PRIVILEGED POSTERIOR CALIBRATION OF MEMORY ROUTING FOR INCOMPLETE MULTIMODAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt memories can compensate for an absent modality only if the remaining evidence retrieves useful prompts. We study how complete training pairs can supervise this memory-access decision without requiring complete inputs at deployment. Privileged Posterior Calibration (PPC) transfers complete-view routing distributions to an observed-view router through a confidence-weighted dense objective, while retaining sparse retrieval for prediction. The goal is not to recover an unknowable complete-view route for every input: a conditional-projection analysis shows that the population objective learns the teacher routing predictable from observed evidence. Complete-view prototypes and label-neighborhood regularization structure the shared memory, and the observed-input task loss remains the prediction objective. Across three vision–language benchmarks under the RAGPT/ANGA-aligned 70% missing-rate protocol, PPC improves over reproduced MemPrompt in all nine settings. On image-missing MM-IMDb, PPC reaches 51.5 test F1-Macro, compared with 49.3 for complete-view feature distillation and 47.9 for the architecture-matched no-calibration control, demonstrating the value of supervising memory access beyond complete-view training alone.

1 INTRODUCTION

Multimodal predictors benefit from complementary evidence, yet often encounter an entirely absent image or text at deployment (Kim et al., 2021; Ma et al., 2022). Prompt tuning adapts frozen backbones to these cases without updating the full model (Lee et al., 2023; Jang et al., 2024). Memory-driven prompting further retrieves instance-specific prompts from a learned bank (Zhao et al., 2025). Its success depends on two distinct questions: what information the memory stores, and whether partial observations can retrieve prompts that help prediction. We address this *routing limitation* by supervising memory access rather than increasing prompt capacity.

Complete training pairs expose task evidence that is absent from the deployed input and provide a privileged reference for memory access. The central challenge is transferring the part of this reference that is predictable from the remaining evidence, rather than enforcing sample-wise recovery of an unavailable view. Probabilistic distillation already recognizes this information asymmetry (Chen et al., 2024); full-modality supervision for missing-modality prompts is also established (Yang et al., 2025). The question is therefore *which decision to supervise and how to transfer that supervision under partial observation*.

We propose **Privileged Posterior Calibration** (PPC), which applies this supervision to distributions over shared prompt-memory slots. The observed input produces a route used for prediction; its complete training counterpart supplies a soft routing target. Rather than enforcing exact representation recovery, a dense, confidence-weighted objective teaches the observed router to approximate complete-view routing averaged over the missing evidence compatible with its input. Our analysis characterizes this learnable target and connects routing discrepancy to prompt-induced prediction loss under explicit stability assumptions. The task loss continues to train the deployed predictor, while prototype initialization and label-neighborhood regularization give the routing target semantic structure. Sparse top-k retrieval is retained throughout, and the complete-view stream is discarded at inference. Figure 1 contrasts this routing supervision with MemPrompt's task-trained retrieval.

Our experiments distinguish full-system performance from the effect of routing supervision. Comparisons to reproduced MemPrompt assess the PPC package; complete-view logit/feature distillation

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.

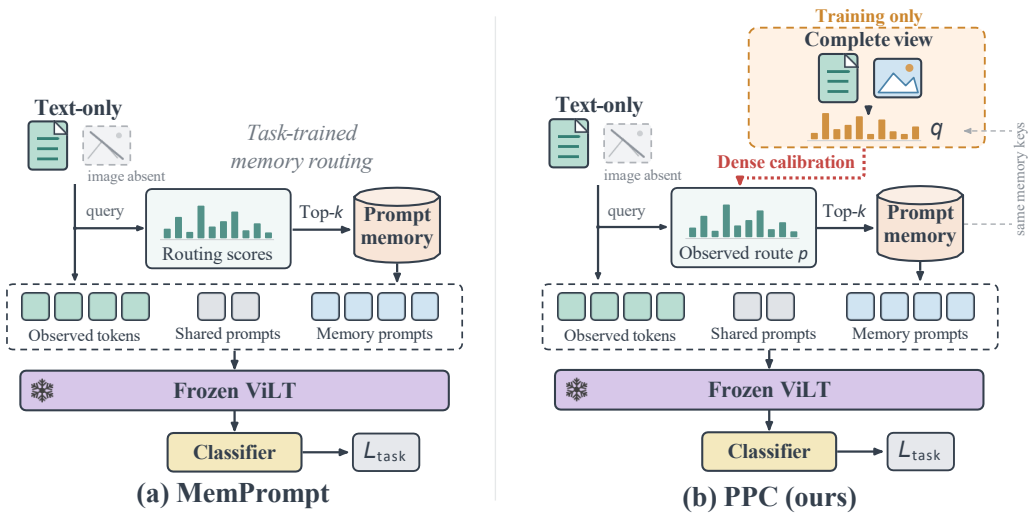

Figure 1: MemPrompt vs. PPC under image missingness. Both methods retrieve prompts from observed evidence. PPC adds a training-only complete view to calibrate the observed route p toward a detached reference q over shared memory keys before sparse top-k selection. The complete-view branch is removed at inference. Routing bars are illustrative.

and an architecture-matched no-calibration control isolate its supervision interface. Comparative routing diagnostics localize the associated change to memory access for the absent modality. Subset evaluation and fixed-predictor route interventions directly assess the predictive utility of this access decision.

Our contributions are summarized as follows:

- We formulate complete-view supervision of shared prompt-memory access and characterize the conditional-projection target available to an incomplete-view router.
- We develop confidence-weighted dense routing calibration with sparse deployment, supported by prototype initialization and semantic-neighborhood regularization.
- We evaluate PPC on three benchmarks and separate full-system gains from the incremental effect of calibration using information-matched ablations and comparative routing diagnostics.

2 Related Work

Learning with missing modalities. Shared representations, disentanglement, and modality experts reduce dependence on individual inputs (Ma et al., 2021; Zuo et al., 2023; Wang et al., 2023a; Yao et al., 2024; Xu et al., 2024). Reconstruction methods instead infer absent information (Woo et al., 2023; Wang et al., 2023b; Yang et al., 2023). RAGPT retrieves contextual substitutes and generates dynamic prompts (Lang et al., 2025). ANGA addresses gradient imbalance between complete and reconstructed examples (Guan et al., 2026). PPC learns access to a parametric prompt memory rather than retrieving external instances to reconstruct a modality.

Parameter-efficient adaptation. MAPs and MSPs condition prompting on modality availability (Lee et al., 2023; Jang et al., 2024). Deep Correlated Prompting models input–prompt, inter-layer, and cross-modal relationships, while SyP combines static and dynamic prompts (Hu et al., 2024; Zhang et al., 2025). MemPrompt retrieves instance-specific prompts from modality memories (Zhao et al., 2025). Missing-modality prompting is also studied with CLIP-based models: AOEPT uses modality-contextualized prompting to counter implicit modality reduction and supports both CLIP and ViLT backbones (Lang et al., 2026). We evaluate its ViLT variant for a backbone-matched comparison. MoRA shares low-rank adaptation parameters across modalities while retaining modality-

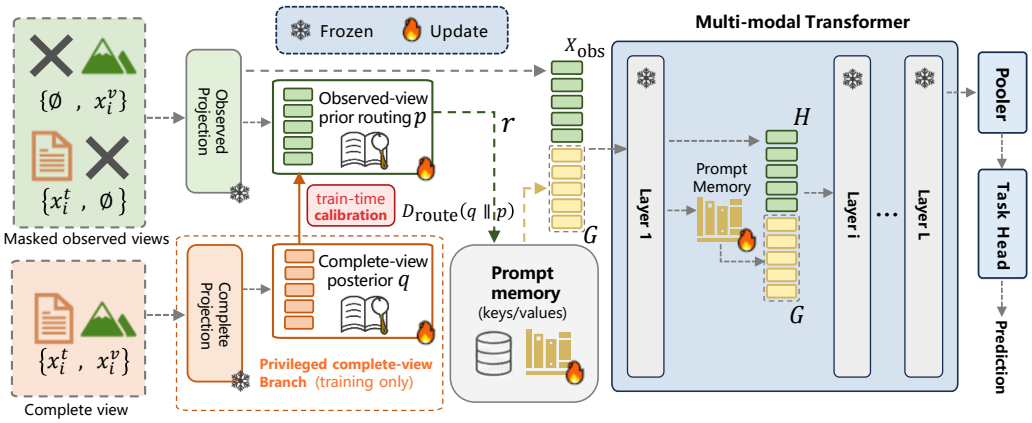

Figure 2: Overview of PPC. The masked observed view produces dense prior routing p for calibration and sparse routing r for memory-prompt retrieval. The complete training pair provides posterior routing q over the same memory keys. Calibration acts before top-k selection; only the observed branch predicts at inference. Appendix A.1 details parameter sharing and gradient paths.

specific parameters (Zhao et al., 2026); PPC instead supervises access to an instance-conditioned prompt memory.

