DUAL-GUIDED PROMPT LEARNING UNDER MISSING MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing prompt-based methods for incomplete multimodal inputs typically use the same prompts for all samples sharing a missing pattern, without adapting them to the sample-specific task-relevant evidence provided by the available modalities. Moreover, they cannot effectively recover the complementary semantic information associated with the missing modality for each sample. To address these limitations, we propose **D**ual-**G**uided **P**rompt Learning (DGP), a framework that combines condition-guided prompt adaptation with prototype-guided semantic compensation. For condition guidance, we develop an evidence-aware hierarchical prompt tree, where the missing pattern determines the routing space and a sample-wise Fisher score dynamically selects and aggregates prompt experts. For prototype guidance, we build a learnable cross-modal prototype memory that uses the available modality representation to retrieve the corresponding prototype and inject it into the missing-modality proxy representation. Together, these two forms of guidance enable DGP to perform sample-wise prompt adaptation and compensate the missing modality with retrieved cross-modal semantic prototypes, without reconstructing the raw missing inputs. Experiments on three visual recognition benchmarks show that DGP achieves the best performance in 24 of 27 test configurations. Further evaluation on two multimodal sentiment analysis benchmarks demonstrates its broader applicability, yielding an average accuracy improvement of 2.03 points over the strongest baselines.

1 INTRODUCTION

Multimodal learning has achieved remarkable success by integrating complementary information from multiple modalities (Baltrušaitis et al., 2018), benefiting a wide range of applications such as visual recognition (Radford et al., 2021), medical diagnosis (Chen et al., 2020), autonomous driving (Chitta et al., 2022), and multimodal sentiment analysis (Tsai et al., 2019). However, real-world deployments often violate the assumption that all modalities are simultaneously available, owing to sensor failures or limitations, communication or data transmission errors, privacy restrictions, and high acquisition costs (Wu et al., 2024). Such incomplete inputs can substantially degrade the performance of multimodal models that are trained and optimized using complete-modality data (Ma et al., 2022).

Existing methods mainly mitigate missing-modality information loss through missing-input reconstruction or pseudo-feature generation (Zhao et al., 2021; Woo et al., 2023), as well as knowledge distillation from complete-modality teachers (Wang et al., 2023b). Although effective, these approaches often require auxiliary generation modules, teacher models, or additional training stages, while inaccurate reconstruction may introduce noise. Prompt learning offers a more parameter-efficient alternative by adapting frozen or partially frozen multimodal backbones with a small set of learnable prompts, without explicitly reconstructing the missing inputs (Lee et al., 2023; Khattak et al., 2023).

Despite their parameter efficiency, many prompt-based methods for incomplete multimodal inputs use prompts primarily determined by the missing pattern (Lee et al., 2023; Guo et al., 2024; Dai et al., 2025). Using task-conditioned sensitivity to distinguish samples sharing the same missing pattern remains less explored. For example, under the same text-missing condition, the observed image may provide clear cues for one sample but be ambiguous for another. Although sharing

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.

the same missing pattern, these samples may therefore benefit from different prompt adaptations. Moreover, prompt adaptation alone does not directly supply complementary information associated with the missing modality. Retrieval-augmented methods, such as RAGPT (Lang et al., 2025) and MISSRAG (Pipoli et al., 2025), use the available input to retrieve auxiliary examples, contextual information, or modality prototypes. However, similarity in the available modality does not necessarily ensure that the retrieved counterpart is useful for the downstream task. Retrieving raw examples or input-level prototypes may also introduce irrelevant details or unreliable cross-modal correspondences.

To address these limitations, we propose Dual-Guided Prompt Learning (DGP), which combines condition-guided prompt adaptation with prototype-guided semantic compensation. For condition guidance, we introduce an evidence-aware hierarchical prompt tree that adapts prompt selection based on sample-specific missing patterns and task-relevant evidence retained by the available modalities. The missing pattern first determines the prompt routing space, while a sample-wise Fisher score derived from the available-modality representation dynamically selects and aggregates prompt experts, enabling differentiated adaptation among samples sharing the same missing pattern. For prototype guidance, we construct a learnable cross-modal semantic prototype memory from complete-modality data. Each memory slot associates available- and missing-modality prototypes, thereby encoding reusable cross-modal semantic correspondences. Using the available-modality representation as a query, the memory retrieves the associated missing-modality prototype and injects it into the missing-modality proxy representation through a confidence-controlled gate. Unlike instance-level retrieval, the memory stores compact prototypes optimized for the downstream task, while the gate suppresses unreliable prototype injection. In this way, condition guidance enables sample-wise prompt adaptation, while prototype guidance provides task-relevant cross-modal semantics without reconstructing the raw missing input.

The main contributions of this work are summarized as follows:

- We propose Dual-Guided Prompt Learning (DGP), a unified framework that integrates condition-guided sample-wise prompt adaptation with prototype-guided cross-modal semantic compensation for missing-modality learning.
- We design an evidence-aware hierarchical prompt tree in which the missing pattern determines the prompt routing space, while a sample-wise Fisher score derived from the available-modality representation dynamically selects and aggregates prompt experts. This enables differentiated prompt adaptation among samples sharing the same missing pattern.
- We develop a learnable cross-modal semantic prototype memory that uses the available-modality representation to retrieve the corresponding missing-modality prototype and inject it into the missing-modality proxy representation through a confidence-controlled gate, without reconstructing the raw missing inputs.
- Extensive experiments on visual recognition and multimodal sentiment analysis benchmarks demonstrate the effectiveness, robustness, and broad applicability of DGP across diverse missing configurations.

