MARKER: MISSING-AWARE LOW-RANK EXPERT ROUTING FOR INCOMPLETE MULTIMODAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Missing modalities commonly arise in real-world applications due to device failures, network instability, or privacy concerns. Recent studies address this issue with prompt-based fine-tuning. However, most of these prompt-based methods struggle with two aspects: (1) they rely on scenario-conditioned prompt sets and require prompt selection for different modality-availability conditions, leading to a less unified adaptation framework; (2) most methods are modality-missing-aware but instance-agnostic. While recent approaches alleviate this limitation through retrieval or query mechanisms, they introduce additional modules and a more complex pipeline. To address these issues, we propose MARKER, a **M**issing-**A**ware Low-**R**an**k** **E**xpert **R**outing framework for incomplete multimodal learning. Specifically, we insert a LoRA-based mixture-of-experts module into a pretrained multimodal backbone. Trained on data with missing modalities, MARKER dynamically routes each input to a sparse combination of LoRA experts, conditioned on both the missingness pattern and instance-specific features, enabling adaptive prediction without maintaining per-scenario prompt banks. Importantly, this design decouples adaptation from per-pattern prompt specification, providing a unified mechanism across varying missingness conditions. Extensive experiments on three real-world datasets demonstrate the effectiveness of MARKER for incomplete multimodal learning.

1 INTRODUCTION

The rapid development of pre-trained multimodal models in recent years Kim et al. (2021); Radford et al. (2021); Fan et al. (2023); Shu et al. (2025) has catalyzed remarkable progress in multimodal learning, driving advancements in tasks such as cross-modal retrieval Wang et al. (2024), image captioning Liu et al. (2025), and multimodal sentiment analysis Wang et al. (2025). While recent advances in multimodal learning have achieved remarkable success, most studies implicitly assume the complete availability of all modalities during both model pre-training and downstream inference. However, in real-world scenarios, device failures, unstable networks, and privacy concerns may lead to missing modalities in the input, which can severely degrade performance and robustness Ma et al. (2022); Hazarika et al. (2022).

Recently, fueled by the growing capabilities of pre-trained models and the advancement of parameter-efficient fine-tuning (PEFT) techniques Wang et al. (2022b); Xu et al. (2025); Hu et al. (2022); Cheng et al. (2025), an increasing number of researchers have turned to prompt-tuning as a strategy to address the challenge of modality missingness. These prompt-based approaches exploit the powerful prior knowledge of pre-trained models to effectively handle missing modality scenarios by optimizing the inserted prompts (shown in Figure 1).

However, existing prompt-based methods typically face two problems: (1) they organize adaptation around scenario-conditioned prompt sets. For example, MAPs Lee et al. (2023) maintains separate prompt parameters for different modality-availability conditions, while MSPs Jang et al. (2024) reduces the number of prompt sets but still relies on modality-dependent prompts. Consequently, each missingness condition must be associated with its corresponding prompt parameters, coupling adaptation to predefined missingness patterns and resulting in a less unified adaptation framework. (2) Most methods employ static prompts shared by all samples under the same modality-missing condition, making them modality-missing-aware but instance-agnostic. Their adaptation therefore

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

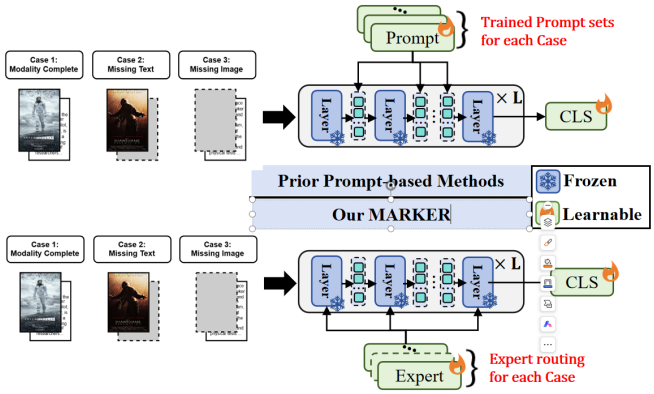


Figure 1: The comparison between existing prompt-based methods and our MARKER. Prompt-based methods address modality-incomplete inputs by reconstructing missing modalities at the data level, inserting specific prompts into data embedding and intermediate hidden state for each modality-missing scenario. MARKER, on the other hand, introduces trainable experts at the model level, adaptively selects experts combination according to the modality-missing scenarios as well as specific instance characteristics.

cannot account for the semantic differences among individual samples that share the same missingness pattern. Recent methods alleviate this limitation through additional input-specific mechanisms. SCP Pipoli et al. (2025) queries a learnable prompt memory using intermediate semantic representations, whereas RAGPT Lang et al. (2025) retrieves relevant information from an external datastore. Although these mechanisms enable instance-level adaptation, they introduce auxiliary query or retrieval modules beyond the learnable prompt parameters, resulting in a more involved adaptation pipeline.

Motivated by these limitations, we develop a MoE-augmented PEFT variant based on low-rank adapters. As illustrated in Figure 2, instead of performing data-level prompt-based adaptation, we insert a mixture of LoRA experts into the pretrained backbone to enable adaptive incomplete multimodal learning. By fine-tuning on datasets with missing modalities, our approach adaptively routes LoRA experts for each input, conditioned on both the current missingness pattern and instance-specific characteristics. The main contributions of this paper are summarized as follows:

- We introduce MARKER, which mitigates performance degradation under incomplete multimodal inputs by selecting a sparse combination of experts conditioned on both modality availability and instance-specific characteristics.
- We investigate different MoE designs, including linear-wise, attention-wise, and their combination, and further dissect expert routing behavior to reveal how missing-modality conditions and instance characteristics impact expert selection
- Extensive experiments on three real-world datasets against 11 competitive baselines demonstrate the effectiveness of MARKER across diverse missing-modality conditions.

2 Related Work

Multimodal Learning with Missing Modalities. Existing approaches for incomplete multimodal learning can be categorized into three types: *(1) Generation-based methods* Ma et al. (2021); Yuan et al. (2021); Woo et al. (2023) aim tro reconstruct the missing modalities. These methods often rely on generative models to infer the latent representation of the absent modality conditioned on the observed modalities. *(2) Joint learning methods* Zuo et al. (2023); Wang et al. (2023); Yao et al. (2024) focus on learning unified representations across modalities during training, with the goal of capturing intrinsic correlations of modalities. *(3) Prompt-based methods* Lee et al. (2023); Jang et al. (2024); Pipoli et al. (2025); Lang et al. (2025) have recently emerged as a more resource-friendly way to handle modality-missing problems. These approaches exploit the knowledge of pre-trained multimodal models and introduce few learnable prompts to encode modality-specific cues, enabling the model to adapt incomplete multimodal learning scenarios.

PEFT for Pre-trained Multimodal Models. Existing multimodal PEFT methods mainly follow prompt- and adapter-based approaches. CoOp Zhou et al. (2022b) learns continuous textual context vectors for CLIP, while CoCoOp Zhou et al. (2022a) conditions these prompts on individual images. MaPLe Khattak et al. (2023) extends prompt learning to both the visual and textual branches, and Prompt-Ladder Cai et al. (2025) reduces back-propagation memory through a lightweight ladder network. Adapter-based methods instead modify the encoded features. CLIP-Adapter Gao et al. (2024b) attaches learnable adapters to the visual and textual branches of CLIP, whereas Tip-Adapter Zhang et al. (2022) constructs a training-free cache-based adapter for efficient downstream transfer. Recent studies combine MoE with PEFT to support multi-task and complex adaptation settings Wang et al. (2022a); Zadouri et al. (2023); Yu et al. (2025). The heterogeneous missingness patterns and instance characteristics in incomplete multimodal inputs motivate the use of MoE-based PEFT, where lightweight routing can assign each input to specialized experts without fine-tuning the entire backbone.

3 METHODOLOGY

In this paper, we introduce MARKER (Shown in Figure 2), which insert Mixture of LoRA experts into the pre-trained model to achieve adaptively incomplete multimodal learning. By fine-tuning on datasets with missing modalities, MARKER can adaptively select a sparse combination of LoRA experts based on specific modality-missing scenario and instance unique characteristics, thereby enhancing the model's inference performance and robustness. This section will provide a detailed introduction of MARKER.

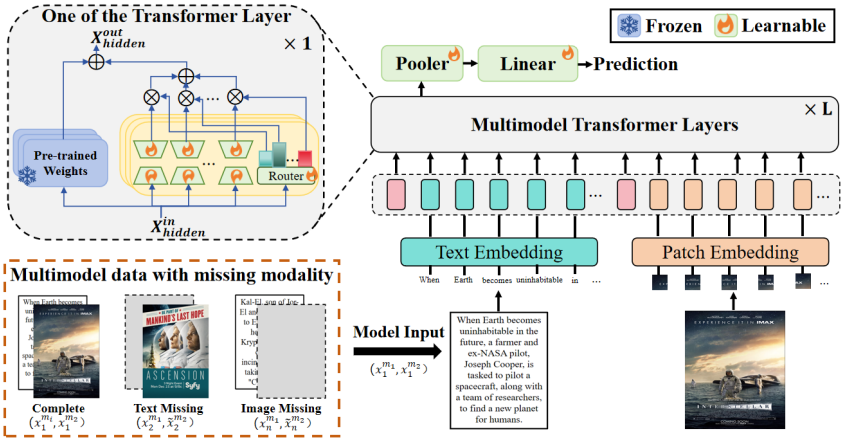


Figure 2: The overview of MARKER. By fine-tuning on datasets with missing modalities, MARKER can adaptively select a sparse combination of LoRA experts based on the current modality missingness and data unique characteristics, thereby enhancing the model's inference performance and robustness.