Privileged distillation. Training-only information can supervise a deployable student (Vapnik & Vashist, 2009; Lopez-Paz et al., 2016; Li et al., 2024). PCD models uncertainty in mapping missing-modality features into a complete-modality space, explicitly avoiding overly restrictive point alignment (Chen et al., 2024). DPMamba transfers full-modality knowledge into missing-modality prompts through a shared text-semantic space and relational constraints (Yang et al., 2025). PPC supervises a complete-view distribution over shared prompt-memory keys, transferring privileged evidence at the decision that selects prompts for a frozen predictor. This memory-access formulation combines a conditional-projection interpretation with dense calibration before sparse retrieval and a semantically structured memory.

3 Methodology

Figure 2 presents the PPC architecture and its training-only complete-view branch.

3.1 Problem Setup

We consider a two-modality vision-language dataset $\mathcal{D} = \{(x_i^t, x_i^v, y_i)\}_{i=1}^N$, where x_i^t and x_i^v denote text and image inputs, and y_i is either a single class label or a multi-hot label vector. Following the standard protocol used in incomplete multimodal learning benchmarks (Lee et al., 2023; Jang et al., 2024; Lang et al., 2025; Guan et al., 2026), complete image-text pairs are available during training, and missingness is simulated by masking one modality for a subset of samples. The original pair $x_i = (x_i^t, x_i^v)$ is the *complete view*; the observed view is $\tilde{x}_i \in \{x_i, (\varnothing, x_i^v), (x_i^t, \varnothing)\}$, where $\varnothing$ denotes an absent modality. Let $\delta_i = \mathbf{1}[\tilde{x}_i \neq x_i]$ indicate whether sample i is incomplete. After embedding and projection, the observed token sequence is denoted as X_{obs}. At inference time, the model receives only $\tilde{x}_i$, without access to the corresponding complete view. PPC therefore studies *complete-pair training with simulated modality absence*, not intrinsically unpaired training or partial image corruption. This information assumption is held fixed in the privileged-supervision controls; it is distinct from a method's choice to use only its masked training view.

3.2 Memory-Driven Prompt Retrieval

PPC freezes a pretrained multimodal Transformer and updates prompt memories, shared prompts, routing projections, and task heads. For readability, we first describe retrieval for a fixed sample and

a fixed prompted layer, and restore sample and layer indices only when defining training losses. For modality $u \in \{t, v\}$, we maintain a prompt memory $\mathcal{M}_u = \{(k_{u,s}, P_{u,s})\}_{s=1}^{S}$, where S is the total number of memory slots, $k_{u,s}$ is a memory key, and $P_{u,s}$ is the corresponding prompt value. We max-pool each modality's layer-wise token features to obtain text and image states. The observed-view query h_u is then generated from the available evidence. For a present modality, a trainable self-modality projection produces the query; for a missing modality, a trainable cross-modal projection uses the remaining modality.

Let $\nu(x) = x/\|x\|_2$ denote ℓ_2 normalization. Routing and prompt retrieval are

$$a_u(s) = \frac{\langle \nu(h_u), \nu(k_{u,s}) \rangle}{\tau}, \qquad r_u = \mathcal{T}(a_u), \qquad G_u = \sum_{s=1}^{S} r_u(s) P_{u,s}, \tag{1}$$

where $\mathcal{T}$ keeps the top k logits and applies softmax over this support; τ is the routing temperature. Following MemPrompt terminology, we call G_u a *generative memory prompt*: it is generated for each instance by retrieving memory values with r_u. These generative memory prompts are inserted near their corresponding modality tokens, together with *shared prompts*, i.e., missing-type-aware prompt tokens shared across samples to encode cross-modal compensation (Zhao et al., 2025).

3.3 Privileged Posterior Calibration

The complete-view stream uses the same frozen backbone without prompt insertion. Its layer-wise states pass through the corresponding self-modality projections to form $\bar{h}_u$, and $\bar{a}_u(s) = \langle \nu(\bar{h}_u), \nu(k_{u,s}) \rangle / \tau$ scores the same memory keys. Thus the teacher supplies an alternative access distribution over the student's memory, not a separate bank of privileged prompt values. For routing diagnostics and semantic regularization, we can also obtain a sparse complete-view routing distribution $\bar{r}_u = \mathcal{T}(\bar{a}_u)$. However, applying the calibration loss only after hard top-k selection may leave posterior-selected but prior-unselected slots without direct gradients. We therefore perform the main routing calibration on dense pre-top-k distributions.

$$p_u = \mathrm{softmax}(a_u), \quad q_u = \mathrm{softmax}(\bar{a}_u). \tag{2}$$

Sparse retrieval r_u remains the prompt-injection rule in training as well as inference. We define the directed route-calibration discrepancy as

$$D_{\mathrm{route}}(q \| p) = \sum_{s=1}^{S} \mathrm{sg}[q(s)] \left(\log \mathrm{sg}[q(s)] - \log p(s) \right), \tag{3}$$

where D_{route} denotes a discrepancy over route distributions, $\|$ indicates the direction from complete-view posterior q to observed-view prior p, and $\mathrm{sg}[\cdot]$ stops gradients through the complete-view target. For a detached target, the logit gradient is $\partial D_{\mathrm{route}} / \partial a(s) = p(s) - q(s)$: a slot omitted by the student's top-k still receives a direct corrective signal. PPC then aligns the observed-view prior with the complete-view posterior by

$$\mathcal{L}_{\mathrm{ppc}} = \frac{1}{2|\mathcal{A}|} \sum_{l \in \mathcal{A}} \sum_{u \in \{t,v\}} \frac{\sum_i \delta_i \bar{c}_{i,u}^{l} D_{\mathrm{route}}(q_{i,u}^{l} \| p_{i,u}^{l})}{\max\left(\sum_i \delta_i \bar{c}_{i,u}^{l}, 1\right)}, \tag{4}$$

where $\mathcal{A}$ contains calibrated layers and the denominator floor prevents amplification of nearly uninformative targets. "Prior" and "posterior" describe conditioning on observed and complete evidence, not a claim of exact Bayesian inference.

A flat posterior provides little routing guidance, so we use a detached confidence weight

$$\bar{c}_{i,u}^{l} = \mathrm{sg}\left[\left(1 - \frac{H(q_{i,u}^{l})}{\log S} \right)^{\gamma} \right] \tag{5}$$

where H is entropy and γ controls weighting sharpness. This weights the routing signal by its concentration; observed-view task supervision supplies the predictive criterion. Stop-gradient is local to this loss: shared keys and projections continue to evolve through the observed task loss and semantic regularization.

3.4 What Can a Partial-Input Router Learn?

Routing calibration transfers a privileged access decision while the task loss trains prediction. To characterize the transferable target, fix the teacher and memory for one layer and modality. Let O denote the observed evidence including its missingness pattern, Q the random complete-view routing target, and $C \geq 0$ its effective calibration weight. The population analogue of a fixed-teacher calibration step is $\mathbb{E}[C\,\mathrm{KL}(Q\|p(O))]$.

Proposition 1 (conditional projection). *Among all routing functions with values in the probability simplex, wherever $\mathbb{E}[C \mid O] > 0$, the minimizer is*

$$p^\star(s \mid O) = \frac{\mathbb{E}[CQ_s \mid O]}{\mathbb{E}[C \mid O]}. \tag{6}$$

The proof is a weighted cross-entropy decomposition (Appendix F). With constant confidence, the target is $\mathbb{E}[Q \mid O]$, not the route of an unrecoverable individual complete input. When several missing views are compatible with the same observation, their teacher routes are averaged rather than required to coincide. This characterizes an unrestricted, fixed-teacher population objective; finite-capacity cosine routers and joint training approximate it.

Connection to prediction. For a fixed routing block with bounded prompt values and a task loss Lipschitz in its retrieved prompt, Appendix F bounds the loss difference between student- and teacher-routed predictions, accounting for top-k truncation and giving a vanishing bound when the selected slots have a separation margin. This gives a conditional mechanism for preserving a useful teacher access decision; observed-view task supervision remains the prediction objective. The fixed-predictor intervention in Section 4.5 directly tests teacher-route utility, separately from agreement.

