BRIDGE THE GAPS: GAP-AWARE PRESERVATION AND EXPANSION FOR CONTINUAL MISSING MODALITY LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual Missing Modality Learning (CMML) requires multimodal models to learn new classes sequentially while retaining old knowledge under incomplete visual or textual inputs. Existing methods mainly recover missing information or reduce interference through efficient adaptation and knowledge preservation, but overlook how missingness alters cross-modal relations before and after fusion. Our stage-wise analysis reveals two coupled challenges: compensated features obscure degradation of the real image-text geometry before fusion, while shared fusion under unequal modality reliability induces misleading modality dependencies after fusion. To address them, we propose GAPE, a Gap-Aware Preservation and Expansion framework. GAPE separately monitors real geometry degradation and compensation-induced shifts to select a safe adaptation stage, and combines counterfactual anchoring, confidence, and modality availability to calibrate fused decisions toward reliable evidence. Experiments on two CMML benchmarks show consistent state-of-the-art performance across diverse missing patterns and ratios.

1 INTRODUCTION

Continual Missing Modality Learning (CMML) requires a multimodal model to incrementally learn disjoint classes while handling incomplete modality observations, which commonly arise from sensor failures, incomplete annotations, privacy constraints, and limited communication bandwidth. Without access to task identity at test time, the model must recognize all previously seen classes from either complete inputs or samples with missing visual or textual modalities. As illustrated in Figure 1, CMML therefore jointly confronts catastrophic forgetting and modality absence. More importantly, their interaction can progressively alter cross-modal representations and fusion behavior as new tasks arrive, making CMML more challenging than conventional class-incremental learning or static missing-modality learning alone.

Existing CMML methods (Zhao et al., 2024; Guo et al., 2025; Liu et al., 2026; Yue et al., 2026) mainly address these challenges through input recovery, parameter-efficient adaptation, or knowledge preservation. Recovery-based methods estimate unavailable modality information before prediction (Ma et al., 2021; Zhao et al., 2024), while adaptation-based methods reduce interference across tasks or modality conditions (Lee et al., 2023; Jang et al., 2024; Hu et al., 2024; Guo et al., 2025; Liu et al., 2026). Knowledge-preservation approaches further retain previously learned information throughout the task sequence (Zhuang et al., 2022; Yue et al., 2026). Although effective, these methods primarily constrain reconstructed inputs, model parameters, or retained knowledge, without explicitly examining how missingness alters cross-modal relations before and after multimodal fusion. This limitation motivates a stage-wise analysis of cross-modal geometry and modality dependency in CMML.

To understand how modality missingness interacts with continual adaptation, we analyze cross-modal relations at two stages of multimodal learning, namely representation formation before fusion and modality-dependent prediction after fusion. Our diagnostic analysis reveals *two coupled effects*. Before fusion, continual adaptation changes the relation between real visual and textual representations, while missing-modality compensation introduces additional relations involving pseudo features. Consequently, the observed cross-modal variation conflates genuine degradation of the

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.

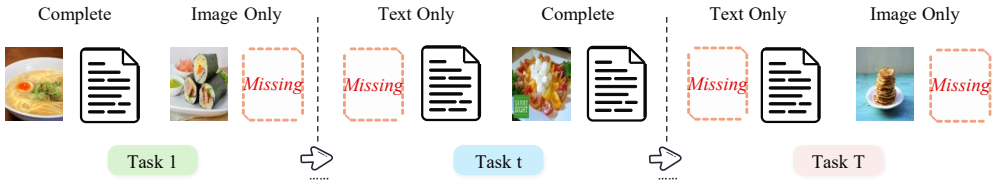


Figure 1: Overview of Continual Missing Modality Learning (CMML). A model sequentially learns disjoint classes from complete or modality-incomplete inputs.

pretrained geometry with compensation-induced shifts, making it unsafe to preserve all observed relations indiscriminately. We refer to this effect as *missing-induced pre-fusion gap ambiguity*. After fusion, complete and incomplete samples update the same fusion pathway despite substantial differences in modality reliability. Compensation errors can therefore make an available modality appear dominant merely because the other modality is missing or poorly compensated, and repeated optimization may encode this pseudo dominance into the shared fusion parameters. We refer to this effect as *fusion-amplified modality dependency bias*. Together, these observations suggest that CMML should preserve the reliable component of pre-fusion geometry while calibrating post-fusion decisions toward genuinely reliable modality evidence.

To address these two challenges, we propose GAPE, a gap-aware preservation and expansion framework for CMML. Before fusion, Missing-Aware Pre-Fusion Gap Preservation separately monitors degradation of the real image-text relation and shifts involving compensated features, thereby selecting a safe adaptation stage before the pretrained geometry is severely distorted. After fusion, Confidence-Aware Post-Fusion Gap Expansion first identifies the potentially bottlenecked modality and applies swap-based counterfactual anchoring to retain its base discriminative capability. It then combines prediction confidence with modality availability to increase the decision influence of genuinely reliable evidence, without assuming that either modality is globally dominant. Finally, logits from available unimodal branches are combined with the calibrated fused prediction to retain direct modality-specific evidence. Experiments on UPMC-Food101-CMML and N24News-CMML show that GAPE consistently improves performance across diverse missing patterns and ratios.

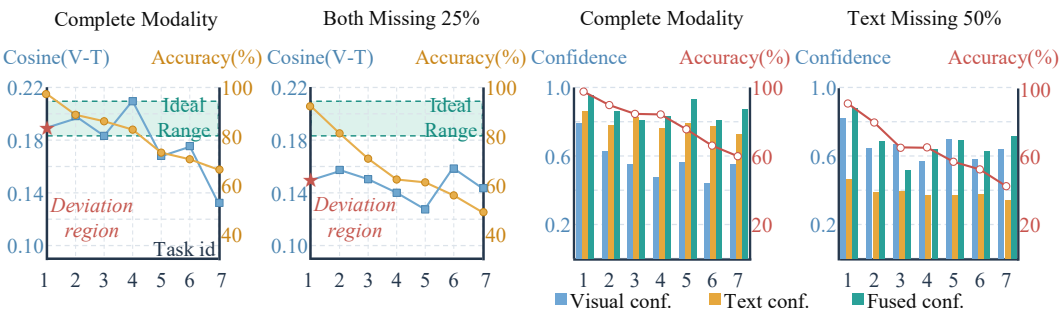
The main contributions are summarized as follows:

- We present a stage-wise analysis of CMML and reveal two coupled challenges: missing-induced pre-fusion gap ambiguity and fusion-amplified modality dependency bias.
- We propose GAPE, which distinguishes real cross-modal degradation from compensation-induced shifts before fusion and calibrates post-fusion modality dependency toward genuinely reliable evidence.
- Extensive experiments on two CMML benchmarks validate the identified challenges and individual components, and demonstrate state-of-the-art performance across diverse missing patterns and ratios with a compact trainable parameter budget.

2 Related Work

Continual Missing Modality Learning. CMML extends multimodal continual learning to task streams with incomplete inputs. Static methods improve missing-modality robustness through fusion and reconstruction (Ma et al., 2022; Wang et al., 2023). Other approaches align incomplete with complete representations, employ distillation, or learn missing-aware and modality-specific prompts (Lin & Hu, 2023; Chen et al., 2024; Lee et al., 2023; Jang et al., 2024). In continual settings, RebQ reconstructs unavailable information before retrieving decomposed prompts, while EffPrompt enables parameter-efficient adaptation to changing modality availability (Zhao et al., 2024; Guo et al., 2025). DeLo isolates interference with dual-decomposed low-rank experts, whereas PAL combines modality-specific prompting with analytic learning (Liu et al., 2026; Yue et al., 2026). Despite reducing forgetting, these methods seldom examine missingness-induced changes in pre-fusion cross-modal geometry or post-fusion modality dependency.

