates; 50% often provides a reasonable trade-off. Overall, MMP [20] remains a solid baseline, and DCP [11] is a competitive method. Our results also confirm that training on "missing-both" generally yields robust models for diverse test settings.

Table 5: Results on tri-modal dataset CMU-MOSI [7].

Dataset	Method	Double-missing		Single-missing	
		ACC	F1	ACC	F1
MOSI	MMP	65.23	66.19	73.72	74.03
	DCP	70.14	70.87	77.53	77.75
	Ours	**72.96**	**73.70**	**80.41**	**80.66**

Scalability. When our method extends to $M > 2$ modalities: (1) HyperPrompt generates layer-wise mapper weights from a shared hypernetwork conditioned on a multi-hot missing mask $\mathbf{z} \in \{0,1\}^M$, so the same formulation covers simultaneous multi-missing cases for any M without architectural changes. (2) For AlignPrompt, we apply mean pooling aggregation to all observed modal-specific factors $\mathcal{O} = \{m \,|\, z_m = 1\}$ and predict each missing modality $u \in \mathcal{U} = \{m \,|\, z_m = 0\}$ factor independently via a completer: $\widehat{\mathbf{C}}^{(u)} = f_u\big(\mathrm{Agg}\big(\{\mathbf{C}^{(m)}\}_{m\in\mathcal{O}}\big)\big)$. We further evaluate on CMU-MOSI [7], a tri-modal benchmark with language, visual, and acoustic streams. Following the single-missing and double-missing settings, we compare with MMP [20] and DCP [11] on this dataset of 0.7 missing rate. Tab. 5 shows that our method consistently achieves the best performance across both missing patterns, verifying stable gains when extended to three modalities.

Table 6: Parameter-matched performance comparison on the MM-IMDb [1] **test** set.

Params	2.8M	4.0M	5.2M	6.4M	7.9M
MMP (%)	47.87	47.58	47.33	47.45	47.21
DCP (%)	50.95	51.18	51.32	51.58	51.37
Ours (%)	**53.18**	**53.34**	**53.49**	**53.85**	**54.32**

Parameter-Matched Comparison. In the Tab. 6, we compare MMP [20], DCP [11], and our method under comparable parameter budgets on the MM-IMDb [1] test set. Although our trainable parameters are mainly introduced by the hypernetwork and completer modules, both are bottleneck-designed and remain lightweight (5.6M trainable parameters, i.e., 3.8% of the 149M frozen CLIP backbone). Across all parameter budgets (2.8M–7.9M), our method consistently outperforms MMP [20] and DCP [11], verifying that the gain mainly stems from the proposed design rather than extra parameters.

5 Conclusion

This work addresses a critical gap in missing modality robustness by replacing costly, unstable feature reconstruction with prompt-level semantic information completion. Furthermore, we apply a missing-type-aware hypernetwork to provide global hierarchical control. Experimental results validate this dual-prompt framework, which achieves state-of-the-art performance on several benchmarks. Thus, this work establishes an efficient and robust direction for adapting VLMs to real-world data.

Acknowledgements

This work was supported by the Australian Research Council (ARC) under the DECRA Fellowship DE240101049.

References

1. Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated multimodal units for information fusion. arXiv preprint arXiv:1702.01992 (2017)
2. Barber, D., Agakov, F.: The im algorithm: a variational approach to information maximization. Advances in neural information processing systems **16**(320), 201 (2004)
3. Bishop, C.M.: Mixture density networks (1994)
4. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101–mining discriminative components with random forests. In: European conference on computer vision. pp. 446–461. Springer (2014)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16×16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
6. Gallo, I., Ria, G., Landro, N., La Grassa, R.: Image and text fusion for upmc food-101 using bert and cnns. In: 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ). pp. 1–6. IEEE (2020)
7. Guo, Z., Jin, T., Zhao, Z.: Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In: Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers). pp. 1726–1736 (2024)
8. Ha, D., Dai, A., Le, Q.V.: Hypernetworks. arXiv preprint arXiv:1609.09106 (2016)
9. Havaei, M., Guizard, N., Chapados, N., Bengio, Y.: Hemis: Hetero-modal image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 469–477. Springer (2016)
10. Hoffman, J., Gupta, S., Darrell, T.: Learning with side information through modality hallucination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 826–834 (2016)
11. Hu, L., Shi, T., Feng, W., Shang, F., Wan, L.: Deep correlated prompting for visual recognition with missing modalities. Advances in Neural Information Processing Systems **37**, 67446–67466 (2024)
12. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 709–727. Springer (2022)
13. Ke, G., He, S., Wang, X., Wang, B., Chao, G., Zhang, Y., Xie, Y., Su, H.: Knowledge bridger: Towards training-free missing modality completion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25864–25873 (2025)
14. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19113–19122 (2023)
15. Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D.: The hateful memes challenge: Detecting hate speech in multimodal memes. Advances in neural information processing systems **33**, 2611–2624 (2020)