3.5 Complete-View Prototype Memory Initialization

Randomly initialized memory keys make early routing unstable, especially when the prior is formed from incomplete observations. To give the prompt memory a data-dependent starting geometry, PPC initializes memory keys from complete-view query prototypes. Before training, we sample a small number of complete training batches and run the frozen backbone with complete image-text inputs. For each layer and modality, identity-initialized projections produce normalized queries, which spherical k-means groups into S centers $\mu^l_{u,s}$. Let $k^{l,0}_{u,s}$ be the randomly initialized key before prototype initialization. The memory keys are initialized once by blending this initial key with the complete-view cluster center:

$$k^l_{u,s} \leftarrow (1-\rho)k^{l,0}_{u,s} + \rho\mu^l_{u,s}, \tag{7}$$

where ρ is the prototype blending ratio. The keys remain trainable after initialization, and all routing similarities normalize both queries and keys. These are label-free query clusters, not class centroids: the number of slots is independent of the number of categories.

3.6 Semantic-Neighborhood Posterior Regularization

While prototype initialization improves the starting geometry of memory, posterior routing can still drift during training. PPC therefore regularizes the posterior space so that semantically related training instances access compatible memory regions. This loss uses labels only during training and never uses validation or test labels. For a mini-batch, i and j index two training samples. The affinity α_{ij} is class agreement for single-label tasks and label-set Jaccard similarity for MM-IMDb. We exclude self-matching by setting $\alpha_{ii} = 0$, and each row is normalized to form a soft target $\bar{\alpha}_{ij}$. Only anchors with a non-self positive neighbor contribute, forming the set $\mathcal{I}_+$.

We represent the sparse complete-view routing distribution by its Hellinger-style embedding $\psi^l_{i,u} = \nu(\sqrt{\max(\bar{r}^l_{i,u}, \varepsilon_{\mathrm{clip}})})$. A softmax of $(\psi^l_{i,u})^\top \psi^l_{j,u}/\tau_s$ over $j \neq i$ defines $\pi^l_{ij,u}$. The semantic-neighborhood loss is then

$$\mathcal{L}_{\mathrm{sem}} = -\frac{1}{2|\mathcal{P}||\mathcal{I}_+|} \sum_{l\in\mathcal{A},u} \sum_{i\in\mathcal{I}_+} \sum_{j\neq i} \bar{\alpha}_{ij} \log \pi^l_{ij,u}. \tag{8}$$

Here $\mathcal{P}$ contains all prompted layers, with zero contributions outside $\mathcal{A}$. Excluding self-matching prevents the loss from being dominated by a trivial self-similarity target. When $\mathcal{I}_+$ is empty, the semantic-neighborhood loss is skipped for that mini-batch. The loss structures sparse routing patterns rather than assigning one slot per class.

3.7 Training Objective and Inference

Training. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \beta(e)\mathcal{L}_{\text{ppc}} + \lambda_{\text{sem}}\mathcal{L}_{\text{sem}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}. \tag{9}$$

For a mini-batch of size B, let z_i and $\bar{z}_i$ denote the logits from the observed branch and complete-view auxiliary branch, respectively. With ℓ_{task} as cross-entropy for Hateful Memes and Food101 and binary cross-entropy for MM-IMDb, we define

$$\mathcal{L}_{\text{task}} = \frac{1}{B}\sum_i \ell_{\text{task}}(z_i, y_i), \quad \mathcal{L}_{\text{aux}} = \frac{1}{B}\sum_i \ell_{\text{task}}(\bar{z}_i, y_i). \tag{10}$$

The calibration coefficient follows a linear warm-up,

$$\beta(e) = \beta_{\max}\min(1, e/E_{\text{warm}}), \tag{11}$$

where e denotes the current training epoch. This warm-up prevents the early, unstructured memory from being over-constrained by posterior targets. The auxiliary classifier shares the task head and is trained on detached complete-view features; it does not directly optimize the route q. Appendix A.1 specifies parameter sharing and loss dependencies.

Inference. At inference time, PPC removes the complete-view stream, posterior routing, prototype collection, and semantic-neighborhood regularization. For each prompted layer and modality, the model computes only the observed-view sparse retrieval distribution $r^l_{i,u} = \mathcal{T}(a^l_{i,u})$, retrieves prompts from the learned memory, and predicts from the incomplete input. There is no teacher forward pass or external-instance lookup at deployment. The additional complete-view computation is confined to training; Appendix A.2 separates this cost from the deployed architecture.

4 Experiments

4.1 Experimental Setup

Datasets. Following prior work (Lee et al., 2023; Jang et al., 2024; Lang et al., 2025; Guan et al., 2026), we evaluate PPC on three vision-language benchmarks. Hateful Memes (Kiela et al., 2020) is a binary hate-detection task, MM-IMDb (Arevalo et al., 2017) is multi-label movie-genre classification, and Food101 (Wang et al., 2015) is 101-class food recognition. Following the RAGPT/ANGA protocol, the split sizes are 8,500/500/1,500 for Hateful Memes, 15,552/2,608/7,799 for MM-IMDb, and 67,972/-/22,716 for Food101.

Evaluation metrics. Following RAGPT and ANGA, we report AUROC for Hateful Memes, F1-Micro for MM-IMDb, and accuracy for Food101, all in percentage.

Missing-modality protocol. We adopt the RAGPT/ANGA-aligned 70% missing-rate protocol. In Table 1, *Text* and *Image* denote the missing modality rather than the observed modality. For example, the Image column means that 70% of samples are image-missing and the remaining 30% are complete. *Both* denotes the mixed-missing setting, where 35% of samples are text-missing, 35% are image-missing, and 30% remain complete. Training uses the same missing-pattern construction as evaluation, while complete pairs remain available during training to construct the privileged posterior branch. No validation or test labels are used for prototype initialization, semantic-neighborhood construction, or memory calibration.

Comparison scope and information access. The shared evaluation protocol does not imply identical use of training information. MemPrompt† uses observed-input training, whereas PPC and the privileged controls in Table 2 can also use each training example's unmasked counterpart. The no-calibration control retains the same complete-view stream, memory architecture, prototypes, semantic regularizer, and auxiliary objective; only $\mathcal{L}_{\text{ppc}}$ is removed. This is the direct mechanism

Table 1: Benchmark point estimates under RAGPT/ANGA-aligned 70% missingness. Text/Image denote the absent modality; Both denotes mixed missingness (35% text-missing, 35% image-missing, 30% complete), not simultaneous absence. Literature results are from ANGA; †: our reproduction. Bold/underline mark the largest/second-largest listed values.

Dataset	Hateful Memes			MM-IMDb			Food101		
Metric	AUROC			F1-Micro			ACC		
Method	Text	Image	Both	Text	Image	Both	Text	Image	Both
Modality-invariant learning methods									
IF-MMIN (Zuo et al., 2023)	57.62	53.44	55.19	39.63	31.95	31.98	66.76	64.36	68.53
ShaSpec (Wang et al., 2023a)	58.75	60.30	60.96	44.04	44.23	44.06	60.99	74.87	70.02
DrFuse (Yao et al., 2024)	57.60	60.66	55.84	47.05	43.58	48.83	66.30	75.09	68.23
CorrKD (Li et al., 2024)	58.74	55.59	57.91	44.82	39.48	41.20	61.37	66.83	62.87
MoMKE (Xu et al., 2024)	63.08	61.35	62.53	50.98	45.67	46.99	66.85	68.40	67.38
VLP-based methods									
ViLT (Kim et al., 2021)	56.61	57.15	56.42	45.33	41.14	40.23	62.92	70.59	65.01
MAPs (Lee et al., 2023)	58.62	60.16	58.89	46.12	44.86	45.48	67.02	75.62	72.52
MSPs (Jang et al., 2024)	59.60	60.05	59.08	49.16	44.62	48.28	71.74	79.09	74.46
IPer (Lu et al., 2025)	62.09	61.68	61.34	53.56	45.82	47.62	72.56	80.23	73.73
Modality reconstruction methods									
AcMAE (Woo et al., 2023)	55.74	59.66	57.25	47.47	43.82	44.05	69.28	73.75	71.15
DiCMoR (Wang et al., 2023b)	58.03	60.27	59.38	47.02	45.19	46.68	70.47	76.54	73.22
MMDS (Yang et al., 2023)	57.14	59.72	59.68	47.47	42.17	42.44	61.15	72.52	63.76
RAGPT (Lang et al., 2025)	64.10	62.57	63.47	55.16	46.44	50.89	75.53	81.98	76.94
ANGA (Guan et al., 2026)	<u>68.54</u>	<u>63.42</u>	<u>65.12</u>	**57.28**	<u>47.75</u>	51.84	<u>77.23</u>	<u>83.03</u>	<u>78.47</u>
Memory-driven prompt methods									
MemPrompt† (Zhao et al., 2025)	62.85	61.34	62.19	54.55	47.50	<u>52.02</u>	74.26	82.37	76.62
PPC (Ours)	**68.93**	**65.31**	**66.76**	<u>55.68</u>	**52.14**	**54.36**	**78.68**	**86.97**	**80.76**

comparison. Literature rows provide benchmark context; the AOEPT comparison uses its ViLT variant with the aligned data and missing-pattern construction.

