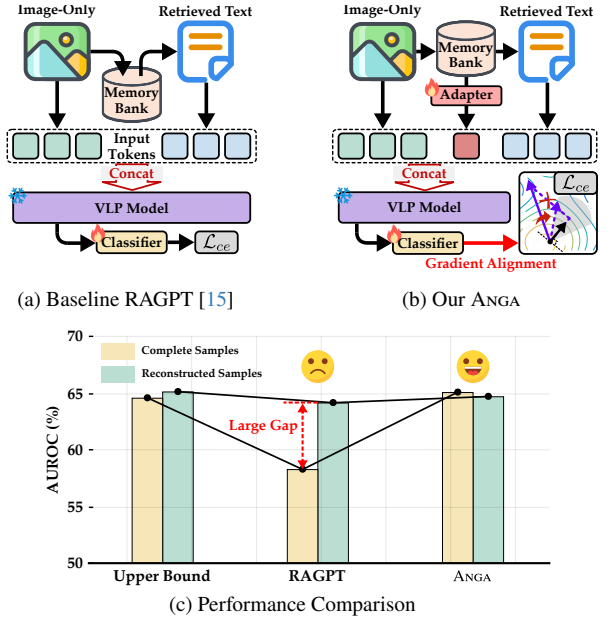
Anchor-Guided Gradient Alignment for Incomplete Multimodal Learning

Zhi-Hao Guan, Longfei Huang, Yang Yang*
Nanjing University of Science and Technology, China
{zhguan, hlf, yyang}@njust.edu.cn

Abstract

Vision-language pre-training (VLP) has achieved remarkable performance across diverse multimodal learning (MML) tasks. Recently, many efforts have focused on reconstructing missing modalities to improve the adaptability of VLP models in incomplete MML scenarios. However, these approaches overlook the learning imbalance under severe missing-modality conditions, i.e., the optimization process is dominated by reconstructed samples, thereby weakening complete-sample representations. In this paper, we propose a novel ANchor-guided Gradient Alignment (ANGA) framework to address this issue. Specifically, we first retrieve similar instances to reconstruct the missing modalities, thereby alleviating information deficiency. We then introduce an entropy-driven curriculum that progressively incorporates reliable reconstructed samples together with complete ones to form an optimization anchor, which guides gradient alignment to mitigate learning imbalance. Furthermore, we design a semantic-enhanced adapter that leverages the retrieved instances to generate dynamic prompts, further enhancing the robustness of the VLP model. Extensive experiments on widely used datasets demonstrate the superiority of ANGA over state-of-the-art (SOTA) baselines across various missing-modality scenarios. The code is available at this repository.



(a) Baseline RAGPT [15] (b) Our ANGA

(c) Performance Comparison

Figure 1. A motivating example of gradient alignment in incomplete MML: **(a).** General retrieval-based reconstruction paradigm (*e.g.*, RAGPT). **(b).** Illustration of our ANGA. **(c).** Performance comparison on HateMemes under a 70% text-missing condition. The baseline RAGPT exhibits a marked learning imbalance between complete and reconstructed samples, while ANGA achieves balanced performance across both subsets.

1. Introduction

Multimodal learning has emerged as a prominent research area in artificial intelligence across diverse domains [4, 41, 43, 46, 53, 55], including speech recognition [10], content retrieval [49], and recommender systems [8]. By integrating complementary information from multiple modalities, MML has become a key paradigm for improving model performance in these applications. Nevertheless, most existing methods assume that all modalities are fully available during both training and inference. In practice, incomplete data are prevalent due to sensor malfunctions, transmission errors, or privacy constraints, posing challenges to model robustness and reliability [19, 42, 54].

Recently, numerous studies have been conducted to address the missing-modality problem, which can be broadly grouped into three categories: (1). modality-invariant learning methods [32, 56], (2). VLP-based methods [6, 11, 16], and (3). modality reconstruction methods [15, 40]. Considering the inherent uncertainty of missing modalities, a straightforward solution is to enforce semantic consistency across modalities. However, such modality-invariant learning methods often overlook modality-specific cues [9], resulting in suboptimal performance. VLP-based methods typically rely on static, input-agnostic prompts to compen-

*Corresponding author.

sate for missing information, which offer limited contextual knowledge and restrict adaptability under mixed missing-modality scenarios. Benefiting from advances in retrieval-augmented [1] and cross-modal generation [30] techniques, modality reconstruction methods have attracted substantial attention and become mainstream in incomplete MML.

Despite the promising results, modality reconstruction methods often overlook an important aspect: *incomplete MML suffers from learning imbalance under high missing ratios.* In such cases, reconstructed samples tend to dominate the optimization process due to the amplified semantic noise they carry, which biases model updates and weakens the representations learned from complete samples. To support our viewpoint, we conduct a toy experiment on the HateMemes dataset [13] under a 70% text-missing condition to examine the performance of complete and reconstructed samples. Following a general retrieval-based reconstruction paradigm (*e.g.*, RAGPT [15]), the missing modality is reconstructed by retrieving semantically similar instances from a memory bank using the available modality (Figure 1a). The results in Figure 1c reveal a striking phenomenon with RAGPT: a large performance gap exists between complete and reconstructed samples. Ideally, both groups should approach their respective upper bounds, which are obtained by training on each subset independently. However, we observe a substantial deviation from this expectation. The performance of complete samples falls markedly below its upper bound, whereas reconstructed samples remain relatively competitive. This finding confirms the existence of learning imbalance and motivates us to design a balanced optimization framework aimed at bridging the performance gap.