3.1 PROBLEM STATEMENT

Assuming the multimodal dataset D contains $M = 2$ modalities, denoted as m_1 and m_2, representing text and image modalities, respectively. Since all modalities missing are meaningless, the model needs to handle a total of $2^M - 1$ possible missing modality cases. Specifically, for dataset within two modalities (text and image), three modality-missing scenarios will arise: $D^c = \{(x_i^{m_1}, x_i^{m_2}, y_i)\}_1^{n_c}$, $D^{m_1} = \{(x_i^{m_1}, y_i)\}_1^{n_1}$, $D^{m_2} = \{(x_i^{m_2}, y_i)\}_1^{n_2}$, where D^c, D^{m_1}, and D^{m_2} are the subset of different modality missing scenarios and $D = \{D^c, D^{m_1}, D^{m_2}\}$. y_i denotes the i-th instance label. $x_i^{m_1}$ and $x_i^{m_2}$ are the instance of two modalities. n_c, n_1, and n_2 represent the total number of instances for different missing modality scenarios. In order to ensure input consistency, missing modalities will be supplemented with dummy data (eg, empty string for text, zeros for image) in subset D^{m_1} and D^{m_2}, which results $\tilde{D}^{m_1} = \{(x_i^{m_1}, \tilde{x_i}^{m_2}, y_i)\}_1^{n_1}$ and $\tilde{D}^{m_2} = \{(\tilde{x_i}^{m_1}, x_i^{m_2}, y_i)\}_1^{n_2}$. The final multimodal dataset with missing modalities can be denote as: $\tilde{D} = \{D^c, \tilde{D}^{m_1}, \tilde{D}^{m_2}\}$, which is a mixture of subset with different modalities-missing scenarios.

3.2 MIXTURE OF LORA EXPERTS

Traditional MoE architectures Shazeer et al. (2017); Fedus et al. (2022) enhance model capacity by substituting model's linear layers with specialized MoE layers. Each MoE layer comprises a set of N independently parameterized linear layers, referred to as experts, denoted by $\{E_i\}_{i=1}^{N}$. A routing mechanism $R(\cdot)$ is employed to determine the contribution of each expert dynamically, based on the input. Given an input x, the MoE layer produces an output y by aggregating the responses of all experts, each weighted by the router's output. The formulations are as follow:

$$y = \sum_{i=1}^{N} R(x)_i \cdot E_i(x) \tag{1}$$

$$R(x)_i = \text{Softmax}(W_r x) \tag{2}$$

Here, $E_i(x)$ represents the i-th linear expert's output, and $R(x)_i$ indicates the weight of associated router. W_r is the trainable weight matrix for router. $R(\cdot)$ typically computes a probability distribution over experts conditioned on x, enabling soft selection or sparse routing depending on the design. Traditional MoE architecture can leverage the router to enable each expert for realizing unique capabilities.

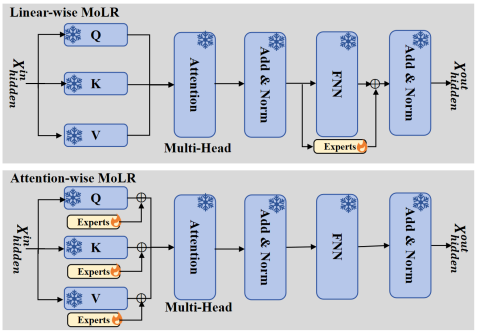


Figure 3: The illustration of two Mixture of LoRA Experts designs.

Unlike traditional MoE approaches that replace the model's linear layers with MoE layers containing multiple linear experts, we propose Mixture of LoRA Experts (MoLR), which uses multiple efficient LoRA experts to substitute the parameter changes during fine-tuning. This significantly reduces the demand for computational resources and learning data. In this paper, we explore the two designs of the mixture of LoRA experts architecture: *Linear-wise MoLR* and *Attention-wise MoLR*.

3.2.1 LINEAR-WISE MOLR.

A common design that target at linear layers within the pre-trained model, as shown on right of the Figure 3. Suppose that the linear layer of the i-th transformer layer is defined as:

$$Linear(x)_i = W_i x + \Delta W_i x \tag{3}$$

By applying Linear-wise MoLR, the weight changes ΔW_i during fine-tuning will be substitute with low-rank format MoE layers, the new linear layer can be formulated as follows:

$$\hat{Linear}(x)_i = W_i x + \frac{\alpha}{r} \sum_{i=1}^{N} R(x)_i B_i A_i x \tag{4}$$

As previously mentioned, r represents LoRA rank, α is the scalling factor, $R(x)_i$ indicates the router, A_i and B_i are the low-rank matrices.

3.2.2 ATTENTION-WISE MOLR.

We also introduce another designs of MoLR that focus on MSA layers Vaswani et al. (2017), as shown on the left of Figure 3. We replace the weight changes of original query, key and value with low-rank experts. Same as Linear-wise MoLR, the formulations can be defined as follow:

$$\hat{Q}(x)_i = W_i^q x + \frac{\alpha}{r} \sum_{i=1}^{N} R(x)_i^q B_i^q A_i^q x \tag{5}$$

$$\hat{K}(x)_i = W_i^k x + \frac{\alpha}{r} \sum_{i=1}^{N} R(x)_i^k B_i^k A_i^k x \tag{6}$$

$$\hat{V}(x)_i = W_i^v x + \frac{\alpha}{r} \sum_{i=1}^{N} R(x)_i^v B_i^v A_i^v x \tag{7}$$

Where $W_i^q, W_i^k, W_i^v \in \mathbb{R}^{d \times d}$ is the projection weights of attention layer. Then, the attention layers can be defined as:

$$\hat{Attention}(x) = Softmax\left(\frac{\hat{Q}(x)_i \hat{K}(x)_i^\top}{\sqrt{d}}\right) \hat{V}(x)_i \tag{8}$$

Since Attention-wise MoLR targets more objectives, it introduces more additional expert parameters compared to Linear-wise MoLR. In the experiments, the variant using linear-wise MoLR is named MARKER-l, and the one using attention-wise MoLR is named MARKER-a. Additionally, we designed a variant that incorporates both two MoLR, named MARKER-b.

3.2.3 Routing Strategies and Expert Shape Allocation

In addition to the types of MoE designs, we also try different routing strategies and expert allocation shapes in the experiment, to investigate their impact under different modality-missing scenarios.

For routing strategies, we adopt Top-K strategy that select the most appropriate K experts for each layer in a discrete way. Besides, we also try a *SoftMerge* strategy Zadouri et al. (2023) that compute a weighted average of experts. The results show that simple Top-K is better than *SoftMerge*, as the latter strategy may introduce more noise in scenarios with different missing modalities.

For expert allocation, prior work Gao et al. (2024a) investigates the impact of different layer-wise expert distributions and finds that a $\bigtriangledown$ allocation performs the best for pretrained models. This allocation assigns more experts to higher layers (the output side) and fewer experts to lower layers (the input side). Following this setting, we consider four expert-allocation shapes: MARKER-$\bigtriangledown$, MARKER-$\bigtriangleup$, MARKER-$\bowtie$, and MARKER-$\square$. MARKER-$\bigtriangledown$ gradually increases the number of experts toward the higher layers, whereas MARKER-$\bigtriangleup$ follows the opposite allocation. The hourglass configuration, MARKER-$\bowtie$, assigns more experts to the lower and higher layers while allocating fewer experts to the intermediate layers. MARKER-$\square$ assigns the same number of experts to every layer. Our experiments show that MARKER-$\square$ achieves the good overall performance under modality-missing conditions.

4 Evaluations

4.1 Experimental Settings

Datasets. Following prior research Lee et al. (2023); Jang et al. (2024), we evaluate our MARKER across three multimodal downstream datasets: (1) *MM-IMDb* Arevalo et al. (2017), a benchmark dataset designed for movie genre classification that leverages both textual and visual information; (2) *UPMC Food-101* Wang et al. (2015), which emphasizes image classification with complementary textual input; and (3) *Hateful Memes* Kiela et al. (2020), which targets hate speech detection within memes by integrating image and text modalities. For the evaluation metrics of each dataset, we report F1-Micro for MM-IMDb dataset, accuracy for UPMC Food-101, and AUROC for Hateful Memes.

Missing data setting. Both the training and test sets contain incomplete multimodal samples. Let η denote the proportion of samples with missing modalities. For **Text** or **Image** missing, η of the samples retain only the image or text modality, respectively, while the remaining $1-\eta$ are complete. For **Both** missing, $\eta/2$ of the samples are text-only, $\eta/2$ are image-only, and $1-\eta$ are complete. More generally, for M modalities, the missing portion can be uniformly distributed across the 2^M-2 nonempty incomplete modality combinations, assigning $\eta/(2^M-2)$ to each combination. We set the default missing rate to $\eta = 70\%$ in all experiments.

Implementation Details. Following prior works Lee et al. (2023); Lang et al. (2025), we utilize the pre-trained ViLT Kim et al. (2021) as the backbone and the parameters of ViLT remain frozen during fine-tuning. Only inserted experts and downstream related pooler and linear layers need to optimize. AdamW optimizer is adopted with a learning rate of $1e-4$. We set each LoRA expert with rank $r = 4$, $\alpha = 8$, and LoRA dropout is 0.1. The top-2 routing strategy is used as default. For default

epxert allocation shape, we construct a square shape (□) MoE, which allocate 4 experts to all layers of ViLT. All experiments are conducted with an NVIDIA RTX 4090 GPU.

Baselines We compare MARKER with 11 competitive baselines from three categories. Generation-based methods include SMIL Ma et al. (2021), TFR-Net Yuan et al. (2021), and AcMAE Woo et al. (2023). Joint learning methods include IF-MMIN Zuo et al. (2023), ShaSpec Wang et al. (2023), DrFuse Yao et al. (2024), and CorrKD Li et al. (2024). Prompt-based methods include MAPs Lee et al. (2023), MSPs Jang et al. (2024), SCP Pipoli et al. (2025), and RAGPT Lang et al. (2025). Among them, RAGPT additionally employs retrieval augmentation. We therefore report RAGPT in the retrieval-based comparison, while the remaining baselines are included in the non-retrieval comparison.

