Preset No More: Pattern-Aware Graph Routing and Meta-Adaptive Prompt Distillation for Incomplete Multimodal Learning

Lisi Mo, Zihan Yang, Ruofan Zhou, Letian Xu, Ruiting Dai†
University of Electronic Science and Technology of China
Email: morris@uestc.edu.cn, 202421011240@std.uestc.edu.cn,
annie_zhou520@163.com, 202521011035@std.uestc.edu.cn, rtdai@uestc.edu.cn

***Abstract*—The efficacy of multimodal learning emanates from the delicate orchestration of cross-modal interactions. However, pervasive modality absence in real-world scenarios shatters established alignment pathways, triggering a structural collapse of the joint representation space. Existing research remains primarily constrained by a preset interaction logic: treating modality absence as a passive feature reduction or a static parameter index, while lacking the capacity to actively reshape reasoning topologies to accommodate specific informational gaps. To this end, we propose Pattern-Aware Graph routing and Meta-Adaptive Prompt Distillation (PAG-MPD), a framework that shifts the missing-modality learning paradigm from static aggregation to a pattern-aware adaptive process. Specifically, we introduce a pattern-aware graph encoder to capture the shifting structural dependencies among available signals, complemented by a differentiable sparse routing mechanism that generates meta-adaptive prompts to dynamically re-calibrate the information flow. Furthermore, we develop an uncertainty-aware distillation strategy to mitigate epistemic uncertainty via confidence-weighted supervision from a complete-modality teacher. Extensive evaluations on MM-IMDb and CMU-MOSEI benchmarks demonstrate that PAG-MPD significantly outperforms state-of-the-art methods, achieving a notable improvement of up to 4.5% on MM-IMDb.**

***Index Terms*—Multimodal Learning, Incomplete Multimodal Learning, Transformer, Teacher-student Learning.**

I. Introduction

Multimodal learning achieves superior reasoning performance by orchestrating complex interactions among heterogeneous signals [1]. The cornerstone of this paradigm lies in establishing stable and accurate cross-modal alignments, enabling different modalities to mutually calibrate and augment each other. However, the pervasive issue of modality absence in real-world deployments shatters these established interaction pathways. Such incompleteness not only leads to significant information loss but also erodes the structural consistency of the joint representation space, posing a formidable challenge to system robustness [2].

Existing researches on missing-modality learning [1], [3], [4] (Fig. 1) primarily follows two paradigms. Joint-learning methods [1], [5] typically employ shared fusion layers or fixed prompting to aggregate features, yet often treat modality absence as a simple subset reduction, lacking the flexibility to dynamically adjust interaction priorities. Generation-based methods [4], [6] attempt to hallucinate missing data to restore a complete input format, but their reasoning remains strictly tethered to the reconstruction quality, which frequently introduces semantic bias and accumulated errors. While recent works [3], [7] explore pattern-specific adaptation, they remain constrained by a preset interaction logic: utilizing missing patterns merely as indices to retrieve fixed parameters, rather than actively re-configuring the underlying reasoning topology to accommodate the current informational vacuum. This rigid coupling between model architecture and input configurations triggers an inherent conflict between static computational graphs and the dynamic structural collapse of multimodal information.

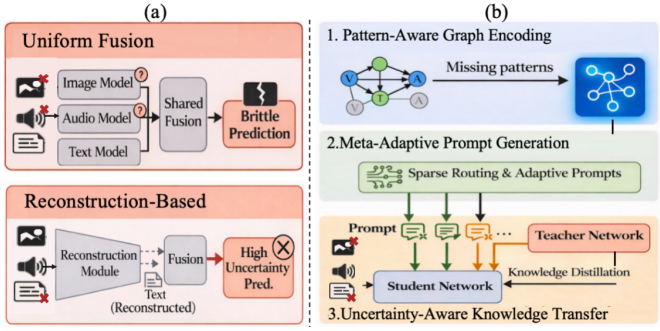

Fig. 1. Comparison of paradigms. (a) Traditional: Static and brittle under modality asymmetry or reconstruction uncertainty. (b) PAG-MPD (Ours): Enables adaptive and robust reasoning via (1) pattern-aware graph encoding, (2) meta-prompt generation, and (3) uncertainty-aware distillation.

Mechanistically, missing modalities induce a non-linear reconfiguration of latent dependencies rather than simple feature loss [šaitis2017multimodalmachinelearningsurvey], [**?**]. The absence of key modalities creates semantic gaps, requiring remaining signals to form compensatory interactions [8]. However, existing methods (Fig. 1) still employ fixed interaction kernels trained on complete data [9], forcing models to rely on non-existent cross-modal cues, which disrupts reasoning and accumulates uncertainty [10].

†Corresponding Author.