To address the aforementioned problem, we propose a balanced optimization framework named ANchor-guided Gradient Alignment (ANGA) for incomplete MML (Figure 1b). Specifically, ANGA constructs an optimization anchor from complete samples as a stable reference for model updates and aligns the gradients of reconstructed samples toward this anchor to mitigate noise-induced bias. To improve the representativeness of the optimization anchor under high missing ratios, we adopt an entropy-driven curriculum [29, 35] that progressively incorporates reliable reconstructed samples. Via this gradient alignment, ANGA effectively rebalances the learning process between complete and reconstructed samples, leading to more consistent performance (Figure 1c). Moreover, since optimization-level balancing alone cannot fully address semantic insufficiency, we introduce a semantic-enhanced adapter that leverages retrieved instances to generate dynamic, context-aware prompts, further enhancing model robustness. To sum up, our contributions are outlined as follows:

- We identify and analyze the learning imbalance problem in existing modality reconstruction methods, which significantly degrades overall performance.
- We propose ANGA to rebalance the optimization between complete and reconstructed samples via gradient alignment, which can also be seamlessly integrated into existing incomplete MML frameworks.
- Extensive experiments demonstrate that ANGA consistently outperforms SOTA baselines across widely used benchmark datasets.

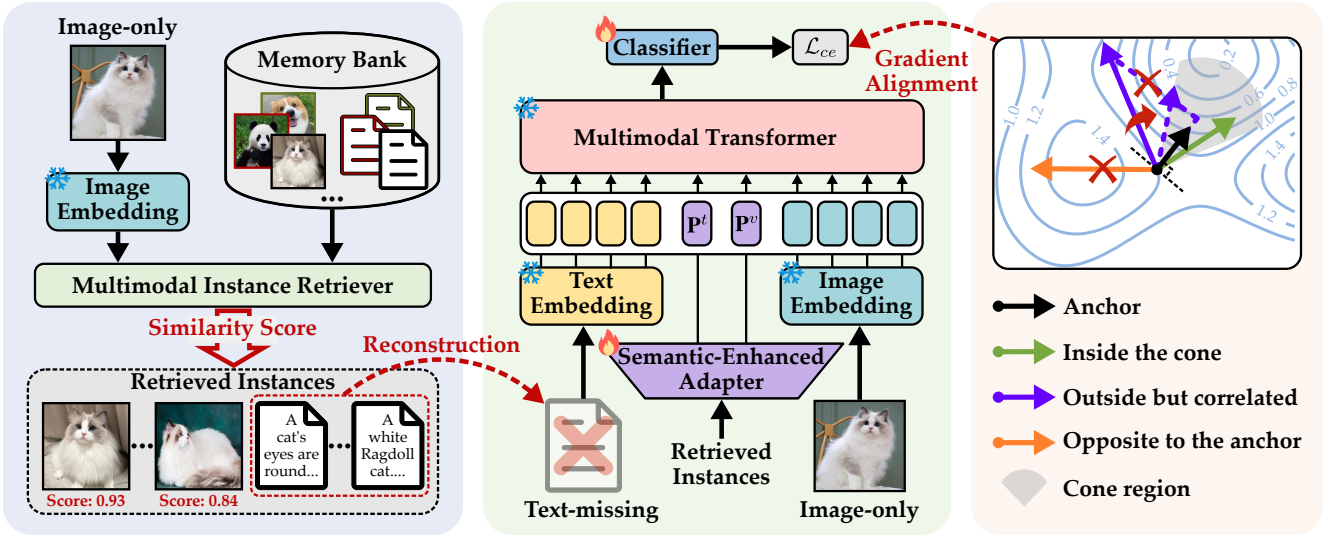
2. Related Work

2.1. Incomplete Multimodal Learning

Incomplete MML aims to achieve robust prediction when certain modalities are missing [12, 17, 32, 42]. Such situations frequently occur in real-world applications, where sensor failures or privacy constraints can cause incomplete multimodal inputs [2, 28, 36]. Existing studies have explored several directions to mitigate the resulting performance degradation. One line of work focuses on modality-invariant learning [18, 56], which extracts inter-modal correlations and maps multimodal features into a shared semantic space, thereby enhancing model robustness under missing-modality conditions. A second direction leverages large-scale pre-training and applies prompt tuning to transfer cross-modal knowledge from VLP models to incomplete multimodal tasks [11, 16, 45]. A third line of research emphasizes modality reconstruction, which restores missing information to improve the quality of joint representations, often through generative [37, 40] or retrieval-based [15, 47] strategies. However, semantic noise in reconstructed samples may cause optimization drift, leading to learning imbalance. To address this issue, we propose an anchor-guided gradient alignment strategy to stabilize optimization between complete and reconstructed samples.

2.2. Learning Imbalance

Learning imbalance is a broader concept that encompasses both implicit and explicit forms of imbalance in MML. The implicit form appears in complete MML scenarios, where modalities with faster convergence tend to dominate the optimization process [33, 52], resulting in modality imbalance [5, 7, 26, 39, 54]. The explicit form occurs in incomplete MML scenarios, where the reconstruction of missing modalities introduces semantic noise that amplifies optimization drift and leads to learning imbalance, which is the main focus of this work. Although prior studies have attempted to mitigate imbalance from various perspectives, such as learning objectives [38, 48], architecture design [3], and optimization strategies [26, 50], these efforts have primarily aimed to balance modality contributions in complete settings. To the best of our knowledge, this is the first attempt to investigate learning imbalance under missing-modality conditions from an optimization perspective.



(a) Multimodal Instance Retriever **(b) Vision-Language Pre-training** **(c) Gradient Alignment**

Figure 2. Illustration of ANGA: **(a).** The multimodal instance retriever identifies the Top-K most similar instances to reconstruct the missing modality. **(b).** The VLP model learns joint representations for final prediction, where a semantic-enhanced adapter generates dynamic prompts to enhance robustness. **(c).** The gradient alignment constructs an optimization anchor to address the learning imbalance problem during training.