2 Related Work

2.1 Multimodal Learning with Missing Modalities

Multimodal learning with missing modalities (Ma et al., 2021; Wu et al., 2026) aims to maintain reliable predictions when one or more modalities are unavailable during training or inference. Existing methods mainly follow three directions. Robust representation learning improves tolerance to incomplete inputs through modality dropout, missing-modality augmentation, shared-specific feature modeling, dynamic fusion, or expert-based adaptation (Lau et al., 2019; Liu et al., 2023; Wang et al., 2023a; Reza et al., 2024; Wu et al., 2025). Missing-information recovery estimates unavailable inputs or latent features from observed modalities using autoencoders, adversarial learning, feature-completion modules, or diffusion models (Cai et al., 2018; Meng et al., 2024; Zhao et al., 2024). Knowledge-transfer methods instead leverage complete multimodal data through teacher–student distillation (Wang et al., 2023b; Li et al., 2024) or retrieval of complete examples (Pipoli et al., 2025). Together, these directions improve robustness to incomplete inputs by strengthening

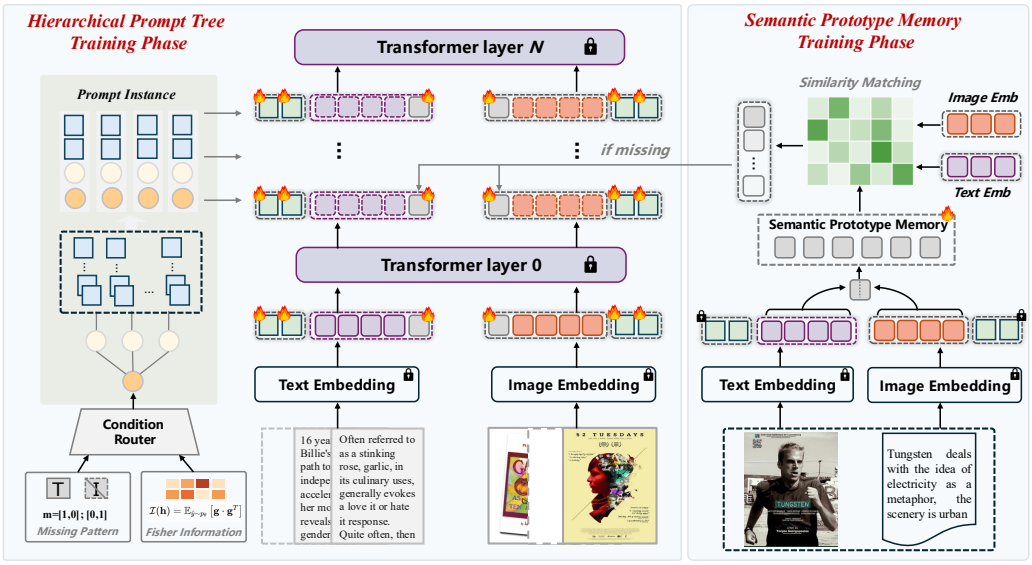


Figure 1: Overview of the proposed training framework. The first stage learns the cross-modal semantic prototype memory from complete multimodal data. The second stage optimizes the evidence-aware hierarchical prompt tree with missing patterns and Fisher information.

observed representations, estimating missing information, or transferring knowledge from complete-modality data. Although effective, these approaches either focus on general robustness to modality absence, require additional recovery modules or teacher models, or retrieve auxiliary content primarily according to semantic similarity, which may not identify the complementary information required by a missing modality for the current prediction.

2.2 Prompt Learning with Missing Modalities

Prompt learning has emerged as a lightweight approach for adapting pretrained backbones to incomplete inputs. Early prompt-based methods (Lee et al., 2023; Guo et al., 2024) typically freeze or partially freeze the pretrained backbone and optimize a small set of prompt parameters to bridge the discrepancy between complete-modality pretraining and incomplete-modality inference. Subsequent studies condition prompts on modality availability, missing indicators, or predefined missing types (Hu et al., 2024; Zhang et al., 2025; Dai et al., 2025; Chen et al., 2026), enabling models to adjust their representations or predictions under different missing patterns. Nevertheless, a missing pattern indicates which modalities are unavailable but does not capture how informative the observed modalities are for the downstream prediction. Consequently, prompts determined primarily by modality availability may not adequately adapt to variations across samples sharing the same missing pattern.

Retrieval-augmented approaches (Zhao et al., 2025; Lang et al., 2025) leverage auxiliary knowledge from complete-modality data to enrich incomplete representations, while AOEPT (Lang et al., 2026) introduces modality-contextualized prompts to alleviate the modality-reduction bottleneck in missing-modality prompt tuning. Orthogonal to prompt adaptation, DPL (Lu et al., 2026) replaces the conventional prediction head with missing-case-aware class prototypes and decomposes each prototype into modality-specific components, enabling the decision process to adapt to the observed modalities. Despite these advances, existing methods generally focus on prompt adaptation based on modality availability, retrieval based on semantic similarity, or prediction-head adaptation to predefined missing cases. They do not jointly adapt prompts to the task-relevant evidence retained by each sample and provide sample-specific semantic compensation for its missing modality. DGP bridges this gap through sample-specific evidence-aware hierarchical prompt routing and a learnable paired cross-modal semantic prototype memory.

3 METHOD

3.1 OVERVIEW

Incomplete multimodal samples with the same missing pattern may still contain substantially different amounts and types of task-relevant information in their observed modalities. Consequently, prompts determined only by modality availability provide the same adaptation strategy for all samples under a given missing pattern, overlooking such sample-specific heterogeneity. Furthermore, modality-aware prompting primarily adapts the processing of available inputs, but does not directly compensate for the semantic information lost with the missing modality.

As illustrated in Fig. 1, DGP addresses these two requirements through a hierarchical prompt tree and a paired cross-modal semantic prototype memory. To capture both the common characteristics associated with each missing pattern and the variations across individual samples, the prompt tree first selects a pattern-specific branch according to the missing indicator m, and then dynamically combines multiple prompt experts within that branch based on the observable content of the current sample. Since prompt adaptation alone cannot restore the complementary semantics carried by an unavailable modality, the available-modality features further query the paired prototype memory for a semantically corresponding prototype, which is injected into the missing-modality representation through a confidence-controlled gate.

3.2 HIERARCHICAL PROMPT TREE

Although the missing pattern indicates which modalities are unavailable, samples sharing the same pattern may retain different amounts of predictive information in their available modalities. We therefore derive a sample-specific routing signal from each incomplete input. Before dynamic prompt injection, the frozen modality encoders extract pre-prompt features from the available modalities, which are pooled and projected into an available-modality representation $\mathbf{h}$. We then compute a Fisher-information-based score with respect to $\mathbf{h}$:

$$\begin{aligned} s(\mathbf{x}) &= \mathrm{Norm}\left(\mathrm{Tr}\left[\mathcal{I}(\mathbf{h})\right]\right), \\ \mathcal{I}(\mathbf{h}) &= \mathbb{E}_{\hat{y}\sim p_\theta(\cdot|\mathbf{h})}\left[\nabla_{\mathbf{h}} \log p_\theta(\hat{y} \mid \mathbf{h}) \nabla_{\mathbf{h}} \log p_\theta(\hat{y} \mid \mathbf{h})^\top\right]. \end{aligned} \tag{1}$$

Here, $p_\theta(\cdot \mid \mathbf{h})$ denotes the predictive distribution produced from the available-modality representation $\mathbf{h}$, with gradients computed with respect to $\mathbf{h}$. Equivalently, $\mathrm{Tr}[\mathcal{I}(\mathbf{h})]$ is the expected squared norm of the predictive log-gradient, characterizing the local sensitivity of the predictive distribution to changes in $\mathbf{h}$. Since prompting modifies the representations used for prediction, this sensitivity provides a task-conditioned signal for prompt routing. The leaf router learns under task supervision how different score values map to expert weights for samples sharing the same missing pattern.

