RETRIEVAL-AUGMENTED LABEL-ATTENTIVE PROMPT LEARNING FOR MULTIMODAL CLASSIFICATION WITH MISSING MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Missing modalities make multimodal classifiers brittle because the available evidence is partial, while common remedies either synthesize the missing modality or rely on static missing-aware prompts. We argue that missing-modality classification should instead be treated as retrieval-augmented evidence selection: a memory bank can provide useful neighbors, but the retrieved labels and embeddings are noisy and should not be trusted uniformly. We introduce Retrieval-Augmented label-attentive prompt Learning for Multimodal Classification with Missing Modalities (RALAP), a frozen-LLM framework that converts retrieved multimodal neighbors into dynamic soft prompts. RALAP first encodes the target self-embeddings and target–neighbor difference embeddings through a sparse autoencoder bottleneck to generate differential soft prompts. It then computes a label-attentive prompt by weighting retrieved instances with a similarity kernel and aggregating their labels into a target-conditioned label prior. The resulting self, difference, and label prompts are injected into a frozen Qwen2-VL model, which jointly generates a class label and a short explanation. We further show theoretically that, with sufficient retrieval coverage and stronger aggregated evidence for ground-truth labels than distractors, ground-truth labels receive higher attention scores. Empirically, the results support the need for target-conditioned reweighting and are consistent with the theoretical behavior of label attention. RALAP-VLM achieves the best reported performance in 10 of 12 evaluation settings across MM-IMDb, HateMemes, and Food101, with a maximum observed improvement of 21.84 percentage points on MM-IMDb. These results support the effectiveness of combining retrieval-enhanced sparse differential prompts with target-conditioned label-attentive prompts under modality missingness.

1 INTRODUCTION

Multimodal models typically assume that all modalities are available at inference time, yet this assumption often fails in practice (Woo et al., 2023; Lee et al., 2023). A movie classifier may receive only a poster or a plot description, a food recognition system only an image or textual description, and a safety classifier incomplete visual or textual evidence. Although recent vision-language models have substantially improved multimodal representation and transfer (Bao et al., 2022; Alayrac et al., 2022; Liu et al., 2024), their performance can still degrade sharply when an entire modality is missing (Wang et al., 2023a; Wei et al., 2023; Dai et al., 2025)(Fig. 1). Prompt-based adaptation is appealing in this setting because it can preserve the pretrained backbone while introducing only a small number of trainable parameters (Liu et al., 2022; Jang et al., 2024; Lee et al., 2023). However, missing-modality prompting faces three key challenges: **(i)** the observed modality may not contain sufficient evidence for prediction (Li et al., 2025); **(ii)** dummy inputs or explicit imputation can introduce distribution shift (Wang et al., 2023b; Dai et al., 2025); and **(iii)** static prompts cannot adapt to instance-specific uncertainty (Zhou et al., 2022a; Zhang et al., 2025).

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.

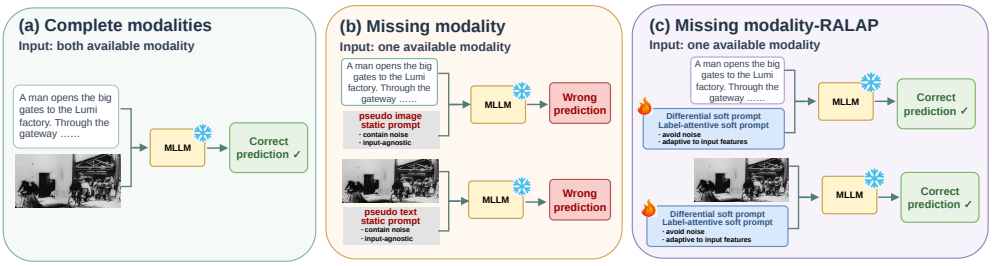


Figure 1: Motivation of RALAP for multimodal classification under modality missingness.

Retrieval offers a natural way to supplement incomplete observations without reconstructing the missing modality. Given an observed modality, a model can retrieve semantically related multimodal samples from an external memory and exploit their representations and labels as auxiliary evidence (Yasunaga et al., 2023; Lin et al., 2023; Pipoli et al., 2025). However, retrieval introduces a different challenge: *relevant neighbors are not necessarily reliable neighbors*. A retrieved sample may be globally similar to the target while containing task-irrelevant or conflicting labels. This issue is particularly pronounced in multi-label settings such as MM-IMDb (Arevalo et al., 2017), where a neighbor can share only a subset of the target genres. Therefore, simply exposing retrieved evidence to a frozen model can propagate retrieval noise rather than resolve missing-modality uncertainty. This raises a central question: *how can a frozen generative model exploit retrieved evidence while distinguishing useful information from noisy retrievals?*

Our key idea is to *select and condition* retrieved evidence rather than directly copy it or reconstruct the missing modality. We decompose retrieval information into two complementary signals. First, **instance-level differential evidence** describes both the target itself and how it differs from its retrieved neighbors. Second, **label-level reliability evidence** estimates which retrieved labels are relevant to the target. This decomposition addresses the three challenges above: retrieval supplements insufficient observations, avoiding explicit reconstruction reduces the distribution shift caused by dummy inputs, and target-conditioned evidence produces instance-adaptive prompts instead of static ones.

Based on this principle, we propose RALAP(Fig. 2), a retrieval-conditioned dynamic prompting framework for multimodal classification under modality missingness. Given a target with only text or only image, RALAP retrieves the top-K multimodal neighbors from an external memory(Lang et al., 2025). An SAE-based sparse bottleneck (Huben et al., 2024; Qiu et al., 2025) encodes the target representation and target–neighbor differences, which are mapped into *self* and *differential* soft prompts. In parallel, a target-conditioned label-attentive module measures target–neighbor relevance, aggregates it into label-level scores, and converts these scores into residual label prompts that emphasize reliable candidate labels. The available observation, retrieved label candidates, self/difference prompts, and label-attentive prompts are then jointly provided to a frozen Qwen2-VL model (Wang et al., 2024), which generates both the class prediction and a concise explanation. Thus, retrieval affects the model at three complementary levels: textual label candidates, instance-level dynamic prompts, and label-level dynamic prompts.

Our contributions are threefold:

- We formulate missing-modality classification as *retrieval-conditioned differential prompting*, using complete multimodal neighbors to provide complementary evidence without explicitly reconstructing the missing modality.
- We propose RALAP, which combines sparse self/differential prompts with target-conditioned label-attentive prompts. The former captures target-specific differential evidence, while the latter filters retrieved information at the label level, enabling instance-adaptive prompting for a frozen vision-language model.
- We theoretically establish target-conditioned label attention through retrieval coverage, attention dominance, and generation-margin enlargement. We further evaluate RALAP on MM-IMDb, HateMemes (Kiela et al., 2020), and Food101 (Wang et al., 2015) when only

one modality is available, with experiments examining the effectiveness and complementarity of differential prompting and label-aware evidence selection.

The code of our work is available at `https://anonymous.4open.science/r/RALAP-7F22`.

2 Related Work

Missing-modality multimodal learning. Existing methods mainly address modality missingness through reconstruction or representation stabilization. SimMLM improves robustness with dynamic modality experts and ranking objectives (Li et al., 2025), while ActionMAE reconstructs dropped modality tokens from the remaining inputs (Woo et al., 2023). Other approaches model shared and modality-specific factors, recover missing distributions, or regularize modality-aware representations (Wang et al., 2023a;b; Dai et al., 2025; Wei et al., 2023; Guan et al., 2026). Related multimodal language-model methods also rely on modality recovery or robust alignment (Pipoli et al., 2025; Zhang et al., 2025; Tsai et al., 2025). In contrast, RALAP avoids missing-modality reconstruction, keeps the generative backbone frozen, and instead converts retrieved multimodal evidence into target-conditioned soft prompts.

Prompting with missing modalities. Prompt learning provides a parameter-efficient alternative to full-model adaptation (Liu et al., 2022; Hu et al., 2022; Zhang et al., 2023). Conditional and multimodal prompting further adapts prompts to the input or both modality branches (Zhou et al., 2022a;b; Khattak et al., 2023). For missing modalities, MAPs introduces missing-aware prompts for frozen multimodal transformers (Lee et al., 2023), followed by methods that improve prompt robustness and adaptation (Lang et al., 2025; Guo & Gu, 2024; Shao et al., 2026; Li et al., 2023; Bang et al., 2024). However, these prompts are typically conditioned on missing patterns or learned task distributions rather than sample-specific retrieved evidence. RALAP instead performs *evidence-aware prompting*, dynamically adapting prompts to each target and its retrieval neighborhood.