Implementation and model selection. Following RAGPT and ANGA, we use ViLT-B/32 pre-trained with masked language modeling and image-text matching as the backbone, resize images to 384×384, and set the maximum text length to 128 for the aligned main comparison. The backbone is frozen; prompt memories, shared prompts, routing modules, and task heads are optimized. We optimize the trainable components with AdamW for 20 epochs using an effective batch size of 64, learning rate 10^{-3}, and weight decay 10^{-5}. PPC uses 16 memory slots, retrieves the top 5 slots, uses prompt length 16, and calibrates the first four prompted layers. Unless otherwise specified, we set $\beta_{\max} = 0.002$, $E_{\text{warm}} = 10$, $\gamma = 2$, $\tau = 1.0$, $\lambda_{\text{aux}} = 0.2$, and use semantic regularization with $\lambda_{\text{sem}} = 5 \times 10^{-4}$ and $\tau_s = 1.0$. When a validation split is available, checkpoints are selected by validation AUROC for Hateful Memes, F1-Micro for the main MM-IMDb comparison, and F1-Macro for the MM-IMDb component analysis. For Food101, which has no validation split under this protocol, we use the checkpoint from the final training epoch. For MM-IMDb, sigmoid predictions are binarized with threshold 0.5.

4.2 Benchmark Comparison

PPC exceeds reproduced MemPrompt and RAGPT in all nine settings in Table 1. For image-missing MM-IMDb, its F1-Micro is 52.14 versus MemPrompt's 47.50. PPC has the largest listed value in eight settings; ANGA leads on text-missing MM-IMDb. These results establish consistent full-system gains over the reproduced memory-prompt baseline without external-instance retrieval at inference. Table 1 reports aggregate-test point estimates; the component study reports three-seed summaries, and Section 4.5 separates complete from incomplete examples.

Table 2: MM-IMDb under 70% image missingness, reported by test F1-Macro. Left: same-protocol baselines and backbone controls. Right: privileged-information controls and PPC ablations. Entries with $\pm$ denote mean and standard deviation over three seeds; † denotes our reproduction. AOEPT uses ViLT, not CLIP.

Variant	F1-Macro
Same-protocol baselines	
RAGPT (Lang et al., 2025)	42.7 ± 0.2
ANGA (Guan et al., 2026)	41.2 ± 0.3
AOEPT† (ViLT) (Lang et al., 2026)	49.07 ± 0.4
Memory-prompt backbone controls	
Shared prompts only	28.1 ± 0.5
MemPrompt† w/o shared prompts	40.6 ± 0.1
MemPrompt†	46.1 ± 0.2

Variant	F1-Macro
Privileged-information baselines	
Complete-view auxiliary only	48.4 ± 0.7
Complete-view logit KD only	47.8 ± 0.5
Complete-view feature KD only	49.3 ± 0.3
PPC ablations	
Full PPC	**51.5** ± 0.2
w/o route calibration	47.9 ± 0.5
w/o confidence weighting	51.0 ± 0.2
w/o prototype initialization	50.7 ± 0.2
w/o semantic regularization	50.6 ± 0.3
w/o complete-view auxiliary loss	50.8 ± 0.2

4.3 Component Ablation

To analyze the contribution of each design, we conduct ablation studies on MM-IMDb under the 70% image-missing setting and report test F1-Macro. This uses the same ANGA/RAGPT-aligned data construction and missing masks; it changes the selection/reporting metric to validation/test F1-Macro. F1-Macro weights the 23 genres equally and complements, rather than replaces, the main F1-Micro comparison. All component variants use validation F1-Macro for checkpoint selection before test evaluation. Since RAGPT and ANGA do not report F1-Macro in their original papers, we additionally evaluate them under the same diagnostic protocol. Table 2 combines these references, matched backbone controls, alternative complete-view supervision, and leave-one-out ablations.

AOEPT comparison. We reproduce AOEPT's ViLT variant under the aligned MM-IMDb protocol as a representative of methods also studied with CLIP. Its test F1-Macro is 49.07 ± 0.4, 46.90 ± 0.3, and 48.24 ± 0.3 for image, text, and mixed (Both) missingness, respectively. These macro scores are separate from the F1-Micro benchmark in Table 1. On image missingness, Full PPC exceeds AOEPT by 2.43 F1-Macro points (Table 2).

Component effects. *Shared prompts only* disables generative memory prompts and keeps only missing-type-aware shared prompts, whereas *MemPrompt w/o shared prompts* retains only the generative memory pathway. Shared prompts alone reach 28.1; adding them to the memory pathway improves MemPrompt from 40.6 to 46.1. The full gain over MemPrompt is 5.4 points, but it cannot all be attributed to calibration: complete-view feature distillation alone improves 46.1 to 49.3. Full PPC adds 2.2 points over feature distillation, the strongest privileged alternative, 3.7 over logit distillation, and 3.6 over the information-matched no-calibration control (47.9). These comparisons answer different questions: the distillation controls test the supervision interface, whereas the no-calibration control tests routing supervision within the full memory system. The remaining leave-one-out ablations support the other components within PPC; each drop measures an effect conditional on the retained components, rather than an additive contribution. Throughout this paper, Full PPC uses dense calibration with semantic self-matches excluded; alternative development configurations are not mixed into its three-seed summary.

4.4 Comparative Routing Diagnostics

On incomplete MM-IMDb test samples under 70% image missingness, Overlap@5 measures the fraction of five selected slots shared by the observed and complete-view routes; Gap measures their sparse route discrepancy. Both statistics average over calibrated layers and incomplete examples (definitions in Appendix E). The observed-text branch retrieves text prompts, while the missing-image branch retrieves visual prompts from the same remaining text evidence.

Table 3: Calibration versus an information-matched control on image-missing MM-IMDb. Routing: point estimates on incomplete test examples, with overlap in percent. F1-Macro: aggregate-test three-seed summaries from Table 2, not paired per-checkpoint routing–accuracy estimates.

Method	Observed-text Overlap@5 ↑	Missing-image Overlap@5 ↑	Missing-image Gap ↓	Test F1-Macro ↑
w/o route calibration	99.68	63.02	5.175	47.9 ± 0.5
Full PPC	98.26	86.10	1.523	51.5 ± 0.2

Calibration increases missing-image overlap by 23.08 points and reduces its gap by 3.652, while slightly reducing the already high observed-text overlap (Table 3). The change is therefore concentrated on routing for the absent modality, rather than a uniform rise in agreement. The 3.6-point classification improvement thus accompanies a targeted improvement in missing-modality routing. The no-calibration control preserves the reference construction; Appendix E gives layer-wise results, and Appendix G.2 reports shuffled-reference overlap and dense routing discrepancy.

4.5 Predictive Utility of Memory Access

Table 4 separates predictive gains by the original test masks and intervenes on visual-memory routes in a fixed, masked-input PPC predictor. PPC improves F1-Macro over MemPrompt by 7.22 points on missing-image examples and 0.87 on complete examples, with the larger gain on incomplete inputs. Replacing only the route with its paired complete-view reference improves missing-image F1-Macro by 1.46 points; random sparse routing reduces it by 7.27. Shuffled complete-view routing scores 0.82 points below observed routing and 2.28 below paired routing. These contrasts support the predictive value of sample-matched complete-view routing for the fixed predictor. Complete images supply reference routes, never predictor inputs (Appendix G.1).