To address this issue, we propose **PAG-MPD** (Fig. 2), a pattern-aware adaptive framework. It enables the model to reconstruct its interaction topology based on the global missing pattern. Specifically, we design a pattern-aware graph encoder to capture dynamic dependencies among available modalities [11], and a differentiable sparse routing mechanism to activate specialized pathways for generating meta-adaptive prompts [12], which guide information flow. Furthermore, we introduce an uncertainty-aware distillation strategy to provide confidence-weighted supervision from a complete-modality teacher, ensuring semantic consistency under partial observations [13].

Our main contributions are summarized as follows:

- We challenge the conventional preset interaction logic by shifting the missing-modality learning paradigm from static aggregation to active topological adaptation.
- We propose a graph-based encoding module coupled with sparse routing to explicitly model structural dependencies and activate instance-specific reasoning pathways.
- We develop a robust knowledge transfer mechanism using meta-prompts and confidence-weighted supervision to mitigate epistemic uncertainty and ensure cross-modal semantic consistency.
- Extensive evaluations on MM-IMDb and CMU-MOSEI demonstrate that PAG-MPD significantly outperforms existing methods, achieving a notable 4.5% F1-score improvement on MM-IMDb. Our source code are available at: https://anonymous.4open.science/r/PARP-9AFD.

II. Method

Problem Definition. We study multimodal classification with incomplete inputs: each sample may observe only a subset of the M modalities, indicated by a binary mask $r_i \in \{0,1\}^M$, where $r_i^m = 1$ if modality m is present and 0 otherwise; the observed set is $\mathcal{O}_i = \{m \mid r_i^m = 1\}$. Given dataset $\mathcal{D} = \{(x_i^{\mathcal{O}_i}, r_i, y_i)\}_{i=1}^N$, the goal is to learn a predictor f_θ robust to diverse missing patterns by $\min_\theta \frac{1}{N}\sum_{i=1}^N \ell(f_\theta(x_i^{\mathcal{O}_i}, r_i), y_i)$, where ℓ is cross-entropy loss.

Overall Framework. PAG-MPD comprises graph encoding, prompt generation, and knowledge transfer. As shown in Fig. 2, graph encoder captures modality dependencies under missing patterns, which guide meta-adaptive prompt generation via sparse routing. Then, confidence-weighted distillation from a full-modality teacher enforces semantic consistency and alleviates uncertainty.

A. Feature Extraction via Pattern-Aware Graph

To handle the heterogeneity of observed modalities and the diversity of missing patterns in incomplete multimodal samples, we propose pattern-aware graph routing, projecting each modality into a unified space and explicitly encoding missing patterns to enable adaptive selective fusion.

Statistical Embedding. First, for each sample i, we encode each modality x_i^m by its modality-specific encoder f_m to $z_i^m = f_m(x_i^m)$ if observed, or replaced with a placeholder otherwise. To aggregate observed semantics and sample-level statistics, we compute

$$u_i = \mathrm{MLP}\left(\left[\frac{1}{\max(|\mathcal{O}_i|,1)}\sum_{m\in\mathcal{O}_i} z_i^m;\ \mathrm{mean}(r_i);\ \mathrm{var}(r_i)\right]\right), \tag{1}$$

where $|\mathcal{O}_i|$ is the number of observed modalities; $\mathrm{mean}(r_i)$ and $\mathrm{var}(r_i)$ summarize the proportion and dispersion of missingness patterns.

Cross-modal Dependencies. Inter-modality relations are modeled via a graph $G = (\mathcal{M}, \mathcal{E})$, with nodes as modalities and edges as dependencies. The adjacency matrix combines prior knowledge and learnable weights: $A = A_{\text{prior}} \odot A_{\text{learned}}$, where $A_{\text{prior}} \in \{0,1\}^{M\times M}$ encodes known semantic couplings (The prior adjacency matrix A_{prior} encodes coarse-grained semantic correlations among modalities and is constructed based on domain knowledge or data statistics. Specifically, $A_{\text{prior}}^{mm'} = 1$ indicates that modality m and m' are considered semantically related, while 0 denotes no assumed dependency), $A_{\text{learned}} \in \mathbb{R}^{M\times M}$ contains learnable weights, and $\odot$ denotes element-wise multiplication. $A_{mm'}$ represents the influence of modality m' on m. Incoming edge weights are normalized via the diagonal degree matrix: $D_{mm} = \sum_{m'} A_{mm'}$.

Modality Embeddings. Each modality is updated via gated graph propagation:

$$\tilde{z}_i^m = \sum_{m'\in\mathcal{M}} (D^{-1/2}AD^{-1/2})_{mm'}\, \gamma_{mm'}(r_i)\, z_i^{m'}, \tag{2}$$

where $\gamma_{mm'}(r_i) = \sigma(a_{mm'}^\top r_i + b_{mm'}) \in (0,1)$ scales the influence of m' on m based on the missing pattern r_i, preventing noisy or absent modalities from dominating the update; σ denotes the sigmoid function. The final routing context is

$$c_i = \left[\mathrm{concat}(\tilde{z}_i^1, \ldots, \tilde{z}_i^M);\ u_i\right] \in \mathbb{R}^{d_c}, \tag{3}$$

where $d_c = M \cdot d + d_u$, and u_i the statistical embedding.