16. Kim, D., Kim, T.: Missing modality prediction for unpaired multimodal learning via joint embedding of unimodal models. In: European Conference on Computer Vision. pp. 171–187. Springer (2024)
17. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 139, pp. 5583–5594 (2021)
18. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM review **51**(3), 455–500 (2009)
19. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: International conference on machine learning. pp. 3519–3529. PMlR (2019)
20. Lee, Y.L., Tsai, Y.H., Chiu, W.C., Lee, C.Y.: 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)
21. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
22. Liang, V.W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.Y.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. Advances in Neural Information Processing Systems **35**, 17612–17625 (2022)
23. Lin, R., Hu, H.: Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. Transactions of the Association for Computational Linguistics **11**, 1686–1702 (2023)
24. Ma, M., Ren, J., Zhao, L., Testuggine, D., Peng, X.: Are multimodal transformers robust to missing modality? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18177–18186 (2022)
25. Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., Peng, X.: Smil: Multimodal learning with severely missing modality. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 2302–2310 (2021)
26. Pipoli, V., Bolelli, F., Sarto, S., Cornia, M., Baraldi, L., Grana, C., Cucchiara, R., Ficarra, E.: Semantically conditioned prompts for visual recognition under missing modality scenarios. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 4968–4977. IEEE (2025)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763 (2021)
28. Shang, C., Palmer, A., Sun, J., Chen, K.S., Lu, J., Bi, J.: Vigan: Missing view imputation with generative adversarial networks. In: 2017 IEEE International conference on big data (Big Data). pp. 766–775. IEEE (2017)
29. Shi, Y., Paige, B., Torr, P., et al.: Variational mixture-of-experts autoencoders for multi-modal deep generative models. Advances in neural information processing systems **32** (2019)
30. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines. Advances in neural information processing systems **25** (2012)
31. Tashiro, Y., Song, J., Song, Y., Ermon, S.: Csdi: Conditional score-based diffusion models for probabilistic time series imputation. Advances in neural information processing systems **34**, 24804–24816 (2021)

32. Wang, H., Chen, Y., Ma, C., Avery, J., Hull, L., Carneiro, G.: 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 (2023)
33. Wang, X., Kumar, D., Thome, N., Cord, M., Precioso, F.: Recipe recognition with large multimodal food dataset. In: 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). pp. 1–6. IEEE (2015)
34. Wang, Y., Cui, Z., Li, Y.: Distribution-consistent modal recovering for incomplete multimodal learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22025–22034 (2023)
35. Wu, R., Wang, H., Chen, H.T., Carneiro, G.: Deep multimodal learning with missing modality: A survey. arXiv preprint arXiv:2409.07825 (2024)
36. Xu, Y., Zhou, F., Zhao, C., Wang, Y., Yang, C., Chen, H.: Distilled prompt learning for incomplete multimodal survival prediction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5102–5111 (2025)
37. Yan, W., Wang, Y., Lin, W., Guo, Z., Zhao, Z., Jin, T.: Low-rank prompt interaction for continual vision-language retrieval. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 8257–8266 (2024)
38. Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6757–6767 (2023)
39. Yao, H., Zhang, R., Xu, C.: Tcp: Textual-based class-aware prompt tuning for visual-language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23438–23448 (2024)
40. Zeng, J., Liu, T., Zhou, J.: Tag-assisted multimodal sentiment analysis under uncertain missing modalities. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1545–1554 (2022)
41. Zhang, J., Wu, S., Gao, L., Shen, H.T., Song, J.: Dept: Decoupled prompt tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12924–12933 (2024)
42. Zhang, Z., Schomaker, L.: Divergan: An efficient and effective single-stage framework for diverse text-to-image generation. Neurocomputing **473**, 182–198 (2022)
43. Zhang, Z., Dai, L., Lin, Q., Diao, Y., Jin, G., Guo, Y., Zhang, J., Hao, X.: Synergistic prompting for robust visual recognition with missing modalities. arXiv preprint arXiv:2507.07802 (2025)
44. Zhao, J., Li, R., Jin, Q.: Missing modality imagination network for emotion recognition with uncertain missing modalities. 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. 2608–2618 (2021)
45. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)
46. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
47. Zhu, B., Niu, Y., Han, Y., Wu, Y., Zhang, H.: Prompt-aligned gradient for prompt tuning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 15659–15669 (2023)