Modality Gap in Vision-Language Models. Vision-language models such as CLIP project heterogeneous modalities into a shared space (Radford et al., 2021), yet their feature distributions remain



(a) **missing-induced pre-fusion gap ambiguity** (b) **fusion-amplified modality dependency bias**

Figure 2: Diagnostic evidence for the two gap-related challenges, evaluated on the UPMC-Food101-CMML validation set across the first seven tasks of the ten-task sequence. See the *Diagnostic Analysis and Our Motivation* section for detailed analysis.

separated because of differences in input statistics, encoder architectures, and semantic abstraction (Liang et al., 2022; Jiang et al., 2023; Ramasinghe et al., 2024). This gap affects retrieval, knowledge transfer, and downstream adaptation. MG-CLIP links preservation of the pretrained gap to forgetting and new-task adaptation under complete observations (Huang et al., 2025). In CMML, real image–text relations coexist with placeholder- or compensation-based relations that primarily reflect missing-induced pseudo shifts. Treating all gap changes equivalently can thus obscure genuine degradation of pretrained multimodal representations.

Multimodal Imbalance and Modality Dependency. Multimodal imbalance arises when modalities contribute unequally during optimization and prediction because of learnability, noise, or class relevance (Wu et al., 2022; Peng et al., 2022; Du et al., 2023). Forcing equal contributions is not always optimal, as it may suppress informative evidence or amplify unreliable inputs (Fan et al., 2023; Wei et al., 2024; Yang et al., 2024). Reliability-aware methods therefore estimate modality importance from confidence, uncertainty, input quality, or mismatched samples (Ma et al., 2023; Zhang et al., 2023; Hwang et al., 2025). With missingness, apparent dominance may result from the other modality being absent or corrupted rather than from improved discriminative ability. Continually optimizing such pseudo-dominance can introduce persistent bias into shared fusion parameters and disrupt previously learned fusion relations.

3 Diagnostic Analysis and Our Motivation

Our diagnostic analysis in Figure 2 provides empirical evidence for both challenges. For pre-fusion gap ambiguity, Figure 2(a)(Left) shows that, even with complete modalities, the real image-text cosine similarity eventually leaves the complete-modality reference range during continual fine-tuning, while average-seen accuracy declines; the construction and interpretation of this ideal range are detailed in Appendix A.3. When both modalities have a 25% missing rate, Figure 2(a)(Right) shows that all seven task-wise gap values fall outside the reference range, and this deviation is accompanied by a more pronounced decline in accuracy. At the decision level, Figure 2(b)(Left) establishes the complete-modality reference, where textual confidence is consistently higher than visual confidence, indicating the stronger base capability of the textual modality, while fused confidence generally exceeds both unimodal branches. More importantly, class-incremental fine-tuning exhibits a persistent preference for the modality with stronger base capability under both complete and missing settings. Consequently, as shown in Figure 2(b)(Right), even with 50% text missingness, the missing textual branch remains competitive with the available visual branch, allowing the missing modality to interfere with the learning of the available modality. Moreover, text missingness reverses the visual–textual confidence ordering, creating a pseudo-dominant visual modality; fused confidence even falls below visual confidence for some tasks. The accompanying decline in average-seen accuracy suggests that this altered dependency reflects missing-induced bias rather than beneficial modality specialization.

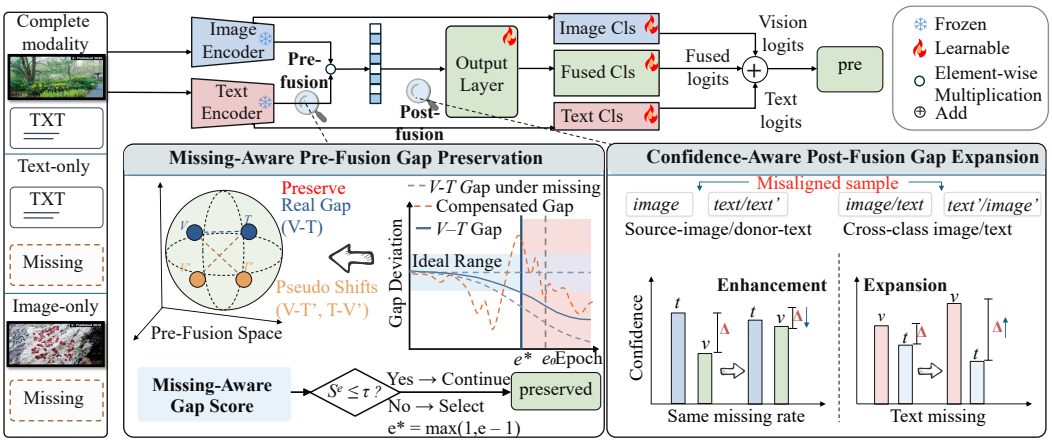


Figure 3: **Overview of GAPE.** Complete or incomplete inputs are processed by the visual, textual, and fused prediction branches. **Missing-Aware Pre-Fusion Gap Preservation** separates compensation-induced pseudo shifts from changes in the real V–T relation and selects a safe adaptation epoch before the real gap leaves its pretrained range. **Confidence-Aware Post-Fusion Gap Expansion** constructs cross-class mismatched pairs and generates confidence- and availability-aware soft targets, enlarging the decision gap in favor of the more reliable modality. Only when a pseudo-dominant modality emerges does it first enhance the base capability of this intrinsically weaker modality before increasing the decision weight of the more reliable modality.

4 Proposed Method

4.1 Problem Formulation

Continual Missing Modality Learning (CMML) consists of K incremental classification tasks $\{\mathcal{D}_t\}_{t=0}^{K-1}$ with disjoint class sets $\{\mathcal{C}_t\}_{t=0}^{K-1}$. Each sample is denoted by $(I_i, T_i, y_i, m_i^v, m_i^t)$, where I_i and T_i are image and text inputs, $y_i \in \mathcal{C}_t$ is the label, and $m_i^v, m_i^t \in \{0,1\}$ indicate modality availability. After task t, the model predicts over $\mathcal{C}_{0:t} = \bigcup_{k=0}^{t} \mathcal{C}_k$ without task identity. Image, text, or both modalities may be missing during training and inference under a missing ratio ρ. Missing text and image inputs are replaced by an empty string and an all-one image, respectively. GAPE preserves cross-modal relations before fusion and calibrates fused predictions using the reliability of available modalities.

We use RGap and CGap to quantify the two effects. RGap is the average cosine similarity between real visual and textual features, $\mathrm{RGap}_t = \frac{1}{|\mathcal{V}_t||\mathcal{T}_t|} \sum_{\mathbf{v}_i \in \mathcal{V}_t} \sum_{\mathbf{t}_j \in \mathcal{T}_t} \frac{\mathbf{v}_i^\top \mathbf{t}_j}{\|\mathbf{v}_i\|_2 \|\mathbf{t}_j\|_2}$, where compensated features are excluded; a lower value indicates a larger modality gap. For a mini-batch $\mathcal{B}_t$, $\mathrm{CGap}_t = \left| \frac{1}{|\mathcal{B}_t|} \sum_{i \in \mathcal{B}_t} c_i^v - \frac{1}{|\mathcal{B}_t|} \sum_{i \in \mathcal{B}_t} c_i^t \right|$, where $c_i^s = [\mathrm{softmax}(\mathbf{z}_i^s)]_{y_i}$ is the true-class confidence of branch $s \in \{v, t\}$. A larger CGap indicates a greater confidence disparity between the two branches.

4.2 Missing-Aware Pre-Fusion Gap Preservation

This module selects the training epoch by monitoring image and text relations before fusion. Under missing inputs, a change in the observed relation comes from two sources. The relation between real image and text features changes during continual adaptation, while compensated inputs introduce additional relation changes. A score that mixes these two sources uses changes from compensated features as an early stopping signal. We therefore measure them separately.