B. Sparse Routing for Prompts Generation

We design K candidate paths, each producing low-rank adapters and T_p meta-adaptive prompts to modulate modality features. Each path encodes a specialized transformation strategy for different missing patterns, and the routing network selects Top-S paths per sample for flexible, pattern-aware fusion, enhancing multimodal robustness.

Differentiable Sparse Routing. Given K candidate paths $\{g_k\}_{k=1}^K$, the routing network maps the context c_i to path logits:

$$\pi = \mathrm{MLP}_{\text{route}}(c_i) \in \mathbb{R}^K, \tag{4}$$

and selects Top-S paths via Gumbel-Top-S sampling [14]:

$$\tilde{\alpha}_k = \frac{\exp((\pi_k + G_k)/\tau)}{\sum_{j=1}^K \exp((\pi_j + G_j)/\tau)}, \quad G_k \sim \mathrm{Gumbel}(0,1) \tag{5}$$

where π_k is the logit score of path k, τ controls sampling smoothness. During training, soft weights $\tilde{\alpha}_k$ enable gradient flow; at inference, only the Top-S paths are activated ($\alpha_k = 1$

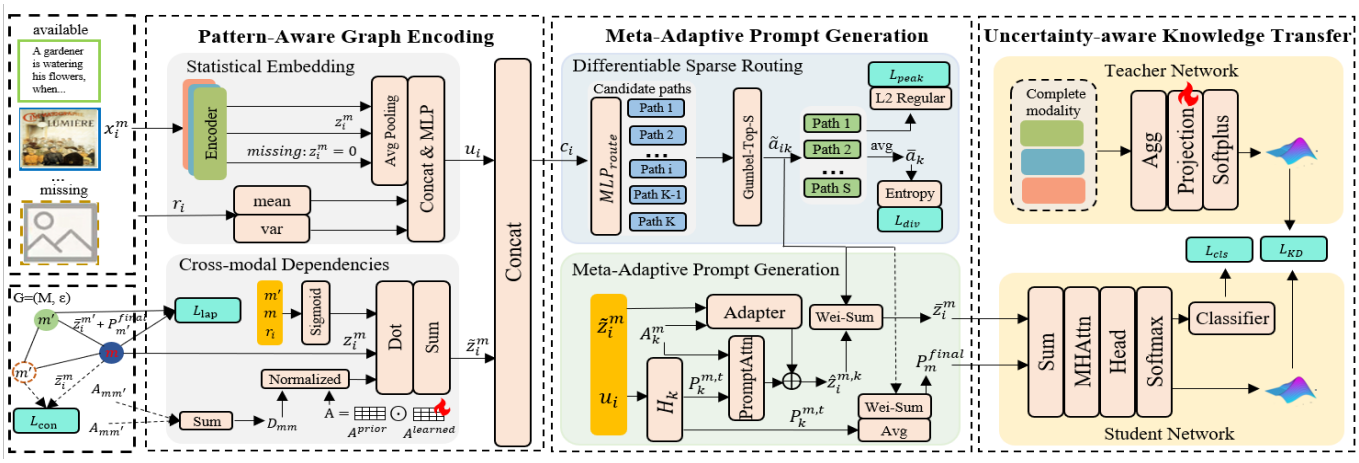

Fig. 2. Overview of **PAG-MPD**. (a) *Pattern-aware graph encoding* maps observed modalities into a unified space with missingness patterns. (b) *Meta-Adaptive Prompt Generation* dynamically generates adapters and prompts conditioned on observed modalities and missing patterns via sparse routing. (c) *Uncertainty-aware knowledge transfer* distills from complete-modality teachers and enforces cross-modal consistency.

for selected paths, 0 otherwise). To further stabilize routing, we impose sparsity and diversity regularization losses:

$$\mathcal{L}_{\text{rount}} = -\sum_{k} \tilde{\alpha}_{ik}^2 - \sum_{k} \bar{\alpha}_k \log \bar{\alpha}_k, \tag{6}$$

where $\bar{\alpha}_k = \frac{1}{B}\sum_{i=1}^{B} \tilde{\alpha}_{ik}$ is the average routing probability of path k over the current batch of size B. The fist part encourages sparse selection for each sample, while the second part encourages diversity across the batch.