3. Preliminaries

In this section, we first define the problem of incomplete MML, then describe the vision-language pre-training framework, and finally present the general modality reconstruction paradigm.

3.1. Problem Definition

Without loss of generality, we consider a multimodal sample consisting of two modalities, t and v (*e.g.*, text and image). Formally, let $\mathcal{D} = \{\mathcal{D}^c, \mathcal{D}^m\}$ denote the dataset, where $\mathcal{D}^c = \{\boldsymbol{x}_i^{(t)}, \boldsymbol{x}_i^{(v)}, \boldsymbol{y}_i\}_{i=1}^{N^c}$ represents the complete-modality subset with N^c training samples, and $\boldsymbol{y}_i \in \{0,1\}^k$ denotes the category labels with a total of k categories. In contrast, $\mathcal{D}^m = \{(\boldsymbol{x}_i^{(t)}, \boldsymbol{y}_i) \vee (\boldsymbol{x}_i^{(v)}, \boldsymbol{y}_i)\}_{i=1}^{N^m}$ corresponds to the missing-modality subset, which contains N^m training samples where only one modality is available. The objective of incomplete MML is to achieve robust multimodal prediction even when certain modalities are missing during both training and inference phases.

3.2. Vision-Language Pre-training

Vision-language pre-training (VLP) aligns visual and textual representations within a shared embedding space. A typical VLP model (*e.g.*, ViLT [14]) first maps text $\boldsymbol{x}_i^{(t)}$ and image $\boldsymbol{x}_i^{(v)}$ into token sequences $\mathcal{T}_i \in \mathbb{R}^{n_t \times d}$ and $\mathcal{V}_i \in \mathbb{R}^{n_v \times d}$ via modality-specific embedding layers, where n_t and n_v denote the corresponding sequence lengths and d is the embedding dimension. The concatenated tokens are then encoded by a multimodal Transformer with B consecutive Multi-head Self-Attention (MSA) [31] layers, yielding a joint representation $\boldsymbol{e}_i \in \mathbb{R}^d$, which is subsequently fed into a classifier for final prediction:

$$\boldsymbol{e}_i = \texttt{Transformer}([\mathcal{T}_i; \mathcal{V}_i]), \quad \hat{\boldsymbol{y}}_i = \sigma(\boldsymbol{W}\boldsymbol{e}_i + \boldsymbol{b}). \quad (1)$$

Here, $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ and $\boldsymbol{b} \in \mathbb{R}^k$ denote the classifier weights and bias, and σ is the softmax function. The objective function is formulated as:

$$\mathcal{L}_{ce} = -\frac{1}{N} \sum\nolimits_{i=1}^{N} \boldsymbol{y}_i^\top \log \hat{\boldsymbol{y}}_i, \quad (2)$$

where $N = N^c + N^m$ is the training sample size.

To efficiently adapt VLP models to downstream tasks, prompt tuning [11, 16, 25] introduces a small set of n_p learnable prompts $\mathbf{P} \in \mathbb{R}^{n_p \times d}$ into the multimodal Transformer, allowing task-specific adaptation while keeping the pre-trained backbone frozen. Let $\mathbf{E}_b \in \mathbb{R}^{L \times d}$, $b = 1, 2, \ldots, B$, denote the input embeddings to the b-th MSA layer, where L is the sequence length. Specifically, $\mathbf{E}_1$ is formed by concatenating $\mathcal{T}_i$ and $\mathcal{V}_i$. The prompts are prepended to the input embeddings as $\mathbf{E}_b^p = [\mathbf{P}; \mathbf{E}_b] \in \mathbb{R}^{(n_p+L) \times d}$, which are fed into the b-th layer and propagated through subsequent layers. The prediction follows the same formulation as Eq. (1), using the final embedding $\mathbf{E}_B^p$. To balance efficiency and effectiveness, the prompts are applied only to selected layers (*e.g.*, the b-th layer) rather than inserted into every layer.

3.3. Modality Reconstruction

To handle incomplete multimodal inputs, prior works [6, 15] commonly adopt a Multimodal Instance Retriever (MIR) that reconstructs missing modalities (Figure 2a). To support this idea, a memory bank $\mathcal{M}$ is built to store unimodal representations as prior knowledge.

Taking the missing-image case as an example, the text token sequences $\mathcal{T}_i$ are encoded by the CLIP textual encoder [27] $\Phi^{(t)}(\cdot)$ to obtain the text embedding $\boldsymbol{e}_i^{(t)} = \Phi^{(t)}(\mathcal{T}_i) \in \mathbb{R}^d$. Subsequently, the MIR employs $\boldsymbol{e}_i^{(t)}$ as a query to retrieve the Top-K most similar instances $\mathcal{N}_i$ from the memory bank $\mathcal{M}$ based on cosine similarity:

$$\mathcal{N}_i = \underset{r \in \mathcal{M}}{\text{Top-}K} \left(\frac{\boldsymbol{e}_i^{(t)\top} \boldsymbol{e}_r^{(t)}}{\|\boldsymbol{e}_i^{(t)}\|_2 \|\boldsymbol{e}_r^{(t)}\|_2} \right). \tag{3}$$

Here, $\|\cdot\|_2$ denotes the L_2 norm, and $\boldsymbol{e}_r^{(t)}$ is the text embedding stored in $\mathcal{M}$. The corresponding modality in $\mathcal{N}_i$ then serves as a semantic reference for reconstructing the missing modality via a mean pooling operation.

For the missing-text case, the image token sequences $\mathcal{V}_i$ are encoded by the CLIP visual encoder [27] $\Phi^{(v)}(\cdot)$ to obtain the image embedding $\boldsymbol{e}_i^{(v)} \in \mathbb{R}^d$, and the retrieval process is performed in the same manner as Eq. (3).