Table 1: Comparison of different methods across MM-IMDb, Hateful Memes, and UPMC Food-101 datasets. Text, Image, Both indicate different Missing Type with default 70% missing rate, as detailed in the Missing data setting section.

Missing Type	**MM-IMDb**			**Hateful Memes**			**UPMC Food-101**		
Methods	**Text**	**Image**	**Both**	**Text**	**Image**	**Both**	**Text**	**Image**	**Both**
(a) Non-retrieval methods									
SMIL	38.32	27.57	35.12	50.32	58.50	54.63	51.83	49.86	46.77
TFR-Net	37.70	39.45	37.24	51.18	55.57	52.12	65.91	67.58	63.41
AcMAE	47.47	43.82	44.05	55.74	59.66	57.25	69.28	73.75	71.15
IF-MMIN	39.63	31.95	31.98	57.62	53.44	55.19	66.76	64.36	68.53
ShaSpec	44.04	44.23	44.06	58.75	60.30	60.96	60.99	74.87	70.02
DrFuse	47.05	44.09	48.83	57.60	60.66	55.84	66.30	75.09	68.23
CorrKD	44.82	39.48	41.20	58.74	55.59	57.91	61.37	66.83	62.87
MAPs	46.12	44.86	45.48	58.62	60.16	58.89	67.02	75.62	72.52
MSPs	49.16	44.62	48.28	59.60	60.05	59.08	71.74	79.09	74.46
SCP	**50.51**	50.62	48.94	52.13	62.33	60.71	71.98	84.55	76.17
MARKER-l	49.84	49.70	**49.43**	59.56	62.54	59.32	73.40	86.41	78.70
MARKER-a	49.43	50.95	49.10	59.80	**63.45**	**61.21**	73.56	86.24	78.80
MARKER-b	49.34	**51.28**	49.38	**60.48**	63.41	59.27	**73.69**	**86.61**	**78.85**
(b) Retrieval-based methods									
RAGPT	55.16	46.44	50.89	**64.10**	62.57	63.47	75.53	81.98	76.94
MARKER-b-r	**58.46**	**49.42**	**51.76**	63.76	**65.22**	**64.68**	**78.79**	**86.86**	**80.09**

4.2 Main Results

We construct 4 MARKER variants for evaluation. MARKER-l, MARKER-a, and MARKER-b employ Linear-wise MoLR, Attention-wise MoLR, and their combination, respectively. We additionally construct a retrieval-augmented variant, MARKER-b-r, to compare with RAGPT under a matched retrieval setting. Built upon MARKER-b, MARKER-b-r adopts the same CLIP-L/14-336 retriever and top-5 neighbor configuration as RAGPT, using the retrieved representations for missing-modality reconstruction and expert-router conditioning, while adaptively incorporating the retrieved-label prior to improve the final prediction.

Table 1 reports the results across three datasets and nine modality-missing settings. Among the non-retrieval methods, MARKER ranks first or second in every setting and achieves the best result in eight of the nine settings. MARKER-b obtains the most first-place results, leading on five settings, while MARKER-a and MARKER-l lead on two and one settings, respectively. Among the retrieval-based methods, MARKER-b-r outperforms RAGPT in eight of the nine settings, including all settings on MM-IMDb and UPMC Food-101. Under these comparison settings, the results demonstrate the effectiveness of MARKER's instance-adaptive expert routing, which selects a sparse combination of LoRA experts according to both the modality-missing pattern and sample-specific features, while also showing that this mechanism can effectively incorporate retrieved representations.

4.3 Ablation Study

Cross-scenario generalization. We evaluate cross-scenario generalization on MM-IMDb by training each method under the default 70% missing-rate protocol and varying the missing rate at test time from 0% to 100%. As shown in Figure 4, the performance of all methods decreases as the missing rate increases, confirming that less available multimodal information makes prediction progressively more

difficult. The frozen ViLT backbone declines rapidly across all three settings, while the adaptation-based methods retain considerably better performance. MAPs also shows a pronounced decrease under severe missingness, particularly in the Missing-both and Missing-text settings. In contrast, SCP and MARKER-b exhibit flatter degradation curves, indicating stronger generalization beyond the missing rate used during training. MARKER-b achieves competitive performance across most settings, with particularly stable behavior under Missing both and Missing image. Under Missing text, SCP performs slightly better at several intermediate missing rates, while MARKER-b remains competitive and performs favorably under more severe missingness.

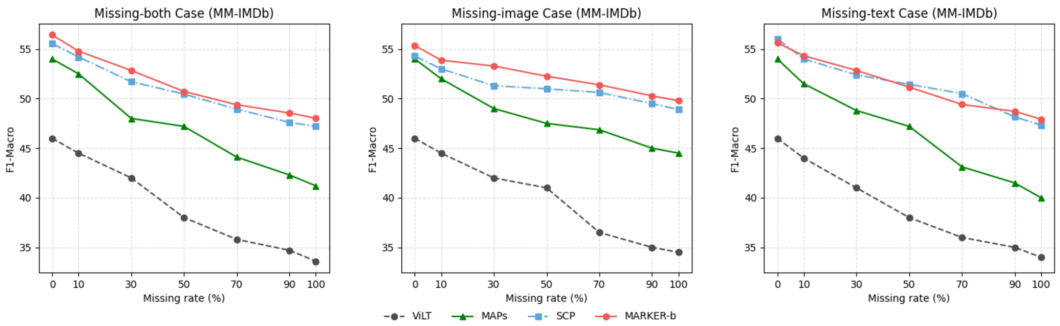


Figure 4: The comparison results of MARKER with several prompt-based methods on the MM-IMDb dataset with different missing rates.

The impact of different expert allocation shapes and routing strategies. To further investigate the impact of different MoE configurations on modality-missing scenarios, we conducted ablation studies on both the routing strategies and expert allocation shapes. Table 2 presents the results of different routing strategies. We find that the default top-2 strategy achieves the best performance. Although previous studies have shown that soft merge can perform well by integrating more experts, in scenarios with modality missing, experts may specialize in different modalities, and merging their outputs could introduce additional noise, ultimately leading to degraded performance. Table 3 presents the results of different expert allocation shapes. Previous studies have found that an $\bigtriangledown$ expert shape tends to yield better performance. However, our results show that a $\square$ expert shape performs better in scenarios with modality missing. We think that this is because missing modalities require the model to involve more experts on the input side during training to compensate for the missing information of original data.

Table 2: Analysis of routing strategies.

Routing strategy	Text	Image	Both
Top-2 (Default)	49.34	51.28	49.38
Top-1	45.28	47.47	44.93
SoftMerge	48.62	50.21	46.74

Table 3: Analysis of expert allocation shapes.

Expert allocation Shape	Text	Image	Both
MARKER-$\bigtriangledown$	49.17	51.26	49.01
MARKER-$\bigtriangleup$	48.32	51.14	48.84
MARKER-$\bowtie$	48.96	51.09	49.07
MARKER-$\square$ (Default)	49.34	51.28	49.38

4.4 Expert Routing Behavior under different Missing Modalities

In this section, we analyze the expert routing behavior of MARKER-b on the MM-IMDb (corresponding analyses on the other datasets are provided in the Appendix). Instead of following the default 70% missing-rate protocol, we evaluate each of the 7,799 test instances under three paired conditions: *Complete*, *Text missing*, and *Image missing*. Because each instance is evaluated under all three conditions with only its modality availability changed, the observed differences can be better attributed to the input missing condition. Specifically, we investigate three questions:

(1) Do different missing-modality conditions induce distinct expert preferences? To answer this question, Figure 5 visualizes the 24 linear-wise MoLR modules (the full visualization covering both linear-wise and attention-wise modules is provided in the Appendix due to space limitations) using the preference deviation $p_{\text{select}} - K/E$, where p_{select} is the expert-selection rate and $K/E = 0.5$ is the uniform top-2 routing baseline. Positive and negative values indicate expert preference and

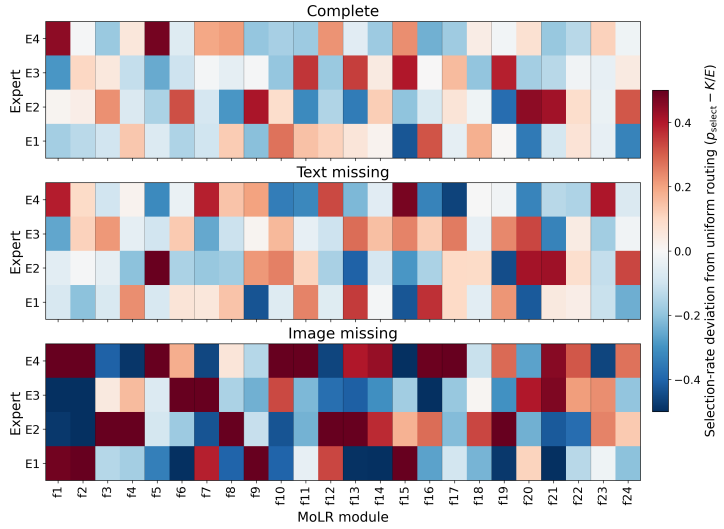


Figure 5: Expert-selection preferences across the 24 linear-wise MoLR modules on MM-IMDb under Complete, Text-missing, and Image-missing conditions. Each cell shows the deviation from the uniform top-2 routing rate, where red indicates preference and blue indicates suppression.

suppression, respectively, while a larger absolute value represents a stronger routing tendency. The heatmap reveals both shared and condition-specific expert preferences. For example, E4 in $f1$ and E2 in $f24$ are preferred under all three conditions, suggesting that they capture transformations generally useful across different modality-availability patterns. In contrast, E2 in $f5$ is strongly preferred under Text missing but suppressed under the other conditions, whereas E2 in $f3$ and $f4$ is preferred under Image missing but suppressed under Text missing. Table 4 further reports the selection rates averaged over all 24 modules. The four expert slots are relatively balanced under Complete inputs. Text missing increases the selection rates of E1 and E3 while decreasing those of E2 and E4. In contrast, Image missing increases the selection rates of E2 and E4 while decreasing those of E1 and E3. Overall, all experts remain actively utilized across the three conditions without global expert collapse, while different missing-modality conditions induce distinct expert preferences.