Let $\mathcal{V}$ and $\mathcal{T}$ denote features from real image and text inputs. Let $\mathcal{V}'$ and $\mathcal{T}'$ denote features from compensated inputs. For an image feature set $\mathcal{A}$ and a text feature set $\mathcal{B}$, we compute the average

cosine similarity at epoch e:

$$c^{(e)}(\mathcal{A},\mathcal{B}) = \frac{1}{|\mathcal{Q}|}\sum_{(\mathbf{a},\mathbf{b})\in\mathcal{Q}} \frac{\mathbf{a}^\top\mathbf{b}}{\|\mathbf{a}\|_2\|\mathbf{b}\|_2}. \tag{1}$$

where $\mathcal{Q}$ contains image and text feature pairs from different classes. These pairs continue to decrease after the similarity of matched pairs stabilizes, thereby more clearly reflecting deviations in the modality gap. The term $c^{(e)}(\mathcal{V},\mathcal{T})$ measures the relation between real image and text features. The terms $c^{(e)}(\mathcal{V},\mathcal{T}')$ and $c^{(e)}(\mathcal{V}',\mathcal{T})$ measure relations involving a compensated feature. We compare each relation with its pretrained value:

$$d^{(e)}(\mathcal{A},\mathcal{B}) = \left[\frac{c^{(0)}(\mathcal{A},\mathcal{B}) - c^{(e)}(\mathcal{A},\mathcal{B})}{\left|c^{(0)}(\mathcal{A},\mathcal{B})\right| + \varepsilon}\right]_+. \tag{2}$$

where $c^{(0)}(\mathcal{A},\mathcal{B})$ denotes the value before optimization. The positive part retains decreases in cosine similarity, and the denominator normalizes the decrease by the pretrained magnitude. We first compute the drift between real image and text features:

$$d^{(e)}_{\text{real}} = d^{(e)}(\mathcal{V},\mathcal{T}). \tag{3}$$

This value records the decrease in the relation between real image and text features. We then combine the two relations involving compensated features:

$$d^{(e)}_{\text{mis}} = w_\rho\left[d^{(e)}(\mathcal{V},\mathcal{T}') + d^{(e)}(\mathcal{V}',\mathcal{T})\right]. \tag{4}$$

This value records the change associated with compensated features. Here, $d(V,T')$ is included only for text-missing samples, whereas $d(V',T)$ is included only for image-missing samples. The inapplicable term is set to zero. Thus, the two terms do not simultaneously apply to the same sample. We set $w_\rho = \rho$ in the image missing and text missing settings. For the both missing setting, we set $w_\rho = \rho/2$.

The real feature term defines the base risk. The compensated feature term adjusts this risk:

$$S^{(e)}_{\text{raw}} = d^{(e)}_{\text{real}}\left(1 + \beta\frac{d^{(e)}_{\text{mis}}}{d^{(e)}_{\text{real}} + d^{(e)}_{\text{mis}} + \varepsilon}\right). \tag{5}$$

Since $d^{(e)}_{\text{real}}$ multiplies the full expression, changes in compensated feature relations alone do not trigger early stopping. They increase the score only when the relation between real features also changes. The raw score has different ranges across missing settings. For the current missing setting π and missing ratio ρ, we use a fixed scale $z_{\pi,\rho} > 0$:

$$S^{(e)} = \frac{S^{(e)}_{\text{raw}}}{S^{(e)}_{\text{raw}} + z_{\pi,\rho} + \varepsilon}. \tag{6}$$

This normalization preserves the ordering of epochs within each setting while reducing scale differences across settings.

We use Task 0 to determine a safe training horizon. Before training, the pretrained real visual–textual relation is recorded as the reference, and the missing-aware deviation score $S^{(e)}$ is computed after each epoch. When $S^{(e)}$ first exceeds the threshold τ, the training horizon is set to $E^\star = \max(e-1, 1)$; otherwise, $E_{\max}$ is used, where $E_{\max}$ denotes the predefined maximum number of calibration epochs. Thus, if the threshold is exceeded after the first epoch, $E^\star$ is set to 1 because the number of training epochs cannot be zero. The resulting $E^\star$ is then applied to all subsequent tasks, preserving the real modality gap while preventing compensation-induced pseudo shifts from dominating the stopping criterion. Appendix A.4 explains why this Task-0-calibrated horizon can be reused across later tasks. This strategy introduces no additional learnable parameters.

Pre-fusion preservation controls representation drift but not how fusion uses the preserved features. It may also restrict plasticity. We therefore use counterfactual anchoring to protect bottleneck-modality evidence and reliability-aware supervision to calibrate the fused dependency.

4.3 CONFIDENCE-AWARE POST-FUSION GAP EXPANSION

PFGP selects a safe adaptation horizon on the first task. Each task then adapts the encoder for this number of epochs before post fusion calibration. The objectives below use features obtained after this adaptation. For readability, we omit epoch indices.

Before calibrating the post-fusion dependency, we protect the modality most likely to become the bottleneck. Because modality dominance can vary across datasets and tasks.

Let $\mathcal{B}$ denote the current mini batch and let $B = |\mathcal{B}|$. To avoid repeating the fusion expression, let $\Phi(\mathbf{v}, \mathbf{t}) = C_f(F(\mathbf{v}, \mathbf{t}))$ denote the fused logits, where F is the fusion module and C_f is the fused classifier. Let C_v and C_t denote the visual and textual classifiers. For sample i, the three branches produce

$$\begin{aligned} \mathbf{z}_i^f &= \Phi(\mathbf{v}_i, \mathbf{t}_i), \\ \mathbf{z}_i^v &= C_v(\mathbf{v}_i), \qquad \mathbf{z}_i^t = C_t(\mathbf{t}_i). \end{aligned} \tag{7}$$

We use $\sigma(\cdot)$ to denote softmax. The auxiliary confidences are $c_i^v = [\sigma(\mathbf{z}_i^v)]_{y_i}$ and $c_i^t = [\sigma(\mathbf{z}_i^t)]_{y_i}$. The fused branch models the joint evidence, while the auxiliary branches provide the unimodal evidence used below.

We compare the auxiliary classifiers under matched availability conditions. At the current missing ratio, we apply the same ratio to the other modality and compare their mean true class confidences. The modality with lower confidence is treated as weak. Consider the case where the visual modality is weak. For sample i, we retain $\mathbf{v}_i$ and replace its text with an available $\mathbf{t}_j$ from a sample $j \in \mathcal{B}$ with $y_j \neq y_i$. We apply the following loss when the image is available and the text is missing:

$$\mathcal{L}_{\text{swap}} = \mathrm{KL}(\sigma(\Phi(\mathbf{v}_i, \mathbf{t}_j)) \,\|\, \mathrm{stopgrad}(\sigma(\mathbf{z}_i^v))) \,. \tag{8}$$

The visual auxiliary prediction provides the target for the fused prediction of the swapped pair. The loss keeps the fused prediction consistent with the weak visual evidence after the text is replaced. It can be minimized through two complementary effects: the fusion module suppresses interference from the mismatched text, while the model relies more strongly on the discriminative features of the source image. The latter can indirectly improve the visual representation and its unimodal classification capability. The textual case exchanges the roles of image and text. We average this loss over selected samples in the mini-batch. The enhancement is activated only when the strong modality is missing or unreliable and the model must rely on the weak modality. Otherwise, strengthening a missing weak modality would mainly amplify compensation-induced signals. This trigger improves fallback capability without forcing equal modality strength.