4. Methodology

Motivated by the observation that incomplete MML suffers from learning imbalance under high missing-modality ratios, we propose ANGA, which addresses this issue through two key components: anchor-guided gradient alignment, which balances optimization between complete and reconstructed samples, and a semantic-enhanced adapter, which performs contextual enrichment to improve model robustness. The overall framework of ANGA is shown in Figure 2.

4.1. Optimization Anchor

Gradient decoupling: Under high missing ratios, noisy reconstructed samples tend to dominate the optimization process, leading to learning imbalance. To address this, we decouple the optimization signals by computing gradients for complete and reconstructed samples independently within each mini-batch $\mathcal{B}$:

$$\mathbf{g}_C = \nabla_\theta \frac{1}{|C|} \sum_{i \in C} \ell_i, \quad \mathbf{g}_M = \nabla_\theta \frac{1}{|M|} \sum_{i \in M} \ell_i, \tag{4}$$

where ℓ_i denotes the per-sample loss (consistent with Eq. (2)), and ∇_θ denotes the gradient with respect to model parameters. C and M represent the complete and reconstructed subsets within $\mathcal{B}$. This decoupling allows $\mathbf{g}_C$ to serve as a clean and basic optimization reference.

Reconstructed sample evaluation: To improve the anchor's representativeness without sacrificing stability, we incorporate a portion of reconstructed samples that are deemed reliable. For each reconstructed sample $\tilde{\boldsymbol{x}}_i = \{\tilde{\boldsymbol{x}}_i^{(t)}, \boldsymbol{x}_i^{(v)}\}$ or $\{\boldsymbol{x}_i^{(t)}, \tilde{\boldsymbol{x}}_i^{(v)}\}$, its reliability is evaluated using prediction entropy, defined as:

$$H(\tilde{\boldsymbol{x}}_i) = -\sum_{j=1}^{k} \hat{y}_{i,j} \log \hat{y}_{i,j}, \tag{5}$$

where $\hat{y}_{i,j}$ is the softmax probability of class j for sample i. Reconstructed samples with lower entropy exhibit higher confidence and are thus preferred for anchor construction.

Anchor construction: Incorporating reliable reconstructed samples can enrich the anchor, but merging them all at once may cause instability as the model's confidence is still evolving. To ensure smooth adaptation, we employ a curriculum strategy [29, 35] that regulates the amount and pace of introducing reliable reconstructions.

Specifically, we first rank all reconstructed samples in ascending order of entropy according to Eq. (5), yielding $\tilde{\mathcal{D}}^m_{\text{rank}}$. We then apply a widely used linear function $\lambda(\cdot)$ to implement this curriculum, formulated as:

$$\lambda(z) = \texttt{min}\left(\lambda_{\max}, \lambda_{\min} + \frac{\lambda_{\max} - \lambda_{\min}}{Z_{\text{grow}}} \cdot z\right), \tag{6}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ denote the lower and upper bounds of the sample ratio, and Z_{grow} is the number of epochs in one curriculum stage. At each epoch z, the top λ fraction of $\tilde{\mathcal{D}}^m_{\text{rank}}$ is progressively incorporated into the anchor set:

$$\mathcal{A}_z = \mathcal{C} \cup \{\tilde{\boldsymbol{x}}_i \mid \tilde{\boldsymbol{x}}_i \in \tilde{\mathcal{D}}^m_{\text{rank}}, i \le \lfloor \lambda(z) \cdot N^m \rfloor\}. \tag{7}$$

Here, $\mathcal{C}$ denotes the overall complete-modality subset over the entire training set, and N^m is the total number of reconstructed samples. Finally, the corresponding anchor gradient within each mini-batch $\mathcal{B}$ is computed as:

$$\mathbf{g}_A = \nabla_\theta \frac{1}{|\mathcal{A}_z \cap \mathcal{B}|} \sum_{i \in (\mathcal{A}_z \cap \mathcal{B})} \ell_i. \tag{8}$$

This batch-wise optimization anchor provides a dynamic yet stable reference for subsequent gradient alignment.

4.2. Gradient Alignment

Given the anchor $\mathbf{g}_A$ defined in Eq. (8), we employ it as a directional reference to guide the optimization of the remaining reconstructed gradient $\mathbf{g}_{M'}$ (where $M' = M - (\mathcal{A}_z \cap \mathcal{B})$) that is not included in the anchor set $\mathcal{A}_z$. To stabilize optimization and address learning imbalance, we perform gradient alignment, where consistent gradients are preserved, deviated ones are projected toward the anchor direction, and conflicting ones are suppressed to prevent interference from noisy updates.

Formally, for each mini-batch $\mathcal{B}$, we compute the cosine similarity between $\mathbf{g}_{M'}$ and $\mathbf{g}_A$, denoted as $\texttt{sim}(\mathbf{g}_{M'}, \mathbf{g}_A)$

following the same cosine formulation as in Eq. (3) but applied in the gradient space. Based on this similarity, $\mathbf{g}_{M'}$ is modulated according to its position within the anchor's cone region, defined by a cosine threshold $\tau = \cos\theta$ with half-angle $\theta \in (0, \frac{\pi}{2})$. As illustrated in Figure 2(c), three distinct cases are considered:

Inside the cone: When the reconstructed gradient $\mathbf{g}_{M'}$ lies inside the cone region, *i.e.*, $\texttt{sim}(\mathbf{g}_{M'}, \mathbf{g}_A) \geq \tau$, we preserve it without modification since it is directionally consistent with the anchor:

$$\tilde{\mathbf{g}}_{M'} \leftarrow \mathbf{g}_{M'}. \tag{9}$$

Outside but correlated: When the reconstructed gradient $\mathbf{g}_{M'}$ lies outside the cone but remains positively correlated with the anchor, *i.e.*, $0 < \texttt{sim}(\mathbf{g}_{M'}, \mathbf{g}_A) < \tau$, we decompose $\mathbf{g}_{M'}$ into parallel and orthogonal components:

$$\mathbf{g}^{\parallel}_{M'} = \frac{\mathbf{g}_{M'} \cdot \mathbf{g}_A}{\|\mathbf{g}_A\|_2^2}\, \mathbf{g}_A, \quad \mathbf{g}^{\perp}_{M'} = \mathbf{g}_{M'} - \mathbf{g}^{\parallel}_{M'}. \tag{10}$$

To minimally adjust the gradient direction, we retain the parallel component while proportionally shrinking the orthogonal one. With $\theta = \arccos\tau$, the maximal lateral ratio at the cone boundary is given by:

$$\kappa_{\max} = \tan\theta = \frac{\sqrt{1-\tau^2}}{\tau}, \tag{11}$$

and the corresponding scaling factor is computed as:

$$\alpha = \frac{\kappa_{\max}\, \|\mathbf{g}^{\parallel}_{M'}\|_2}{\|\mathbf{g}^{\perp}_{M'}\|_2}. \tag{12}$$

Here, $\alpha < 1$ ensures that the orthogonal component is reduced to fit within the cone boundary. The modulated gradient is then obtained as:

$$\tilde{\mathbf{g}}_{M'} \leftarrow \mathbf{g}^{\parallel}_{M'} + \alpha\, \mathbf{g}^{\perp}_{M'}. \tag{13}$$

By construction, $\texttt{sim}(\tilde{\mathbf{g}}_{M'}, \mathbf{g}_A) = \tau$, ensuring that $\tilde{\mathbf{g}}_{M'}$ lies exactly on the safe-cone boundary and constraining gradient deviation while preserving the anchor-parallel component.

Opposite to the anchor: When the reconstructed gradient $\mathbf{g}_{M'}$ produces conflicting optimization signals with respect to the anchor, *i.e.*, $\texttt{sim}(\mathbf{g}_{M'}, \mathbf{g}_A) < 0$, it may cause optimization drift. We therefore suppress this gradient to prevent interference with the anchor gradient $\mathbf{g}_A$:

$$\tilde{\mathbf{g}}_{M'} \leftarrow \mathbf{0}. \tag{14}$$

After obtaining the aligned reconstructed gradient $\tilde{\mathbf{g}}_{M'}$, the model parameters θ are updated using the aggregated gradient, formulated as:

$$\theta \leftarrow \theta - \eta\,(\mathbf{g}_C + \tilde{\mathbf{g}}_{M'}), \tag{15}$$

where η denotes the learning rate of the optimizer.

4.3. Semantic-Enhanced Adapter

To explicitly enrich contextual knowledge and enhance the robustness of ANGA under missing-modality scenarios, we propose a Semantic-Enhanced Adapter (SEA). For each target instance, including both complete and reconstructed samples, SEA generates dynamic, context-aware prompts from the retrieval pool $\mathcal{N}_i$. Specifically, given the target text sequences $\mathcal{T}_i \in \mathbb{R}^{n_t \times d}$, SEA performs cross-attention [31] where $\mathcal{T}_i$ serves as the query and the Top-K retrieved text sequences $\mathcal{T}_i^{\mathcal{R}} = \{\mathcal{T}_i^{r_j}\}_{j=1}^{K} \in \mathbb{R}^{K \times n_t \times d}$ serve as the key and value. This operation yields textual-aware prompts $\tilde{\mathbf{P}}_i^t \in \mathbb{R}^{n_t \times d}$, formulated as:

$$\begin{aligned} \tilde{\mathbf{P}}_i^t &= \texttt{CrossAttn}\Big(f_t^Q(\mathcal{T}_i), f_t^K(\mathcal{T}_i^{\mathcal{R}}), f_t^V(\mathcal{T}_i^{\mathcal{R}})\Big) \\ &= \texttt{softmax}\left(\frac{f_t^Q(\mathcal{T}_i) f_t^K(\mathcal{T}_i^{\mathcal{R}})^\top}{\sqrt{d}}\right) f_t^V(\mathcal{T}_i^{\mathcal{R}}), \end{aligned} \tag{16}$$

where $f_t^Q(\cdot)$, $f_t^K(\cdot)$, and $f_t^V(\cdot)$ denote the respective projection functions (*e.g.*, linear layers), and $\sqrt{d}$ is the scaling factor in dot-product attention.

For visual-aware prompts, SEA applies the same operation as in Eq. (16) between the target image sequence $\mathcal{V}_i \in \mathbb{R}^{n_v \times d}$ and its retrieved counterparts $\mathcal{V}_i^{\mathcal{R}} = \{\mathcal{V}_i^{r_j}\}_{j=1}^{K} \in \mathbb{R}^{K \times n_v \times d}$, generating $\tilde{\mathbf{P}}_i^v \in \mathbb{R}^{n_v \times d}$. Finally, an adaptive pooling operation produces the dynamic prompts $\mathbf{P}_i^t, \mathbf{P}_i^v \in \mathbb{R}^d$ for each target instance, which are injected into the MSA layers of the multimodal Transformer (Figure 2b).

5. Experiments

In this section, we present comprehensive experiments on three datasets to compare ANGA with SOTA baselines. Ablation studies and further analyses are conducted to demonstrate the effectiveness and robustness of our method.

5.1. Experimental Setup

Datasets: Following prior works [11, 15, 16], we evaluate our ANGA on three benchmark datasets: (1). HateMemes [13], a hateful meme detection benchmark with 10,000 image-text pairs. (2). MM-IMDb [24], a movie genre classification dataset with 25,959 pairs, formulated as a multi-label task since each movie may belong to multiple genres. (3). Food-101 [34], a large-scale food classification dataset with 90,688 image-text pairs across 101 categories.