Retrieval-augmented prompting and difference modeling. Retrieval-augmented methods introduce external evidence (Izacard et al., 2023), with multimodal extensions exploring retrieved demonstrations, late interaction, and context compression (Yasunaga et al., 2023; Lin et al., 2023; Cheng et al., 2024). Recent missing-modality approaches further combine retrieval with multimodal prompting or LLM reasoning (Ma et al., 2025; Zhao et al., 2025; Pipoli et al., 2025), but retrieved neighbors may contain noisy or irrelevant labels. DEP shows that latent difference modeling can generate compact personalized prompts (Qiu et al., 2025). Building on this idea, RALAP models target–neighbor differences and uses target-conditioned label attention to selectively reweight retrieved labels, mitigating retrieval noise in missing-modality classification.

3 Method

3.1 Problem Setup

We study retrieval-augmented classification under modality missingness with a frozen generative language model (Fig. 2). Each target sample x originally contains text and image modalities, while only one modality $m \in \text{text}, \text{image}$ is available at inference, represented by $x_q^m \in \mathbb{R}^D$. Text-only and image-only correspond to image-missing and text-missing cases, respectively, while both-modality serves as a complete-modality reference. We construct a multimodal memory bank $\mathcal{M} = \{(r_i, y_i)\}_{i=1}^N$, where $r_i \in \mathbb{R}^D$ is a fused text–image representation and $y_i \in \{0,1\}^C$ is a label indicator over $\mathcal{V}_{\text{label}} = \{c_1, \ldots, c_C\}$.

Given x_q^m, the retriever returns

$$\mathcal{R}_K(x_q^m) = \{(r_i, y_i, \rho_i)\}_{i=1}^K \tag{1}$$

where ρ_i denotes query–neighbor similarity. The frozen LLM generates $\hat{o} =$ `Label:` $\hat{\ell} \parallel$ `Explanation:` $\hat{e}$. The retrieved labels induce

$$\mathcal{C}_K = \bigcup_{i=1}^{K} \{c_j \mid y_{i,j} = 1\}. \tag{2}$$

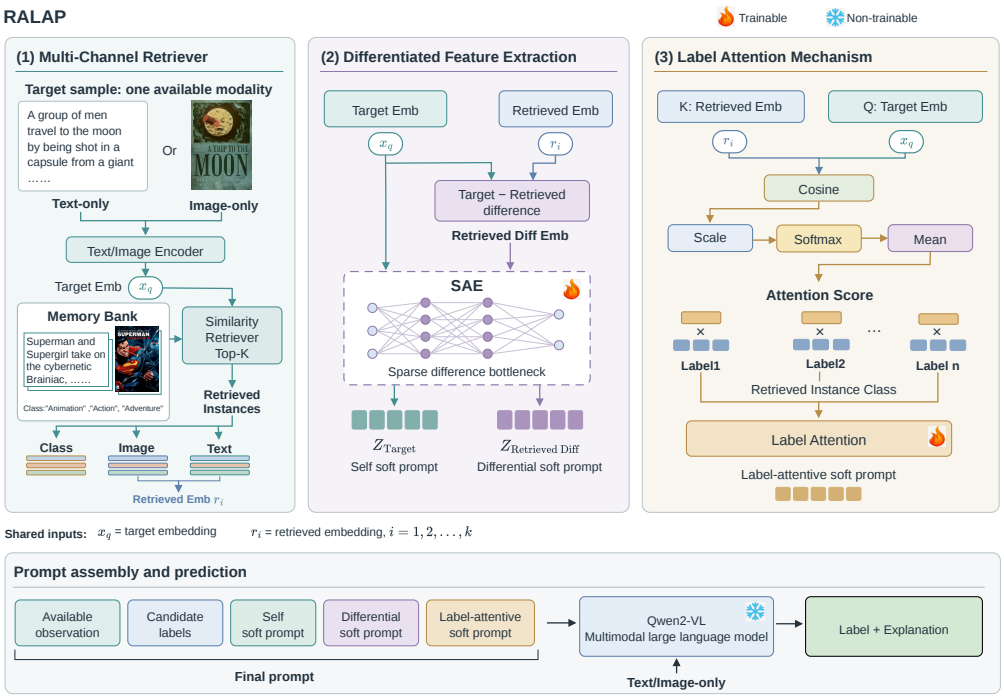


Figure 2: Overview of RALAP . Given a target sample with one available modality, the framework retrieves complete multimodal neighbors, extracts sparse self and target-neighbor differential prompts, and constructs target-conditioned label-attentive prompts. These dynamic prompts guide a frozen Qwen2-VL model to generate the classification labels and a concise explanation.

which should cover the target labels but may also introduce distractors $\mathcal{C}_K^- = \mathcal{C}_K \setminus \{c_j \mid y_j^* = 1\}$ as K increases. This creates a coverage–noise trade-off. To selectively exploit retrieved evidence, RALAP combines sparse target–neighbor differential prompting with target-conditioned label attention. With frozen parameters θ_{LLM} and trainable prompt parameters ϕ, the model learns

$$p\theta_{\text{LLM}}\big(\hat{o} \mid x, \mathcal{R}_K(x_q^m), P_{\text{diff}}(x, \mathcal{R}_K; \phi), P_{\text{label}}(x, \mathcal{R}_K; \phi)\big), \tag{3}$$

where only ϕ is optimized.

3.2 Sparse Difference Soft Prompts

For each target, RALAP combines the available-modality embedding x_q with the top-K retrieved multimodal embeddings. Let r_i denote the fused embedding of the i-th neighbor and w_i its normalized retrieval-similarity weight, with $\tilde{r}_i = w_i r_i$. We construct the self–difference sequence

$$H = \big[\underbrace{x_q, \ldots, x_q}_{K \text{ times}}; x_q - \tilde{r}_1, \ldots, x_q - \tilde{r}_K\big] \in \mathbb{R}^{2K \times D}. \tag{4}$$

The first half preserves target-specific evidence, while the second explicitly captures target–neighbor differences.

An SAE (Qiu et al., 2025) compresses this sequence as

$$Z = f_{\text{enc}}(H), \qquad \hat{H} = f_{\text{dec}}(Z), \tag{5}$$

where $Z \in \mathbb{R}^{2K \times d_s}$ and $\hat{H} \in \mathbb{R}^{2K \times D}$. Its objective combines reconstruction and sparsity:

$$\mathcal{L}_{\text{SAE}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{sparse}} \mathcal{L}_{\text{sp}}, \tag{6}$$

where $\mathcal{L}_{\text{rec}}$ is the element-wise mean Smooth L1 reconstruction loss(Huben et al., 2024), and $\mathcal{L}_{\text{sp}}$ is the mean absolute latent activation. This sparse bottleneck preserves discriminative target–retrieval information while suppressing diffuse retrieval noise.

The two halves of Z are projected separately into self and differential prompts:

$$P_{\text{self}} = \text{MLP}_{\text{self}}(Z_{1:K}), \qquad P_{\text{diff}} = \text{MLP}_{\text{diff}}(Z_{K+1:2K}). \tag{7}$$

These vectors replace the corresponding special-token embeddings, allowing retrieval-conditioned evidence to adapt generation without updating the backbone.

3.3 Theory of Label Attention

We characterize how retrieval and target-conditioned label attention jointly affect prediction through three stages: retrieval coverage, attention dominance, and generation-margin enlargement.

Definition (Effective Labels). For a target x with ground-truth labels $\mathcal{Y}^*$ and top-K retrieved neighbors $\mathcal{R}_K$, $M_{i,c} \in \{0, 1\}$ indicate whether the i-th retrieved neighbor contains label c.

The retrieval-induced candidate set is defined as

$$\mathcal{C}_K = \left\{ c \,\middle|\, \sum_{i=1}^{K} M_{i,c} > 0 \right\}. \tag{8}$$

A ground-truth label $y \in \mathcal{Y}^*$ is *retrieval-effective* if

$$y \in \mathcal{C}_K, \tag{9}$$

and *attention-effective* if

$$s_y > \max_{c \in \mathcal{C}_K \setminus \mathcal{Y}^*} s_c. \tag{10}$$

Let ℓ_c^{base} and ℓ_c^{attn} denote the generation logits before and after label-attentive prompting. Label y is *generation-effective* if

$$\ell_y^{\text{attn}} - \max_{c \notin \mathcal{Y}^*} \ell_c^{\text{attn}} > \ell_y^{\text{base}} - \max_{c \notin \mathcal{Y}^*} \ell_c^{\text{base}}. \tag{11}$$

These notions respectively capture label availability, label-level discrimination, and improvement in the generation space.

Theorem (Retrieval Coverage and Attention Dominance). Assume that each retrieved instance contains a ground-truth label $y \in \mathcal{Y}^*$ with probability at least $p_y > 0$, and let

$$p_{\min} = \min_{y \in \mathcal{Y}^*} p_y. \tag{12}$$

Under conditional independence given the target,

$$\Pr\left(\mathcal{Y}^* \nsubseteq \mathcal{C}_K\right) \leq |\mathcal{Y}^*|(1 - p_{\min})^K. \tag{13}$$

Hence, for $\delta \in (0, 1)$,

$$K_{\min}(\delta) := \left\lceil \frac{\log(|\mathcal{Y}^*|/\delta)}{-\log(1 - p_{\min})} \right\rceil. \tag{14}$$

guarantees retrieval effectiveness for all ground-truth labels with probability at least $1 - \delta$ whenever $K \geq K_{\min}(\delta)$.