We next form mismatched image and text pairs for fusion calibration. Let $\mathcal{P}$ contain all pairs $(i, j) \in \mathcal{B}^2$ with $y_i \neq y_j$, $m_i^v = 1$, and $m_j^t = 1$. Their fused logits are

$$\mathbf{z}_{ij}^{\text{mis}} = \Phi(\mathbf{v}_i, \mathbf{t}_j). \tag{9}$$

The two inputs in each pair support different labels. A hard target would select one label and discard the evidence from the other modality. We instead construct a target from their relative reliability. For pair (i, j), we write $c_v = c_i^v$ and $c_t = c_j^t$. We combine pair level confidence with mini batch availability:

$$a_s = \frac{1}{B} \sum_{k \in \mathcal{B}} m_k^s, \qquad r_s = c_s a_s^{\kappa}, \qquad s \in \{v, t\}. \tag{10}$$

Here, a_s is the availability rate of modality s, and κ controls the availability penalty. The reliability score is high when the modality is confident for the current pair and frequently available in the mini batch.

For each pair, we set $(w_v, w_t) = (2, 1)$ when $r_v \geq r_t$, and set $(w_v, w_t) = (1, 2)$ otherwise. The larger coefficient identifies the modality that receives stronger supervision. Both modalities remain in the target. Let $\mathbf{e}_y$ denote the one hot vector for class y. The visual share and the soft target are

$$\begin{aligned} \alpha_{ij} &= \frac{w_v c_v^{\gamma}}{w_v c_v^{\gamma} + w_t c_t^{\gamma}}, \\ \mathbf{q}_{ij} &= \alpha_{ij} \mathbf{e}_{y_i} + (1 - \alpha_{ij}) \mathbf{e}_{y_j}, \end{aligned} \tag{11}$$

Table 1: **Main results on UPMC-Food101-CMML and N24News-CMML**. AP and FG are computed from classification accuracy, and the best AP results are highlighted in bold.

Dataset	ρ	#Image	#Text	MAP		MSP		L2P		DualPrompt		RebQ		DeLo		ACIL		PAL		GAPE	
				AP↑	FG↓	AP↑	FG↓	AP↑	FG↓	AP↑	FG↓	AP↑	FG↓	AP↑	FG↓	AP↑	FG↓	AP↑	FG↓	AP↑	FG↓
UPMC-Food101		100%	90%	20.66	82.50	21.45	80.03	34.09	5.80	59.56	3.38	68.67	9.50	77.19	10.21	78.62	6.24	82.45	4.35	**86.29**	5.61
	10%	90%	100%	21.53	82.20	23.29	78.98	35.21	5.51	59.90	1.72	72.46	7.64	80.23	7.88	79.78	5.15	84.66	3.74	**88.70**	4.23
		95%	95%	22.84	80.13	22.12	79.35	34.00	6.55	58.18	6.32	71.06	8.22	78.12	10.32	79.24	5.83	83.50	3.72	**87.92**	4.81
		100%	70%	16.57	83.77	21.37	79.95	30.40	6.53	51.86	6.46	62.06	10.76	74.08	9.69	73.31	7.76	76.80	5.24	**79.90**	6.37
	30%	70%	100%	24.00	77.12	22.21	76.02	34.75	4.66	52.66	5.11	71.62	6.44	78.20	6.34	78.49	5.57	83.09	3.53	**87.25**	4.04
		85%	85%	20.66	79.81	21.89	79.01	29.76	6.87	48.56	8.73	66.37	8.07	73.63	8.89	74.84	6.56	79.14	4.32	**83.75**	6.10
		100%	50%	18.18	79.77	17.88	77.31	28.95	6.33	47.70	6.09	55.87	12.07	71.41	14.77	68.10	8.64	71.92	5.76	**73.19**	9.12
	50%	50%	100%	23.85	75.59	21.57	77.99	32.18	4.55	50.22	5.03	69.23	5.84	76.38	7.39	77.49	5.18	81.56	3.32	**85.18**	4.20
		75%	75%	18.66	79.80	18.10	78.04	25.30	6.17	43.67	8.52	62.40	8.24	70.59	8.27	71.71	7.48	75.76	4.59	**80.23**	6.88
		100%	30%	17.68	77.82	19.29	78.17	26.57	5.80	43.09	8.96	50.00	12.47	70.04	14.91	63.84	9.88	66.78	6.21	**71.41**	9.24
	70%	30%	100%	22.48	75.21	21.22	75.21	30.43	3.99	50.28	5.19	69.41	3.73	74.80	5.32	76.44	5.35	80.51	3.00	**83.61**	4.35
		65%	65%	20.00	75.95	19.76	76.54	24.62	5.66	40.69	7.26	59.92	8.56	67.42	9.41	68.35	7.63	72.59	4.92	**75.76**	7.93
		100%	10%	16.92	76.60	18.62	75.14	24.94	5.60	37.26	11.46	48.15	12.76	67.30	15.68	61.34	10.42	64.34	7.01	**68.62**	8.97
	90%	10%	100%	24.89	70.84	21.69	71.23	29.77	4.54	51.16	4.80	67.71	4.71	74.21	4.83	76.93	5.33	80.68	2.67	**83.42**	3.41
		55%	55%	18.41	75.19	20.45	74.31	24.62	5.66	40.69	7.29	54.67	8.78	63.37	10.97	65.93	4.33	69.69	5.32	**73.13**	8.37
N24News		100%	90%	17.34	60.23	19.12	59.11	27.23	11.23	36.67	10.42	42.37	12.92	48.92	11.45	51.50	9.50	54.66	8.18	**65.65**	21.04
	10%	90%	100%	16.98	59.40	20.69	55.39	25.69	10.87	34.07	8.19	42.14	13.26	49.35	10.86	50.97	9.54	54.68	6.62	**67.65**	21.68
		95%	95%	18.18	61.34	21.13	57.18	27.18	11.12	35.98	9.93	44.19	14.02	49.84	11.72	50.63	10.37	54.52	8.07	**66.61**	21.17
		100%	70%	16.23	58.12	19.34	56.89	25.32	10.28	34.29	10.41	41.66	12.17	46.84	11.31	48.86	10.20	51.31	8.17	**60.21**	23.08
	30%	70%	100%	15.30	55.48	18.21	54.45	24.11	11.38	33.76	10.02	39.75	13.37	47.63	10.42	49.74	9.56	52.82	7.56	**67.85**	21.80
		85%	85%	17.14	56.72	19.37	55.72	22.37	12.51	34.19	10.93	39.44	11.85	45.92	11.46	48.46	10.53	50.37	8.89	**63.59**	22.83
		100%	50%	16.24	57.09	17.91	57.62	23.47	11.82	32.24	9.32	40.48	15.48	44.31	12.34	45.89	10.57	48.76	7.02	**54.42**	24.71
	50%	50%	100%	14.97	56.81	18.76	53.37	24.29	10.86	33.01	11.18	38.42	15.40	45.86	10.71	47.36	10.23	50.50	7.01	**67.59**	22.15
		75%	75%	16.83	57.92	18.01	55.76	22.06	12.42	32.36	11.23	38.38	14.86	44.73	11.83	46.71	10.17	48.05	6.98	**62.30**	21.56
		100%	30%	15.43	55.02	17.13	53.14	21.43	10.38	30.48	9.28	37.44	16.63	41.96	12.76	43.28	11.21	45.34	9.92	**48.93**	27.47
	70%	30%	100%	14.32	52.84	16.15	52.57	23.81	15.36	31.91	13.12	36.84	17.44	43.71	11.48	44.99	10.95	48.15	8.75	**66.74**	22.92
		65%	65%	17.34	53.17	17.24	53.27	20.51	13.91	30.73	12.61	37.27	14.64	42.12	12.05	43.74	10.29	46.02	8.66	**56.90**	25.15
		100%	10%	14.12	54.10	16.88	51.03	20.78	8.03	29.94	13.04	38.06	14.08	41.35	11.82	42.66	10.39	45.50	8.27	**49.30**	20.47
	90%	10%	100%	13.12	55.27	15.19	52.89	21.32	10.12	30.13	11.82	35.44	15.05	41.86	11.21	43.30	10.64	47.57	8.20	**68.13**	20.61
		55%	55%	14.08	53.19	14.78	51.38	19.83	12.26	29.27	13.51	34.56	15.93	39.74	12.63	41.19	11.19	44.03	7.23	**54.28**	26.35

where γ controls confidence sharpness. The coefficients select which modality receives stronger supervision, while the confidence terms determine the final share of each label. We train the fused logits of mismatched pairs with

$$\mathcal{L}_{\text{mis}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \text{CE}\big(\mathbf{z}_{ij}^{\text{mis}}, \mathbf{q}_{ij}\big)\,. \tag{12}$$

Gradients do not pass through α_{ij} or $\mathbf{q}_{ij}$, so the auxiliary branches do not change the target through their confidence values.