Comparison methods: We compare our ANGA with three types of SOTA baselines: (1). Modality-invariant learning methods: IF-MMIN [56], ShaSpec [32], DrFuse [51], CorrKD [18], and MoMKE [44]. (2). VLP-based methods: ViLT [14], MAPs [16], MSPs [11], and IPer [22]. (3). Modality reconstruction methods (including both generative- and retrieval-based approaches): AcMAE [40], DiCMoR [37], MMDS [47], and RAGPT [15].

Table 1. Comparison with SOTA baselines under a 70% missing rate across various missing-modality scenarios. The best results are highlighted in **bold**, and the second-best ones are underlined. Higher AUROC, F1-Micro, and ACC indicate better performance.

Dataset →	HateMemes			MM-IMDb			Food101		
Metric →	AUROC (%)	AUROC (%)	AUROC (%)	F1-Micro (%)	F1-Micro (%)	F1-Micro (%)	ACC (%)	ACC (%)	ACC (%)
Method ↓	Text	Image	Both	Text	Image	Both	Text	Image	Both
Modality-invariant learning methods									
IF-MMIN [56]	57.62	53.44	55.19	39.63	31.95	31.98	66.76	64.36	68.53
ShaSpec [32]	58.75	60.30	60.96	44.04	44.23	44.06	60.99	74.87	70.02
DrFuse [51]	57.60	60.66	55.84	47.05	43.58	48.83	66.30	75.09	68.23
CorrKD [18]	58.74	55.59	57.91	44.82	39.48	41.20	61.37	66.83	62.87
MoMKE [44]	63.08	61.35	62.53	50.98	45.67	46.99	66.85	68.40	67.38
VLP-based methods									
ViLT [14]	56.61	57.15	56.42	45.33	41.14	40.23	62.92	70.59	65.01
MAPs [16]	58.62	60.16	58.89	46.12	44.86	45.48	67.02	75.62	72.52
MSPs [11]	59.60	60.05	59.08	49.16	44.62	48.28	71.74	79.09	74.46
IPer [22]	62.09	61.68	61.34	53.56	45.82	47.62	72.56	80.23	73.73
Modality reconstruction methods									
AcMAE [40]	55.74	59.66	57.25	47.47	43.82	44.05	69.28	73.75	71.15
DiCMoR [37]	58.03	60.27	59.38	47.02	45.19	46.68	70.47	76.54	73.22
MMDS [47]	57.14	59.72	59.68	47.47	42.17	42.44	61.15	72.52	63.76
RAGPT [15]	64.10	62.57	63.47	55.16	46.44	50.89	75.53	81.98	76.94
ANGA (Ours)	**68.54**	**63.42**	**65.12**	**57.28**	**47.75**	**51.84**	**77.23**	**83.03**	**78.47**

Evaluation metrics: Following [11, 15, 16], we adopt AUROC, F1-Micro, and accuracy (ACC) as evaluation metrics. AUROC measures a model's ability to distinguish positive from negative samples. F1-Micro is computed from the aggregated true positives, false positives, and false negatives across all categories, while ACC denotes the proportion of correct predictions.

Setting of missing patterns: We define the missing rate δ as the proportion of samples with missing modalities relative to the entire dataset. For each dataset, we consider two widely used missing patterns: (1). Single-modality missing, where a missing rate of δ indicates that δ of the samples are image-only or text-only, while the remaining $(1-\delta)$ are complete. (2). Both-modality missing, where $\frac{\delta}{2}$ of the samples are image-only, $\frac{\delta}{2}$ are text-only, and the remaining $(1-\delta)$ are complete. Unless otherwise specified, we set $\delta = 70\%$ in all experiments.

Implementation details: Following [11, 15, 16], we adopt the pre-trained ViLT [14] as the backbone. For each task, the memory bank $\mathcal{M}$ is constructed from the corresponding training set, and the number of retrieved instances K is chosen from $\{1, 3, 5, 7, 9\}$. For anchor construction, the curriculum stage Z_{grow} and the lower and upper bounds of the sample ratio, λ_{min} and λ_{max}, are set to 5, 0.1, and 0.3, respectively. For prompt design, we insert the prompt into the second MSA layer of ViLT (*i.e.*, $b = 2$). The cosine threshold τ is set to 0.2. We train the model using the AdamW [21] with a learning rate of $\eta = 10^{-3}$, a weight decay of 10^{-5}, and a batch size of 64. All experiments are conducted on an NVIDIA GeForce RTX 4090 GPU.

5.2. Comparison with SOTA Baselines

We conduct comprehensive comparisons to evaluate the effectiveness of ANGA in addressing the incomplete MML problem. Based on the classification results under a 70% missing rate across all datasets, as reported in Table 1, we summarize the following key observations:

Firstly, both modality-invariant learning and VLP-based methods achieve competitive results but still fall short of ANGA due to their inherent limitations. For instance, modality-invariant methods tend to overemphasize shared representations while overlooking modality-specific information, which ultimately weakens discriminative performance. In contrast, VLP-based methods generally outperform the vanilla Transformer (*i.e.*, ViLT [14]), verifying the benefit of prompt learning in improving model robustness. Nevertheless, as the prompts used in these methods are typically input-agnostic, they fail to incorporate instance-specific knowledge, thereby limiting their adaptability under missing-modality conditions.

Secondly, modality reconstruction methods typically recover the missing modality through generation [37, 40] or retrieval [15, 47]. However, these methods often introduce semantic noise, which amplifies optimization drift and results in learning imbalance under high missing rates. In contrast, ANGA constructs an optimization anchor that aligns the gradients of reconstructed samples with those of complete ones, effectively suppressing this drift and improving overall performance.