Based on the Fisher information score, we construct a two-level hierarchical prompt tree. At the first level, the missing pattern m deterministically selects a coarse adaptation branch $\mathcal{B}_m$, separating inputs with different modality-availability conditions. At the second level, a sample-dependent leaf router aggregates heterogeneous prompt experts within the selected branch according to $s(\mathbf{x})$.

Each branch $\mathcal{B}_m$ contains E learnable prompt experts,

$$\mathcal{P}_m = \{\mathbf{P}_{m,e}\}_{e=1}^{E}, \tag{2}$$

where the experts are designed to realize complementary prompt-modulation behaviors. In our implementation, we use three experts: a transformation-oriented expert based on a nonlinear projection $f_{\mathrm{trans}}(\cdot)$, a robustness-oriented expert whose normalized response is controlled by an evidence-dependent gate $g_{\mathrm{rob}}(s(\mathbf{x}))$, and an alignment-oriented expert that exchanges information among modality-specific prompt tokens. Their transformed prompts are written compactly as

$$\widehat{\mathbf{P}}_{m,e} = \mathcal{F}_e\left(\mathbf{P}_{m,e}; s(\mathbf{x})\right), \qquad e = 1, \ldots, E, \tag{3}$$

where $\mathcal{F}_e$ denotes the expert-specific transformation.

The leaf router projects the scalar sensitivity score into a routing representation and predicts a distribution over the E experts:

$$\boldsymbol{\pi}_m = \mathrm{Softmax}\left(\mathbf{W}_m^{\ell}\phi(s(\mathbf{x})) + \mathbf{b}_m^{\ell}\right), \tag{4}$$

where $\phi : \mathbb{R} \to \mathbb{R}^{d_r}$ is a learnable projection and $\mathbf{W}_m^{\ell} \in \mathbb{R}^{E \times d_r}$ is the leaf-router parameter of branch $\mathcal{B}_m$. The final dynamic prompt is obtained by weighted aggregation:

$$\mathbf{D}_m = \sum_{e=1}^{E} \pi_{m,e} \widehat{\mathbf{P}}_{m,e}. \tag{5}$$

The resulting prompt $\mathbf{D}_m$ is then inserted into the designated layers of the frozen backbone. This routing strategy preserves missing pattern-specific adaptation while allowing prompts within each branch to respond differently to the predictive sensitivity of individual samples.

3.3 Semantic Prototype Memory

The hierarchical prompt tree adapts prompt selection to the missing pattern and the observable evidence of each sample, but it does not explicitly enrich missing-modality representations with complementary cross-modal information. We therefore introduce a paired cross-modal semantic prototype memory, which learns modality-specific retrieval keys and semantic prototypes from complete-modality training samples.

The memory contains K paired slots:

$$\mathcal{M} = \left\{ \left(\mathbf{k}_i^v, \mathbf{k}_i^t, \mathbf{M}_i^v, \mathbf{M}_i^t \right) \right\}_{i=1}^{K}, \tag{6}$$

where $\mathbf{k}_i^v$ and $\mathbf{k}_i^t$ are the visual and textual retrieval keys of the i-th slot, and $\mathbf{M}_i^v$ and $\mathbf{M}_i^t$ are their corresponding semantic prototypes. Sharing the same slot index associates modality-specific prototypes that represent a reusable cross-modal correspondence.

Given pre-prompt visual and textual features $\mathbf{z}_0^v$ and $\mathbf{z}_0^t$, we compute modality-specific retrieval distributions as

$$\begin{aligned} \mathbf{w}^v &= \mathrm{Softmax}\left(\frac{\mathrm{sim}\left(\psi_v(\mathbf{z}_0^v), \mathbf{K}^v\right)}{\tau} \right), \\ \mathbf{w}^t &= \mathrm{Softmax}\left(\frac{\mathrm{sim}\left(\psi_t(\mathbf{z}_0^t), \mathbf{K}^t\right)}{\tau} \right), \end{aligned} \tag{7}$$

where ψ_v and ψ_t project the modality features into the corresponding key spaces, $\mathbf{K}^{v/t}$ denotes the modality-specific key bank, and τ is a temperature parameter.

The visual retrieval distribution is used to aggregate the paired textual prototypes, while the textual retrieval distribution is used to aggregate the paired visual prototypes:

$$\mathbf{S}^{t \leftarrow v} = \sum_{i=1}^{K} w_i^v \mathbf{M}_i^t, \qquad \mathbf{S}^{v \leftarrow t} = \sum_{i=1}^{K} w_i^t \mathbf{M}_i^v. \tag{8}$$

During training, each complete image-text pair is used to construct two reciprocal retrieval directions. The visual modality retrieves its paired textual counterpart for the missing-text view, while the textual modality retrieves its paired visual counterpart for the missing-image view. This bidirectional construction enables the retrieval keys and semantic prototypes to learn cross-modal correspondences under task supervision. During inference, only the retrieval direction from the available modality r to the missing modality $\bar{r}$ is activated. To suppress unreliable retrievals, we compute a confidence-controlled gate:

$$g_{\bar{r} \leftarrow r} = \mathbb{I}_{\bar{r}\text{-miss}} \cdot \sigma\left(\frac{\max_i w_i^r - \delta}{\gamma} \right), \tag{9}$$

where $\mathbb{I}_{\bar{r}\text{-miss}}$ indicates whether modality $\bar{r}$ is unavailable, $\max_i w_i^r$ measures the retrieval confidence, and δ and γ control the activation threshold and smoothness of the gate, respectively.

The retrieved counterpart is injected into the missing-modality representation as

$$\widetilde{\mathbf{h}}^{\bar{r}} = \mathbf{h}_{\mathrm{miss}}^{\bar{r}} + g_{\bar{r} \leftarrow r} \rho_{\bar{r}} \left(\mathbf{S}^{\bar{r} \leftarrow r} \right), \tag{10}$$

where $\mathbf{h}_{\mathrm{miss}}^{\bar{r}}$ is the representation of the missing-modality branch before memory enhancement, and $\rho_{\bar{r}}$ projects the retrieved prototype into the corresponding representation space. Thus, the available modality serves as the retrieval anchor, while only the paired counterpart associated with the missing modality is injected. The mechanism enriches missing-modality representations without reconstructing the original inputs or modifying complete-modality samples.