4.4 Decision-Level Prediction and Training Objective

We next combine the fused logits with logits from the available unimodal branches for normal samples.

$$\mathbf{z}_i = \mathbf{z}_i^f + m_i^v \mathbf{z}_i^v + m_i^t \mathbf{z}_i^t. \tag{13}$$

The availability masks remove the contribution from a missing modality. The model predicts $\hat{y}_i = \arg\max_c [\mathbf{z}_i]_c$ at inference. Using the mini batch $\mathcal{B}$ defined above, the classification loss and the total objective are

$$\begin{aligned} \mathcal{L}_{\text{cls}} &= \frac{1}{B} \sum_{i\in\mathcal{B}} \text{CE}(\mathbf{z}_i, y_i), \\ \mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{swap}} + \mathcal{L}_{\text{mis}}. \end{aligned} \tag{14}$$

The three terms respectively train normal samples, preserve weak modality evidence, and calibrate fusion under mismatched inputs. PFGP selects the adaptation horizon and does not add a loss term.

5 Experiments

5.1 Experimental Setup

Datasets and evaluation. We evaluate GAPE on UPMC-Food101-CMML and N24News-CMML, derived from UPMC Food-101 (Wang et al., 2015) and N24News (Wang et al., 2022a). Following

prior CMML protocols (Zhao et al., 2024; Yue et al., 2026), UPMC-Food101-CMML contains 100 classes with 61,142/6,846/22,716 train/validation/test samples across 10 tasks, while N24News-CMML contains 24 classes with 48,988/6,123/6,124 samples across 6 tasks. We evaluate text-, image-, and both-missing settings with $\rho \in \{10\%, 30\%, 50\%, 70\%, 90\%\}$ during training and testing, using the same setting at task-agnostic inference. We report Average Performance (AP), Average Forgetting (FG), and last-task accuracy (Last); higher AP and Last and lower FG are better for overall continual learning performance.

Comparison methods. We compare with static missing-modality methods MAP and MSP (Lee et al., 2023; Jang et al., 2024); rehearsal-free prompt-based continual learners L2P and DualPrompt (Wang et al., 2022c;b); CMML methods RebQ and DeLo (Zhao et al., 2024; Liu et al., 2026); and exemplar-free analytic methods ACIL and PAL (Zhuang et al., 2022; Yue et al., 2026). All methods use the same data splits and missing-modality settings.

Implementation details. GAPE uses CLIP ViT-B/16 (Radford et al., 2021) with rank-8 LoRA adapters (Hu et al., 2022) on the key and value projections; all other backbone parameters are frozen. Normalized image and text features are fused by element-wise multiplication and classified by a cosine classifier with scale 30. We use AdamW with learning rate 10^{-4}, cosine scheduling, batch size 4. Task 0 calibrates the pre-fusion horizon for up to ten epochs using $\beta = 0.5$ and $\tau = 0.5$, and reuses the selected epoch thereafter. Post-fusion calibration uses an epoch of mismatched-pair augmentation with coefficients $2.0/1.0$, availability exponent $\kappa = 0.5$, confidence-sharpness exponent $\gamma = 0.3$, and masked unimodal logits. Hyperparameter sensitivity analyses are presented in Appendix B.3.

5.2 Main Results

Table 1 shows that GAPE achieves the highest AP under every evaluated missing-modality configuration on both benchmarks. On UPMC-Food101-CMML, GAPE outperforms the best competing method in each setting by 1.27–4.61 AP points. The gains are larger on N24News-CMML, ranging from 3.59 to 20.56 AP points, indicating that the proposed calibration is particularly effective for the more challenging news classification setting. The results also reveal a clear modality asymmetry. When text remains available at $\rho = 90\%$, GAPE reaches 83.42% AP on UPMC-Food101 and 68.13% on N24News, compared with 68.62% and 49.30% when text is missing. This trend is consistent with GAPE's reliability-aware design, which increases reliance on the more informative available modality rather than enforcing equal contributions. Although GAPE does not always obtain the lowest FG, its consistent AP advantage yields the strongest final classification performance across diverse missingness conditions. Additional results across different random seeds and continual-learning task orders are reported in Appendices B.1 and B.2, respectively.

5.3 Gap-Aware Method Validation

Pre-Fusion Gap Preservation. To verify that GAPE preserves pretrained cross-modal geometry rather than fitting compensation features, we examine real image–text relations before fusion. As shown in Figure 4, GAPE achieves higher average RGap, a smaller deviation from the complete-modality reference, and better AP than LoRA-CLIP (Radford et al., 2021; Hu et al., 2022). The reduced deviation suggests that missing-induced pseudo shifts are less likely to distort the real image–text gap, thereby preserving pretrained capability and improving continual classification.

Post-Fusion Gap Expansion. Pre-fusion preservation alone does not guarantee reliable modality use after fusion.

As shown in Table 2, RebQ (Zhao et al., 2024) reconstructs the missing-modality query with prompts, but its negative Confidence Gain suggests that the available branch is not sufficiently strengthened. ARL (Wei et al., 2025) reweights modality optimization using unimodal prediction variance, yet its modest CGap shows that the missing branch still keeps non-negligible influence. In contrast, GAPE produces a larger CGap and positive Confidence Gain by enhancing available-modality evidence while reducing missing-modality interference, yielding the best AP.

Table 2: Post-fusion gap expansion on UPMC-Food101-CMML (Text-50%).

Metric	LoRA-CLIP	RebQ	ARL	GAPE
CGap ↑	0.2303	0.1953	0.2582	**0.5032**
Confidence Gain	0.0856	-0.0057	0.1175	0.1082
AP (%) ↑	35.52	56.11	44.33	**73.19**

Table 3: Ablation of PFGP and PFGE under Both-70% missingness.

PFGP	PFGE	Food101		N24News	
		AP (%) ↑	Last (%)	AP (%) ↑	Last (%)
×	×	28.73	81.53	24.38	78.33
✓	×	47.08	61.91	25.69	68.84
×	✓	69.13	80.96	37.69	86.95
✓	✓	**75.76**	74.16	**56.90**	59.85

5.4 Ablation Study

To isolate the contribution of each gap-aware stage, we independently enable Pre-Fusion Gap Preservation (PFGP) and Post-Fusion Gap Expansion (PFGE), yielding four variants on both benchmarks.

Besides AP, we also report Last, the accuracy on the most recently learned task, to assess current-task plasticity. As shown in Table 3, removing both modules gives high Last but poor AP, indicating strong bias toward newly learned classes. PFGP improves retention by preserving pretrained cross-modal geometry, while PFGE brings a larger AP gain by calibrating the fused prediction toward reliable modalities. Combining both modules achieves the best AP on both datasets, showing that pre-fusion preservation and post-fusion calibration are complementary.

Training Cost Comparison. In continual learning, improved robustness should not depend on repeatedly updating a large portion of the model, as this increases training cost and limits scalability across long task sequences. We therefore compare trainable-parameter efficiency to assess whether the gains of GAPE require a large adaptation budget and whether its efficiency generalizes across datasets. As summarized in Figure 5, GAPE consistently achieves the strongest AP with a compact update budget, outperforming parameter-heavy CMML baselines while requiring substantially fewer trainable parameters. Prompt-based methods are lighter but considerably less effective. The same trade-off is observed on both datasets, indicating that the efficiency of GAPE is not specific to a particular benchmark.