Meta-Adaptive Prompt Generation. Each activated path g_k contains a Prompt Network H_k that, conditioned on the sample's modality embeddings and missing-pattern vector u_i, generates: 1) a low-rank adapter per modality,

$$A_k^m = U_k^m V_k^{m\top}, \quad U_k^m, V_k^m \in \mathbb{R}^{d\times d_r},\ d_r \ll d \tag{7}$$

applied as $\text{Adapter}(\tilde{z}_i^m; A_k^m) = \tilde{z}_i^m + A_k^m \tilde{z}_i^m$, and 2) T_p meta-adaptive prompt tokens per modality,

$$\{P_k^{m,t}\}_{t=1}^{T_p} = H_k(\tilde{z}_i^m, u_i), \quad P_k^{m,t} \in \mathbb{R}^d, \tag{8}$$

where T_p is the number of prompt tokens per modality. These tokens are injected into modality encoders or cross-modal Transformers, guiding attention via a prompt-based cross-attention operator $\text{PromptAttn}(\cdot)$. The updated representation for modality m under path k is:

$$\hat{z}_i^{m,k} = \text{Adapter}(\tilde{z}_i^m; A_k^m) + \text{PromptAttn}(\tilde{z}_i^m, \{P_k^{m,t}\}). \tag{9}$$

The final representation of modality m is computed as $\bar{z}_i^m = \sum_{k=1}^{K} \tilde{\alpha}_k \hat{z}_i^{m,k}$, and the corresponding meta-adaptive prompts is obtained by aggregating the prompts over activated paths:

$$p_{\text{final}}^m = \sum_{k=1}^{K} \tilde{\alpha}_k \frac{1}{T_p} \sum_{t=1}^{T_p} P_k^{m,t}. \tag{10}$$

C. Uncertainty-Aware Knowledge Transfer

A full-modality teacher provides uncertainty-weighted supervision, guiding a student reasoning network to learn stable and generalizable representations under partial-modality observations via prompt-guided cross-modal consistency.

We first compute the fused representation $z_{i,\text{fused}} = \text{MHAttn}(Q, K, V)$, with Q, K, V as linear projections of $\{\tilde{z}_i^m + p_{\text{final}}^m\}$. The student predicts $p_{\text{stu}}(y \mid x_i^{\mathcal{O}_i}) = \text{softmax}(\text{Head}(z_{i,\text{fused}}))$, and its classification loss over N training samples is $\mathcal{L}_{\text{cls}} = \frac{1}{N}\sum_{i=1}^{N} \ell(p_{\text{stu}}(y \mid x_i^{\mathcal{O}_i}), y_i)$, where $\ell(\cdot)$ is cross-entropy loss.

Uncertainty-Aware Distillation. To provide more stable supervision, a teacher model f_{tea} is trained on complete-modality samples, producing a Dirichlet distribution [15] to capture predictive uncertainty:

$$\boldsymbol{\alpha}_{\text{tea}} = \text{softplus}\Big(W_\alpha\, Agg(\{z_i^m\}_{m\in M})\Big) + \mathbf{1} \in \mathbb{R}^C_{>0}, \tag{11}$$

where z^m is the embedding of modality m, $\text{Agg}(\cdot)$ is a modality aggregator (e.g., mean or attention pooling), W_α is a learnable projection, $\text{softplus}(x) = \log(1 + e^x)$ ensures positive Dirichlet parameters;$+\mathbf{1}$ guarantees all entries are strictly > 0, as required by Dirichlet distributions [15]. The teacher's predictive probability for class c is computed as

$$p_{\text{tea}}(y = c) = \mathbb{E}[\pi_c], \quad \pi \sim \text{Dir}(\boldsymbol{\alpha}_{\text{tea}}) \tag{12}$$

where π is a random probability vector drawn from a Dirichlet distribution, and $\mathbb{E}[\pi_c]$ denotes the expected value (mean) of π_c. It encodes both class confidence and epistemic uncertainty. The student matches the teacher's distribution via the Dirichlet–Dirichlet KL divergence [16]:

$$\mathcal{L}_{\text{KD}} = \text{KL}\big(\text{Dir}(\boldsymbol{\alpha}_{\text{tea}}) \,\|\, \text{Dir}(\boldsymbol{\alpha}_{\text{stu}})\big), \tag{13}$$

where $\boldsymbol{\alpha}_{\text{stu}}$ is inferred from the student's visible modalities.

Each sample is weighted by the inverse of the teacher's predictive uncertainty to emphasize reliable supervision, yielding the uncertainty-aware distillation loss:

$$w_i = \frac{1}{1 + \text{Tr}(\text{Cov}[\pi_{\text{tea},i}])}, \quad \tilde{\mathcal{L}}_{\text{KD}} = \frac{1}{N}\sum_{i=1}^{N} w_i\, \mathcal{L}_{\text{KD}}(i), \tag{14}$$

TABLE I

MAIN RESULTS UNDER 60% MISSING RATE. BOLD AND UNDERLINE DENOTE THE BEST AND SECOND-BEST RESULTS, RESPECTIVELY. ± DENOTES THE STANDARD DEVIATION OVER 5 INDEPENDENT RUNS. Δ REPRESENTS THE ABSOLUTE IMPROVEMENT OF PAG-MPD OVER THE SECOND-BEST BASELINE.