3.4 TRAINING PHASE

DGP is trained in two stages. In the first stage, we learn the paired cross-modal semantic prototype memory from complete-modality samples. Each modality is used to retrieve its paired counterpart under task supervision, while prototype regularization encourages cross-modal consistency, prototype diversity, and balanced memory utilization. Further training details are provided in Appendix A. The objective is

$$\mathcal{L}_{\text{mem}} = \mathcal{L}_{\text{task}} + \lambda_{\text{proto}}\mathcal{L}_{\text{proto}}, \tag{11}$$

where $\mathcal{L}_{\text{proto}}$ collectively denotes the regularization terms used to organize the paired prototype memory.

In the second stage, we simulate incomplete inputs and optimize the hierarchical prompt tree using the learned prototype memory. The missing pattern determines the adaptation branch, while the Fisher information score guides sample-specific prompt aggregation within the selected branch. The training objective is

$$\mathcal{L}_{\text{prompt}} = \mathcal{L}_{\text{task}} + \lambda_{\text{route}}\mathcal{L}_{\text{route}}, \tag{12}$$

where $\mathcal{L}_{\text{route}}$ encourages balanced expert utilization and prevents routing collapse.

4 EXPERIMENTS

In this section, we conduct extensive experiments to answer the following research questions: **RQ1:** How does DGP compare with state-of-the-art missing-modality learning methods? **RQ2:** How robust is DGP to changes in missing configurations? **RQ3:** Does DGP provide consistent benefits across different multimodal tasks under missing-modality settings? **RQ4:** What is the contribution of each key component in DGP? Additional analyses of computational efficiency and performance stability are provided in Appendix B.

Datasets and metrics. We evaluate visual recognition on three image–text benchmarks: MM-IMDb (Arevalo et al., 2017), UPMC Food101 (Wang et al., 2015), and Hateful Memes (Kiela et al., 2020), using F1-Macro, accuracy, and AUROC, respectively. To assess the applicability of DGP beyond visual recognition, we further evaluate it on the multimodal sentiment analysis benchmarks CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Zadeh et al., 2018), reporting accuracy and F1-score.

Missing-modality protocol. We adopt a general and challenging setting in which modality absence occurs during both training and testing. Let η denote the percentage of incomplete samples. In the missing-text setting, $\eta\%$ of the samples contain only images; in the missing-image setting, $\eta\%$ contain only text. The remaining $(100-\eta)\%$ samples are complete in both settings. In the missing-both setting, $\eta/2\%$ of the samples are image-only and another $\eta/2\%$ are text-only, while the remaining $(100-\eta)\%$ are complete.

Implementation details. For visual recognition, we use CLIP (Radford et al., 2021) with a ViT-B/16 image encoder and keep its encoders frozen. We optimize only the prompt parameters, paired semantic prototype memory, and task-specific prediction heads. DGP uses three prompt experts in each branch and a prototype memory containing 16 paired slots. For multimodal sentiment analysis, we follow MPLMM (Guo et al., 2024) in preprocessing and feature extraction, and instantiate condition-guided prompt adaptation and prototype-guided semantic compensation for text, audio, and video. Missing patterns and available-modality features condition prompt routing; retrieval distributions from available modalities are averaged over shared-index prototype slots, and the resulting prompts enrich modality representations before multimodal Transformer fusion.

Baselines. For visual recognition, we compare DGP with CoOp (Zhou et al., 2022), MMP (Lee et al., 2023), MaPLe (Khattak et al., 2023), DePT (Zhang et al., 2024), DCP (Hu et al., 2024), RAGPT (Lang et al., 2025), SyP (Zhang et al., 2025), AOEPT (Lang et al., 2026), and DPL (Lu et al., 2026). For multimodal sentiment analysis, we compare with MPLMM (Guo et al., 2024), KB-MMC (Ke et al., 2025), and DPMSA (Meng et al., 2026). All methods are evaluated using the same data splits, missing masks, and backbone configurations. AOEPT is reimplemented under this

Table 1: Comparison results under different missing rates on the testing set.

Datasets	Missing rate η	Image	Text	CoOp	MMP	MaPLe	DePT	DCP	RAGPT	SyP	AOEPT	DPL	Ours
MM-IMDb (F1-Macro)	50%	100%	50%	48.06	48.88	49.58	50.64	52.13	47.48	53.90	53.87	55.11	**55.46** (+0.35)
		50%	100%	49.89	51.46	52.32	52.78	54.32	45.02	56.27	55.93	57.79	**58.46** (+0.67)
		75%	75%	48.37	49.32	49.56	50.87	52.32	46.45	55.02	54.74	56.07	**57.18** (+1.11)
	70%	100%	30%	44.13	45.64	45.52	46.38	48.52	47.16	51.37	51.70	53.30	**53.53** (+0.23)
		30%	100%	48.82	50.52	50.64	52.13	53.14	48.16	54.20	54.73	56.56	**57.02** (+0.46)
		65%	65%	46.84	48.12	49.16	50.32	51.42	43.81	52.90	53.31	**55.16**	55.14 (-0.02)
	90%	100%	10%	44.76	45.32	46.84	47.56	49.26	45.98	50.21	50.55	50.92	**51.64** (+0.72)
		10%	100%	48.32	49.12	50.13	50.88	52.22	38.44	53.72	54.09	54.36	**55.50** (+1.14)
		55%	55%	44.12	44.87	45.12	46.54	48.04	41.09	49.63	50.13	52.17	**52.35** (+0.18)
Food101 (Accuracy)	50%	100%	50%	77.45	77.89	79.64	80.16	82.11	84.13	83.20	84.50	84.28	**85.84** (+1.34)
		50%	100%	87.02	87.16	87.35	82.14	89.12	85.11	89.64	89.44	89.53	**89.78** (+0.14)
		75%	75%	81.24	81.72	82.34	83.12	85.24	83.51	86.17	85.90	86.64	**87.74** (+1.10)
	70%	100%	30%	76.34	76.52	77.02	77.34	78.87	75.53	79.56	80.77	81.51	**82.23** (+0.72)
		30%	100%	84.78	85.64	85.89	86.12	87.32	81.89	88.67	88.86	87.89	**88.97** (+0.11)
		65%	65%	78.87	79.12	79.84	81.46	81.87	76.94	82.45	83.24	83.38	**85.07** (+1.69)
	90%	100%	10%	71.87	73.14	73.46	74.12	75.26	76.20	76.33	77.47	78.62	**78.64** (+0.02)
		10%	100%	81.67	82.14	83.12	83.56	85.78	85.13	86.41	87.03	86.23	**87.52** (+0.49)
		55%	55%	76.46	76.58	77.85	78.12	79.87	80.02	81.03	81.67	81.01	**82.58** (+0.91)
Hateful Memes (AUROC)	50%	100%	50%	60.56	60.31	60.87	61.87	62.32	66.68	68.25	69.16	70.75	**71.06** (+0.31)
		50%	100%	62.41	62.35	63.13	63.88	64.46	64.20	66.80	66.58	65.72	**68.76** (+1.96)
		75%	75%	64.87	65.84	65.46	65.86	66.02	64.19	68.16	68.58	64.29	**68.78** (+0.20)
	70%	100%	30%	60.74	61.12	61.26	61.56	62.82	64.10	68.94	69.03	69.13	**70.40** (+1.27)
		30%	100%	62.74	63.24	63.14	63.48	64.12	62.57	66.98	66.40	63.85	**67.54** (+0.56)
		65%	65%	64.82	65.04	65.23	65.48	66.08	63.47	68.42	**69.78**	63.06	67.83 (-1.95)
	90%	100%	10%	60.03	57.21	60.74	61.14	62.08	66.26	69.70	69.12	68.53	**69.84** (+0.14)
		10%	100%	61.46	61.52	61.87	62.42	63.87	61.36	64.54	66.18	62.80	**66.93** (+0.75)
		55%	55%	64.32	63.34	64.85	65.37	66.78	62.38	**68.93**	68.67	65.01	67.13 (-1.80)