Table 4: Average expert-selection rate (%) across the 24 linear-wise MoLR modules on the MM-IMDb test set.

Input Condition	E1	E2	E3	E4
Complete	46.73	**51.74**	50.52	51.01
Text missing	48.31	50.12	**53.49**	48.08
Image missing	41.15	53.38	44.61	**60.86**

(2) Do different missing-modality conditions form distinct global routing structures? We concatenate the router probabilities from all MoLR modules for each test instance and project the resulting representation into a shared two-dimensional PCA space. A larger distance in this space indicates greater differences in routing preferences and generally corresponds to more distinct Top-K expert combinations. As shown in Figure 6, the Image-missing representations form a compact cluster that is clearly separated from the Complete and Text-missing representations. The Complete and Text-missing representations partially overlap, but their centroids remain distinct. This difference may arise from ViLT's token composition: under the default setting, each input contains up to 40 text tokens but approximately 145 visual tokens. Consequently, removing the text perturbs a smaller portion of the mean-pooled sequence than removing the image, keeping its routing representations and expert combinations closer to those under the Complete condition. These results demonstrate that different missing-modality conditions form distinct global routing structures. However, the degree of separation among these structures may depend on factors such as the backbone architecture and dataset characteristics.

(3) Do different samples under the same missing-modality condition still activate different expert combinations? To evaluate instance-specific routing, we compare the Top-K expert combinations selected for different samples under the same missing condition and average their Jaccard distances across all MoLR modules. Figure 7 visualizes the resulting distributions using violin plots, where

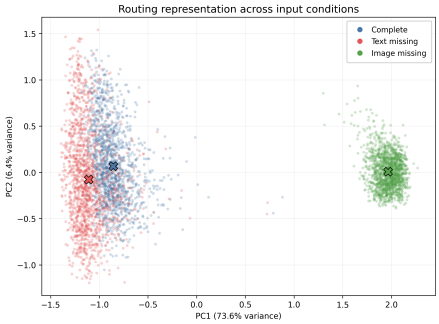


Figure 6: Routing representations on MM-IMDb projected into a shared PCA space.

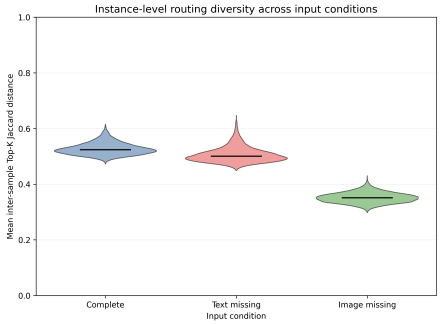


Figure 7: Distribution of the mean inter-sample Top-K Jaccard distance averaged across all MoLR modules on MM-IMDb.

zero denotes identical expert combinations and larger values indicate greater routing diversity. The average distances are 0.53, 0.51, and 0.35 for Complete, Text missing, and Image missing, respectively. Complete and Text missing exhibit similar levels of diversity, whereas Image missing shows more concentrated routing, consistent with its compact PCA cluster and stronger expert preferences. However, all three distributions remain clearly above zero, showing that MARKER combines condition-dependent expert preferences with instance-specific routing rather than assigning a fixed expert combination to each missing condition.

4.5 Efficiency and Resource Overhead Analysis

We evaluate the parameter and computational efficiency of MARKER on MM-IMDb. GFLOPs are measured per sample for end-to-end inference, while latency and throughput use batch sizes of 1 and 16, respectively. RAGPT originally precomputes its memory bank and retrieval results. For a fair comparison, we include online query encoding, similarity search, Top-K selection, and feature lookup in its efficiency measurements.

As shown in Table 5, MARKER introduces a moderate number of trainable parameters compared with MAPs and SCP, while remaining smaller than RAGPT. MARKER also achieves the lowest GFLOPs, indicating that sparse Top-K routing and low-rank experts effectively limit computation without extending the input sequence. However, this reduction in GFLOPs does not directly translate into the lowest latency or highest throughput. Dynamic routing, expert dispatch, and numerous small expert operations introduce indexing and kernel-launch overhead that is not fully reflected by FLOPs, whereas MAPs and SCP benefit from more regular dense computation. RAGPT incurs the highest computational cost and the lowest throughput because its online pipeline additionally performs query encoding and retrieval.

Table 5: Efficiency comparison between MARKER and representative baselines.

Method	Trainable Params	Trainable Ratio	GFLOPs	Latency (ms)	Throughput (Samples/s)
MAPs	2.03M	1.79%	61.50	16.95	274.86
RAGPT	7.25M	6.13%	390.18	57.29	63.53
SCP	2.06M	1.77%	66.07	19.36	221.46
MARKER	3.26M	2.86%	33.89	70.34	124.35

5 Conclusion

In this work, we propose MARKER, a Mixture of LoRA Experts for Adaptive Incomplete Multimodal Learning. Unlike existing prompt-based approaches that suffer from static prompts and scalability issues, MARKER dynamically selects expert combinations based on both modality-missing scenarios and instance-specific features. Through extensive experiments on three real-world datasets, MARKER outperform baselines in many cases, demonstrating its effectiveness in handling incomplete multimodal scenarios.

AI USE STATEMENT

Generative AI tools were used to improve the clarity and readability of the manuscript. The authors reviewed all AI-assisted edits and take responsibility for the final content of this work.

REFERENCES

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and FabioA. González. Gated multimodal units for information fusion. *Learning,Learning*, Feb 2017.

Siqi Cai, Xuan Liu, Jingling Yuan, and Qihua Zhou. Prompt-ladder: Memory-efficient prompt tuning for vision-language models on edge devices. *Pattern Recognition*, 163:111460, 2025.

Jiaxiang Cheng, Pan Xie, Xin Xia, Jiashi Li, Jie Wu, Yuxi Ren, Huixia Li, Xuefeng Xiao, Shilei Wen, and Lean Fu. Resadapter: Domain consistent resolution adapter for diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2438–2446, 2025.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36:35544–35575, 2023.

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

Chongyang Gao, Kezhen Chen, Jinmeng Rao, Baochen Sun, Ruibo Liu, Daiyi Peng, Yawen Zhang, Xiaoyuan Guo, Jie Yang, and VS Subrahmanian. Higher layers need more lora experts. *arXiv preprint arXiv:2402.08562*, 2024a.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024b.

Devamanyu Hazarika, Yingting Li, Bo Cheng, Shuai Zhao, Roger Zimmermann, and Soujanya Poria. Analyzing modality robustness in multimodal sentiment analysis. *arXiv preprint arXiv:2205.15465*, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Jaehyuk Jang, Yooseung Wang, and Changick Kim. Towards robust multimodal prompting with missing modalities. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8070–8074. IEEE, 2024.

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, Firooz Hamed, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Cornell University - arXiv,Cornell University - arXiv*, May 2020.

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

Jian Lang, 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.

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.

An-An Liu, Quanhan Wu, Ning Xu, Hongshuo Tian, and Lanjun Wang. Enriched image captioning based on knowledge divergence and focus. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

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.

Vittorio Pipoli, Federico Bolelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, Costantino Grana, Rita Cucchiara, and Elisa Ficarra. 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.

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.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Fangxun Shu, Yue Liao, Lei Zhang, Le Zhuo, Chenning Xu, Guanghao Zhang, Haonan Shi, Long Chan, Tao Zhong, Zhelun Yu, Wanggui He, Siming Fu, Haoyuan Li, Si Liu, Hongsheng Li, and Hao Jiang. Llava-mod: Making llava tiny via moe-knowledge distillation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

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

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

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. Dlf: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 21180–21188, 2025.

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.

Yabing Wang, Fan Wang, Jianfeng Dong, and Hao Luo. Cl2cm: Improving cross-lingual cross-modal retrieval via cross-lingual knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5651–5659, 2024.

Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. AdaMix: Mixture-of-adaptations for parameter-efficient model tuning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 139–149, 2022b.

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.

Chen Xu, Yuhan Zhu, Haocheng Shen, Boheng Chen, Yixuan Liao, Xiaoxin Chen, and Limin Wang. Progressive visual prompt learning with contrastive feature re-formation. *International Journal of Computer Vision*, 133(2):511–526, 2025.

Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 16416–16424, 2024.

Beiming Yu, Zhenfei Yang, and Xiushuang Yi. MoKA:parameter efficiency fine-tuning via mixture of kronecker product adaption. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 10172–10182. Association for Computational Linguistics, January 2025.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 4400–4407, 2021.

Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermiş, Acyr Locatelli, and Sara Hooker. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. *arXiv preprint arXiv:2309.05444*, 2023.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pp. 493–510. Springer, 2022.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022a.

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

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

A APPENDIX

The appendix is organized as follows:

- Appendix A.1: Additional Experiments;
- Appendix A.1.1: Detailed Implementation Details;
- Appendix A.1.2: Parameter Scalability with Increasing Modalities;
- Appendix A.1.3: Expert Routing Behavior on Hateful Memes and UPMC Food-101;
- Appendix A.1.4: Layer-wise Routing Dynamics and Specialization;
- Appendix A.2: Theoretical Analysis;
- Appendix A.2.1: Impact of Modality Missingness on Multimodal Models;
- Appendix A.2.2: Why Prompt/LoRA Can Mitigate Missingness-Induced Degradation;
- Appendix A.2.3: MARKER vs. Static Prompt Methods: Conditional Updates via MoLR.