Datasets	$\mathcal{M}$issing	MCTN	MMIN	MPVR	EPE-P	P-EA	Ours (PAG-MPD)	Δ
MM-IMDb	L	43.79±0.42	44.88±0.35	46.45±0.27	44.59±0.31	<u>46.98</u>±0.28	**49.99**±0.24	+3.01
	V	44.53±0.38	46.64±0.32	49.89±0.25	<u>51.32</u>±0.25	50.45±0.31	**53.46**±0.22	+2.14
	L,V	45.07±0.29	46.19±0.27	47.72±0.22	48.91±0.20	<u>49.95</u>±0.19	**51.25**±0.26	+1.30
	Mismatch	44.16±0.51	45.76±0.44	47.17±0.39	47.66±0.42	<u>48.13</u>±0.38	**50.30**±0.45	+2.17
CMU-MOSEI	L	54.05±0.39	54.16±0.35	54.81±0.31	55.27±0.32	<u>55.56</u>±0.29	**58.20**±0.25	+2.64
	V	60.71±0.34	62.70±0.28	63.52±0.26	<u>63.92</u>±0.22	63.80±0.27	**67.86**±0.19	+3.94
	A	60.93±0.31	61.87±0.27	64.85±0.24	<u>65.21</u>±0.28	64.95±0.24	**68.21**±0.21	+3.00
	L,V	53.18±0.41	54.59±0.38	56.08±0.32	<u>56.58</u>±0.35	56.31±0.31	**59.48**±0.28	+2.90
	L,A	55.74±0.38	54.49±0.35	54.56±0.33	<u>55.17</u>±0.26	54.91±0.30	**58.66**±0.23	+3.49
	V,A	62.21±0.33	62.22±0.29	63.98±0.24	<u>64.86</u>±0.24	64.43±0.27	**67.79**±0.22	+2.93
	L,V,A	62.51±0.29	64.43±0.25	64.83±0.21	64.86±0.19	<u>64.87</u>±0.21	**67.73**±0.18	+2.86
	Mismatch	53.13±0.53	60.93±0.49	63.71±0.46	63.72±0.48	<u>63.81</u>±0.45	**66.70**±0.52	+2.89

where $\mathrm{Cov}[\pi_{\mathrm{tea},i}]$ is the covariance matrix of $\pi_{\mathrm{tea},i}$, capturing total predictive uncertainty (epistemic + aleatoric). The trace $\mathrm{Tr}(\cdot)$ sums the variance across all classes, so larger w_i places greater emphasis on confident, reliable distillation signals.

Prompt-Guided Cross-Modal Consistency. To encourage inter-modal coherence, we apply a graph Laplacian regularizer [17] and cross-modal contrastive learning [18]:

$$\mathcal{L}_{\mathrm{sync}} = \sum_{(m,m')} A_{mm'} \left\| (\bar{z}_i^m + p_{\mathrm{final}}^m) - (\bar{z}_i^{m'} + p_{\mathrm{final}}^{m'}) \right\|_2^2 - \sum_{(m,m')} \log \frac{\exp(\langle \phi(\bar{z}_i^m), \phi(\bar{z}_i^{m'}) \rangle / \tau_c)}{\sum_j \exp(\langle \phi(\bar{z}_i^m), \phi(\bar{z}_j^{m'}) \rangle / \tau_c)}, \tag{15}$$

where $\phi(\cdot)$ is a non-linear projection head, $\tau_c > 0$ is a temperature, and $A_{mm'}$ is the modality graph adjacency. $\mathcal{L}_{\mathrm{lap}}$ encourages semantically similar modalities to align, while $\mathcal{L}_{\mathrm{con}}$ enforces instance-level cross-modal alignment.

Combined Objective. The final multi-task objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{KD}} \tilde{\mathcal{L}}_{\mathrm{KD}} + \lambda_{\mathrm{sync}} \mathcal{L}_{\mathrm{sync}} + \lambda_{\mathrm{rount}} \mathcal{L}_{\mathrm{route}}, \tag{16}$$

where hyper-parameters λ are coefficients used to balance the contribution of each term.

III. EXPERIMENTS

Datasets and Baselines. We evaluate PAG-MPD on MM-IMDb [19] and CMU-MOSI [20] with F1-Macro and three-class accuracy as evaluation metrics. We compare it with state-of-the-art models, including MCTN [21], MMIN [13], MPVR [22], EPE-P [23], P-EA [24].

Implementation Details. Following [22], [25], [26], we employ ViLT [27], BERT [28], and COVAREP [27] as the modality-specific encoders for image, text, and audio, respectively. Extracted features are projected into a unified dimension of $d = 512$, while audio features are mapped to a 128-dimensional space. Missing modalities are represented by zero-padded placeholders or learned embedding tokens.

TABLE II

ABLATION OF CORE MODULES UNDER 60% MISSING RATE ($\mathcal{M}$: ALL MODALITIES). Δ ↓ INDICATES THE PERFORMANCE DROP.