Table 4: Predictive checks on 70% image-missing MM-IMDb (test F1-Macro, mean $\pm$ standard deviation). Left: subset evaluation; All repeats Table 2. Right: visual-memory route interventions with the PPC predictor fixed, evaluated only on missing-image examples. Complete-view routes are offline diagnostic references, not deployment inputs.

Method	All	Missing	Complete
MemPrompt†	46.1 ± 0.2	42.43 ± 0.18	54.94 ± 0.5
Full PPC	51.5 ± 0.2	49.65 ± 0.25	55.81 ± 0.2

Route in Full PPC	Missing
Observed r	49.65 ± 0.3
Complete-view $\bar{r}$	51.11 ± 0.5
Shuffled $\bar{r}$	48.83 ± 0.3
Random sparse	42.38 ± 0.4

5 Conclusion

PPC uses complete training views to supervise a deployable memory-access decision. Its conditional-projection interpretation accommodates irrecoverable ambiguity, while dense calibration supplies gradients beyond the student's sparse support. Matched privileged controls identify an additional benefit from routing supervision beyond complete-view training alone. The resulting model retains sparse prompting at inference, connecting privileged training evidence to memory-based prediction without requiring the missing modality at deployment.

AI use statement

AI tools were used to assist with literature retrieval and discovery, including identifying related work, as well as manuscript writing and language polishing. The authors take full responsibility for the final content and all claims in this paper.

Reproducibility statement

Section 3 defines the model and training objectives, and Section 4.1 specifies the datasets, missing-modality protocol, optimization settings, and evaluation metrics. Appendices A–C provide implementation details, parameter-sharing and optimization paths, the hyperparameter selection protocol, and sensitivity results. Table 2 and Appendix D describe the ablation controls; Section 4.4 and Appendix E provide the routing diagnostics. Appendix F states the assumptions and proofs for the decision-level analysis. Appendix G details subset evaluation, fixed-predictor route interventions, shuffled-reference diagnostics, and cross-task controls. Imported benchmark values and information-matched component comparisons are identified separately; point estimates and repeated-run summaries are distinguished, and no statistical significance is inferred from point estimates.

References

John Arevalo, Thamar Solorio, Manuel Montes-y Gomez, and Fabio A. Gonzalez. Gated multimodal units for information fusion. In *Proceedings of the International Conference on Learning Representations Workshop*, 2017.

Mengxi Chen, Fei Zhang, Zihua Zhao, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Probabilistic conformal distillation for enhancing missing modality robustness. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1142.

Zhi-Hao Guan, Longfei Huang, and Yang Yang. Anchor-guided gradient alignment for incomplete multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37896–37905, 2026.

Lianyu Hu, Tongkai Shi, Wei Feng, Fanhua Shang, and Liang Wan. Deep correlated prompting for visual recognition with missing modalities. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2153.

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

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2611–2624, 2020.

Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5583–5594, 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*, pp. 18035–18043, 2025.

Jian Lang, Rongpei Hong, Ting Zhong, and Fan Zhou. AOEPT: Breaking the implicit modality-reduction bottleneck in modality-missing prompt tuning. In *Proceedings of the International Conference on Machine Learning*, 2026.

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.

David Lopez-Paz, Leon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *Proceedings of the International Conference on Learning Representations*, 2016.

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

Mengmeng Ma, Jian Ren, Long Zhao, 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.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multimodal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15878–15887, 2023a.

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

Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21968–21977, 2023b.

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*, pp. 2776–2784, 2023.

Wenxin Xu, Hexin Jiang, and Xuefeng Liang. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the ACM International Conference on Multimedia*, pp. 438–446, 2024.

Yanwu Yang, Hairui Chen, Zhikai Chang, Yang Xiang, Chenfei Ye, and Heather Ting Ma. Incomplete learning of multimodal connectome for brain disorder diagnosis via modal-mixup and deep supervision. In *Proceedings of the Medical Imaging with Deep Learning Conference*, pp. 1006–1018, 2023.

Yueguang Yang, Jiahui Qu, Ling Huang, and Wenqian Dong. DPMamba: Distillation prompt mamba for multimodal remote sensing image classification with missing modalities. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 2224–2232, 2025. doi: 10.24963/ijcai.2025/248.

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*, pp. 16416–16424, 2024.

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

Shu Zhao, Nilesh Ahuja, Tan Yu, Tianyi Shen, and Vijaykrishnan Narayanan. MoRA: Missing modality low-rank adaptation for visual recognition. In *Proceedings of the International Conference on Learning Representations*, 2026.

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

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

A Implementation Details

Training and inference scope. All results in this manuscript use the ANGA/RAGPT-aligned benchmark construction, not a separate native MemPrompt data protocol. Paired training examples are available before masking; PPC uses the unmasked training view for routing targets, prototype initialization, semantic regularization, and the auxiliary classifier. Validation/test labels never enter these operations. At deployment, the complete-view stream and training-only losses are removed. Complete test inputs are accessed only for explicitly identified offline routing diagnostics, not for the benchmark predictions.

Backbone and optimization. PPC and its memory-prompt controls use ViLT-B/32 pretrained with masked language modeling and image-text matching. The Transformer backbone is frozen, and only prompt memories, routing projections, shared prompts, and task heads are updated. Following the RAGPT/ANGA-aligned protocol used in the main paper, we resize images to 384×384, set the maximum text length to 128, and train for 20 epochs with AdamW, effective batch size 64, learning rate 10^{-3}, and weight decay 10^{-5}. The MM-IMDb component experiments keep this construction and optimization setting but use validation/test F1-Macro for selection/reporting. Per-GPU batch size and GPU count may vary across servers; gradient accumulation preserves the effective batch size. Table 5 summarizes the PPC settings, rather than prescribing the optimization configuration of external methods such as AOEPT.

Table 5: Core implementation settings.

Item	Setting
Backbone	ViLT-B/32 with MLM+ITM pretraining; backbone frozen
Input	384×384 images, 32×32 patches, max text length 128
Optimization	AdamW, 20 epochs, batch size 64, lr 10^{-3}, wd 10^{-5}
Prompt memory	$S = 16$, $k = 5$, prompt lengths 16/16, prompted layers 0–5
PPC routing	calibrated layers 0–3, $\beta_{\max} = 2{\times}10^{-3}$, warm-up 10, $\gamma = 2$
Auxiliary loss	complete-view auxiliary weight $\lambda_{\text{aux}} = 0.2$
Prototype init.	16 complete batches, 8 k-means iterations, key blend 0.5
Semantic reg.	$\lambda_{\text{sem}} = 5{\times}10^{-4}$, $\tau_s = 1.0$
Evaluation	MM-IMDb sigmoid threshold fixed to 0.5

Prototype initialization. For each prompted layer and modality, PPC collects complete-view queries from a small number of training batches and runs spherical k-means to initialize memory keys. This initialization is performed once before training. Labels are not used in this step, the number of prototypes equals the memory size S, and memory keys remain trainable afterwards. Therefore, the procedure does not degenerate into class prototypes, including on binary Hateful Memes, and it naturally supports the multi-label structure of MM-IMDb.

Checkpoint selection. Checkpoints are selected by validation AUROC for Hateful Memes, validation F1-Micro for the main MM-IMDb benchmark, and validation F1-Macro for the MM-IMDb component analysis. The test split is evaluated only after model configuration and checkpoint selection. Food101 has no validation split in this protocol and uses the final training epoch with globally fixed hyperparameters. No test-set threshold search is performed.

A.1 Optimization Graph and Information Access

The complete-view stream is an unprompted forward pass of the shared frozen backbone. Its states define routing queries and detached auxiliary-classification features. The auxiliary head is shared with the observed branch. In particular, auxiliary classification is not a loss on teacher-selected prompts: it regularizes the shared classifier, whereas the routing target is shaped by shared memory parameters and semantic neighborhoods. Table 6 makes this separation explicit.

The shared evaluation protocol fixes the prediction problem, not every method's information budget. The full PPC–MemPrompt comparison includes the effect of complete-view training. The w/o-calibration comparison instead holds complete-view access, architecture, initialization, semantic

Table 6: Optimization paths in the stated PPC configuration. A frozen backbone can transmit gradients to observed prompts without updating its own weights. Stop-gradient in calibration does not freeze the shared memory across optimization steps.