A.1 Additional Experiments

A.1.1 Detailed Implementation Details

We initialize the backbone from `dandelin/vilt-b32-mlm`. The embedding and Transformer parameters of ViLT remain frozen, while the inserted MoLR modules, ViLT pooler, and task-specific classifier are optimized. MARKER-l applies MoLR to the feed-forward linear layers, MARKER-a applies MoLR to the query, key, and value projections of self-attention, and MARKER-b combines both placements. By default, MoLR is inserted into all 12 Transformer layers using the square expert-allocation shape, with four experts per layer. Each LoRA expert uses rank $r = 4$ and scaling factor $\alpha = 8$. The router operates at the sample level by mean-pooling token representations and selects the top-2 experts in each MoLR module. The maximum text length is 40, and input images are resized to 384×384.

All models are trained for at most 20 epochs using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The per-device batch size is 16. We adopt a linear learning-rate schedule, FP16 training, and gradient clipping with a maximum norm of 1.0. For MM-IMDb and UPMC Food-101, the learning rate, weight decay, warm-up ratio, LoRA dropout, and load-balancing loss weight are set to 1×10^{-4}, 0.02, 0.1, 0.1, and 0.01, respectively. For Hateful Memes, these values are set to 7.5×10^{-5}, 0.05, 0.15, 0.15, and 0.02, respectively. MM-IMDb is formulated as a 23-label classification task using the plot-outline text representation, a fixed threshold of 0.5, and F1-Micro as the evaluation metric. UPMC Food-101 is formulated as a 101-class classification task and evaluated using accuracy. Hateful Memes is formulated as a binary classification task and evaluated using AUROC.

MARKER-b-r extends MARKER-b with the retrieval configuration used by RAGPT. We use CLIP-L/14-336 as the retriever, store the top-20 candidates, and select the top-5 neighbors for each query. The retrieved representations are used for missing-modality reconstruction and expert-router conditioning. Their labels are converted into a similarity-weighted prior with a temperature of 0.07 and label smoothing of 0.05, which is incorporated into the final logits through a learned gate initialized to 0.1. The retrieval-module dropout is set to 0.2, and no learnable prompt parameters are introduced. Early stopping begins after five epochs and uses a patience of three with a minimum improvement of 0.001. The reported test results are taken from the epoch achieving the best validation metric.

A.1.2 Parameter Scalability with Increasing Modalities

For a dataset with M modalities, there are $2^M - 1$ valid modality-availability conditions. The increasing diversity of missing-modality patterns may require greater adaptation capacity. MARKER can expand this capacity by increasing either the number of experts in each MoLR module or the LoRA rank of each expert, without introducing condition-specific adaptation modules.

Figure 8 reports the trainable parameter count under different expert numbers and LoRA ranks. The default configuration contains four rank-4 experts in each of the 60 MoLR modules, resulting in 3.26M trainable parameters. Increasing either dimension produces approximately linear parameter growth, in contrast to the exponential growth in the number of missing-modality conditions. For

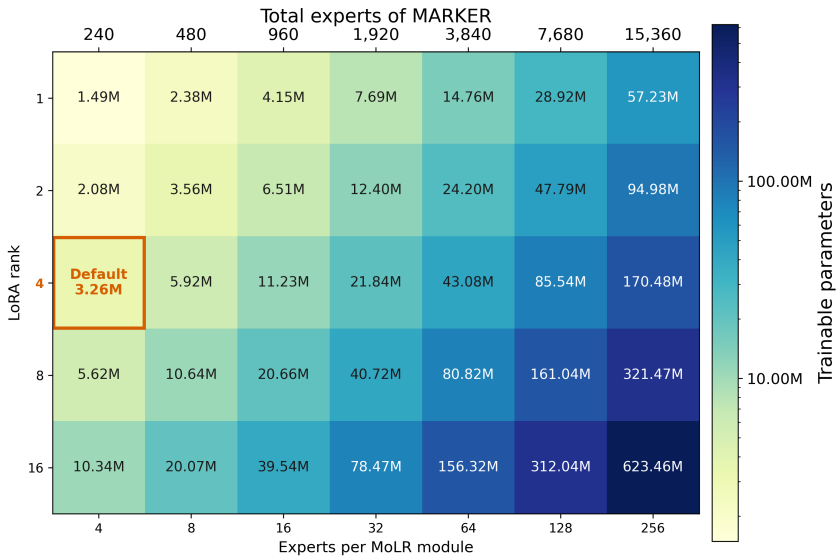


Figure 8: Trainable parameter scaling of MARKER under different numbers of experts per MoLR module and LoRA ranks. Each cell reports the total number of trainable parameters. The upper axis shows the total number of expert instances across the 60 MoLR modules within MARKER, and the orange box indicates the default configuration with four rank-4 experts per module.

example, assigning 256 rank-4 experts to each module yields 15,360 expert instances across the model, while the total number of trainable parameters reaches 170.48M. This configuration provides up to $\binom{256}{2} = 32{,}640$ possible Top-2 expert combinations within each module. Meanwhile, sparse routing activates only two experts per module for each input, corresponding to 120 active expert instances across all 60 modules, regardless of the total expert-pool size. MARKER can therefore increase its routing capacity substantially while retaining sparse, shared adaptation rather than allocating an independent parameter set to every missing-modality condition.

A.1.3 Expert Routing Behavior on Hateful Memes and UPMC Food-101

The main text analyzes the expert-routing behavior of MARKER on MM-IMDb. This section extends the same paired-input protocol to Hateful Memes and UPMC Food-101 to examine whether the observed routing properties persist across datasets with different task characteristics. For each test instance, we construct *Complete*, *Text missing*, and *Image missing* inputs while keeping its identity and label unchanged, and record the routing decisions of all MoLR modules. We examine module-wise expert preferences through expert-selection heatmaps, global routing structures through PCA projections of router probabilities, and instance-specific routing diversity through Top-K Jaccard distances. These analyses evaluate whether MARKER consistently adapts its expert utilization to modality availability while retaining sample-dependent routing within each missing condition.

Expert preferences under different missing conditions. The main text reports only the 24 linear-wise MoLR modules due to space limitations. Here, we extend the visualization to all MoLR modules, including the query, key, and value projections in self-attention and the two linear projections in each FFN layer. Figures 9-11 use the same preference deviation $p_{\text{select}} - K/E$ as the main text, where positive and negative values indicate that an expert is preferred or suppressed relative to uniform top-2 routing, respectively.

The complete heatmaps reinforce the observations from the linear-wise modules reported in the main text. Across the three datasets, several module-specific experts remain preferred under all missing conditions. For example, E4 in $f1$ on MM-IMDb, E1 in $f7$ on UPMC Food-101, and E4 in $f5$ on Hateful Memes consistently exhibit positive preference deviations under Complete, Text-missing, and Image-missing inputs. These stable preferences suggest that some experts capture transformations that remain useful regardless of modality availability. Such shared preferences appear alongside

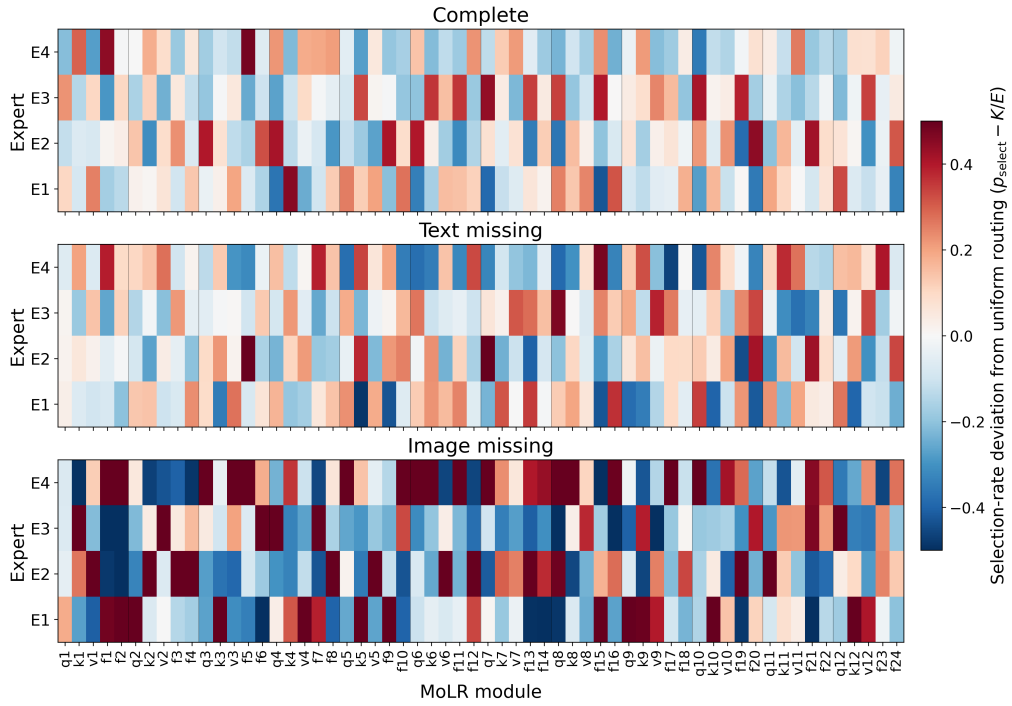


Figure 9: Expert-selection preferences across all MoLR modules on MM-IMDb. The columns cover the query, key, value, and two FFN projections in the 12 Transformer layers. Each cell shows the deviation from the uniform top-2 routing rate, where red indicates preference and blue indicates suppression.