Variant	MM-IMDb	Δ ↓	CMU-MOSEI	Δ ↓
Full PAG-MPD	**51.25**±0.26	-	**67.73**±0.18	-
w/o PAGR	48.48±0.35	2.77	64.82±0.31	2.91
w/o DSR	48.89±0.28	2.36	65.44±0.27	2.29
w/o UA-KD	50.21±0.24	1.04	66.14±0.22	1.59
w/o LRA	50.47±0.19	0.78	66.41±0.20	1.32

To maintain parameter efficiency, the backbone encoders are frozen, and only the task-specific heads and prompt networks ($T_p = 16, L = 5$) are optimized. The dynamic sparse routing component is configured with $K = 4$ candidate paths and Top-$S = 2$ selection. Low-rank adapters utilize a bottleneck dimension of $d_r = 32$. For the joint optimization objective, the temperature for Gumbel sampling is set to $\tau = 1.0$, and the contrastive alignment temperature is $\tau_c = 0.07$. The balancing coefficients are set to $\lambda_{\mathrm{KD}} = 0.01$, $\lambda_{\mathrm{sync}} = 0.1$, and $\lambda_{\mathrm{route}} = 0.05$. All experiments are implemented in PyTorch and optimized using the Adam optimizer with a learning rate of 1×10^{-3}, weight decay of 2×10^{-2}, and a batch size of 32. Training is conducted for 50 epochs on a single NVIDIA A100 GPU.

Following [29], [30], we evaluate PAG-MPD under two protocols: **Fix**ed Incompleteness and **Mismatch**ing Observation. We quantify data absence via the Missingness Intensity (η): $\eta = 1 - \frac{\sum_{i=1}^{N} M_i}{N \times M}$, where $M_i = \sum_{m=1}^{M} r_i^m$ denotes the number of observed modalities for sample i. We ensure $M_i \geq 1$ for all samples to prevent empty observations. In this paper, we set a default $\eta = 60\%$ to evaluate robustness under severe absence. For Fixed Incompleteness, the pattern r_i is identical for all samples; for Mismatching Observation, patterns are randomized per instance to examine model adaptability.

Main Results. Table I presents a comprehensive performance comparison. PAG-MPD consistently outperforms all state-of-

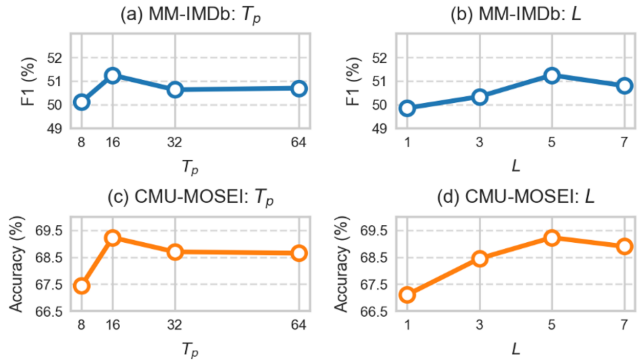

Fig. 3. Ablation study on prompt T_p and L across different datasets under a 60% missing rate ($\mathcal{M}$: all modalities, train-test consistent).

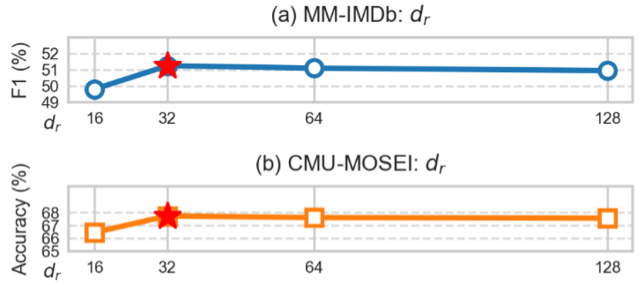

Fig. 4. Ablation study on the low-rank adapter dimension d_r under a 60% missing rate ($\mathcal{M}$: all modalities, train–test consistent).

the-art baselines across all missing-modality scenarios. The key observations are as follows:

(1) Significant Performance Gains. Our framework achieves a substantial improvement over the second-best baseline (P-EA or EPE-P), with a maximum Δ of +3.01 on MM-IMDb and +3.94 on CMU-MOSEI. This confirms that modeling conditional inter-modal dependencies via pattern-aware graphs is superior to static prompt mechanisms.

(2) Superiority in Multi-modality Complexity. The performance edge is more pronounced in the three-modality CMU-MOSEI dataset compared to the two-modality MM-IMDb. For instance, when the visual (V) or audio (A) modality is absent in CMU-MOSEI, PAG-MPD maintains high accuracy with gains of 3.94% and 3.00%, respectively, demonstrating its ability to capture rich multimodal dependencies through meta-adaptive prompt generation.