Let

$$z_i = \cos(q, k_i), \qquad \mathcal{I}_c(K) = \{i \leq K \mid M_{i,c} = 1\}, \qquad n_c(K) = |\mathcal{I}_c(K)| \tag{15}$$

Define the aggregated semantic evidence for label c as

$$G_c(K) = \tau \log \left[\frac{\sum_{i \in \mathcal{I}_c(K)} \exp(z_i/\tau)}{n_c(K) + \epsilon} \right], \tag{16}$$

The corresponding attention score satisfies

$$s_c = \frac{\exp(G_c(K)/\tau)}{Z_K}, \qquad Z_K = \sum_{j=1}^{K} \exp(z_j/\tau), \tag{17}$$

Suppose the retrieved neighborhood preserves a positive evidence margin

$$\Gamma_K := \min_{y \in \mathcal{Y}^*} G_y(K) - \max_{c \in \mathcal{C}_K \setminus \mathcal{Y}^*} G_c(K) \geq \Delta, \qquad \Delta > 0, \tag{18}$$

and define

$$K_{\max}(\Delta) := \max\{K \mid \Gamma_K \geq \Delta\}. \tag{19}$$

For

$$K_{\min}(\delta) \leq K \leq K_{\max}(\Delta), \tag{20}$$

all ground-truth labels are covered with probability at least $1-\delta$ and, conditioned on coverage,

$$\frac{s_y}{s_c} \geq \exp(\Delta/\tau) > 1 \tag{21}$$

for any $y \in \mathcal{Y}^*$ and $c \in \mathcal{C}_K \setminus \mathcal{Y}^*$. Thus, every covered ground-truth label is attention-effective.

Proof. The detailed derivation is provided in Appendix B.3.

Takeaway. The theorem identifies a retrieval-size trade-off. If $K < K_{\min}(\delta)$, relevant labels may be absent; if $K > K_{\max}(\Delta)$, additional noisy neighbors may weaken the evidence margin. The desirable regime is therefore

$$K \in [K_{\min}(\delta), K_{\max}(\Delta)], \tag{22}$$

where retrieval provides sufficient label coverage while target-conditioned attention preserves separation between relevant labels and distractors.

Corollary (Generation-Margin Enlargement). Assume that the label-attentive contribution to the generation logit is locally monotonic in its attention score:

$$\ell_c^{\text{attn}} = \ell_c^{\text{base}} + \beta s_c + \xi_c, \qquad \beta > 0, \tag{23}$$

where ξ_c denotes the residual approximation error associated with label c. For a ground-truth label y and a distractor label c, applying the above approximation to both labels gives

$$\Delta m_{y,c} = \beta(s_y - s_c) + (\xi_y - \xi_c). \tag{24}$$

If

$$|\xi_y - \xi_c| < \beta(s_y - s_c), \tag{25}$$

then

$$\ell_y^{\text{attn}} - \ell_c^{\text{attn}} > \ell_y^{\text{base}} - \ell_c^{\text{base}}. \tag{26}$$

Hence, attention dominance increases the ground-truth–distractor generation margin under bounded residual error.

Proof. The detailed derivation is provided in Appendix B.4.

Takeaway. Under the assumptions above and within the valid retrieval interval, label effectiveness follows the successive mechanism

$$\text{retrieval-effective} \;\Rightarrow\; \text{attention-effective} \;\Rightarrow\; \text{generation-effective}. \tag{27}$$

Retrieval first makes relevant labels available, target-conditioned attention separates them from distractors, and label prompting transfers this advantage to the generation space.

3.4 Frozen LLM Generation Objective

The frozen LLM jointly generates the label and explanation as $\hat{o} = \texttt{Label:}\ \hat{\ell} \parallel \texttt{Explanation:}\ \hat{e}$. Only target output tokens are supervised, while prompt and padding positions are masked. We separately optimize the label and explanation spans:

$$\mathcal{L}_{\text{LLM}} = \alpha\mathcal{L}_{\text{label}} + (1-\alpha)\mathcal{L}_{\text{exp}}. \tag{28}$$

Here, $\mathcal{L}_{\text{label}}$ and $\mathcal{L}_{\text{exp}}$ are the span-normalized autoregressive cross-entropy losses for the label and explanation tokens, respectively.

The complete training objective combines generation, SAE regularization, and label-attentive supervision:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{label}} + (1-\alpha)\mathcal{L}_{\text{exp}} + \beta\left(\mathcal{L}_{\text{rec}} + \lambda_{\text{sparse}}\mathcal{L}_{\text{sp}}\right) + \gamma\mathcal{L}_{\text{attn}}. \tag{29}$$

We use $\alpha = 0.8$, $\beta = 0.1$, $\lambda_{\text{sparse}} = 0.001$, and $\gamma = 0.05$. Only the SAE, alignment MLPs, and label-attentive module are trainable; gradients from $\mathcal{L}_{\text{LLM}}$ propagate through the frozen Qwen2-VL(Wang et al., 2024) backbone to these prompt modules.

Table 1: Comparison of different methods on MM-IMDb, HateMemes, and Food101.

	MM-IMDb						HateMemes			Food101		
Input Type	Text		Image		Both		Text	Image	Both	Text	Image	Both
Method	F1-M	F1-S	F1-M	F1-S	F1-M	F1-S	AUROC	AUROC	AUROC	ACC	ACC	ACC
SMIL	38.32	38.55	27.57	35.27	35.12	31.87	50.32	58.50	54.63	51.83	49.86	46.77
TFR-Net	37.70	38.82	38.14	39.45	37.24	38.11	51.18	55.57	52.12	65.91	67.58	63.41
AcMAE	47.47	46.73	43.82	42.20	44.05	43.75	55.74	59.66	57.25	69.28	73.75	71.15
IF-MMIN	39.63	38.10	31.95	26.89	31.98	29.33	57.62	53.44	55.19	66.76	64.36	68.53
ShaSpec	44.04	42.05	44.23	42.53	44.06	42.13	58.75	60.30	60.96	60.99	74.87	70.02
DrFuse	47.05	45.22	43.58	42.19	48.83	47.15	57.60	60.66	55.84	66.30	75.09	68.23
CorrKD	44.82	45.27	39.48	39.11	41.20	40.51	58.74	55.59	57.91	61.37	66.83	62.87
MAPs	46.12	45.47	44.86	43.19	45.48	44.30	58.62	60.16	58.89	67.02	75.62	72.52
MSPs	49.16	48.81	44.62	43.06	48.28	46.71	59.60	60.05	59.08	71.74	79.09	74.46
RAGPT	55.16	55.00	46.44	45.12	50.89	50.22	**64.10**	62.57	63.47	75.53	**81.98**	76.94
RALAP	**60.78**	**61.69**	**65.64**	**66.96**	**63.19**	**64.33**	62.47	**67.32**	**64.89**	**82.18**	79.30	**80.74**
Δ (pp)	+5.62	+6.69	+19.20	+21.84	+12.30	+14.11	−1.63	+4.75	+1.42	+6.65	−2.68	+3.80

Table 2: Ablation study of DSP variants in RALAP.

Module	MM-IMDb						HateMemes			Food101		
	Text		Image		Both		Text	Image	Both	Text	Image	Both
	F1-M	F1-S	F1-M	F1-S	F1-M	F1-S	AUROC	AUROC	AUROC	ACC	ACC	ACC
without DSP	46.93	49.06	53.79	58.05	51.35	53.56	57.88	63.47	60.67	74.54	64.10	69.32
SAE-based DSP	50.97	50.11	55.71	54.79	53.36	52.45	60.35	63.19	62.27	79.00	72.53	75.76
Label-based DSP	44.52	46.83	58.00	59.61	51.33	53.22	59.15	64.05	61.60	74.58	68.83	71.71
Both DSP	60.78	61.69	65.64	66.96	63.19	64.33	61.15	66.89	64.02	82.18	79.30	80.74

4 EXPERIMENTS

4.1 OVERALL PERFORMANCE

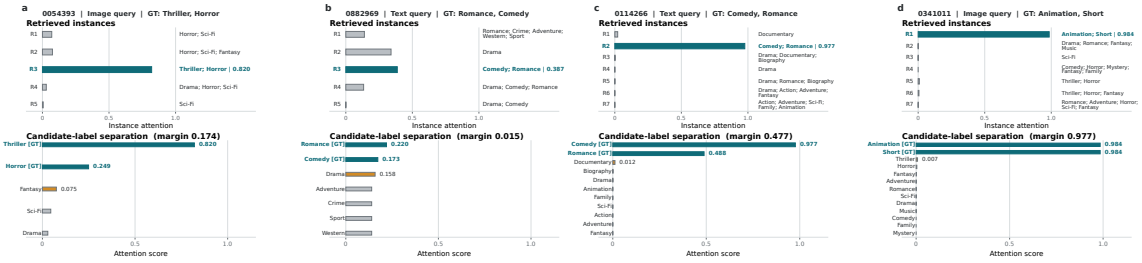
Table 1 compares RALAP with ten competitive baselines (Lang et al., 2025) on MM-IMDb(Arevalo et al., 2017), HateMemes(Kiela et al., 2020) , and Food101(Wang et al., 2015). RALAP achieves the best performance in 10 of the 12 reported metric–input combinations, demonstrating strong robustness to modality missingness. RALAP is evaluated at a 100% missing rate, while the reported baseline results use a 70% missing rate. Therefore, the absolute performance differences should be interpreted with this difference in evaluation setting in mind.