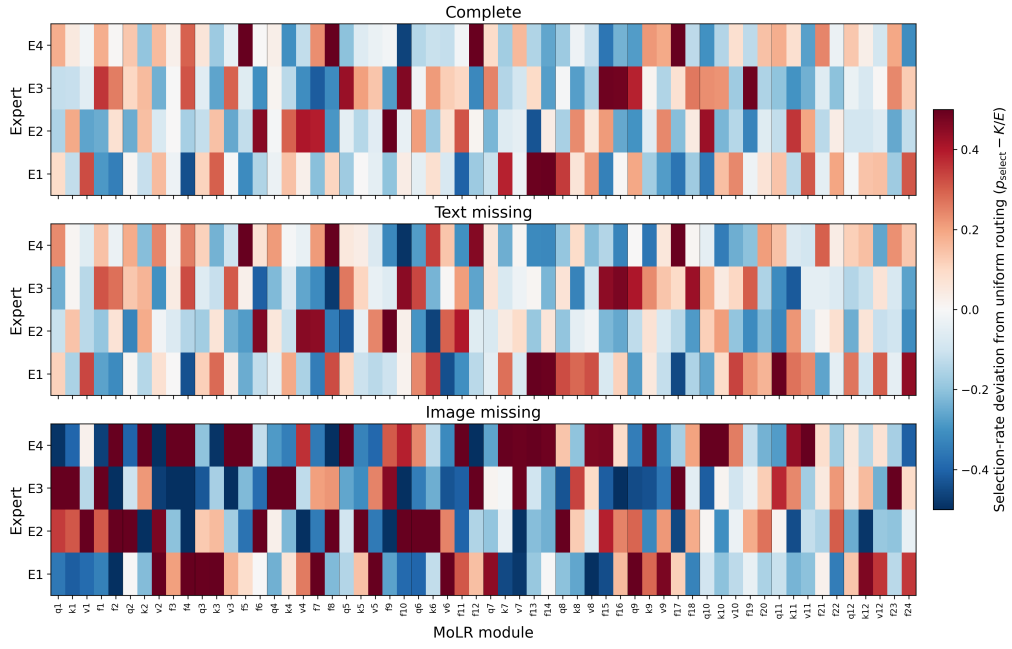


Figure 10: Expert-selection preferences across all MoLR modules on Hateful Memes under the Complete, Text-missing, and Image-missing conditions.

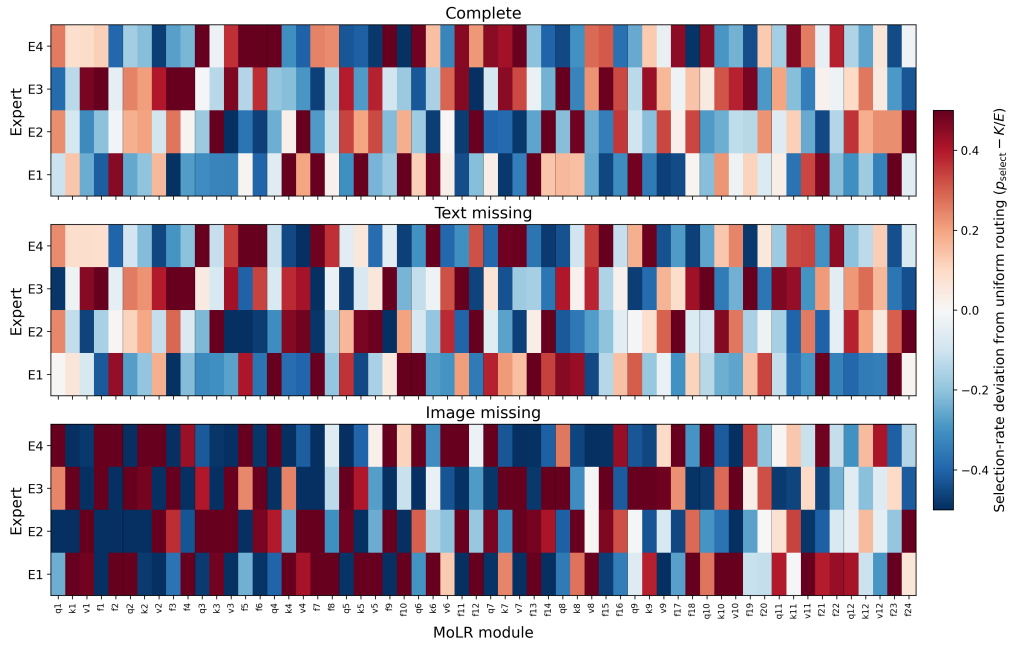


Figure 11: Expert-selection preferences across all MoLR modules on UPMC Food-101 under the Complete, Text-missing, and Image-missing conditions.

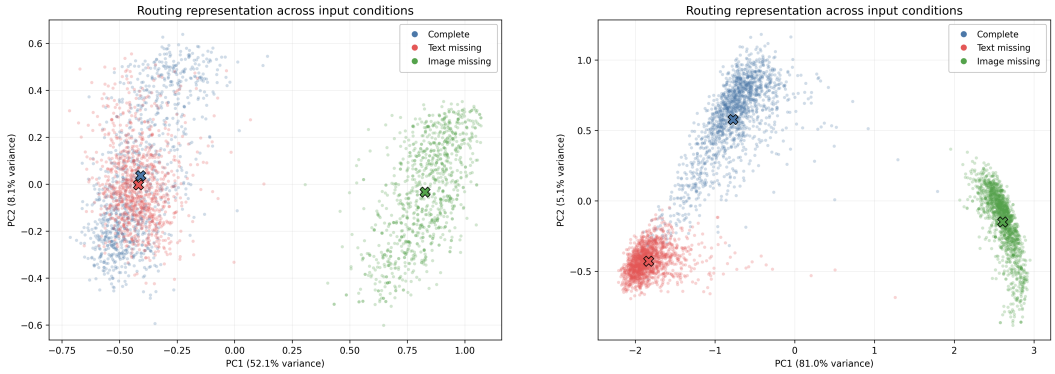
heterogeneous activation patterns across the remaining experts, indicating that no single expert dominates all MoLR modules.

Other module-specific experts exhibit clear condition-dependent preference reversals. On MM-IMDb, E1 in $k5$ is suppressed under Text missing but strongly preferred under Image missing, whereas E3 in $q8$ shows the opposite behavior. On UPMC Food-101, E2 in $v3$ changes from strong suppression under Complete and Text missing to strong preference under Image missing, while E3 in f_9 is preferred under Text missing but suppressed under Image missing. A similar pattern appears on Hateful Memes: E2 in $k6$ is suppressed under Text missing but preferred under Image missing, whereas E2 in $v4$ exhibits the reverse tendency. These examples show that condition-dependent specialization occurs in both attention-wise and linear-wise MoLR modules. Moreover, Image missing generally produces more polarized preference and suppression patterns than Text missing across all three datasets, although the affected module–expert pairs differ by dataset. Overall, the complete heatmaps extend the main-text findings beyond the FFN modules and show that MARKER learns both shared and missingness-dependent expert preferences throughout the Transformer backbone.

Global routing structures. We further examine whether the condition-specific expert preferences form distinct global routing structures. Following the analysis in the main text, we concatenate the router probabilities from all MoLR modules for each test instance and project the resulting representation into a shared two-dimensional PCA space within each dataset. A larger distance in this space indicates a greater difference in overall routing preferences and generally corresponds to more distinct Top-K expert combinations.

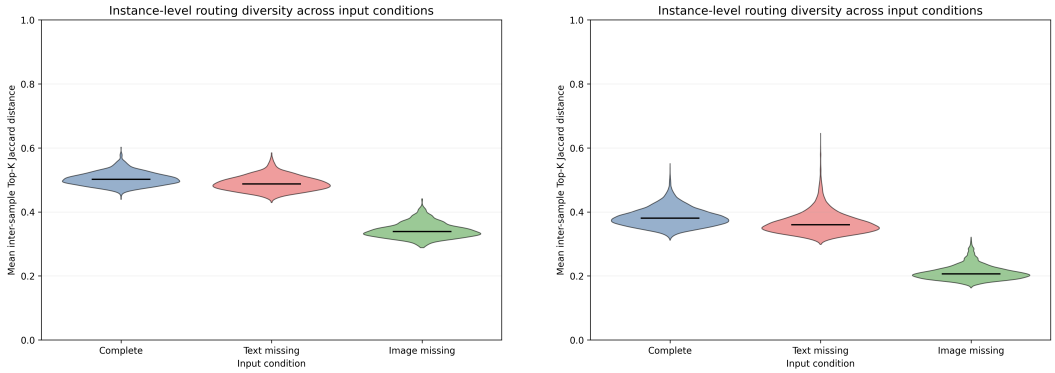
As shown in Figure 12, UPMC Food-101 exhibits three well-separated routing clusters. Text missing shifts the routing representations away from the Complete condition, while Image missing produces a substantially larger displacement and forms a compact cluster on the opposite side of the PCA space. This separation indicates that the router learns distinct global expert-allocation patterns for all three modality-availability conditions on UPMC Food-101.

On Hateful Memes, the Image-missing representations are also clearly separated from the other two conditions, primarily along the first principal component. The Complete and Text-missing representations overlap more substantially, but their centroids remain different, indicating a smaller yet observable change in their global routing preferences. This pattern is similar to MM-IMDb, where



(a) Hateful Memes. (b) UPMC Food-101.

Figure 12: PCA projections of the routing representations under different missing conditions. Each point represents one test instance, and crosses denote the corresponding condition centroids. PCA is performed separately within each dataset.



(a) Hateful Memes. (b) UPMC Food-101.

Figure 13: Distributions of the mean inter-sample Top-K Jaccard distance across all MoLR modules under different missing conditions. Larger distances indicate greater diversity in the expert combinations activated by different samples under the same condition, while the black horizontal lines denote the medians.

Image missing forms a separate cluster while Complete and Text missing remain closer in the routing space.