Lastly, compared with all three types of baselines, our ANGA consistently achieves superior performance across

Table 2. Ablation study of different components in ANGA under the 70% text-missing condition.

Variant			HateMemes	MM-IMDb	Food101
MIR	GA	SEA	AUROC (%)	F1-Micro (%)	ACC (%)
✗	✗	✗	59.38	49.62	68.53
✓	✗	✗	64.63	54.17	74.86
✓	✓	✗	66.82	<u>56.32</u>	<u>77.01</u>
✓	✗	✓	<u>67.31</u>	55.74	76.89
✓	✓	✓	**68.54**	**57.28**	**77.23**

all evaluation metrics. In particular, ANGA yields significant improvements on both the HateMemes and Food101 datasets. Through gradient alignment, ANGA surpasses the best baseline RAGPT [15] by 4.44%/0.85%/1.65% in AUROC and 1.70%/1.05%/1.53% in ACC across text-, image-, and both-missing scenarios, respectively. These consistent gains confirm the effectiveness of ANGA in tackling the learning imbalance challenge, which has been largely overlooked by previous incomplete MML approaches.

5.3. Ablation Study

We conduct ablation studies to verify the effectiveness of each component in ANGA under a 70% text-missing condition. The results presented in Table 2 reveal the following observations: (1). Without the multimodal instance retriever (MIR), the missing modality is replaced with dummy values, leading to severe information deficiency and the worst performance across all datasets. By retrieving semantically similar instances, MIR effectively compensates for the missing information and yields notable improvements (*e.g.*, +5.25% AUROC on HateMemes). (2). Based on MIR, gradient alignment (GA) refines the optimization by aligning the gradients of reconstructed samples with those of reliable ones. This process alleviates optimization drift caused by semantic noise and helps maintain balanced learning between complete and reconstructed samples. GA brings additional improvements over MIR alone (*e.g.*, +2.19% AUROC on HateMemes), confirming its effectiveness in stabilizing training under severe missing-modality conditions. (3). The semantic-enhanced adapter (SEA) generates dynamic prompts from retrieved samples, enabling the model to incorporate instance-related contextual knowledge and thereby further improve model robustness. By integrating MIR, GA, and SEA components, ANGA achieves the best overall performance, highlighting their complementary contributions to incomplete MML.

5.4. Further Analysis

Sensitivity to hyperparameters: In calibrating our ANGA, we identify two key hyperparameters: the number of retrieved instances K in Eq. (3) and the cosine threshold τ that defines the cone region. Results in Figure 3 on the

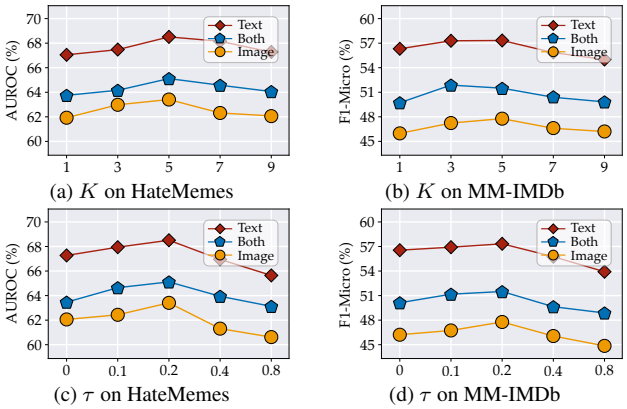


Figure 3. Sensitivity to hyperparameters K and τ on HateMemes and MM-IMDb across various missing-modality scenarios.

HateMemes and MM-IMDb datasets lead to two observations: (1). The performance of our ANGA is only marginally affected by K, highlighting its insensitivity to the retrieval cardinality. Despite some fluctuations, ANGA still demonstrates excellent effectiveness, *i.e.*, being consistently better than the dummy-padding baseline (Variant-1 in Table 2). (2). As τ increases, performance first improves and then declines. This indicates that a moderately sized cone region helps alleviate the learning imbalance, but over-considering gradient alignment may hinder model generalization.

Robustness to varying missing rates: To assess the robustness of ANGA under varying degrees of modality absence, we conduct experiments on the HateMemes dataset across text- and both-missing scenarios. We compare ANGA with three representative SOTA baselines: IF-MMIN, MSPs, and RAGPT. As shown in Figure 4a and Figure 4b, ANGA consistently outperforms all baselines across a wide range of missing rates, achieving the highest AUROC. Figures 4c and 4d further illustrate the relative performance degradation with respect to the complete-modality condition. Notably, ANGA exhibits the smallest degradation at every missing rate, demonstrating its strong robustness to incomplete MML. These results indicate that ANGA leverages dynamic prompts guided by retrieved instances to effectively alleviate the impact of missing information, thereby maintaining robust overall performance.

Generalizability to severe missing conditions: We evaluate the generalizability of ANGA by training with missing ratios ranging from 10% to 50% and testing under a 90% missing rate. The experiments are conducted on the HateMemes dataset, comparing ANGA with four SOTA baselines: IF-MMIN, MSPs, AcMAE, and RAGPT. As shown in Figure 5a and Figure 5b, increasing the missing ratio during training consistently improves performance for all models across text- and both-missing scenarios, suggesting that exposure to more incomplete data enhances generalizability to

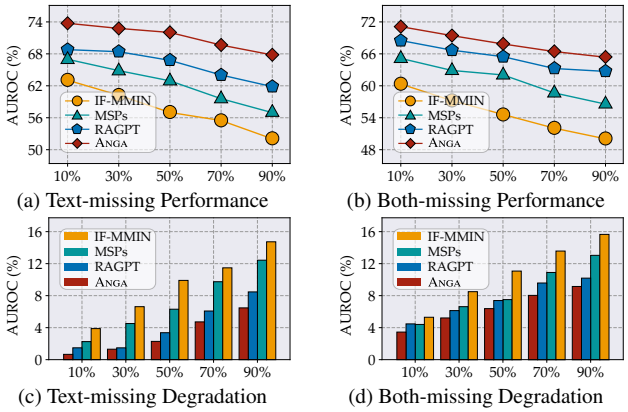


Figure 4. Robustness of ANGA and baselines on HateMemes across various missing-modality scenarios. **(a).** and **(b).** show the performance under different missing rates, while **(c).** and **(d).** illustrate the relative performance degradation with respect to the complete-modality condition.