unified protocol, which differs from its original experimental setup; therefore, its results may not exactly match those reported in the original paper.

4.1 Main Results

This subsection addresses RQ1 by comparing DGP with state-of-the-art missing-modality learning methods. As shown in Table 1, DGP achieves the strongest overall performance across MM-IMDb, Food101, and Hateful Memes. The improvements are particularly consistent on MM-IMDb and Food101, where DGP maintains its advantage across different missing patterns and rates. This performance is consistent with the intended role of the evidence-aware hierarchical prompt tree in adapting prompt selection to both missing patterns and sample-wise differences in the task-relevant information provided by the available modalities. Compared with DPL, which adapts the prediction head using missing-case-aware class prototypes, DGP generally obtains higher scores across datasets and missing settings, suggesting the benefit of introducing adaptive prompts and cross-modal enrichment before final prediction. On Hateful Memes, DGP performs favorably across different missing settings, particularly under the missing-text and missing-image conditions, while showing some performance variation under severe missing-both conditions. Overall, these findings support the effectiveness of combining sample-adaptive prompting with prototype-guided semantic compensation across diverse datasets and missing configurations.

4.2 Robustness Analysis

This subsection answers RQ2 and RQ3 by evaluating the robustness of DGP across missing rates and modality-availability conditions, and further examining its applicability to different multimodal tasks. We first examine robustness to mismatches between the training and test missing rates. Specifically, models are trained under the missing-both setting at rates of 50%, 70%, and 90%, and then evaluated across test missing rates from 10% to 90%, as shown in Fig. 2. DGP consistently achieves the best performance on Food101 across test missing rates from 10% to 90% when trained under the missing-both setting at rates of 50%, 70%, and 90%. Although the performance of all methods decreases as the test missing rate increases, DGP maintains its advantage throughout the evaluated range. At the 90% test missing rate, DGP outperforms the strongest baselines by 1.28, 1.38, and 0.91 points under the three training settings, respectively. These results indicate that DGP remains effective when the test missing rate differs from the training condition, demonstrating its robustness to varying degrees of modality incompleteness.

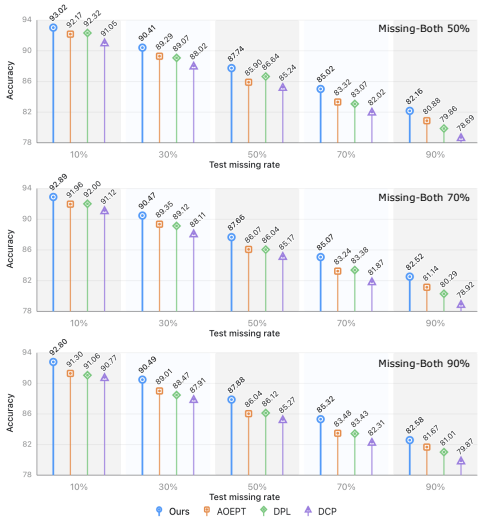


Figure 2: Performance across test missing rates after training with missing-both 50%, 70%, and 90% .



Figure 3: Performance under 50%, 70%, and 90% missing-image and missing-text after training with missing-both 50%, 70%, and 90% .

We next examine whether models trained under missing-both can generalize to modality-specific missing conditions, as shown in Fig. 3. DGP achieves the best performance in all 18 combinations of training missing rates, missing modalities, and test missing rates. Its gains are particularly pronounced under missing-text conditions, where the available image modality retains less task-relevant information than the available text modality. For example, under 90% missing-both training and 90% missing-text testing, DGP achieves 79.11, exceeding the strongest baseline, AOEPT, by 2.73 points. This behavior is consistent with the intended roles of evidence-aware prompt routing and paired semantic retrieval in adapting to different incomplete inputs and enriching missing-modality representations.

To answer RQ3, we further evaluate DGP on CMU-MOSI and CMU-MOSEI, which involve text, audio, and video modalities. As shown in Table 2, DGP achieves the best accuracy across all six modality-availability settings on both datasets, while obtaining the best F1-score in all six settings on CMU-MOSI and five on CMU-MOSEI. In terms of average performance, DGP surpasses the strongest baseline by 2.28 points in accuracy and 1.92 points in F1-score on CMU-MOSI, and by 1.77 and 0.70 points on CMU-MOSEI, respectively. Together with the results on the visual recognition benchmarks, these findings demonstrate the broader applicability of DGP across different multimodal tasks and modality combinations. This broader applicability is consistent with its evidence-aware prompt adaptation and semantic compensation, both of which operate according to the information available in each incomplete input.

Table 2: CMU-MOSI and CMU-MOSEI results by available modality (accuracy/F1). T, A, and V denote text, audio, and video; Avg. averages the six cases.