Component	Sharing / status	Direct training signals
Backbone weights	Shared, frozen	None
Self-modality query projections	Shared, trainable	Task and calibration via observed routes; semantic loss via complete routes
Cross-modal query projections	Observed branch, trainable	Task and calibration
Memory keys	Shared, trainable	Task, calibration via p, and semantic loss via $\bar{r}$
Memory values / shared prompts	Observed prediction, trainable	Observed task loss
Task classifier	Shared, trainable	Observed task and complete-view auxiliary losses
Complete target q	Detached in D_{route}	No gradient through q or its confidence in calibration

regularization, and auxiliary training fixed. Alternative logit/feature distillation controls receive the same paired training data, but change the supervision interface. We do not infer an isolated calibration effect from differences against imported literature results.

Numerical and semantic conventions. We use $\varepsilon_{\text{clip}} = 10^{-8}$ for the sparse routing embedding. The calibration denominator is $\max(\sum_i \delta_i \bar{c}_i, 1)$; in particular, a batch of nearly uniform teachers is downweighted rather than rescaled to unit total weight. For single-label tasks $\alpha_{ij} = \mathbf{1}[y_i = y_j]$; for MM-IMDb,

$$\alpha_{ij} = \frac{y_i^\top y_j}{\max(\|y_i\|_1 + \|y_j\|_1 - y_i^\top y_j, 1)}.$$

The final configuration excludes self-matches both from affinity targets and from the contrastive denominator, and uses non-self positive anchors. With $\bar{\alpha}_{ij} = \alpha_{ij} / \sum_{b \neq i} \alpha_{ib}$,

$$\pi^l_{ij,u} = \frac{\exp((\psi^l_{i,u})^\top \psi^l_{j,u} / \tau_s)}{\sum_{b \neq i} \exp((\psi^l_{i,u})^\top \psi^l_{b,u} / \tau_s)}, \qquad j \neq i.$$

An empty positive-anchor set contributes zero. The semantic term averages over all prompted layers with zero contributions outside the calibrated layers, as defined in the main text.

A.2 Computational Accounting

Let L_p be the number of prompted layers, d the key/query dimension, S the slots per modality, k the retrieved slots, and b the prompt length. Two modality banks contain $2L_pSd$ key parameters and $2L_pSbd$ value parameters. Two self- and two cross-modality bias-free $d \times d$ projections per layer add $4L_pd^2$ parameters, besides shared prompts and the classifier. These are architectural counts, not measured latency claims. For $L_p = 6$, $d = 768$, $S = 16$, and $b = 16$, the respective counts are 147,456 key parameters, 2,359,296 value parameters, and 14,155,776 projection parameters. The shared complete-view route reuses the self-modality projections and keys rather than duplicating them; these counts do not include shared prompt tokens, pooling, or classification heads.

Per inference, scoring the banks costs $O(2L_pSd)$ and aggregating selected values costs $O(2L_pkbd)$, in addition to query projections and prompted-backbone computation. Neither operation searches stored training instances. Relative to the same routed predictor without privileged training, PPC adds no complete-view forward pass at inference. Training adds unprompted complete-view encoding, dense route losses, pairwise semantic comparisons, and one-time prototype collection/clustering. Semantic comparisons scale quadratically with the contrastive batch size. This separates the unchanged deployed architecture from the additional training computation; hardware profiling measures the corresponding wall-clock and memory costs.

B HYPERPARAMETER SELECTION PROTOCOL

The main hyperparameters of PPC were selected on MM-IMDb under the 70% image-missing setting using validation F1-Macro. This development setting is intentionally separate from the main benchmark table, which reports MM-IMDb F1-Micro under the RAGPT/ANGA-aligned protocol. After selecting the configuration, we keep the same PPC-specific settings across Hateful Memes, MM-IMDb, and Food101, while using the standard evaluation metric of each dataset. Table 7 lists the search space.

Table 7: Hyperparameter search space. Selected values are bold.

Parameter	Candidates
S	8, 12, **16**, 20, 24
k	1, 3, **5**, 8
$\beta_{\max}$	5×10^{-4}, 10^{-3}, $\mathbf{2\times10^{-3}}$, 5×10^{-3}, 10^{-2}
E_{warm}	0, 5, **10**
γ	0, 1, **2**, 4
λ_{aux}	0, 0.1, **0.2**, 0.5
λ_{sem}	0, 10^{-4}, $\mathbf{5\times10^{-4}}$, 10^{-3}
τ_s	0.5, **1.0**, 2.0
Prototype batches	0, 8, **16**, 32
Prototype blend ρ	0, 0.25, **0.5**, 1.0

C HYPERPARAMETER SENSITIVITY

This section studies whether PPC depends on a narrow set of hyperparameters. Unless specified otherwise, we vary one parameter at a time around the final configuration and report validation F1-Macro on MM-IMDb under 70% image missingness. The goal is not to retune each dataset, but to verify that the selected values are stable and interpretable. Tables 8 and 9 report the memory-size, retrieval-size, and calibration-strength sweeps.

Table 8: Sensitivity to memory size and sparse retrieval size on MM-IMDb. Moderate memory and retrieval sizes perform best, suggesting that PPC benefits from structured memory access rather than simply increasing capacity.

Setting	Value	Val F1-Macro
Memory slots S	8	50.43
	12	50.86
	16	**51.48**
	20	51.05
	24	51.29
Retrieved slots k	1	49.61
	3	51.17
	5	**51.48**
	8	51.35

Memory size and top-k. Increasing the number of memory slots improves performance from $S = 8$ to $S = 16$, but larger memories do not yield further gains. This indicates that the memory should be expressive enough to represent diverse compensatory cues, but overly large memories make routing harder under incomplete evidence. The top-k sweep shows a similar pattern: retrieving only one slot is too restrictive, while retrieving too many slots slightly dilutes the prompt signal. We therefore use $S = 16$ and $k = 5$ as a balanced setting.

Calibration strength. The sweep over positive $\beta_{\max}$ values favors moderate calibration strength: larger weights reduce validation performance relative to 2×10^{-3}. The separate no-calibration ablation in Table 2 reports test F1-Macro and is not a numerical anchor for this validation sweep.

Table 9: Sensitivity to route-calibration strength. A small but nonzero calibration weight is important.

Setting	Value	Val F1-Macro
$\beta_{\max}$	5×10^{-4}	50.64
	10^{-3}	51.07
	2×10^{-3}	**51.48**
	5×10^{-3}	50.94
	10^{-2}	50.12

The selected weight balances the task objective against route calibration, which is averaged over modalities and calibrated layers.

Table 10: Sensitivity to auxiliary and semantic regularization parameters. Both losses are helpful in a reasonable range, but PPC does not depend on a single fragile value.

Setting	Value	Val F1-Macro
λ_{aux}	0	50.80
	0.1	51.19
	0.2	**51.48**
	0.5	51.04
λ_{sem}	0	50.60
	10^{-4}	51.03
	5×10^{-4}	**51.48**
	10^{-3}	51.21
τ_s	0.5	51.21
	1.0	**51.48**
	2.0	51.14

Auxiliary and semantic regularization. Table 10 shows that the complete-view auxiliary loss is useful but should remain modest: $\lambda_{\text{aux}} = 0.2$ performs best, while a larger weight slightly hurts the observed-branch objective. Semantic regularization follows the same pattern. Removing it drops performance, but an overly strong weight weakens the main routing objective. The contrastive temperature $\tau_s = 1.0$ gives the best validation result among the tested values, indicating a balanced neighborhood sharpness.

Table 11: Sensitivity to complete-view prototype initialization. Moderate blending gives the memory a semantic starting geometry while preserving trainable flexibility.

Prototype setting	Description	Val F1-Macro
No prototype init.	random memory keys	50.41
8 batches, $\rho = 0.5$	fewer complete-view samples	51.25
16 batches, $\rho = 0.5$	selected setting	**51.48**
32 batches, $\rho = 0.5$	more complete-view samples	50.70
16 batches, $\rho = 1.0$	no random-key blending	51.02

Prototype initialization. Table 11 shows that the tested prototype initializations outperform random keys. The 16-batch setting performs best; increasing to 32 batches lowers performance, so more initialization data does not guarantee better routing. The selected blend $\rho = 0.5$ also exceeds full replacement with $\rho = 1.0$, supporting moderate prototype blending in this experiment.

D Additional Ablation Details

The main paper reports component ablations on MM-IMDb under the 70% image-missing evaluation setting. Here we clarify the ablation variants.