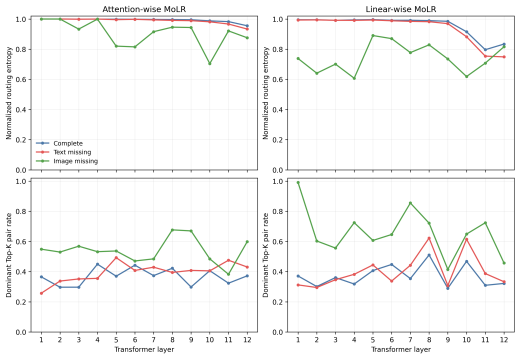
Across the three datasets, Image missing consistently induces the largest shift from the Complete condition, whereas the magnitude of the Text-missing shift varies across datasets. It is most pronounced on UPMC Food-101, moderate on MM-IMDb, and relatively small on Hateful Memes. This difference suggests that the extent of routing reorganization depends not only on the missing modality but also on the task and data characteristics. Nevertheless, the consistent separation of the condition centroids confirms the main-text observation that different missing conditions form distinct global routing structures.

Instance-specific expert combinations. We further extend the instance-level routing analysis to Hateful Memes and UPMC Food-101. Following the setting used for MM-IMDb, we compute the Top-K Jaccard distance between the expert combinations selected for different samples under the same missing condition. The distance is calculated for each MoLR module and then averaged across all modules. A larger distance indicates that samples activate more diverse expert combinations, whereas a distance of zero indicates identical routing decisions. The resulting distributions are visualized using violin plots in Figures 13a and 13b.

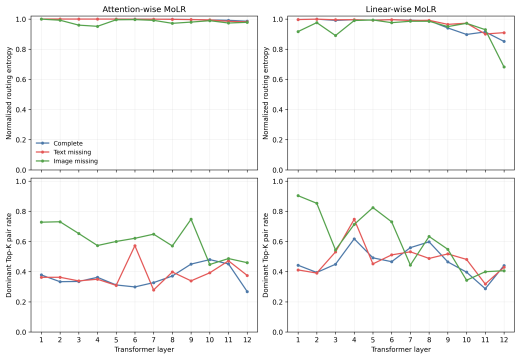
Together with the MM-IMDb results in Figure 7, the distances remain clearly above zero under all missing conditions across the three datasets. On Hateful Memes, the mean distances are approximately 0.50, 0.49, and 0.34 for Complete, Text missing, and Image missing, respectively. The corresponding values on UPMC Food-101 are approximately 0.38, 0.37, and 0.21. Although the absolute degree of routing diversity varies across datasets, Complete and Text-missing inputs consistently exhibit similar and relatively diverse routing decisions, whereas Image missing produces more concentrated expert combinations. Importantly, the distances under Image missing also remain nonzero. These consistent results further demonstrate that MARKER does not assign a fixed expert combination to each missing condition; instead, different samples under the same condition continue to activate different expert combinations.

A.1.4 Layer-wise Routing Dynamics and Specialization

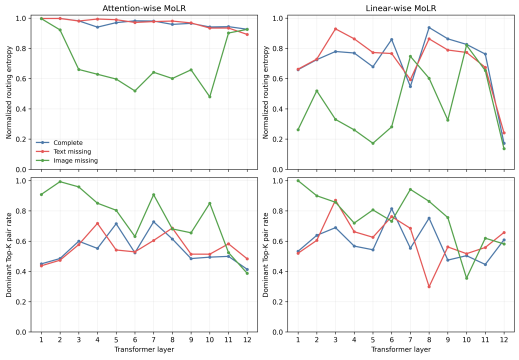
The preceding analyses characterize expert preferences and instance-specific routing at the global level. We next examine how these behaviors vary across the depth of MARKER on all three datasets. First, we measure the layer-wise changes in Top-K expert combinations when complete inputs are converted into Text-missing or Image-missing inputs, revealing where missing-modality information has the greatest effect on routing decisions. We then analyze the relationship between routing specialization and missingness sensitivity to determine whether modules with more concentrated expert preferences respond differently to modality removal. Together, these experiments characterize both the depth-wise dynamics of expert routing and the specialization of individual MoLR modules.



(a) MM-IMDb.



(b) Hateful Memes.



(c) UPMC Food-101.

Figure 14: Layer-wise routing diagnostics across the three datasets. For each dataset, the upper panels report normalized routing entropy, while the lower panels report the dominant Top-K expert-pair rate across the 12 Transformer layers. Higher entropy indicates more distributed routing probabilities, whereas a higher dominant-pair rate indicates that more samples activate the same expert combination.

Layer-wise routing changes. Figure 14 reports two complementary routing diagnostics across the 12 Transformer layers. Within each layer, the attention-wise results are averaged over the query, key,

and value MoLR modules, while the linear-wise results are averaged over the intermediate and output projections. Normalized routing entropy measures how evenly each router distributes its probabilities across experts, with lower values indicating more concentrated routing. The dominant Top-K pair rate measures the proportion of samples assigned to the most frequently selected expert combination. Because the former uses the complete probability distribution while the latter uses the discrete Top-K selection, they capture related but not identical aspects of routing specialization.

On MM-IMDb, the attention-wise routers under Complete and Text missing maintain entropy values close to one across most layers, indicating distributed routing probabilities. Image missing produces larger fluctuations, particularly at Layers 5, 6, and 10, where the entropy decreases while the dominant-pair rate increases. The distinction is stronger in the linear-wise modules: Image missing yields lower entropy throughout most of the network and causes the dominant-pair rate to approach one at Layer 1. By comparison, Complete and Text missing remain similar until the later layers, where both become more concentrated. These results indicate that removing the image reorganizes routing across multiple depths, with a stronger effect on the FFN projections.

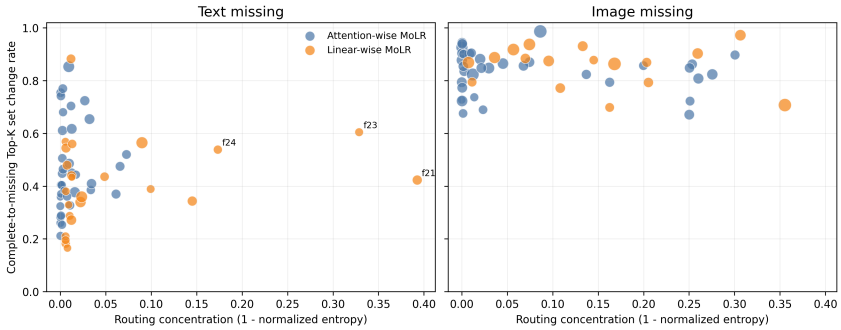
On Hateful Memes, the attention-wise routing entropy remains close to one under all three conditions. Nevertheless, Image missing substantially increases the dominant-pair rate, especially in Layers 1, 2, and 9. This shows that router probabilities can remain relatively smooth while their Top-K ordering consistently favors the same expert combination. A similar pattern appears in the linear-wise modules: Image missing produces high dominant-pair rates in the early layers, whereas its routing behavior becomes closer to the other conditions in later layers. Complete and Text missing exhibit largely overlapping trends across both module types, suggesting that text removal causes a smaller reorganization of the layer-wise routing structure on this dataset.

UPMC Food-101 exhibits the most pronounced condition-dependent specialization. Under Image missing, attention-wise entropy decreases considerably from Layers 3 to 10, accompanied by high dominant-pair rates in several layers. The routing distribution becomes more balanced again in Layers 11 and 12, indicating that this concentration is localized mainly to the middle of the attention stack. The linear-wise modules display stronger and less uniform specialization: Image missing produces very low entropy in several early and middle layers, while the dominant-pair rate reaches or approaches one at Layers 1 and 7. Complete and Text missing also show greater layer-wise variation in the linear-wise modules than on the other datasets, particularly at the final layer, suggesting that Food-101 induces stronger FFN specialization even when both modalities are available.

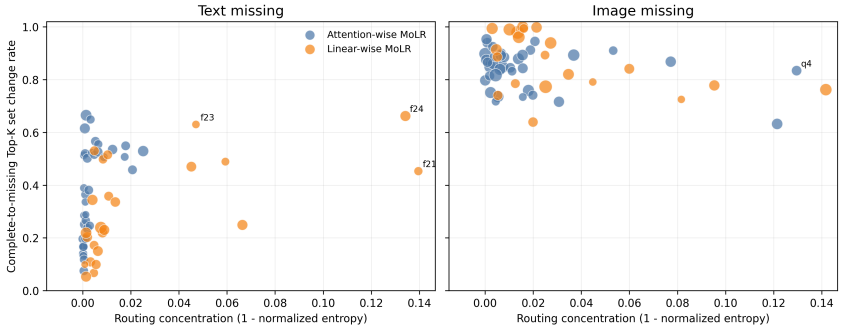
Across the three datasets, Text missing generally preserves routing patterns closer to those of Complete inputs, whereas Image missing causes more concentrated and condition-specific routing. However, the affected layers and the magnitude of these changes vary across datasets and between attention-wise and linear-wise modules. The routing response is therefore distributed across the Transformer depth rather than confined to a single layer, while its precise layer-wise organization depends on both the missing modality and the downstream task.

Routing specialization and missingness sensitivity. Figure 15 examines whether a module's routing specialization is associated with its sensitivity to modality removal. Each point represents one MoLR module. The horizontal axis reports routing concentration under the corresponding missing condition, measured as one minus normalized routing entropy, while the vertical axis reports the proportion of samples whose Top-K expert combination changes from the Complete condition. Marker size encodes the dominant Top-K pair rate, with larger markers indicating that more samples select the same expert combination. Attention-wise and linear-wise MoLR modules are shown separately by color.