Dataset	Available	MPLMM	KB-MMC	DPMSA	Ours
CMU-MOSI	{T}	80.12 / 80.31	82.43 / 82.58	80.79 / 81.13	**84.26 / 84.29**
	{A}	62.71 / 63.65	65.46 / 65.02	63.21 / 63.98	**68.98 / 68.74**
	{V}	63.12 / 63.74	65.38 / 65.49	63.53 / 64.17	**66.67 / 65.86**
	{T,A}	80.76 / 81.09	82.84 / 82.27	80.99 / 81.49	**83.79 / 83.80**
	{T,V}	81.12 / 81.19	82.23 / 82.36	81.41 / 81.56	**84.26 / 84.22**
	{A,V}	65.02 / 65.41	66.31 / 66.78	65.55 / 65.62	**70.37 / 69.06**
	Avg.	72.14 / 72.57	74.11 / 74.08	72.59 / 72.98	**76.39 / 76.00**
CMU-MOSEI	{T}	79.12 / 79.17	80.74 / 80.89	80.21 / 80.53	**81.55 / 81.48**
	{A}	67.33 / 68.71	69.62 / 69.94	67.91 / 69.12	**72.70 / 71.17**
	{V}	67.29 / 69.40	68.48 / 69.53	67.34 / 69.79	**71.80 / 71.43**
	{T,A}	80.45 / 80.43	81.63 / 81.52	80.49 / 80.60	**82.10 / 82.03**
	{T,V}	80.11 / 80.13	81.79 / 81.96	81.32 / 81.55	**82.17 / 82.09**
	{A,V}	68.21 / 69.91	68.52 / 70.05	69.08 / **70.24**	**71.10** / 69.91
	Avg.	73.75 / 74.68	75.13 / 75.65	74.73 / 75.29	**76.90 / 76.35**

4.3 Ablation Study

This subsection answers RQ4 by examining the key design choices in DGP, including evidence guidance, adaptive prompt routing, semantic memory construction, and the number of prototype

Table 3: Ablation of evidence-aware prompt design under the missing-both 70%.

Method	MM-IMDb (F1-Macro)	Food101 (Accuracy)	Hateful Memes (AUROC)
DGP w/o Evidence Guidance	54.27 (-0.87)	84.05 (-1.02)	66.11 (-1.72)
DGP w/o Route Selection	54.60 (-0.54)	84.31 (-0.76)	66.36 (-1.47)
Full DGP	**55.14**	**85.07**	**67.83**

Table 4: Comparison of semantic compensation strategies under the missing-both 70%, including no compensation, instance retrieval, and semantic prototype memory.

Method	MM-IMDb (F1-Macro)	Food101 (Accuracy)	Hateful Memes (AUROC)
DGP w/o Semantic Prototype Memory	54.54 (-0.60)	83.94 (-1.13)	66.75 (-1.08)
DGP w/ Instance Memory Bank	54.43 (-0.71)	84.27 (-0.80)	66.78 (-1.05)
Full DGP	**55.14**	**85.07**	**67.83**

slots. Specifically, RQ4 investigates how these components contribute to the overall performance of DGP. Unless otherwise specified, both training and testing are conducted under the missing-both 70% setting.

Table 3 shows the complementary roles of evidence guidance and route selection. The largest reductions occur on Hateful Memes, where removing evidence guidance and replacing adaptive routing with fixed aggregation decrease performance by 1.72 and 1.47 points, respectively. These results show that both the sensitivity-based signal and adaptive expert aggregation contribute to DGP under this evaluation setting. The former characterizes the predictive distribution's local sensitivity to changes in the available-modality representation, while the latter maps this signal to sample-specific prompt mixtures. Table 4 examines both the contribution and the design of the semantic prototype memory. The first variant removes the semantic prototype memory, thereby eliminating prototype-based enrichment of missing-modality representations. The second replaces it with a RAGPT-style (Lang et al., 2025) Instance Memory Bank constructed from the same complete-modality training data. Given an incomplete sample, the bank retrieves instances similar to the available modality and aggregates their paired features from the missing-modality side for compensation. Although instance retrieval slightly improves over the memory-free variant on Food101 and Hateful Memes, it remains inferior to the full DGP across all three datasets. These results suggest that instance similarity does not always identify the task-relevant cross-modal cues needed to compensate for the missing modality, whereas task-supervised paired prototypes capture such correspondences more effectively. Table 5 studies the sensitivity to the number of prototype slots K. Performance improves as K increases from 4 to 16 but decreases at $K = 32$, indicating a trade-off between semantic capacity and prototype redundancy. We therefore set $K = 16$ in all experiments.

Table 5: Prototype memory size sensitivity at 70% missing-both.

Bank Size	MM-IMDb (F1-Macro)	Food101 (Accuracy)	Hateful Memes (AUROC)
size = 4	54.86	84.55	66.41
size = 8	55.09	84.90	66.72
size = 16 (Ours)	**55.14**	**85.07**	**67.83**
size = 32	54.79	84.06	67.34

5 Conclusion

In this paper, we presented Dual-Guided Prompt Learning (DGP) for multimodal learning with missing modalities. DGP combines an evidence-aware hierarchical prompt tree for sample-adaptive prompt routing with a paired cross-modal semantic prototype memory for enriching missing-modality representations. Experiments on visual recognition and multimodal sentiment analysis benchmarks demonstrate its effectiveness across varying missing configurations. DGP introduces modest parameter overhead due to its prompt experts and prototype memory. Future work will explore more compact designs and extend DGP to open-world scenarios with dynamic modality quality and diverse modality combinations.

AI USE STATEMENT

Generative AI tools, including ChatGPT, were used for language polishing, grammar checking, LaTeX formatting, and feedback on the presentation of the method and experimental results. The authors reviewed the AI-assisted content and take full responsibility for the manuscript.

ETHICS STATEMENT

This work evaluates DGP on the publicly available MM-IMDb, UPMC Food101, Hateful Memes, CMU-MOSI, and CMU-MOSEI benchmarks. It does not introduce a new dataset, recruit participants, or collect new personal data. Some benchmark content, particularly Hateful Memes, may be offensive, while the sentiment datasets include recordings of individuals. The data are used for academic evaluation subject to their applicable terms. DGP may inherit biases and coverage limitations from these datasets and pretrained encoders, and missing modalities may further affect performance across different conditions. Benchmark results should not be interpreted as a guarantee of fairness or reliability in deployment.