The gains are most pronounced on MM-IMDb. Compared with the strongest baseline, RAGPT(Lang et al., 2025), RALAP improves F1-M/F1-S from 55.16/55.00 to 60.78/61.69 under Text, from 46.44/45.12 to 65.64/66.96 under Image, and from 50.89/50.22 to 63.19/64.33 under Both. On HateMemes, RALAP performs best under Image and Both, while RAGPT remains stronger under Text. On Food101, RALAP leads under Text and Both but not Image. These results suggest that retrieval-conditioned prompting is particularly effective when missing modalities increase ambiguity, with gains depending on how well the available modality retrieves target-relevant evidence.

4.2 ABLATION STUDY

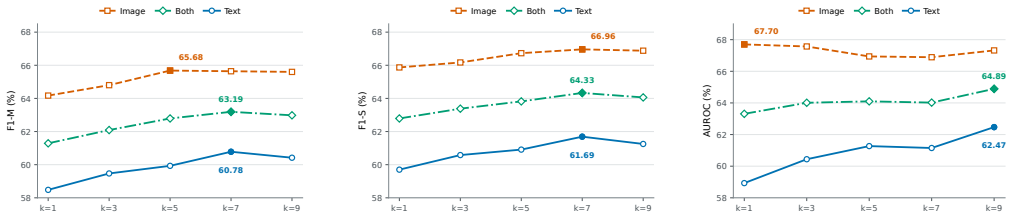
We conduct ablation experiments to assess the two dynamic soft-prompting components in RALAP: the SAE-based differential soft prompt (*SAE-based DSP*) and the label-attentive soft prompt (*Label-based DSP*). All variants are evaluated with a 100% missing rate and retrieval size $K = 7$, with results summarized in Table 2.

SAE-based DSP generally improves performance by encoding target–neighbor differences, with particularly strong gains on Food101 and several MM-IMDb and HateMemes settings, although it is not uniformly beneficial across all metrics. Label-based DSP provides more stable but modality-dependent improvements by reweighting retrieved labels according to target relevance, especially for image-based inputs. Importantly, combining both components consistently yields the strongest re-



(a) Top-$k = 5$. *(b) Top-$k = 7$.*

Figure 3: Qualitative label-attention analysis on MM-IMDb under image-only and text-only queries with different values of K.



(a) MM-IMDb: F1-M. *(b) MM-IMDb: F1-S.* *(c) HateMemes: AUROC.*

Figure 4: Sensitivity analysis of K on MM-IMDb and HateMemes.

sults across the three datasets and input conditions. Relative to *without DSP*, the full model improves MM-IMDb by up to 13.85 F1-M and 12.63 F1-S points, HateMemes by up to 3.42 AUROC points, and Food101 by up to 15.20 accuracy points. These results support the complementary roles of the two modules: SAE-based DSP captures instance-level differential evidence, while Label-based DSP filters retrieval information at the label level.

4.3 Label Attention Qualitative Behavior

Fig. 3(a) and Fig. 3(b) illustrate representative MM-IMDb cases for $K = 5$ and $K = 7$ under both image and text queries. Across both retrieval sizes, target-conditioned attention assigns the largest weights to neighbors whose labels best match the ground truth, while the aggregated label scores preserve a positive true–distractor margin. At $K = 5$, the exact-match neighbors receive the highest attention weights of 0.820 and 0.387, but competing labels reduce the margins to 0.174 and 0.015, respectively. At $K = 7$, attention becomes more concentrated on the most relevant neighbors, with dominant weights of 0.977 and 0.984 and corresponding margins of 0.477 and 0.977. These cases indicate that label-attentive prompts adaptively reweights instance-level evidence rather than uniformly propagating retrieved labels, allowing ground-truth labels to remain dominant even as the retrieval neighborhood becomes larger and more heterogeneous.

4.4 Hyper-Parameter Analysis

Fig. 4(a), Fig. 4(b) and Fig. 4(c) analyze the sensitivity of RALAP to the number of retrieved instances $K \in \{1, 3, 5, 7, 9\}$ on MM-IMDb and HateMemes, respectively. The two datasets exhibit different trends due to their label structures: MM-IMDb is a multi-label multi-class task, whereas HateMemes is a single-label binary task.

On MM-IMDb, performance improves consistently as K increases from 1 to 7 across all three input settings. Under Text, F1-M/F1-S increase from 58.48/59.70 to 60.78/61.69; under Image, they rise from 64.17/65.87 to 65.64/66.96; and under Both, they improve from 61.29/62.79 to 63.19/64.33. Thus, $K = 7$ achieves the best result on all six reported metrics. Further increasing K to 9 causes a slight decline across all metrics, indicating that the benefit of broader retrieval eventually saturates

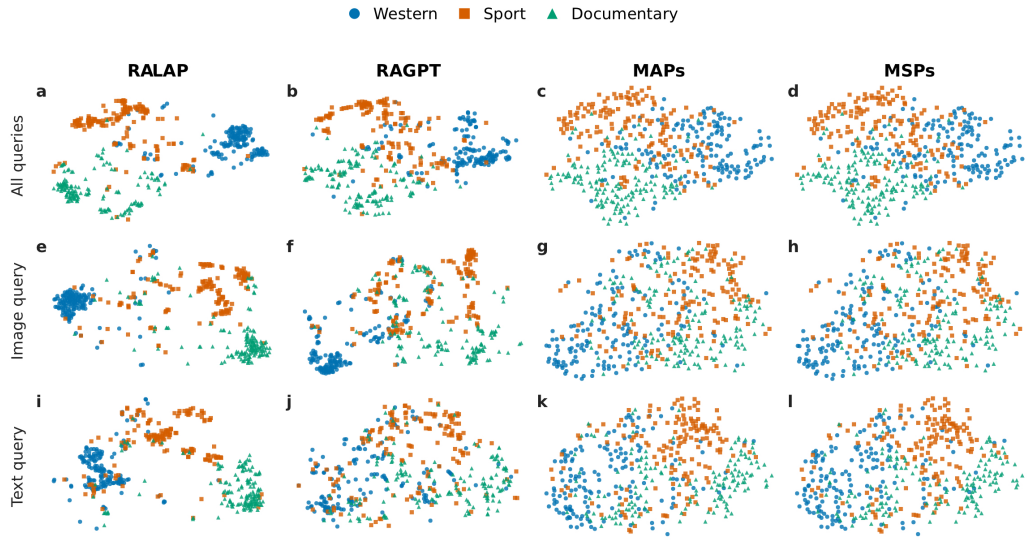


Figure 5: t-SNE comparison of different methods across query modalities on MM-IMDb.

as additional neighbors introduce more noisy or ambiguous evidence. This trend is consistent with the coverage–noise trade-off discussed in Sec. 3.3.

HateMemes is less sensitive to K. Because its binary label space does not expand with additional retrieved instances, larger neighborhoods mainly provide repeated evidence for the same two classes. Accordingly, Text and Both AUROC improve gradually with K, while Image performance remains relatively stable. These results indicate that the optimal retrieval size depends on both label-space structure and the informativeness of the available modality.

4.5 Model Prediction Visualization

Fig. 5 presents the t-SNE visualization of prediction representations for three MM-IMDb genres, i.e., Western, Sport, and Documentary, comparing RALAP with RAGPT(Lang et al., 2025), MAPs(Lee et al., 2023), and MSPs(Jang et al., 2024) under All, Image, and Text query settings. Across the three settings, MAPs and MSPs exhibit relatively diffuse distributions with substantial overlap among different labels, while RAGPT achieves partial class separation but still retains noticeable inter-class mixing. In contrast, RALAP forms more compact and clearly separated class regions, particularly for Western and Documentary, and this pattern remains visible for both image and text based queries.