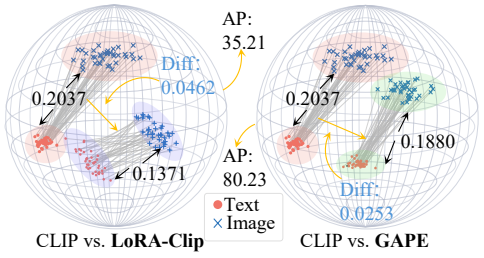


Figure 4: Pre-fusion gap preservation on UPMC-Food101-CMML (Both-50%).

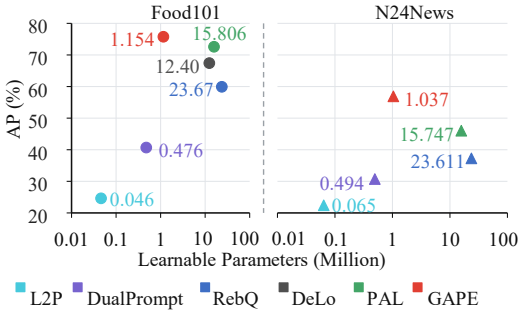


Figure 5: Trainable parameters and AP on both benchmarks under Both-70% missingness.

6 Conclusion

GAPE addresses pre- and post-fusion modality gaps in Continual Missing Modality Learning. It preserves the real visual–textual relation within its pretrained range by excluding compensation-induced pseudo shifts, while using modality confidence and availability to favor reliable modalities after fusion and reduce pseudo-dominance. Experiments on two CMML benchmarks show consistent state-of-the-art performance with strong parameter efficiency. **Nevertheless**, our gap characterization mainly relies on average cosine similarity and classifier confidence, which may not fully capture fine-grained cross-modal geometry or instance-level uncertainty. The evaluation also focuses on image–text class-incremental learning with predefined missing ratios. Future work will explore richer gap and reliability measures, extend GAPE to additional modalities and dynamic missing conditions, and integrate it with broader continual learning strategies.

AI USE STATEMENT

Generative AI tools, including ChatGPT, were used to assist with translation, literature search, manuscript drafting and editing, LaTeX formatting, and the presentation of the methodology, experimental analyses, figures, and tables. They were not used to generate benchmark data or conduct the reported experiments.

All AI-assisted outputs were reviewed by the authors. Technical claims, equations, citations, and numerical results were independently verified against the original sources, implementation, and experimental records. The authors take full responsibility for the content of this work, including any text or artifacts produced with AI assistance.

ETHICS STATEMENT

This work uses the publicly available UPMC-Food101-CMML and N24News-CMML benchmarks and does not involve the collection of new personal data or human-subject research. The datasets are used for academic research with appropriate attribution to their original sources and in accordance with their applicable terms of use.

GAPE may inherit biases from the underlying datasets and pretrained encoders, and its behavior may vary across modalities, missing-data patterns, and deployment domains. The reported results should therefore not be interpreted as guarantees of fairness, robustness, or safety in real-world applications. Deployment in sensitive or safety-critical settings would require additional domain-specific evaluation. This work adheres to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Section 4 provides a complete description of GAPE, including its formulation, architecture, objectives, and training procedure. Section 5.1 specifies the datasets, data splits, missing-modality protocols, evaluation metrics, baselines, optimization settings, and hyperparameters.

Appendix A provides additional analyses, while Appendix B reports robustness across random seeds and task orders, together with the ranges and selection criteria used in the hyperparameter analysis. The anonymized supplementary material includes the preprocessing, training, and evaluation code, experimental configurations, and execution scripts. All datasets used in this work are publicly available, and the complete code will be released upon publication.

REFERENCES

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, pp. 36218–36242, 2024.

Chenzhuang Du, Jiaye Teng, Tingle Li, Yichen Liu, Tianyuan Yuan, Yue Wang, Yang Yuan, and Hang Zhao. On uni-modal feature learning in supervised multi-modal learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 8632–8656, 2023.

Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20029–20038, 2023.

Zirun Guo, Shulei Wang, Wang Lin, Weicai Yan, Yangyang Wu, and Tao Jin. Efficient prompting for continual adaptation to missing modalities. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4317–4327, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.219.

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.

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, pp. 67446–67466, 2024. doi: 10.52202/079017-2153.

Linlan Huang, Xusheng Cao, Haori Lu, Yifan Meng, Fei Yang, and Xialei Liu. Mind the gap: Preserving and compensating for the modality gap in CLIP-based continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3777–3786, 2025. doi: 10.1109/ICCV51701.2025.00360.

Seong-Hyeon Hwang, Soyoung Choi, and Steven Euijong Whang. MIDAS: Misalignment-based data augmentation strategy for imbalanced multimodal learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Jaehyuk Jang, Yooseung Wang, and Changick Kim. Towards robust multimodal prompting with missing modalities. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8070–8074, 2024. doi: 10.1109/ICASSP48485.2024.10447257.

Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Understanding and constructing latent modality structures in multimodal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7661–7671, 2023.

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*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5583–5594. PMLR, 2021.

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. doi: 10.1109/CVPR52729.2023.01435.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17612–17625, 2022.

Ronghao Lin and Haifeng Hu. MissModal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702, 2023.

Xiwei Liu, Yulong Li, Feilong Tang, and Imran Razzak. DeLo: Dual decomposed low-rank experts collaboration for continual missing modality learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(28):23855–23863, 2026. doi: 10.1609/aaai.v40i28.39561.

Huan Ma, Qingyang Zhang, Changqing Zhang, Bingzhe Wu, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. Calibrating multimodal learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23429–23450. PMLR, 2023.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. SMIL: Multimodal learning with severely missing modality. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):2302–2310, 2021. doi: 10.1609/aaai.v35i3.16330.

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. 18156–18165, 2022. doi: 10.1109/CVPR52688.2022.01764.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8237, 2022. doi: 10.1109/CVPR52688.2022.00806.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Thalaiyasingam Ajanthan. Accept the modality gap: An exploration in the hyperbolic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27253–27262, 2024.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2556–2565, 2018.

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

Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17117–17128, 2023.

Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. N24News: A new dataset for multimodal news classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 6768–6775, Marseille, France, 2022a. European Language Resources Association.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer G. Dy, and Tomas Pfister. DualPrompt: Complementary prompting for rehearsal-free continual learning. In *Computer Vision – ECCV 2022*, pp. 631–648, 2022b.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer G. Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 139–149, 2022c.

Shicai Wei, Chunbo Luo, and Yang Luo. Improving multimodal learning via imbalanced learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2250–2259, 2025. doi: 10.1109/ICCV51701.2025.00217.

Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27328–27337, 2024.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 24043–24055, 2022.

Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. Facilitating multimodal classification via dynamically learning modality gap. In *Advances in Neural Information Processing Systems*, volume 37, pp. 62108–62122, 2024.

Xianghu Yue, Yiming Chen, Xueyi Zhang, Xiaoxue Gao, Mengling Feng, Mingrui Lao, Huiping Zhuang, and Haizhou Li. PAL: Prompting analytic learning with missing modality for multi-modal class-incremental learning. *Pattern Recognition*, 179:113467, 2026. doi: 10.1016/j.patcog.2026.113467.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41753–41769. PMLR, 2023.

Shu Zhao, Xiaohan Zou, Tan Yu, and Huijuan Xu. Reconstruct before query: Continual missing modality learning with decomposed prompt collaboration. *arXiv preprint arXiv:2403.11373*, 2024.

Huiping Zhuang, Zhenyu Weng, Hongxin Wei, Renchunzi Xie, Kar-Ann Toh, and Zhiping Lin. ACIL: Analytic class-incremental learning with absolute memorization and privacy protection. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11602–11614, 2022.