- **Shared prompts only** removes generative memory prompts and keeps only missing-type-aware shared prompts.
- **MemPrompt w/o shared prompts** keeps generative memory prompts but removes shared prompts from the reproduced MemPrompt backbone.
- **Complete-view auxiliary only** keeps the complete-view branch with an auxiliary task loss but removes route calibration.
- **Complete-view logit KD only** distills complete-view logits to the observed branch, replacing route-level supervision with prediction-level supervision.
- **Complete-view feature KD only** aligns hidden representations from the complete and observed branches, replacing route-level supervision with representation-level supervision.
- **PPC w/o route calibration** keeps the complete-view stream, memory architecture, prototype initialization, semantic regularization, and auxiliary task loss, removing only the routing-calibration loss.

The backbone controls distinguish shared prompting from generative memory prompting within MemPrompt. The privileged-information controls compare different uses of complete-view training information, while PPC w/o route calibration isolates the effect of routing supervision within the full architecture. The latter comparison retains both prompt pathways and does not change the inference architecture or prompt capacity.

E Routing Diagnostics

We further report routing diagnostics for PPC on MM-IMDb under 70% image missingness. The observed-text branch retrieves prompts for the present text modality from text evidence. The missing-image branch retrieves visual compensatory prompts from the remaining text evidence because the image input is masked. Overlap@5 measures the top-5 slot agreement between observed-view routing and complete-view teacher routing. The gap is the sparse diagnostic route discrepancy used only for analysis. Table 12 reports layer-wise PPC routing, and Table 13 isolates the effect of route calibration.

Diagnostic definitions. For a sample, modality, and calibrated layer, define

$$\mathrm{Overlap@}k = \frac{|\mathrm{TopK}(r) \cap \mathrm{TopK}(\bar{r})|}{k}, \qquad k = 5, \tag{12}$$

and

$$\mathrm{Gap}(\bar{r}, r) = \sum_s \hat{\bar{r}}_s \big(\log \hat{\bar{r}}_s - \log \hat{r}_s\big), \quad \hat{r}_s = \max(r_s, 10^{-8}), \quad \hat{\bar{r}}_s = \max(\bar{r}_s, 10^{-8}). \tag{13}$$

These sparse vectors are not renormalized after clipping. Gap is a support-sensitive diagnostic, not a KL divergence between normalized distributions and not the optimized dense loss. Statistics first evaluate each sample/layer, then average over incomplete examples and calibrated layers; overall statistics also average the two branches. The dense, unclipped $\mathrm{KL}(q\|p)$ is specified separately in Appendix G.2.

Table 12: Layer-wise routing diagnostics for PPC. Overlap@5 is shown in percentage. The missing-image branch is harder because it must retrieve visual prompts using only text evidence.

Layer / branch	Overlap@5	Gap
Layer 0, missing-image	80.63	0.630
Layer 1, missing-image	95.17	0.179
Layer 2, missing-image	79.29	2.342
Layer 3, missing-image	89.30	2.942
All layers, missing-image	86.10	1.523
All layers, observed-text	98.26	0.062
All layers, overall	92.18	0.793

Table 13: Effect of route calibration. Overlap@5 is reported in percentage.

Metric	w/o calib.	Full PPC	Δ
Observed-text Overlap@5 ↑	**99.68**	98.26	-1.42
Missing-image Overlap@5 ↑	63.02	**86.10**	$+23.08$
Missing-image Gap ↓	5.175	**1.523**	-3.652
Overall Overlap@5 ↑	81.35	**92.18**	$+10.83$
Overall Gap ↓	2.593	**0.793**	-1.800

The comparative diagnostic localizes the larger alignment change to the missing-image branch: overlap rises by 23.08 points and the gap falls by 3.652, while observed-text overlap falls slightly. Routing entries are point estimates; each model is compared with its own complete-view route over its own slots. Cross-model slot identities are not equated. Prediction interventions and instance-specificity checks complement this within-model agreement measurement in Appendix G.

Why the control is not raw MemPrompt. The matched no-calibration model retains the complete-view stream and the same memory architecture, so its reference construction is unchanged. MemPrompt has no formally trained privileged posterior. A separate MemPrompt diagnostic would have to define an offline reference, for example by passing the held-out complete input through that checkpoint's frozen backbone and native query rule over its own keys. Such a reference is a diagnostic construction, not a teacher used by the original method. We do not treat the no-calibration PPC result as a measured MemPrompt routing result.

Agreement and collapse. Independent uniformly chosen k-subsets have expected Overlap@k equal to k/S (31.25% for $k = 5, S = 16$). A stronger control preserves each model's slot preferences while breaking sample correspondence: compare r_i with $\bar{r}_{\pi(i)}$ for a shuffled sample index π. Appendix G.2 reports this control alongside dense routing discrepancy. High shuffled overlap reflects shared access patterns and does not, by itself, rule out concentration on globally popular slots. The uniform expectation is an analytical reference, not a measured shuffled baseline.

F Decision-Level Analysis

This analysis concerns the supervision interface, not global convergence of the jointly trained model. The teacher, memory values, and remaining predictor are held fixed when comparing routing functions. A route is a probability vector in Δ^{S-1}, and logarithms are natural. Statements about task loss below are not statements about non-decomposable F1.

F.1 Conditional Projection of a Privileged Route

Let O be the observed evidence (including the missingness pattern), $Q \in \Delta^{S-1}$ the complete-view target, and $C \geq 0$ an integrable detached weight. Consider

$$J(p) = \mathbb{E}[C\,\mathrm{KL}(Q\|p(O))]\,. \tag{14}$$

Define $A(O) = \mathbb{E}[C \mid O]$ and $\mu_s(O) = \mathbb{E}[CQ_s \mid O]/A(O)$ wherever $A(O) > 0$. Because $\sum_s Q_s = 1$, $\mu(O)$ is a probability vector. Conditioning on O yields

$$\mathbb{E}[C\,\mathrm{KL}(Q\|p(O)) \mid O] = \mathbb{E}\left[C\sum_s Q_s \log Q_s \mid O\right] - A(O)\sum_s \mu_s(O)\log p_s(O) \tag{15}$$

$$= K(O) + A(O)\,\mathrm{KL}(\mu(O)\|p(O)), \tag{16}$$

where $K(O)$ is independent of p. Nonnegativity of KL proves Proposition 1. If $A(O) = 0$, calibration imposes no constraint there. Zero target components use the usual $0 \log 0 = 0$ convention.

With constant $C > 0$, $p^\star(O) = \mathbb{E}[Q \mid O]$. More generally, confidence reweights the conditional teacher distribution. For the finite-batch objective, C may absorb the detached batch normalizer as well as the incomplete-example mask. The simple formula with raw entropy confidence alone describes the corresponding population-weighted idealization, not an assertion that random minibatch

normalization has no effect. Restricted cosine routers, the simultaneous task loss, sparse retrieval, and an evolving teacher further distinguish actual optimization from this unrestricted minimizer.

Ambiguity is not an optimization failure. Suppose two equally likely complete views share the same observation and produce $Q = (1, 0)$ and $Q = (0, 1)$ with equal confidence. The projected route is $(1/2, 1/2)$. No deterministic observed-input router can reproduce both complete-view decisions. A top-1 rule must break the tie, illustrating why dense-target predictability and sparse deployment are separate considerations. Entropy-based confidence cannot remove this ambiguity: both individual teachers are confident.

F.2 When Route Transfer Preserves a Useful Prediction

Consider one routing block with fixed observed context o, label y, and prompt values $P_1, \dots, P_S$ satisfying $\|P_s\| \le M$. Define $G(r) = \sum_s r_s P_s$. Let F_o be the remaining predictor and assume the composed task loss is L-Lipschitz in the injected prompt:

$$|\ell(F_o(g), y) - \ell(F_o(g'), y)| \le L\|g - g'\|. \tag{17}$$

The comparison holds the observed context fixed; a teacher-routed intervention is not a complete-input classifier.