(3) Robustness under Mismatching Protocols. Even when training and testing missing patterns are randomized (Mismatch), PAG-MPD retains strong leads of +2.17 and +2.89 points. Despite the higher standard deviations in these challenging scenarios, our framework achieves the lowest variance among top performers, proving that uncertainty-aware distillation effectively stabilizes knowledge transfer under stochastic observations.

Ablation Study on Core Modules. We evaluate PAG-MPD with four variants: without pattern-aware graph routing (w/o PAGR), dynamic sparse routing (w/o DSR), uncertainty-aware knowledge distillation (w/o UA-KD), and low-rank adapters

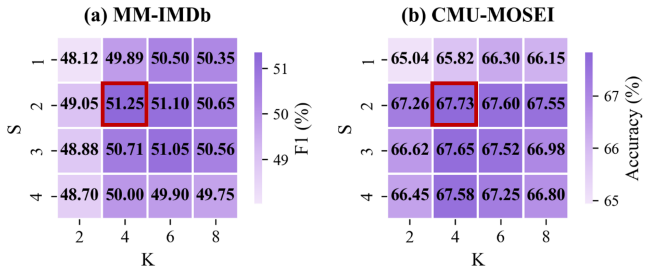

Fig. 5. Ablation on dynamic sparse routing K (a), S (b) under 60% missing rate ($\mathcal{M}$: all modalities, train-test consistent).

TABLE III
SENSITIVITY ANALYSIS UNDER 60% MISSING RATE. THE SELECTED DEFAULT SETTINGS ARE HIGHLIGHTED.

Hyper-parameter	Value	MM-IMDb	CMU-MOSEI
λ_{sync}	0.001	48.94±0.45	64.28±0.38
	0.050	50.56±0.32	66.85±0.27
	0.100	**51.25**±0.26	**67.73**±0.18
	0.200	50.82±0.29	67.24±0.24
	0.500	49.75±0.38	65.91±0.35
λ_{KD}	0.001	49.24±0.41	65.05±0.32
	0.005	50.68±0.28	67.02±0.24
	0.010	**51.25**±0.26	**67.73**±0.18
	0.050	50.81±0.32	67.08±0.26
	0.100	50.15±0.37	66.42±0.31
Uncertainty w_i	w/o w_i	49.64±0.48	65.71±0.39
	with w_i	**51.25**±0.26	**67.73**±0.18

(w/o LRA). TableII shows that PAGR and DSR are most critical, especially on CMU-MOSI, underscoring their roles in modeling inter-modal dependencies and adaptive fusion. UA-KD and LRA yield moderate gains by improving stability and efficiency. Overall, all modules jointly contribute to PAG-MPD's robustness.

Ablation on Prompt Configuration. We vary prompt length ($T_p \in \{8, 16, 32, 64\}$) and injection layers ($L \in \{1, 3, 5, 7\}$). Results (Fig. 4 (a)) show $T_p = 16$ and $L = 5$ achieve the best trade-off between expressiveness, interaction, and stability. Shorter prompts or fewer layers underfit, while longer prompts or deeper injections bring marginal gains with slight instability, underscoring the importance of prompt design for effective adaptation.

Ablation on Low-Rank Adapter (LRA) Dimension. As shown in Fig. 4, We test $d_r \in \{16, 32, 64, 128\}$ and find $d_r = 32$ offers the best trade-off. Smaller sizes underfit, while larger ones yield marginal gains with higher cost, confirming the need for properly sized LRAs for efficient adaptation.

Ablation on Dynamic Sparse Routing. We vary candidate paths $K \in \{2, 4, 6\}$ and selections $S \in \{1, 2, 3\}$ (Fig. 5). The best trade-off is at $K = 4, S = 2$: too few paths hinder adaptivity, while larger settings bring only slight gains with higher cost. This confirms the importance of balanced sparse routing for robust multimodal reasoning.

Hyper-parameter Sensitivity. Table III evaluates the stability of PAG-MPD regarding λ_{sync}, λ_{KD}, and the uncertainty weight w_i. The robust bell-shaped trend across five magnitudes demonstrates performance resilience, with optima at $\lambda_{\text{sync}} = 0.1$ and $\lambda_{\text{KD}} = 0.01$ validating balanced structural-semantic alignment and distillation. The significant degradation without w_i confirms the necessity of prioritizing reliable teacher signals for robust knowledge transfer under partial observations.

IV. Conclusion

In this paper, we present PAG-MPD, a framework designed to overcome the rigidity of preset interaction logic in missing-modality learning. By synthesizing pattern-aware graph routing and meta-prompting, PAG-MPD achieves a dynamic reconfiguration of reasoning topologies, while an uncertainty-aware distillation strategy ensures cross-modal semantic consistency. Extensive evaluations demonstrate its superior robustness under extreme missingness intensity and mismatched protocols. Future research will explore the integration of causal reasoning for enhanced interpretability, the scalability to multimodal foundation models, and the extension to complex downstream tasks such as embodied planning and empathetic dialogue.