5 Conclusion

We introduced RALAP, a retrieval-conditioned dynamic prompting framework for multimodal classification under modality missingness. RALAP combines sparse target–neighbor differential prompts with target-conditioned label-attentive prompts to selectively exploit retrieved evidence using a frozen Qwen2-VL backbone. Across MM-IMDb, HateMemes, and Food101, RALAP achieves the best reported performance in 10 of 12 evaluation settings, with the largest gain reaching 21.84 points on MM-IMDb. Its advantage is especially pronounced on the multi-label MM-IMDb task, while competitive gains are also observed on HateMemes and Food101. Ablation results further confirm that sparse differential prompts and label-attentive prompts are complementary, showing the effectiveness of retrieval-conditioned evidence selection.

REPRODUCIBILITY STATEMENT

The main paper specifies the problem formulation, retrieval-conditioned sparse differential prompting, target-conditioned label-attentive prompting, frozen-LLM generation objective, datasets, evaluation protocol, and principal results and ablations. RALAP uses the pre-trained Qwen2-VL-7B-Instruct model as a frozen multimodal generative backbone. We evaluate retrieval sizes $K \in \{1, 3, 5, 7, 9\}$ and use the retrieved instances to construct the self and differential prompts together with the target-conditioned label-attentive prompt. Only the RALAP-specific modules are optimized, including the sparse autoencoder, prompt-alignment networks, and label-attentive module, while all Qwen2-VL backbone parameters remain frozen. Training uses AdamW with a learning rate of 1×10^{-5} for five epochs and DeepSpeed ZeRO Stage 2. All experiments are conducted on two NVIDIA RTX 4090 GPUs.

The appendix provides further implementation details on prompt construction and trainable/frozen modules, complete proofs of the retrieval-coverage and generation-margin results, and additional experiments on retrieval size, generative backbones, explanation quality, dynamic-prompt ablations, and modality missingness. The supplementary material will include an anonymous source-code release, environment and configuration files, data-preprocessing and memory-bank construction scripts, and evaluation scripts for reproducing the reported tables and figures. Together, these materials document the model, retrieval, training, inference, and evaluation procedures required to reproduce the experiments.

AI USE STATEMENT

We used generative AI tools to assist with the implementation and debugging of selected components, to translate parts of the manuscript, to edit the paper for improved readability, and to create or modify software code. We did not use generative AI to generate synthetic datasets, formulate research hypotheses, design or critique the methodology or experiments, perform data cleaning or reformatting, conduct mathematical derivations or proofs, carry out qualitative or thematic data analysis, or interpret experimental results. All AI-assisted code was validated through experiments. The authors reviewed all AI-assisted text and outputs and take full responsibility for the content of the final manuscript.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eldon Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Shreya Samangooei, Mario Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

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

Jihwan Bang, Sumyeong Ahn, and Jae-Gil Lee. Active prompt learning in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27004–27014, 2024.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Xin Cheng, Xun Wang, Xingxing Zhang, Tao Ge, Si-Qing Chen, Furu Wei, Huishuai Zhang, and Dongyan Zhao. xRAG: Extreme context compression for retrieval-augmented generation with one token. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Ruiting Dai, Chenxi Li, Yandong Yan, Lisi Mo, Ke Qin, and Tao He. Unbiased missing-modality multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24507–24517, 2025.

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.

Yuncheng Guo and Xiaodong Gu. JoAPR: Cleaning the lens of prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28695–28705, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Robert Huben, Hoagy Cunningham, Logan Riggs, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *Proceedings of the 12th International Conference on Learning Representations*, ICLR 2024, May 2024.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251): 1–43, 2023.

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

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023.

Douwe Kiela, 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.

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

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14943–14952, June 2023.

Juncheng Li, Minghe Gao, Longhui Wei, Siliang Tang, Wenqiao Zhang, Mengze Li, Wei Ji, Qi Tian, Tat-Seng Chua, and Yueting Zhuang. Gradient-regulated meta-prompt learning for generalizable vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2551–2562, 2023.

Sijie Li, Chen Chen, and Jungong Han. SimMLM: A simple framework for multi-modal learning with missing modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24068–24077, 2025.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 61–68, 2022.

Hongchi Ma, Guanglei Yang, Debin Zhao, Yanli Ji, and Wangmeng Zuo. ReMP-AD: Retrieval-enhanced multi-modal prompt fusion for few-shot industrial visual anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20425–20434, 2025.

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

Yilun Qiu, Tianhao Shi, Xiaoyan Zhao, Fengbin Zhu, Yang Zhang, and Fuli Feng. Latent inter-user difference modeling for LLM personalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10599–10617, 2025.

Maoyuan Shao, Yutong Gao, Xinyang Huang, Lijuan Sun, Guoshun Nan, and Chuang Zhu. CAPT: Confusion-aware prompt tuning for reducing vision-language misalignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3154–3164, 2026.

Yun-Da Tsai, Ting-Yu Yen, Keng-Te Liao, and Shou-De Lin. Enhance modality robustness in text-centric multimodal alignment with adversarial prompting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27740–27747, 2025.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frédéric Precioso. Recipe recognition with large multimodal food dataset. In *IEEE International Conference on Multimedia & Expo Workshops (ICME)*, pp. 1–6, 2015. doi: 10.1109/ICMEW.2015.7169757.

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. 22025–22034, 2023b.

Shicai Wei, Chunbo Luo, and Yang Luo. MMANet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20039–20049, 2023.

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*, 2023.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. Retrieval-augmented multimodal language modeling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 39755–39769, 2023.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023.

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, October 2025.

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 Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 2458–2466, 2025. doi: 10.24963/ijcai.2025/274.

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

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

A Implementation Details

The implementation uses a Qwen2-VL-compatible `RALAP Model`. The trainable modules are the sparse autoencoder, the self/difference alignment MLPs, and the label-attentive prompt module. The non-RALAP parameters of the LLM are frozen. The prompt template inserts `[HIS_TOKEN_i]`, `[DIFF_TOKEN_i]`, and `[LAB_ATTN_i]` special tokens, whose embeddings are replaced or residually adjusted at the input embedding layer.

B Additional Method and Theoretical Details

This appendix provides the task-output formulation, complete SAE objectives, generation-supervision details, and theoretical derivations omitted from the main text for brevity.

B.1 Task Output Formulation

The formulation supports binary, multi-class single-label, and multi-class multi-label classification. For binary and multi-class single-label tasks, $y^* \in \{0,1\}^C$ satisfies

$$\sum_{c=1}^{C} y_c^* = 1, \tag{30}$$

and the target label string is

$$\ell(y^*) = c_j, \qquad j = \arg\max_c y_c^*. \tag{31}$$

For multi-class multi-label classification, $y^* \in \{0,1\}^C$ satisfies

$$\sum_{c=1}^{C} y_c^* \geq 1, \tag{32}$$

with

$$\ell(y^*) = \text{Join}\left(c_j \mid y_j^* = 1\right). \tag{33}$$

The generated labels are evaluated as an unordered set. In all cases, the frozen LLM produces

$$\hat{o} = \texttt{Label:} \quad \hat{\ell} \,\|\, \texttt{Explanation:} \quad \hat{e}. \tag{34}$$

B.2 Sparse Autoencoder Objectives

As defined in the main text, the SAE encodes the self–difference sequence $H \in \mathbb{R}^{2K\times D}$ into $Z \in \mathbb{R}^{2K\times d_s}$ and reconstructs it as $\hat{H} \in \mathbb{R}^{2K\times D}$. The reconstruction loss is the element-wise mean Smooth L1 error(Huben et al., 2024):

$$\mathcal{L}_{\text{rec}} = \frac{1}{2KD} \sum_{i=1}^{2K} \sum_{j=1}^{D} \text{SmoothL1}\Big(H_{ij}, \hat{H}_{ij}\Big), \tag{35}$$

Sparsity is imposed through the mean absolute latent activation:

$$\mathcal{L}_{\text{sp}} = \frac{1}{2Kd_s} \sum_{i=1}^{2K} \sum_{j=1}^{d_s} |Z_{ij}|. \tag{36}$$

The complete SAE(Qiu et al., 2025) objective is

$$\mathcal{L}_{\text{SAE}} = \mathcal{L}_{\text{rec}} + \lambda_{\text{sparse}} \mathcal{L}_{\text{sp}}. \tag{37}$$

Reconstructing all $2K$ self–difference vectors encourages the sparse bottleneck to retain information required to distinguish the target from its retrieved neighborhood, while the sparsity term suppresses diffuse latent activations.