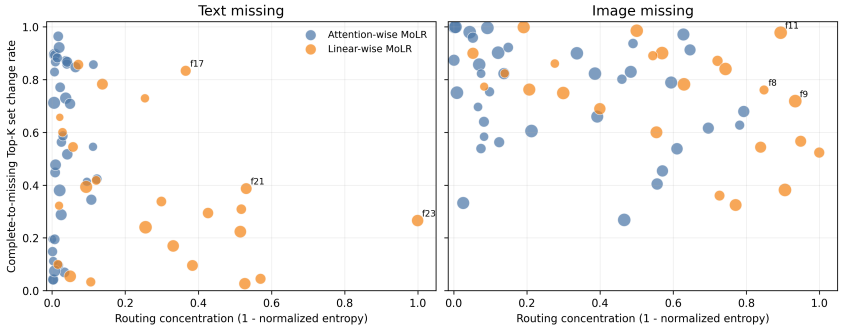
On MM-IMDb, most attention-wise modules remain close to zero routing concentration under Text missing, although their expert-combination change rates vary considerably. The linear-wise modules cover a wider concentration range, with several later FFN modules, such as $f21$, $f23$, and $f24$, exhibiting stronger specialization but only moderate routing changes. Under Image missing, nearly all modules move to the upper part of the plot, indicating that their Top-K expert combinations differ substantially from those under Complete inputs. The linear-wise modules are generally more concentrated and have larger dominant-pair rates than the attention-wise modules. Thus, image removal causes widespread rerouting across both module types, rather than affecting only modules that are already highly specialized.



(a) MM-IMDb.



(b) Hateful Memes.



(c) UPMC Food-101.

Figure 15: Relationship between routing specialization and sensitivity to missing modalities across the three datasets. Each point denotes one MoLR module. Routing concentration is measured as one minus normalized routing entropy under the corresponding missing condition, while the vertical axis gives the Complete-to-missing Top-K expert-combination change rate. Marker size represents the dominant Top-K pair rate.

Hateful Memes exhibits a sharper contrast between the two missing conditions. Under Text missing, most modules have low routing concentration and relatively small or moderate expert-combination changes. A few linear-wise modules become more concentrated, but their sensitivity does not increase proportionally. Under Image missing, the change rates rise to approximately 0.6-1.0 across almost all modules. Many attention-wise modules remain close to zero concentration despite their high change rates, showing that a router can retain a relatively distributed probability assignment while changing the identities of its selected Top-K experts. The linear-wise modules contain more concentrated routing decisions, but the scattered distribution again indicates no monotonic relationship between specialization and missingness sensitivity.

UPMC Food-101 displays the broadest range of routing concentration. Its linear-wise modules frequently show stronger specialization than the attention-wise modules under both missing conditions. Under Text missing, highly concentrated modules can exhibit either large or small expert-combination changes. For example, some later FFN modules lie far to the right but remain in the lower half of the plot, whereas other moderately concentrated modules show much higher change rates. Under Image missing, most modules exhibit both higher routing concentration and larger marker sizes, indicating more concentrated probability distributions and greater agreement on a dominant expert combination. Their change rates also remain generally high, although considerable variation persists across modules.

Across the three datasets, Image missing consistently induces larger Complete-to-missing expert-combination changes than Text missing and generally produces stronger routing concentration, particularly in the linear-wise modules. However, routing concentration and missingness sensitivity do not follow a universal monotonic relationship: highly specialized modules are not necessarily the most sensitive, and modules with diffuse routing probabilities can still substantially change their Top-K selections. These results show that expert specialization and responsiveness to missing modalities describe complementary aspects of MARKER's routing behavior.

A.2 Theoretical Analysis

This section provides a unified theoretical perspective for MARKER and connects directly to the problem setup and the MoLR/LoRA designs. We organize the analysis into three parts: (1) how modality missingness perturbs multimodal models under the mixed incomplete dataset $\tilde{D}$; (2) why lightweight adaptation methods such as prompt tuning and LoRA can mitigate missingness-induced degradation via compact corrections; and (3) why MARKER offers stronger robustness than prior *static* prompt-based methods by enabling instance- and scenario-adaptive parameter updates.

A.2.1 Impact of Modality Missingness on Multimodal Models

Following the settings of manuscript, we consider $M = 2$ modalities: m_1 (text) and m_2 (image), and the mixed dataset $\tilde{D} = \{D^c, \tilde{D}^{m_1}, \tilde{D}^{m_2}\}$ that combines complete inputs and dummy-filled incomplete inputs. Note that in Sec. 3.1, $\tilde{D}^{m_1} = \{(x^{m_1}, \tilde{x}^{m_2}, y)\}$ corresponds to missing m_2 (image is dummy-filled), while $\tilde{D}^{m_2} = \{(\tilde{x}^{m_1}, x^{m_2}, y)\}$ corresponds to missing m_1 (text is dummy-filled). Dummy filling keeps the input interface unchanged but introduces a *structured perturbation* rather than random noise: an entire modality stream is removed and replaced by a near-constant placeholder.

This perturbation affects multimodal prediction in two coupled ways. **(i) Information removal.** Missing a modality discards complementary evidence, increasing intrinsic ambiguity for many samples. **(ii) Fusion-context mismatch.** Multimodal backbones pretrained on complete inputs typically exploit cross-modal interactions and correlations. When one modality is replaced by dummy values, these interactions are systematically altered, which can shift intermediate representations and decision statistics. Consequently, a pretrained model may behave unevenly across complete and incomplete subsets in $\tilde{D}$, leading to large performance gaps among scenarios and sensitivity to test-time missingness composition.

A.2.2 Why Prompt/LoRA Can Mitigate Missingness-Induced Degradation

A natural remedy is to adapt the pretrained model on $\tilde{D}$ so that it performs well not only on complete inputs but also on dummy-filled incomplete inputs. While full fine-tuning can achieve this, it is expensive and can overfit when incomplete data are limited. Lightweight adaptation methods address

missingness by learning *compact corrections* on top of a frozen backbone, aiming to reduce scenario disparity with limited trainable parameters.

Prompt tuning. Prompt-based methods introduce a small number of trainable prompt parameters (e.g., prompt tokens) to modulate internal activations. From the viewpoint above, prompts provide an activation-level compact correction that can partially compensate fusion-context mismatch under dummy-filled inputs. However, in most *static* designs, prompts are shared across all instances within the same missingness type, which limits their ability to address intra-scenario heterogeneity.

LoRA as parameter-level correction. LoRA performs adaptation directly on model projections via low-rank updates while freezing pretrained weights (see Sec. 3.2). Compared with activation-level prompting, LoRA provides a parameter-level compact correction that is particularly suitable when missingness induces *systematic* shifts (e.g., reliance shift and fusion mismatch), since structured additive updates can restore decision statistics without rewriting the entire backbone. Overall, both prompt tuning and LoRA aim to reduce scenario disparity by learning compact corrections, while differing in where the correction is applied (activation-level vs. parameter-level).

A.2.3 MARKER vs. Static Prompt Methods: Conditional Updates via MoLR

Despite being effective, a single shared prompt (or a single shared LoRA) still applies an essentially *static* correction. Incomplete multimodal learning is inherently conditional: the appropriate correction differs across scenarios (complete vs. different missingness types), and even within the same scenario, instances may require different adaptations (e.g., text-dominant vs. image-dominant samples). This conditional heterogeneity motivates making the update itself dependent on the current input.

MARKER addresses this by modeling the fine-tuning update as a *mixture of LoRA experts* routed by an input-conditioned gate (MoLR). Following Sec. 3.2, we index experts by $i \in \{1, \dots, N\}$ and denote router weights by $R(x)_i$ (computed via a Softmax router with parameters W_r). For clarity within this section, we use ℓ to index a transformer layer (or a projection) being adapted. Instead of applying a single low-rank update to a layer weight, MARKER uses multiple LoRA experts and forms an input-dependent mixture update. At layer ℓ, the effective update can be written as

$$\Delta W_\ell(x) = \sum_{i=1}^{N} R(x)_i \, \Delta W_{\ell,i}, \tag{9}$$

where each $\Delta W_{\ell,i}$ follows the LoRA-style low-rank parameterization introduced in Sec. 3.2. This makes the correction *instance-adaptive* through $R(x)$ and *scenario-aware* since the missingness pattern is reflected by dummy-filled inputs and can also be explicitly provided to the router. With sparse routing (e.g., Top-K in Sec. 3.2), expert specialization can be further encouraged, reducing interference across conditions.

An intuitive condition-wise decomposition for MoLR. Let u denote a condition variable capturing both the current missingness scenario and instance characteristics. For each expert i, define its conditional expected loss under condition u as

$$L_i(u) = \mathbb{E}[\ell(f_{\theta,\phi_i}(x), y) \mid u], \tag{10}$$

where ϕ_i denotes the parameters of the i-th LoRA expert and $\ell(\cdot,\cdot)$ is the task loss. Let $L^\star(u) = \min_i L_i(u)$ be the condition-wise best achievable loss among experts. Since MoLR forms a convex combination of experts via the Softmax router (so $R(u)_i \geq 0$ and $\sum_{i=1}^{N} R(u)_i = 1$), its conditional loss satisfies

$$L_{\mathrm{MoLR}}(u) = \sum_{i=1}^{N} R(u)_i \, L_i(u) = L^\star(u) + \underbrace{\sum_{i=1}^{N} R(u)_i \big(L_i(u) - L^\star(u)\big)}_{\text{routing regret } \mathcal{G}(u) \geq 0}. \tag{11}$$

This decomposition highlights that MoLR approaches the condition-wise best expert up to a nonnegative routing regret $\mathcal{G}(u)$, which becomes small when the router assigns higher weights to experts with lower $L_i(u)$ under that condition. Moreover, MoLR generalizes single-LoRA adaptation as a special case (e.g., when the router consistently concentrates on one expert). Thus, under conditional heterogeneity, specialization and routing provide a natural mechanism to better match different

missingness conditions and instance characteristics than a static correction, which we further study empirically in our ablations.

Why this improves robustness over static prompts. Static prompt tuning typically shares the same prompt parameters across all instances under a given modality-missing scenario, yielding a pattern-wise constant correction that cannot fully resolve intra-scenario heterogeneity. In contrast, MARKER performs *parameter-level* conditional adaptation: Eq. equation 9 allows different instances (and different missingness scenarios) to activate different low-rank experts, providing a richer family of compact corrections while remaining parameter-efficient. As a result, MARKER is better aligned with the mixed training distribution $\tilde{D}$ and tends to yield stronger robustness under varying missingness compositions at inference.