REPRODUCIBILITY STATEMENT

Section 3 describes the sensitivity-based prompt routing, paired cross-modal prototype memory, and training objectives. Section 4 specifies the datasets, evaluation metrics, missing-modality settings, backbone, and main implementation choices, including the tri-modal variant used for sentiment analysis. Appendix A provides additional training details and visual-recognition hyperparameters. Appendix B reports computational cost and stability across three random seeds under a representative missing-both setting. Selected implementation code for the visual-recognition experiments is provided in the supplementary materials.

REFERENCES

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

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

Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. Deep adversarial learning for multi-modality missing data completion. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1158–1166, 2018.

Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE transactions on medical imaging*, 41 (4):757–770, 2020.

Xin Chen, Huanjie Tao, and Benran Li. Towards robust incomplete multimodal open-set domain generalization with uncertain missing modalities. *Knowledge-Based Systems*, pp. 115777, 2026.

Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):12878–12895, 2022.

Ruiting Dai, Yuqiao Tan, Lisi Mo, Tao He, Ke Qin, and Shuang Liang. Robustpt: Dynamic disentanglement prompt tuning in vision-language models with missing modalities. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pp. 164–172, 2025.

Zirun Guo, Tao Jin, and Zhou Zhao. 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.

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.

Guanzhou Ke, Shengfeng He, Xiaoli Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Yi Xie, and Hexing Su. Knowledge bridger: Towards training-free missing modality completion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25864–25873, 2025.

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.

Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18035–18043, 2025.

Jian Lang, Rongpei Hong, Ting Zhong, and Fan Zhou. Aoept: Breaking the implicit modality-reduction bottleneck in modality-missing prompt tuning. *arXiv preprint arXiv:2605.24816*, 2026.

Kenneth Lau, Jonas Adler, and Jens Sjölund. A unified representation network for segmentation with missing modalities. *arXiv preprint arXiv:1908.06683*, 2019.

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.

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

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

Jueqing Lu, Yuanyuan Qi, Xiaohao Yang, Shuaicheng Niu, Fucai Ke, Shujie Zhou, Wei Tan, Jionghao Lin, Wray Buntine, Hamid Rezatofighi, et al. Dpl: Decoupled prototype learning for enhancing robustness of vision-language transformers to missing modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39219–39229, 2026.

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

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.

Jing Meng, Zhenfang Zhu, Qiang Lu, Hongli Pei, Kefeng Li, Yuzhi Ren, Meng Li, Xiaowen Sun, and Dawei Zhao. A prompt-driven framework for compensation and fusion in multimodal sentiment analysis with missing modalities. *Knowledge-Based Systems*, pp. 115438, 2026.

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.

Vittorio Pipoli, Alessia Saporita, Federico Bolelli, Marcella Cornia, Lorenzo Baraldi, Costantino Grana, Rita Cucchiara, and Elisa Ficarra. Missrag: Addressing the missing modality challenge in multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3215–3224, 2025.

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.

Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 47(2):742–754, 2024.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6558–6569, 2019.

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

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

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 (ICMEW)*, pp. 1–6. IEEE, 2015.

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

Chenwei Wu, Zitao Shuai, Zhengxu Tang, Luning Wang, and Liyue Shen. Dynamic modeling of patients, modalities and tasks via multi-modal multi-task mixture of experts. In *The thirteenth international conference on learning representations*, 2025.

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

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88, 2016.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018.

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.

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.

Jinming Zhao, Ruichen Li, and Qin Jin. 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.

Yihan Zhao, Wei Xi, Xiao Fu, and Jizhong Zhao. Enhancing multimodal model robustness under missing modalities via memory-driven prompt learning. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence, IJCAI 2025*, pp. 2458–2466, 2025.

Yunpeng Zhao, Cheng Chen, Qing You Pang, Quanzheng Li, Carol Tang, Beng-Ti Ang, and Yueming Jin. Dealing with all-stage missing modality: Towards a universal model with robust reconstruction and personalization. *arXiv preprint arXiv:2406.01987*, 2024.

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, 2022.

A Additional Training Details

A.1 Two-Stage Optimization

As described in the main paper, DGP is optimized in two stages. The first stage organizes paired cross-modal semantic prototypes using complete-modality training samples. The second stage simulates incomplete inputs and learns the evidence-aware hierarchical prompt tree while using the learned prototype memory for semantic compensation. This separation prevents unstable prompt routing at the beginning of training, when the correspondence between the two modality-specific prototype banks has not yet been established.

Stage I: semantic prototype-memory learning. Let $\mathcal{C}$ denote the complete-modality samples in a mini-batch. For a sample $n \in \mathcal{C}$, the visual and textual queries produce retrieval distributions $\mathbf{w}_n^v$ and $\mathbf{w}_n^t$ over the K paired memory slots. We encourage the two modalities of a complete sample to identify the same semantic slot using the symmetric retrieval-consistency loss

$$\mathcal{L}_{\mathrm{ret}} = \frac{1}{2|\mathcal{C}|} \sum_{n \in \mathcal{C}} \left[D_{\mathrm{KL}}\left(\mathbf{w}_n^v \| \mathbf{w}_n^t\right) + D_{\mathrm{KL}}\left(\mathbf{w}_n^t \| \mathbf{w}_n^v\right) \right]. \tag{13}$$

For the paired visual and textual keys $(\mathbf{k}_i^v, \mathbf{k}_i^t)$ in slot i, their cross-modal correspondence is regularized by

$$\mathcal{L}_{\mathrm{pair}} = \frac{1}{K} \sum_{i=1}^{K} \left[1 - \cos(\mathbf{k}_i^v, \mathbf{k}_i^t) - m_p \right]_+, \tag{14}$$

where m_p is a margin and $[\cdot]_+ = \max(0, \cdot)$. To avoid redundant slots, we penalize correlations among keys from the same modality:

$$\mathcal{L}_{\mathrm{div}} = \frac{1}{2K(K-1)} \sum_{r \in \{v,t\}} \sum_{i \neq j} \cos^2(\mathbf{k}_i^r, \mathbf{k}_j^r). \tag{15}$$