APPENDIX

A ADDITIONAL ANALYSES

A.1 VISUALIZATION OF THE PRE-FUSION MODALITY GAP

We visualize the image and text feature distributions to examine how modality missingness changes the pre-fusion geometry. We randomly sample 512 image–caption pairs from CC3M (Sharma et al., 2018), extract normalized visual and textual features using the pretrained CLIP ViT-B/16 encoders (Radford et al., 2021), and project the features into two dimensions with UMAP (McInnes et al., 2018). We compare complete inputs with a setting in which 30% of the images are missing. The missing images are represented using the same placeholder protocol as in the main experiments.

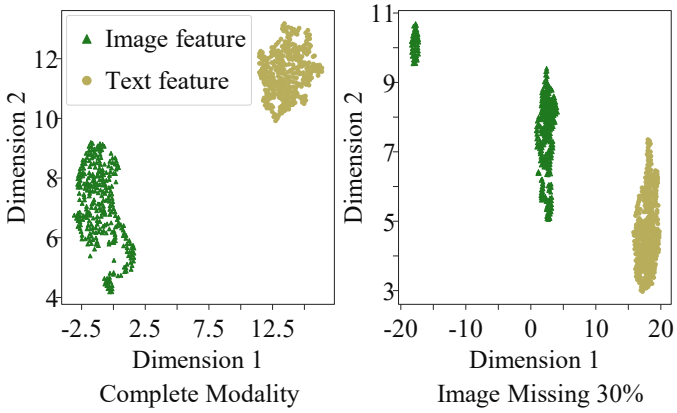


Figure 6: Pre-fusion feature distributions on CC3M. Left: complete modality. Right: 30% image missingness. Each point denotes a visual or textual feature projected by UMAP.

As shown in Figure 6, complete inputs produce compact within-modality clusters while the visual and textual distributions retain a clear separation, consistent with the modality gap of a dual-encoder model. Under image missingness, the visual features form an additional displaced cluster associated with placeholder inputs, while the real visual and textual features preserve a different relation. Consequently, a single aggregate gap mixes two effects: the evolution of the real visual–textual relation and a missingness-induced pseudo shift. This observation supports the design of GAPE's Missing-Aware Pre-Fusion Gap Preservation, which evaluates the real visual–textual geometry separately from relations involving compensated or placeholder features.

A.2 BACKBONE SELECTION AND FAIRNESS

GAPE uses CLIP ViT-B/16 as its backbone, whereas RebQ adopts ViLT by default. These backbones embody different multimodal interaction strategies. ViLT performs token-level fusion within a joint transformer (Kim et al., 2021), which provides deep cross-modal interaction. CLIP instead uses separate image and text encoders trained contrastively on large-scale image–text pairs (Radford et al., 2021); cross-modal interaction is shallow, but its modality-specific representations transfer strongly and remain directly accessible before fusion.

RebQ's own appendix compares CLIP and ViLT under the same method (Zhao et al., 2024). Its reported results are reproduced in Table 4. ViLT obtains higher AP than CLIP for all three 70% missing patterns, showing that CLIP is not an inherently favorable backbone for this CMML benchmark.

The comparison between GAPE and ViLT-based methods is relatively fair because the two backbones offer different and counterbalancing advantages in CMML. ViLT performs early token-level fusion, enabling more explicit and fine-grained cross-modal interaction, whereas CLIP relies on separate encoders and late fusion, which may limit its ability to exploit multimodal interactions. Conversely, CLIP benefits from large-scale contrastive pretraining and strong transferability, which

Table 4: Backbone comparison reported by RebQ on UPMC-Food101-CMML at a 70% missing ratio (Zhao et al., 2024).

Backbone	#Image	#Text	AP ↑	FG ↓
CLIP	100%	30%	39.69	15.26
CLIP	30%	100%	60.44	14.08
CLIP	65%	65%	53.06	6.53
ViLT	100%	30%	50.00	12.47
ViLT	30%	100%	69.41	3.73
ViLT	65%	65%	59.92	8.56

is the main reason why we choose CLIP as our backbone, providing a robust prior under continual and missing-modality shifts. GAPE effectively preserves and exploits this pretrained knowledge while updating only lightweight components, but it does not benefit from the tightly coupled fusion architecture of ViLT. In the Both-70% setting, GAPE achieves 75.76 AP with 1.154 million trainable parameters, compared with 59.92 AP and 23.67 million trainable parameters for the ViLT-based RebQ. Although this is not a controlled same-backbone ablation, the result demonstrates that GAPE achieves stronger performance despite using a less fusion-oriented architecture and a substantially smaller adaptation budget. Therefore, CLIP's stronger pretrained prior and ViLT's stronger explicit fusion capability represent a reasonable trade-off, making the backbone comparison relatively balanced rather than systematically favoring GAPE.

A.3 Construction and Interpretation of the Ideal Range

The ideal range in Figure 2 is an operational reference interval constructed from the complete-modality trajectory. It is used only for diagnostic visualization and does not serve as a training objective or introduce any test-set information.

Complete-modality reference trajectory. Let

$$g_t^{\mathrm{cmp}} = \mathrm{RGap}_t^{\mathrm{cmp}}$$

denote the real image–text cosine similarity after learning task t under complete modalities, where RGap is computed using the same definition as in the main paper. We evaluate g_t^{cmp} on the validation split for the first T_d tasks and obtain the complete-modality reference trajectory

$$\mathcal{G}^{\mathrm{cmp}} = \left\{ g_1^{\mathrm{cmp}}, g_2^{\mathrm{cmp}}, \ldots, g_{T_d}^{\mathrm{cmp}} \right\}.$$

We define the relative decrease between two consecutive tasks as

$$r_t = \frac{\left[g_{t-1}^{\mathrm{cmp}} - g_t^{\mathrm{cmp}} \right]_+}{\left| g_{t-1}^{\mathrm{cmp}} \right| + \epsilon}, \qquad t = 2, \ldots, T_d, \tag{15}$$

where $[x]_+ = \max(x, 0)$ and ϵ is a small constant for numerical stability. An increase or a minor fluctuation therefore does not indicate degradation.

The first pronounced degradation point is defined as

$$t_{\mathrm{drop}} = \min \left\{ t \in \{2, \ldots, T_d\} \mid r_t \geq \eta_{\mathrm{drop}} \right\}, \tag{16}$$

where η_{drop} is a predefined minimum relative-drop threshold. We set $\eta_{\mathrm{drop}} = 0.1$ and use this same value for all missing-modality patterns and ratios. Importantly, this value is determined using only the complete-modality validation trajectory and is not selected according to the results of any missing-modality experiment or test-set performance.

If no task satisfies Equation (16), we set $t_{\mathrm{drop}} = T_d + 1$, meaning that no pronounced degradation is detected within the analyzed task sequence. The stable task set is then defined as

$$\mathcal{T}_{\mathrm{stable}} = \left\{ t \mid 1 \leq t < t_{\mathrm{drop}} \right\}.$$

Based on this stable portion of the complete-modality trajectory, the ideal range is defined as

$$\mathcal{I}_{\text{ideal}} = [g_{\min}^{\text{cmp}}, g_{\max}^{\text{cmp}}], \tag{17}$$

where

$$g_{\min}^{\text{cmp}} = \min_{t \in \mathcal{T}_{\text{stable}}} g_t^{\text{cmp}}, \qquad g_{\max}^{\text{cmp}} = \max_{t \in \mathcal{T}_{\text{stable}}} g_t^{\text{cmp}}.$$

Therefore, the ideal range corresponds to the minimum and maximum complete-modality cosine similarities observed before the first pronounced degradation. It should be understood as a complete-modality stable reference range rather than a theoretically optimal interval or a confidence interval around a single pretrained value.

A.4 Why Can the Horizon Calibrated on Task 0 Be Reused?

We do not assume that Task 0 and the subsequent tasks have identical data distributions, nor do we regard $E^\star$ as the task-wise optimal number of training epochs. Instead, $E^\star$ serves as a conservative, task-shared update budget that restricts the representation drift introduced by each task.