For a probability vector v, write $T_k(v)$ for its top-k renormalization and $\epsilon_k(v) = 1 - \sum_{s \in \mathrm{TopK}(v)} v_s$ for discarded probability mass. Ties follow any fixed rule. Since $\|T_k(v) - v\|_1 = 2\epsilon_k(v)$, the triangle inequality gives

$$\|G(T_k(p)) - G(T_k(q))\| \le M\|T_k(p) - T_k(q)\|_1 \tag{18}$$

$$\le M\big(\|p - q\|_1 + 2\epsilon_k(p) + 2\epsilon_k(q)\big). \tag{19}$$

Pinsker's inequality and Eq. (17) therefore yield

$$|\ell(F_o(G(T_k(p))), y) - \ell(F_o(G(T_k(q))), y)| \le LM\left[\sqrt{2\mathrm{KL}(q\|p)} + 2\epsilon_k(p) + 2\epsilon_k(q)\right]. \tag{20}$$

Under the stated fixed-block assumptions, this is a conditional stability bound. With constant bounds L, M, expectation and Jensen's inequality give

$$R(p) \le R(q) + LM\sqrt{2\mathbb{E}\mathrm{KL}(q\|p)} + 2LM\,\mathbb{E}[\epsilon_k(p) + \epsilon_k(q)]. \tag{21}$$

Here $R(q)$ is the risk of the non-deployable teacher-route intervention on the same observed predictor. If that intervention improves on an uncalibrated route by a margin larger than the error terms, this bound is sufficient for student improvement relative to that route. Table 4 specifies the corresponding empirical intervention. Confidence-weighted calibration controls a weighted discrepancy; applying this bound to that objective requires taking all expectations under the corresponding reweighted distribution.

A sharper bound away from selection ties. The general bound can be loose when much dense probability mass is discarded. A local bound removes this non-vanishing truncation term when the selected slots are separated. Let $q_{(j)}$ be the jth largest teacher probability, $m(q) = q_{(k)} - q_{(k+1)} > 0$, and $\eta = \sqrt{2\mathrm{KL}(q\|p)}$. If $\eta < m(q)/2$, then $\|p - q\|_\infty \le \|p - q\|_1 \le \eta$ ensures that the top-k support A is the same for both routes. Write $Z_p = \sum_{s \in A} p_s$ and $Z_q = \sum_{s \in A} q_s$. On this fixed support,

$$\left\| \frac{p_A}{Z_p} - \frac{q_A}{Z_q} \right\|_1 \le \frac{\|p_A - q_A\|_1 + |Z_q - Z_p|}{Z_p} \le \frac{2\eta}{\min(Z_p, Z_q)}.$$

Consequently,

$$|\ell(F_o(G(T_k(p))), y) - \ell(F_o(G(T_k(q))), y)| \le \frac{2LM\sqrt{2\mathrm{KL}(q\|p)}}{\min(Z_p, Z_q)}. \tag{22}$$

Unlike the general bound, this expression vanishes with the routing discrepancy. Its margin condition is sufficient, not a property claimed for every trained sample.

Dense distributions can be close while their top-k supports differ near a tie, and prompt values can encode different information despite equal slot overlap. Applying these arguments to a deep routed network additionally requires controlling each intervention's downstream sensitivity; the one-block results are not global optimization theorems. In PPC, task supervision and information-matched experiments supply the predictive criterion, while route diagnostics describe the associated access behavior.

Limits of the teacher assumption. A complete view can emphasize nuisance content, a confident route can be wrong, and many observations can be compatible with multiple complete views. If all teacher queries select one popular slot, agreement can be perfect without instance-specific retrieval. PPC's confidence weighting and semantic regularization are design responses, not proofs that these cases never occur. This is why prediction quality, information access, and routing behavior must be evaluated separately.

G Predictive and Routing Checks

These experiments complement aggregate benchmark scores with subset evaluation, route interventions, and controls for the interpretation of routing agreement. They distinguish the benefit of the full PPC system from the effect of calibration and from sample-specific agreement with a complete-view reference.

G.1 Subset Evaluation and Route-Only Interventions

Subset evaluation. Table 4 partitions the original image-missing test set into missing-image and complete subsets, without resampling missingness. F1-Macro is computed separately within each subset over all 23 MM-IMDb genres, using the same sigmoid threshold of 0.5. The aggregate score is computed over all test predictions, not as an average of subgroup F1 scores: F1 is nonlinear in the confusion counts. PPC's gain over MemPrompt is 7.22 points on missing-image examples and 0.87 points on complete examples. The 5.4-point aggregate gain is therefore accompanied by improvements on both subsets, with the larger improvement on incomplete inputs.

Fixed-predictor intervention. The intervention uses validation-F1-Macro-selected Full PPC checkpoints and holds model parameters, masked inputs, classifier, memory values, and shared prompts fixed. It replaces only visual-memory routing at calibrated layers 0–3 for missing-image examples; text-memory routing and other layers retain their deployed behavior. The four conditions are observed sparse routing r_i, the same sample's complete-view sparse routing $\bar{r}_i$, another sample's complete-view routing, and a uniformly sampled k-slot support with weights $1/k$, with $k = 5$ throughout. A complete-view pass supplies reference routing, not complete features to the prediction stream. Shuffling changes the sample supplying the reference, not the identities of memory slots.

The reported changes relative to observed routing are $+1.46 \pm 0.36$ points for paired complete-view routes, -0.82 ± 0.1 for shuffled routes, and -7.27 ± 0.2 for random sparse routes. These are reported summaries of within-model changes, not standard deviations obtained by subtracting the row-level uncertainties in Table 4. The paired route exceeds the shuffled route by 2.28 points in mean F1-Macro; this contrast is descriptive, without a paired-seed significance test. The loss from random routing shows that the selected memory content matters, while the loss after shuffling supports the predictive value of matching the complete-view route to the observed sample. This intervention assesses the utility of a reference route in a fixed predictor; it is not a full-input classifier or a universal performance ceiling.

G.2 Shuffled References and Dense Routing Discrepancy

Shuffling complete-view reference routes preserves their marginal slot popularity while breaking their correspondence with observed samples. It therefore provides a more informative agreement control than independent uniform slot selection. Table 14 reports shuffled overlap alongside the paired point estimates from Table 3. It also reports the dense discrepancy $N^{-1} \sum_i \mathrm{KL}(q_i \| p_i)$ before top-k selection, averaged over calibrated layers and missing-image examples. This unweighted discrepancy differs from both the confidence-weighted training objective and the clipped sparse Gap in Appendix E.

The dense discrepancy decreases from 0.26 to 0.15 with calibration, while shuffled overlap changes from 58.01 to 60.86. The paired anchors exceed the shuffled means by 5.01 and 25.24 points, respectively. This larger excess over shuffled agreement is consistent with stronger sample-aligned memory access and with the paired route's predictive advantage in Table 4. These are descriptive

Table 14: Visual-memory routing on missing-image MM-IMDb examples. Paired Overlap@5 repeats the point estimates in Table 3; shuffled overlap and dense KL are supplementary mean $\pm$ standard deviation summaries. The paired column is not a paired-seed mean and must not be used with the shuffled column to infer significance. Overlap is in percent.

Method	Paired Overlap@5	Shuffled Overlap@5	Dense KL
w/o calibration	63.02	58.01 ± 0.3	0.26 ± 0.03
Full PPC	86.10	60.86 ± 0.2	0.15 ± 0.03

contrasts between paired point estimates and shuffled means, not paired-seed effect estimates. Predictive utility is assessed by the classification and fixed-predictor comparisons, not by overlap alone; these overlap statistics do not rule out concentration on commonly selected slots.

G.3 Cross-Task Controls

Table 15 extends the privileged controls to text-missing Hateful Memes and Food101. The additional controls use the corresponding architecture, missingness construction, and validation metric, with complete-view training access. Existing MemPrompt and PPC benchmark anchors are retained as point estimates, while the additional controls are summarized by mean and standard deviation; these are descriptive comparisons, not paired-seed significance tests.

Table 15: Cross-task controls under 70% text missingness. MemPrompt and Full PPC are the benchmark point estimates from Table 1; privileged controls are supplementary mean $\pm$ standard deviation summaries. The distinct reporting forms are retained rather than assigning uncertainty to the original point estimates.

Method	Hateful Memes / AUROC	Food101 / Accuracy
MemPrompt†	62.85	74.26
Complete-view logit KD	67.77 ± 0.16	76.86 ± 0.25
w/o route calibration	67.95 ± 0.11	77.52 ± 0.13
Full PPC	68.93	78.68

The PPC anchors exceed the logit-KD means by 1.16 AUROC and 1.82 accuracy points, respectively. They also exceed the no-calibration means of 67.95 and 77.52 by 0.98 AUROC and 1.16 accuracy points. These descriptive comparisons favor Full PPC in both complementary settings, extending the positive direction of the MM-IMDb calibration comparison to other tasks and metrics without implying equal effect sizes or paired-seed significance. This distinction preserves the broader full-system comparison: PPC exceeds reproduced MemPrompt in all nine settings of Table 1, whereas the additional gain attributable to calibration is evaluated separately.