B.3 Proof of Retrieval Coverage and Attention Dominance

For each ground-truth label $y \in \mathcal{Y}^*$, assume that an individual retrieved instance contains y with probability at least $p_y > 0$. Under conditional independence given the target,

$$\Pr(y \notin \mathcal{C}_K) \leq (1 - p_y)^K. \tag{38}$$

Applying the union bound over all ground-truth labels gives

$$\Pr\left(\exists y \in \mathcal{Y}^* : y \notin \mathcal{C}_K\right) \leq \sum_{y \in \mathcal{Y}^*} (1 - p_y)^K \leq |\mathcal{Y}^*|(1 - p_{\min})^K. \tag{39}$$

where

$$p_{\min} = \min_{y \in \mathcal{Y}^*} p_y. \tag{40}$$

Therefore,

$$K \geq K_{\min}(\delta) := \left\lceil \frac{\log(|\mathcal{Y}^*|/\delta)}{-\log(1 - p_{\min})} \right\rceil \tag{41}$$

ensures

$$\Pr\left(\mathcal{Y}^* \subseteq \mathcal{C}_K\right) \geq 1 - \delta. \tag{42}$$

For attention dominance,

$$s_c = \frac{\exp(G_c(K)/\tau)}{Z_K}. \tag{43}$$

If the retrieved neighborhood satisfies

$$\Gamma_K = \min_{y \in \mathcal{Y}^*} G_y(K) - \max_{c \in \mathcal{C}_K \setminus \mathcal{Y}^*} G_c(K) \geq \Delta, \qquad \Delta > 0, \tag{44}$$

then, for any ground-truth label y and retrieved distractor c,

$$\frac{s_y}{s_c} = \exp\left(\frac{G_y(K) - G_c(K)}{\tau}\right) \geq \exp(\Delta/\tau) > 1. \tag{45}$$

Hence $s_y > s_c$. Together with the coverage bound, retrieval effectiveness and attention effectiveness hold for

$$K_{\min}(\delta) \leq K \leq K_{\max}(\Delta). \tag{46}$$

The lower bound ensures sufficient ground-truth coverage, whereas the upper bound prevents additional retrieval noise from violating the ground-truth–distractor evidence margin.

B.4 Proof of Generation-Margin Enlargement

Assume that the label-attentive contribution to the generation logit is locally approximated by

$$\ell_c^{\text{attn}} = \ell_c^{\text{base}} + \beta s_c + \xi_c, \qquad \beta > 0, \tag{47}$$

where ξ_c denotes the residual approximation error associated with label c. For a ground-truth label y and a distractor label c, applying the above approximation to both labels gives

$$\Delta m_{y,c} = \beta(s_y - s_c) + (\xi_y - \xi_c). \tag{48}$$

If

$$|\xi_y - \xi_c| < \beta(s_y - s_c), \tag{49}$$

then $\Delta m_{y,c} > 0$, yielding

$$\ell_y^{\text{attn}} - \ell_c^{\text{attn}} > \ell_y^{\text{base}} - \ell_c^{\text{base}}. \tag{50}$$

Thus, under the local monotonicity and bounded-residual assumptions, attention dominance transfers to the generation space. Together with the previous result,

$$\text{retrieval-effective} \;\Rightarrow\; \text{attention-effective} \;\Rightarrow\; \text{generation-effective}. \tag{51}$$

B.5 Generation Supervision Details

The input template contains the dataset name, available modality, active retrieval query, retrieved candidate labels, target observation, self prompt tokens, differential prompt tokens, and the label-attentive table. Prompt and padding positions are masked, so only the target output contributes to the autoregressive objective.

Let $\mathcal{T}_{\text{label}}$ denote the supervised target positions before the first `Explanation:` marker, and let $\mathcal{T}_{\text{exp}}$ contain the positions from that marker through the explanation and end-of-sequence token. For token-level autoregressive cross-entropy

$$\ell_t = -\log p(y_t \mid y_{<t}, x), \tag{52}$$

the span-normalized losses are

$$\mathcal{L}_{\text{label}} = \frac{1}{|\mathcal{T}_{\text{label}}|} \sum_{t \in \mathcal{T}_{\text{label}}} \ell_t, \qquad \mathcal{L}_{\text{exp}} = \frac{1}{|\mathcal{T}_{\text{exp}}|} \sum_{t \in \mathcal{T}_{\text{exp}}} \ell_t. \tag{53}$$

The generation objective is therefore

$$\mathcal{L}_{\text{LLM}} = \alpha \mathcal{L}_{\text{label}} + (1 - \alpha)\mathcal{L}_{\text{exp}}. \tag{54}$$

Combining generation, SAE regularization, and label-attentive supervision gives

$$\mathcal{L} = \alpha \mathcal{L}_{\text{label}} + (1 - \alpha)\mathcal{L}_{\text{exp}} + \beta(\mathcal{L}_{\text{rec}} + \lambda_{\text{sparse}} \mathcal{L}_{\text{sp}}) + \gamma \mathcal{L}_{\text{attn}}. \tag{55}$$

We use $\alpha = 0.8$, $\beta = 0.1$, $\lambda_{\text{sparse}} = 0.001$, and $\gamma = 0.05$, so the effective coefficient of $\mathcal{L}_{\text{sp}}$ in the total objective is 10^{-4}. Only the SAE, alignment MLPs, and label-attentive module are trainable; gradients from $\mathcal{L}_{\text{LLM}}$ propagate through the frozen Qwen2-VL(Wang et al., 2024) backbone to these modules.

C Experiment Details

C.1 Effect of Retrieval Size with Different Generative Backbones

Table 3 examines the effect of retrieval size K with two generative backbones: Qwen2.5-7B-Instruct for RALAP-LLM and Qwen2-VL-7B-Instruct for RALAP-VLM. We vary $K \in \{3, 5, 7\}$ on MM-IMDb (Arevalo et al., 2017) and report F1-M/F1-S under Text, Image, and Both settings.

RALAP-LLM shows only mild and non-monotonic sensitivity to K. Text performs best at $K = 3$ (58.42/58.93), whereas Image and Both peak at $K = 5$ with 52.03/52.12 and 54.63/55.15, respectively. The average over all six metrics remains nearly unchanged at 54.75, 54.93, and 54.79 for $K = 3, 5, 7$, suggesting that additional retrieved neighbors provide limited benefit to the text-only backbone once sufficient evidence is available.

In contrast, RALAP-VLM improves consistently as K increases. From $K = 3$ to $K = 7$, Text increases from 59.47/60.58 to 60.78/61.69, Image from 64.80/66.17 to 65.64/66.96, and Both from 62.09/63.38 to 63.19/64.33. Accordingly, the average score rises from 62.75 to 63.31 and 63.77 for $K = 3, 5, 7$. This trend indicates that the multimodal backbone exploits the additional retrieval evidence more effectively.

RALAP-VLM consistently outperforms RALAP-LLM across all values of K and input settings, with the largest gap under Image. At $K = 7$, the gains are 14.48/15.33 points for Image, compared with 2.95/3.14 for Text and 8.73/9.24 for Both. This larger Image advantage suggests that retaining multimodal modeling capacity is particularly important when retrieval-conditioned evidence is strongly tied to visual information.

Table 3: Effect of Retrieval Size with Different Generative Backbones on MM-IMDb.

Method	K	Text		Image		Both	
		F1-M	F1-S	F1-M	F1-S	F1-M	F1-S
RALAP-LLM	3	58.42	58.93	50.85	50.84	54.54	54.89
	5	57.48	58.18	52.03	52.12	54.63	55.15
	7	57.83	58.55	51.16	51.63	54.46	55.09
RALAP-VLM	3	59.47	60.58	64.80	66.17	62.09	63.38
	5	59.93	60.91	**65.68**	66.73	62.79	63.82
	7	**60.78**	**61.69**	65.64	**66.96**	**63.19**	**64.33**

Table 4: Performance of RALAP with $K = 7$ on MM-IMDb, HateMemes, and Food101.

Dataset	Text					Image					Both				
	R-1	R-L	MET.	BL.	BERT	R-1	R-L	MET.	BL.	BERT	R-1	R-L	MET.	BL.	BERT
MM-IMDb	0.96	0.96	0.95	91.20	0.96	0.38	0.35	0.41	23.55	0.68	0.67	0.66	0.68	57.96	0.82
HateMemes	0.99	0.99	0.99	99.60	0.99	0.98	0.98	0.98	97.06	0.98	0.99	0.99	0.98	98.33	0.99
Food101	0.99	0.99	0.99	99.74	0.99	0.50	0.49	0.51	25.98	0.71	0.75	0.74	0.75	63.87	0.85

C.2 Explanation Quality

Table 4 reports the quality of explanations generated by RALAP on MM-IMDb(Arevalo et al., 2017), HateMemes(Kiela et al., 2020), and Food101(Wang et al., 2015) under a 100% missing rate with $K = 7$. We evaluate the generated explanations using ROUGE-1 (R-1), ROUGE-L (R-L), METEOR (MET.), BLEU (BL.), and BERTScore (BERT). Overall, RALAP achieves strong reference alignment across all three datasets, while the performance varies noticeably with the available input modality.