References

[1] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *Proc. ICML*, ser. ICML '11, L. Getoor and T. Scheffer, Eds. New York, NY, USA: ACM, June 2011, pp. 689–696.

[2] L. Xing, M. Chen, J. Yao, Y. Zhang, and Y. Wang, "Pre-post interaction learning for brain tumor segmentation with missing mri modalities," in *Proc. ICASSP*, 2024, pp. 1711–1715.

[3] S. Qian, Z. Fan, Q. He, and C. Wang, "Timm: A type-induced missing-aware mixture-of-experts model for imputing missing modalities," in *IJCNN*, 2025, pp. 1–8.

[4] N. Wu and L. Cai, "Ncmi: A framework for missing modality generation in multi-modal recommendation," in *IJCNN*, 2025, pp. 1–7.

[5] N. Neverova, C. Wolf, and G. W. T. et al., "Moddrop: adaptive multi-modal gesture recognition," 2015.

[6] Z. W. L. Cai and H. G. et al., "Deep adversarial learning for multi-modality missing data completion," in *Proc. KDD*, ser. KDD '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1158–1166.

[7] W. Huang, Y. Chen, X. Jiang, C. Gao, T. Zhang, Q. Chen, and Y. Wang, "Mitigating pervasive modality absence through multimodal generalization and refinement," in *Proc. AAAI*, 2025, pp. 26 796–26 804.

[8] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 177–18 186.

[9] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6558–6569.

[10] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, " Trusted Multi-View Classification With Dynamic Evidential Fusion ," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 45, no. 02, pp. 2551–2566, Feb. 2023.

[11] W. Zhao, K. Yang, P. Ding, C. Na, and W. Li, "Graph attention contrastive learning with missing modality for multimodal recommendation," *Knowledge-Based Systems*, vol. 311, p. 113035, 2025.

[12] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton, and J. 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*. OpenReview.net, 2017.

[13] J. Zhao, R. Li, and Q. Jin, "Missing modality imagination network for emotion recognition with uncertain missing modalities," in *Proc. ACL/IJCNLP*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 2608–2618.

[14] W. Kool, H. van Hoof, and M. Welling, "Ancestral gumbel-top-k sampling for sampling without replacement," *Journal of Machine Learning Research*, vol. 21, no. 47, pp. 1–36, 2020.

[15] J. E. Mosimann, "On the compound multinomial distribution, the multivariate β- distribution, and correlations among proportions," *Biometrika*, vol. 49, no. 1/2, pp. 65–82, 1962.

[16] T. Huillet and C. Paroissin, "Estimations of the parameter of a dirichlet distribution using residual allocation model representations and sampling properties," *Statistical Methodology*, vol. 2, no. 2, pp. 95–110, 2005.

[17] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Adv. Neural Inf. Process. Syst.*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds., vol. 14. MIT Press, 2001.

[18] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," 2015.

[19] J. Arevalo, T. Solorio, M. Montes-y Gmez, and F. A. Gonzlez, "Gated multimodal units for information fusion," *arXiv preprint arXiv:1702.01992*, 2017.

[20] A. Zadeh, R. Zellers, E. Pincus, and L. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[21] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: learning robust joint representations by cyclic translations between modalities," in *Proc. AAAI*, ser. AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019.

[22] Y.-L. Lee, Y.-H. Tsai, W.-C. Chiu, and C.-Y. Lee, "Multimodal prompting with missing modalities for visual recognition," in *Proc. CVPR*, 2023, pp. 14 943–14 952.

[23] Z. Liu, Z. Wang, X. Li, Y. Zhang, Y. Zhang, and F. Wu, "Epe-p: Evidence-based parameter-efficient prompting for multimodal learning with missing modalities," in *Proc. ICML*, vol. 235, 2024, pp. 28 975–28 986.

[24] Y. Zhang, Z. Liu, Z. Wang, X. Li, and F. Wu, "Robust multimodal learning with missing modalities via parameter-efficient adaptation," in *Proc. ACL*, vol. 1, 2024, pp. 11 778–11 792.

[25] G. Kwon, Z. Cai, A. Ravichandran, E. Bas, R. Bhotika, and S. Soatto, "Masked vision and language modeling for multi-modal representation learning," in *The Eleventh International Conference on Learning Representations*, 2023.

[26] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[27] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *Proc. ICML*. PMLR, 2021, pp. 5583–5594.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguist.: Human Lang. Technol., Vol. 1 (Long Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[29] R. Wu, H. Wang, H.-T. Chen, and G. Carneiro, "Deep multimodal learning with missing modality: A survey," 2024.

[30] Y. Zhan, R. Yang, J. You, M. Huang, W. Liu, and X. Liu, "A systematic literature review on incomplete multimodal learning: techniques and challenges," *Systems Science & Control Engineering*, vol. 13, no. 1, p. 2467083, 2025.