Let $\theta_{t,0}$ denote the trainable parameters at the beginning of task t, and let $\theta_{t,e}$ denote the parameters after e optimization updates. Under the fixed training schedule, one update unit can be interpreted as the optimization performed within one epoch. The update is written as

$$\theta_{t,k+1} = \theta_{t,k} - \eta_{\text{lr}} \nabla \mathcal{L}_t(\theta_{t,k}), \tag{18}$$

where η_{lr} is the effective learning rate. We assume that the gradient norm is uniformly bounded over tasks within the relevant optimization region:

$$\|\nabla \mathcal{L}_t(\theta)\|_2 \le G, \qquad \forall t. \tag{19}$$

It follows that the parameter displacement after e updates satisfies

$$\|\theta_{t,e} - \theta_{t,0}\|_2 \le \eta_{\text{lr}} e G. \tag{20}$$

Let $\mathcal{G}_t(\theta)$ denote the missing-aware cross-modal gap for task t. Assuming that $\mathcal{G}_t$ is $L_{\mathcal{G}}$-Lipschitz continuous with respect to the trainable parameters, we have

$$\begin{aligned} &|\mathcal{G}_t(\theta_{t,e}) - \mathcal{G}_t(\theta_{t,0})| \\ &\quad \le L_{\mathcal{G}} \|\theta_{t,e} - \theta_{t,0}\|_2 \le L_{\mathcal{G}} \eta_{\text{lr}} e G. \end{aligned} \tag{21}$$

Therefore, an update horizon satisfying

$$L_{\mathcal{G}} \eta_{\text{lr}} E^\star G \le \tau_{\text{safe}} \tag{22}$$

provides a uniform upper bound on the incremental representation drift caused by each task, as long as the tasks share the same optimization configuration and satisfy the bounded-gradient condition above.

Task 0 provides a clean calibration stage because it starts directly from the pretrained model and is unaffected by accumulated drift from previous tasks. PFGP empirically estimates the largest admissible horizon before a pronounced gap deviation is observed and selects the epoch immediately preceding the detected crossing as $E^\star$. The same $E^\star$ is subsequently reused as a common plasticity budget. Hence, transferring $E^\star$ relies on shared and bounded optimization dynamics rather than on the stronger assumption that all tasks have identical data distributions or identical task-wise optimal stopping times.

B Additional Experimental Details

B.1 Robustness Across Random Seeds

We repeat the UPMC-Food101-CMML experiments with random seeds $\{1, 2, 3\}$ under text-, image-, and both-missing settings at missing ratios of 30% and 70%. The seed used for the main experiments (42) is not included in this aggregation. Table 5 reports the mean and sample standard deviation across the three additional seeds. For a metric x, the sample standard deviation is

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}, \qquad n = 3.$$

Table 5: Main results under three additional random seeds on UPMC-Food101-CMML. Values are reported as mean ± sample standard deviation across seeds 1, 2, and 3.

Missing Setting	Ratio	AP ↑	FG ↓	Last ↑
Text Missing	30%	80.07 ± 0.53	6.30 ± 0.10	78.62 ± 0.62
Image Missing	30%	86.99 ± 0.34	4.23 ± 0.32	86.79 ± 0.13
Both Missing	30%	83.83 ± 0.33	5.98 ± 0.34	84.07 ± 0.39
Text Missing	70%	69.83 ± 0.55	11.36 ± 0.58	69.61 ± 1.83
Image Missing	70%	83.80 ± 0.27	4.06 ± 0.35	83.69 ± 0.31
Both Missing	70%	76.79 ± 0.20	7.58 ± 0.39	76.92 ± 0.80

Table 6: Results averaged over the original, reverse, and random task orders on UPMC-Food101-CMML. "±" denotes the sample standard deviation across the three task orders.

Missing Setting	Ratio	AP ↑	FG ↓	Last ↑
Text Missing	30%	80.66 ± 0.66	6.28 ± 0.26	78.67 ± 3.74
Image Missing	30%	86.52 ± 0.72	3.90 ± 0.26	83.33 ± 5.76
Both Missing	30%	83.73 ± 0.04	5.59 ± 0.61	82.15 ± 3.95
Text Missing	70%	71.29 ± 0.11	10.06 ± 0.72	70.08 ± 2.83
Image Missing	70%	83.73 ± 0.50	3.88 ± 0.46	81.67 ± 4.33
Both Missing	70%	76.09 ± 0.69	7.86 ± 0.41	73.47 ± 3.08

The AP standard deviations are at most 0.55 points, and the FG standard deviations are at most 0.58 points. The largest variation occurs for Last under Text-70%, but its standard deviation remains below two points. These results show that the improvements of GAPE are not tied to the seed used in the main paper. They also reproduce the same modality asymmetry observed in the main results: text missingness is the most difficult setting, whereas image missingness yields the highest AP.

B.2 Robustness to Task Order

Class-incremental performance may depend on which classes are encountered early or late in the stream. We therefore evaluate three task sequences: the original class order used in the main paper, its reverse, and a random permutation. Table 6 reports the mean and sample standard deviation across these three task orders.

AP remains stable across task orders, with standard deviations no larger than 0.72 points. FG shows similarly limited variation. Last is more sensitive because changing the order also changes the classes assigned to the final task; thus, its variation reflects both order sensitivity and the intrinsic difficulty of the terminal class group. The consistently small AP variation indicates that the overall continual-learning benefit of GAPE does not depend on a favorable class sequence.

B.3 Hyperparameter Sensitivity

We study three hyperparameters: the availability exponent κ, the confidence exponent γ, and the mismatched-loss weight λ_{mis}. Each sweep varies one hyperparameter while holding the others fixed. The availability sweep fixes $(\gamma, \lambda_{\text{mis}}) = (0.3, 1.0)$, the confidence sweep fixes $(\kappa, \lambda_{\text{mis}}) = (0.5, 1.0)$, and the loss-weight sweep fixes $(\kappa, \gamma) = (0.5, 0.3)$. All runs use seed 42 and a 70% missing ratio on UPMC-Food101-CMML.

For compact presentation, Table 7 reports the equal-weight average over the Text-70%, Image-70%, and Both-70% settings. For $M \in \{\text{AP}, \text{FG}\}$,

$$\overline{M}_{\text{miss}}(\theta) = \frac{M_{\text{Text70}}(\theta) + M_{\text{Image70}}(\theta) + M_{\text{Both70}}(\theta)}{3},$$

where θ denotes the value of the hyperparameter being swept.

The availability exponent has the smallest effect: its average AP varies by 0.97 points over the tested range, indicating relative robustness to κ. A moderate confidence exponent of 0.30 performs best,

Table 7: Hyperparameter sensitivity averaged over Text-70%, Image-70%, and Both-70% missing-modality settings. Bold values denote the best result within each one-dimensional sweep.

Hyperparameter	Value	Avg. AP ↑	Avg. FG ↓
Availability exponent	0.00	76.06	**7.08**
	0.25	75.96	7.14
	0.50	**76.93**	7.17
	0.75	76.88	7.09
	1.00	76.82	7.13
Confidence exponent	0.00	76.64	**6.48**
	0.10	76.46	6.68
	0.30	**76.93**	7.17
	0.50	76.08	7.04
	1.00	75.86	7.15
Mismatched-loss weight	0.25	74.26	7.46
	0.50	74.83	7.73
	1.00	**76.93**	7.17
	1.50	76.34	7.16
	2.00	76.19	**6.85**

while larger values gradually reduce AP. The mismatched-loss weight is the most influential hyperparameter. A weight of 1.00 yields the highest average AP, whereas 2.00 achieves the lowest average FG at the expense of classification performance. After evaluating all candidate hyperparameter configurations based on their corresponding AP values, we select the configuration that achieves the highest AP and adopt it as the final setting in our experiments.