For MM-IMDb, the Text setting achieves the strongest explanation quality, with 0.9648 R-1, 0.9645 R-L, 0.9545 METEOR, 91.2010 BLEU, and 0.9645 BERTScore. Performance decreases under Image and Both, suggesting that explanation generation for this multi-label task is more challenging when textual semantics are unavailable or must be inferred indirectly from retrieved evidence. HateMemes shows consistently high scores under all three settings, with R-1 and BERTScore remaining above 0.98, indicating that RALAP can generate explanations highly aligned with the references even under modality missingness. On Food101, the Text setting again achieves near-saturated scores, whereas Image and Both obtain lower but still substantially positive semantic alignment, with BERTScore values of 0.7139 and 0.8566, respectively.

C.3 Ablation of Dynamic Prompting for Explanation Quality

C.3.1 MM-IMDb

Table 5 evaluates the contribution of the two dynamic soft-prompting components in RALAP to the quality of generated explanations on MM-IMDb. We report ROUGE-1 (R-1), ROUGE-L (R-L), METEOR (MET.), BLEU (BL.), and BERTScore (BERT) under the Text, Image, and Both input settings. The ablation includes four configurations: removing both dynamic soft-prompting modules (*without DSP*), using only the SAE-based differential soft prompts (*SAE DSP only*), using only the label-attentive prompts (*Label DSP only*), and combining the two components (*Both DSP*).

The most important observation is that combining the two prompting mechanisms yields the best performance on all reported metrics and under all three input settings. Compared with *without DSP*, *Both DSP* improves R-1/R-L/METEOR from 0.2671/0.2403/0.2300 to 0.9648/0.9645/0.9545 under the Text setting, while BLEU increases from 8.1249 to 91.2010 and BERTScore from 0.5285 to 0.9645. The improvement remains substantial under the more challenging Image setting, where R-1 increases from 0.1483 to 0.3884, METEOR from 0.1297 to 0.4140, BLEU from 2.3750 to 23.5524,

Table 5: Ablation study on MM-IMDb. Best results are shown in **bold**.

Method	Text					Image					Both				
	R-1	R-L	MET.	BL.	BERT	R-1	R-L	MET.	BL.	BERT	R-1	R-L	MET.	BL.	BERT
without DSP	0.26	0.24	0.23	8.12	0.52	0.14	0.13	0.12	2.37	0.42	0.20	0.18	0.17	4.34	0.47
SAE DSP only	0.60	0.60	0.59	54.35	0.69	0.0708	0.0702	0.0536	0.9747	0.2521	0.34	0.33	0.32	26.92	0.47
Label DSP only	0.28	0.26	0.27	6.77	0.51	0.15	0.14	0.14	2.59	0.42	0.22	0.20	0.21	4.58	0.47
Both DSP	**0.96**	**0.96**	**0.95**	**91.20**	**0.96**	**0.38**	**0.35**	**0.41**	**23.55**	**0.68**	**0.68**	**0.66**	**0.68**	**57.97**	**0.82**

and BERTScore from 0.4208 to 0.6895. Under the Both setting, the combined model similarly raises R-1 from 0.2077 to 0.6766 and BERTScore from 0.4746 to 0.8270. The simultaneous improvements in both lexical-overlap metrics and BERTScore indicate that the benefit of the combined prompting mechanism is not restricted to surface-form matching, but also extends to reference-based semantic similarity.

The two individual components, however, exhibit markedly different behaviors. *SAE DSP only* is highly effective under the Text setting, increasing R-1 from 0.2671 to 0.6096 and BLEU from 8.1249 to 54.3522. It also substantially improves most metrics under the Both setting. Nevertheless, its performance deteriorates under the Image setting: R-1 decreases from 0.1483 to 0.0708, BLEU from 2.3750 to 0.9747, and BERTScore from 0.4208 to 0.2521. This modality-dependent behavior suggests that the sparse self-and-difference representations provide rich instance-level evidence, but such continuous evidence alone may not always be sufficiently constrained to guide explanation generation when the available retrieval signal is dominated by visual information.

In contrast, *Label DSP only* produces smaller but generally more stable changes. Its clearest benefit appears under the Image setting, where R-1/R-L/METEOR improve from 0.1483/0.1322/0.1297 to 0.1591/0.1415/0.1435, with corresponding improvements in BLEU and BERTScore. Under Text and Both, the gains are modest and not uniform across all metrics; for example, Text BLEU and BERTScore slightly decrease relative to *without DSP*. This result suggests that label attention alone is insufficient to provide the fine-grained semantic content required for high-quality explanations. Instead, its main role is to provide target-conditioned label-level guidance that identifies which retrieved label evidence should receive greater emphasis.

The strong performance of *Both DSP* therefore reveals a clear complementarity between the two modules rather than a simple additive effect. The SAE-based differential prompts encode sample-specific continuous evidence from the target and its retrieved neighbors, whereas the label-attentive prompts provide an explicit class-aware constraint on how such evidence should be interpreted. This interaction is particularly evident in the Image setting: although *SAE DSP only* performs substantially worse than the no-DSP variant, adding Label DSP raises R-1 from 0.0708 to 0.3884 and BERTScore from 0.2521 to 0.6895. Similarly, under the Both setting, *SAE DSP only* substantially improves lexical metrics but slightly decreases BERTScore from 0.4746 to 0.4733; after introducing Label DSP, BERTScore increases to 0.8270. These results suggest that label-level conditioning helps regularize and disambiguate the continuous retrieval evidence encoded by the SAE prompts, making the resulting explanations better aligned with the predicted label semantics.

Overall, the ablation results support the joint design of RALAP. SAE-based differential prompting serves as the primary carrier of fine-grained retrieval-conditioned semantic information, while label-attentive prompting acts as a target-conditioned selector that anchors this information to class-relevant evidence. Neither component alone consistently achieves strong explanation quality across modalities, whereas their combination produces substantial and consistent improvements over all reference-based explanation metrics. We note, however, that these automatic metrics measure similarity to reference explanations and do not by themselves establish factual faithfulness; the results should therefore be interpreted as evidence of improved reference alignment rather than a complete measure of explanation correctness.

C.3.2 HATEMEMES

Table 6 reports the ablation results for explanation generation on HateMemes. We evaluate four variants of RALAP under the Text, Image, and Both settings using ROUGE-1 (R-1), ROUGE-L (R-L),

Table 6: Ablation study on HateMemes. Best results are shown in **bold**.

Method	Text					Image					Both				
	R-1	R-L	MET.	BL.	BERT	R-1	R-L	MET.	BL.	BERT	R-1	R-L	MET.	BL.	BERT
without DSP	0.24	0.22	0.24	7.83	0.46	0.18	0.15	0.16	3.40	0.43	0.21	0.19	0.20	5.60	0.44
SAE DSP only	0.15	0.12	0.09	1.66	0.41	0.10	0.09	0.10	6.30	0.23	0.12	0.11	0.10	4.35	0.32
Label DSP only	0.23	0.21	0.22	7.73	0.45	0.21	0.19	0.20	4.63	0.44	0.22	0.20	0.21	6.05	0.44
Both DSP	**0.99**	**0.99**	**0.99**	**99.60**	**0.99**	**0.98**	**0.98**	**0.98**	**97.06**	**0.98**	**0.99**	**0.99**	**0.98**	**98.33**	**0.99**

METEOR (MET.), BLEU (BL.), and BERTScore (BERT). The four configurations are: removing both dynamic soft-prompting modules (*without DSP*), retaining only the SAE-based differential soft prompts (*SAE DSP only*), retaining only the label-attentive dynamic soft prompts (*Label DSP only*), and combining both components (*Both DSP*).

The most striking result is that *Both DSP* achieves the best performance on every metric and under every input setting, with a substantially larger margin than either individual component. Under the Text setting, R-1/R-L/METEOR increase from 0.2495/0.2288/0.2432 without DSP to 0.9976/0.9976/0.9971 with Both DSP, while BLEU rises from 7.8309 to 99.60 and BERTScore from 0.4635 to 0.9979. The same pattern holds for Image and Both. Under Image, R-1 increases from 0.1805 to 0.9843 and BERTScore from 0.4302 to 0.9851; under Both, R-1 increases from 0.2152 to 0.9906 and BERTScore from 0.4468 to 0.9915. The consistently high scores across both lexical-overlap metrics and BERTScore indicate that the joint prompting mechanism substantially improves the alignment between generated explanations and the reference explanations.

In contrast, using only the SAE-based differential prompts consistently degrades explanation quality on HateMemes. Under the Text setting, R-1 decreases from 0.2495 to 0.1506 and BLEU from 7.8309 to 1.6674. Similar degradation is observed under Image and Both, where BERTScore drops from 0.4302 to 0.2348 and from 0.4468 to 0.3245, respectively. This result differs from the behavior observed on MM-IMDb and suggests that instance-level self-and-difference representations alone are insufficient for stable explanation generation on HateMemes. Since HateMemes is a single-label binary classification task, the explanation is strongly tied to distinguishing between only two competing semantic decisions. In such a setting, unconstrained continuous retrieval evidence may introduce instance-specific variations that are not necessarily aligned with the final class decision, thereby weakening the correspondence between classification and explanation.