We further regularize the sharpness and mini-batch utilization of the memory. Let $\mathcal{H}(\cdot)$ denote entropy, $\bar{\mathbf{w}}^r$ the average retrieval distribution for modality r, and $\mathbf{u}$ the uniform distribution over K slots. The corresponding losses are

$$\mathcal{L}_{\mathrm{ent}} = \frac{1}{2|\mathcal{C}|} \sum_{n \in \mathcal{C}} \sum_{r \in \{v,t\}} \left(\mathcal{H}(\mathbf{w}_n^r) - \rho \log K \right)^2, \tag{16}$$

$$\mathcal{L}_{\mathrm{bal}} = \frac{1}{2} \sum_{r \in \{v,t\}} \| \bar{\mathbf{w}}^r - \mathbf{u} \|_2^2, \tag{17}$$

where ρ controls the target retrieval entropy. These terms organize different parts of the paired memory: $\mathcal{L}_{\mathrm{ret}}$ encourages the two modalities of a complete sample to assign similar weights to the same slots, while $\mathcal{L}_{\mathrm{pair}}$ regularizes the visual and textual retrieval keys within each slot. The diversity, entropy, and balance terms regulate slot separation and utilization. The prototype values contribute to predictions through the retrieved counterparts in the two missing-modality views and are optimized by $\mathcal{L}_{\mathrm{task}}$. The prototype-memory regularizer used in Eq. (11) of the main paper is therefore

$$\begin{aligned} \mathcal{L}_{\mathrm{proto}} = {} & \alpha_{\mathrm{ret}} \mathcal{L}_{\mathrm{ret}} + \alpha_{\mathrm{pair}} \mathcal{L}_{\mathrm{pair}} + \alpha_{\mathrm{div}} \mathcal{L}_{\mathrm{div}} \\ & + \alpha_{\mathrm{ent}} \mathcal{L}_{\mathrm{ent}} + \alpha_{\mathrm{bal}} \mathcal{L}_{\mathrm{bal}}. \end{aligned} \tag{18}$$

The complete Stage-I objective is $\mathcal{L}_{\mathrm{mem}} = \mathcal{L}_{\mathrm{task}} + \lambda_{\mathrm{proto}} \mathcal{L}_{\mathrm{proto}}$. Only complete samples contribute to $\mathcal{L}_{\mathrm{ret}}$, because both retrieval directions are required to form a valid correspondence target.

Stage II: evidence-aware prompt-tree learning. After Stage I, we retain the learned paired prototype memory and simulate the missing patterns described in the main experimental protocol. For each sample, the missing pattern selects the corresponding coarse branch. The normalized Fisher score computed from the available-modality representation controls the leaf distribution $\boldsymbol{\pi}_n$ over the

Table 6: Default optimization and model settings for visual recognition.

Setting	Value
Backbone	CLIP ViT-B/16
Optimizer	AdamW
Batch size	128
Base learning rate	1×10^{-3}
Weight decay	0.02
Maximum epochs	100
Prompt depth / length	6 / 36
Experts per branch E	3
Memory slots K	16
Retrieval temperature τ	0.25
Pair margin m_p	0.20

E prompt experts in that branch. To discourage routing collapse while preserving sample-dependent expert selection, we use

$$\mathcal{L}_{\text{ent}}^{\text{route}} = -\frac{1}{B}\sum_{n=1}^{B}\mathcal{H}(\boldsymbol{\pi}_n), \tag{19}$$

$$\mathcal{L}_{\text{bal}}^{\text{route}} = \left\| \frac{1}{B}\sum_{n=1}^{B}\boldsymbol{\pi}_n - \frac{1}{E}\mathbf{1} \right\|_2^2, \tag{20}$$

$$\mathcal{L}_{\text{route}} = \beta_{\text{ent}}\mathcal{L}_{\text{ent}}^{\text{route}} + \beta_{\text{bal}}\mathcal{L}_{\text{bal}}^{\text{route}}. \tag{21}$$

The Stage-II objective corresponding to Eq. (12) is $\mathcal{L}_{\text{prompt}} = \mathcal{L}_{\text{task}} + \lambda_{\text{route}}\mathcal{L}_{\text{route}}$. We use binary cross-entropy for the multilabel MM-IMDb task and cross-entropy for Food-101 and Hateful Memes. For CMU-MOSI and CMU-MOSEI, we follow the MPLMM preprocessing, feature extraction, data splits, and multimodal Transformer backbone, and optimize the sentiment classification objective under the six incomplete modality-availability cases reported in the main paper.

A.2 Optimization and Reproducibility Settings

For visual recognition, all methods use the same CLIP ViT-B/16 backbone, dataset splits, missing masks, and input preprocessing. The CLIP encoders are frozen; only the prompts, paired semantic prototype memory, routing modules, and task-specific prediction head are optimized. Unless otherwise specified, we use the settings in Table 6 for all three visual benchmarks. Model selection is performed using the validation metric associated with each dataset, and the selected checkpoint is evaluated once on the test set.

B Efficiency and Stability

B.1 Parameter and FLOP Comparison

Table 7 compares the trainable parameter count and the computational cost of representative missing-modality prompt-learning methods. All methods use the same backbone and input configuration. GFLOPs are measured for one forward pass of a single image–text sample and exclude data loading and preprocessing. DGP contains 8.55M trainable parameters, remaining below 10M, but incurs more computation than the compared methods because it performs sample-dependent expert routing and paired prototype retrieval. This result shows that DGP trades additional computation for the consistent robustness gains reported across missing patterns and rates; reducing this overhead is a useful direction for future work.

B.2 Stability across Random Seeds

To evaluate the stability of DGP with respect to training randomness, we repeat each experiment using $R = 3$ independent random seeds under the representative Missing-Both 70% setting. The

Table 7: Comparison of trainable parameters and inference computation. Lower is better for both columns.

Method	Trainable parameters (M)	GFLOPs
DCP	5.80	51.31
RAGPT	7.32	53.32
AOEPT	**1.81**	**44.32**
DGP (Ours)	8.55	91.61

Table 8: Stability of DGP across three independent random seeds under the Missing-Both 70% setting. Results are reported as mean $\pm$ sample standard deviation.

Dataset	Metric	Mean $\pm$ std.
MM-IMDb	F1-Macro	55.20 ± 0.37
Food-101	Accuracy	85.05 ± 0.26
Hateful Memes	AUROC	67.90 ± 0.53

dataset splits are fixed across runs, while parameter initialization, mini-batch order, and stochastic missing-mask sampling vary according to the random seed. For run-level scores $q_1, \ldots, q_R$, we report

$$\bar{q} = \frac{1}{R}\sum_{r=1}^{R} q_r, \qquad \sigma_q = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(q_r - \bar{q})^2}. \tag{22}$$

As shown in Table 8, DGP achieves standard deviations of 0.37, 0.26, and 0.53 on MM-IMDb, Food-101, and Hateful Memes, respectively. The limited variation across runs indicates that the performance of DGP is stable with respect to random initialization, data order, and stochastic missing-mask sampling.