The *Label DSP only* variant exhibits a markedly different behavior. Its performance remains close to, and in several Image and Both metrics slightly exceeds, the *without DSP* baseline. For example, under Image, R-1/R-L/METEOR improve from 0.1805/0.1569/0.1605 to 0.2135/0.1903/0.2011, while BERTScore increases from 0.4302 to 0.4419. Under Both, R-1 and BERTScore improve from 0.2152 and 0.4468 to 0.2261 and 0.4490, respectively. These comparatively modest gains indicate that label-level conditioning alone can provide useful class-aware guidance, but does not supply sufficient fine-grained semantic content to construct a high-quality explanation.

The large gap between *Both DSP* and the two single-component variants therefore points to a strong interaction between SAE-based differential prompting and label-attentive prompting. The SAE module provides continuous target–retrieval evidence, while the label-attentive module constrains this evidence toward class-relevant directions. This interaction appears to be particularly important for HateMemes: although SAE DSP alone degrades all three input settings, introducing Label DSP together with SAE DSP changes the behavior dramatically. For instance, under Image, R-1 improves from 0.1052 with SAE DSP only to 0.9843 with Both DSP, and BERTScore increases from 0.2348 to 0.9851. Under Both, R-1 similarly increases from 0.1279 to 0.9906. These results suggest that label-level conditioning does not merely contribute an additional prompt signal; rather, it helps disambiguate and regularize the continuous differential evidence produced by the SAE, allowing the VLM to associate retrieved semantic differences with the correct binary decision.

Table 7: Ablation study on Food101. Best results are shown in **bold**.

Method	Text					Image					Both				
	R-1	R-L	MET.	BL.	BERT	R-1	R-L	MET.	BL.	BERT	R-1	R-L	MET.	BL.	BERT
without DSP	0.2551	0.23	0.30	8.56	0.50	0.12	0.10	0.09	0.09	0.42	0.19	0.17	0.20	5.40	0.46
SAE DSP only	0.90	0.90	0.90	87.41	0.91	0.41	0.41	0.36	10.98	0.54	0.65	0.65	0.63	39.99	0.73
Label DSP only	0.21	0.23	0.28	7.38	0.48	0.12	0.10	0.10	0.08	0.42	0.18	0.16	0.19	4.75	0.45
Both DSP	**0.99**	**0.99**	**0.99**	**99.74**	**0.99**	**0.50**	**0.49**	**0.51**	**25.98**	**0.71**	**0.75**	**0.74**	**0.75**	**63.87**	**0.85**

C.3.3 FOOD101

Table 7 reports the ablation results of RALAP on explanation generation for Food101. We evaluate four variants, including removing both dynamic soft-prompting modules (*without DSP*), using only SAE-based differential soft prompts (*SAE DSP only*), using only label-attentive prompts (*Label DSP only*), and combining both components (*Both DSP*). The generated explanations are evaluated under the Text, Image, and Both settings using ROUGE-1 (R-1), ROUGE-L (R-L), METEOR (MET.), BLEU (BL.), and BERTScore (BERT).

The results demonstrate that combining the two dynamic prompting components consistently achieves the best explanation quality across all input conditions. Under the Text setting, *Both DSP* improves R-1/R-L/METEOR from 0.2551/0.2382/0.3056 without DSP to 0.9991/0.9990/0.9985, while BLEU increases from 8.5638 to 99.74 and BERTScore increases from 0.5016 to 0.9992. Similar improvements are observed under the Image and Both settings. Specifically, under the Image setting, R-1 and BERTScore increase from 0.1279 and 0.4227 to 0.5072 and 0.7139, respectively. Under the Both setting, R-1 increases from 0.1915 to 0.7531, and BERTScore improves from 0.4622 to 0.8566. The consistent improvements across both lexical similarity metrics and semantic similarity metrics indicate that the joint dynamic prompting strategy substantially enhances the alignment between generated explanations and reference explanations.

Unlike HateMemes, the SAE-based differential prompts alone already provide significant improvements on Food101. Under the Text setting, *SAE DSP only* increases R-1 from 0.2551 to 0.9080 and BLEU from 8.5638 to 87.41. Under the Both setting, it improves R-1 from 0.1915 to 0.6598 and BERTScore from 0.4622 to 0.7300. This phenomenon is consistent with the characteristics of Food101, where the classification objective relies heavily on fine-grained visual and semantic attributes. The retrieved instances provide rich instance-level similarities, and the SAE-based differential representation can effectively encode the visual-semantic differences between the target sample and retrieved neighbors, thereby providing informative evidence for explanation generation.

However, SAE DSP alone remains insufficient compared with the full model, particularly under the Image setting. Although *SAE DSP only* achieves strong performance under Text and Both, its Image BERTScore is 0.5467, which is lower than the 0.7139 achieved by *Both DSP*. This gap indicates that continuous retrieval evidence alone may not fully determine which semantic attributes should be emphasized in the generated explanation. The label-attentive module provides complementary class-aware guidance by explicitly modeling the relevance between retrieved labels and the target sample, allowing the model to select more discriminative evidence from retrieved neighbors.

The contribution of Label DSP alone is relatively limited but complementary. Under the Image setting, *Label DSP only* improves R-1/R-L/METEOR from 0.1279/0.1081/0.0998 to 0.1282/0.1090/0.1003 and maintains a similar BERTScore. Under the Both setting, it also provides slight improvements over the no-DSP baseline. These results suggest that label-level guidance alone cannot provide sufficient semantic information for generating detailed explanations, but it can offer useful category constraints that help organize and interpret retrieval evidence.

C.4 COMPARISON UNDER COMPLETE MODALITY MISSINGNESS

Table 8 compares RALAP with MAPs(Lee et al., 2023), MSPs(Jang et al., 2024), and RAGPT(Lang et al., 2025) under the same experimental setting, where the missing rate is fixed to 100% and the retrieval size is set to $K = 7$. This setting enables a controlled comparison in the complete absence of one modality, eliminating performance differences caused by different missing-rate configurations.

Table 8: Comparison of different methods on MM-IMDb, HateMemes, and Food101.

	MM-IMDb		HateMemes	Food101
Input Type	Both		Both	Both
Method	**F1-M**	**F1-S**	**AUROC**	**ACC**
MAPs	42.27	40.85	56.45	72.17
MSPs	42.51	41.41	55.31	72.50
RAGPT	51.69	51.04	60.42	78.30
RALAP	**63.19**	**64.33**	**64.02**	**80.74**

RALAP consistently achieves the best performance across all three datasets and all reported metrics. On MM-IMDb, RALAP reaches 63.19 F1-M and 64.33 F1-S, outperforming the strongest baseline, RAGPT, by 11.50 and 13.29 percentage points, respectively. The margin is substantially larger than those observed on HateMemes and Food101. On HateMemes, RALAP improves AUROC from 60.42 to 64.02, corresponding to a gain of 3.60 points. On Food101, accuracy increases from 78.30 to 80.74, yielding a 2.44-point improvement. These results show that the advantage of RALAP is maintained across multi-label classification, binary classification, and fine-grained multi-class recognition under the same complete-missing-modality condition.

The largest improvement is observed on MM-IMDb, which is a multi-label multi-class task. In this setting, retrieval must not only recover semantically relevant evidence but also distinguish multiple relevant labels from noisy labels introduced by neighboring instances. RALAP explicitly addresses this issue through target-conditioned label attention, which reweights retrieved labels according to their compatibility with the target instance, while the sparse self-and-difference prompts encode instance-specific retrieval evidence. The large gains over RAGPT therefore suggest that directly incorporating retrieved instances is insufficient when the retrieved neighborhood contains heterogeneous label evidence; explicitly modeling label relevance and target–neighbor differences provides a more effective way to exploit the retrieved memory.

The improvements on HateMemes and Food101 are smaller but remain consistent. HateMemes is a binary classification task, where the candidate label space is considerably smaller and label coverage saturates quickly. Consequently, retrieval mainly contributes additional discriminative evidence rather than introducing new candidate labels, leaving less room for label-level reweighting to improve performance. Similarly, on Food101, the 2.44-point gain over RAGPT indicates that retrieval-conditioned prompting remains beneficial, while the smaller margin suggests that the contribution of retrieved evidence depends on the task structure and the informativeness of the available modality.

Notably, the comparison with RAGPT is particularly informative because both methods exploit retrieval, whereas RALAP additionally introduces sparse differential prompting and target-conditioned label attention. RALAP improves over RAGPT on every reported dataset and metric under the identical 100% missing-rate and $K = 7$ setting. Although this comparison alone does not isolate the contribution of each individual component, it provides evidence that explicitly modeling how retrieved instances differ from the target and how reliable their associated labels are is more effective than treating retrieved evidence uniformly. We further disentangle these contributions through the ablation studies reported in the main paper.