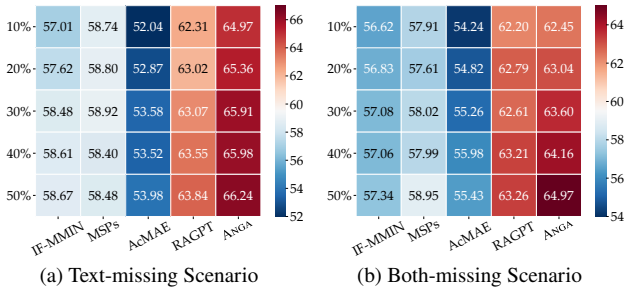


Figure 5. Generalizability of ANGA and baselines on HateMemes under various training missing rates and a 90% testing missing rate, evaluated by AUROC.

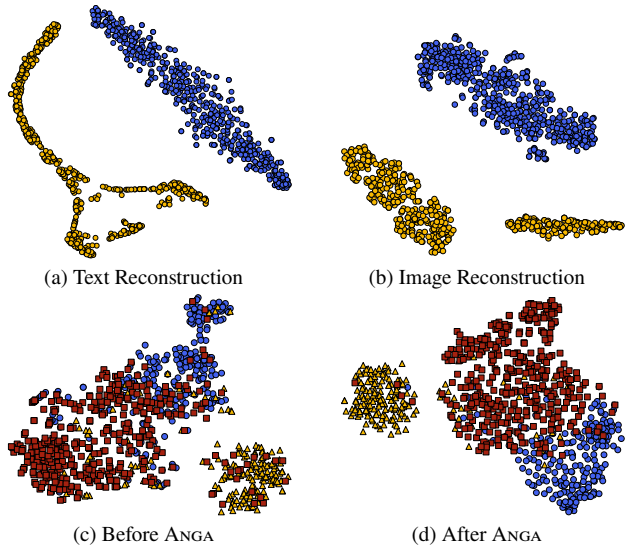


Figure 6. t-SNE visualization: **(a).** and **(b).** illustrate the reconstructed (yellow points) and ground-truth (blue points) modalities for text and image on HateMemes. **(c).** and **(d).** show the final-layer embeddings for the Sport (●), Thriller (▲), and Musical (■) genres on MM-IMDb.

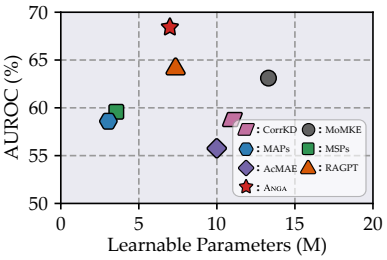


Figure 7. Comparison of learnable parameters (M) on the HateMemes dataset under a 70% text-missing condition.

severe missing-modality conditions at inference time. Remarkably, ANGA achieves the best overall performance, benefiting from its gradient alignment strategy, which effectively addresses learning imbalance and promotes stable optimization under high missing ratios.

Visualizations: We visualize the learned embedding spaces under a 70% missing-modality condition to illustrate the effectiveness of ANGA, and conduct two types of analysis: (1). *Reconstruction Visualization.* Figure 6a and Figure 6b show the t-SNE [23] distributions of the reconstructed and ground-truth modalities, where 500 testing samples from HateMemes are randomly selected for visualization. As observed, the reconstructed text and image embeddings still deviate from their ground-truth counterparts, revealing a clear semantic gap [20] under a general retrieval-based reconstruction paradigm (*e.g.*, RAGPT [15]). (2). *Decision Space Visualization.* Figure 6c and Figure 6d present the t-SNE plots of the final-layer embeddings for three movie genres (Sport, Thriller, and Musical) from the MM-IMDb testing set. While the embedding space before ANGA shows entangled representations with poor class separability, applying ANGA yields compact and well-separated clusters.

Efficiency comparison: We analyze the number of learnable parameters on the HateMemes dataset. As shown in Figure 7, ANGA achieves the best performance while maintaining competitive parameter efficiency, demonstrating its effectiveness in incomplete MML.

6. Conclusion

In this paper, we propose a novel framework for incomplete multimodal learning, termed ANchor-guided Gradient Alignment (ANGA). By retrieving semantically similar instances, we reconstruct missing modalities and generate dynamic prompts to incorporate retrieved knowledge, effectively alleviating information deficiency. Building upon this, an entropy-driven curriculum integrates reliable reconstructed samples with complete ones into an optimization anchor, which guides gradient alignment and addresses learning imbalance under severe missing-modality conditions. Extensive experiments verify ANGA's effectiveness.

Acknowledgements

This work was supported in part by the NSFC (62276131), and in part by the Natural Science Foundation of Jiangsu Province of China under Grant (BK20240081).

References

[1] Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. In *ACL*, pages 16776–16809, 2025. 2

[2] Yuzhuo Dai, Jiaqi Jin, Zhibin Dong, Siwei Wang, Xinwang Liu, En Zhu, Xihong Yang, Xinbiao Gan, and Yu Feng. Imputation-free and alignment-free: Incomplete multi-view clustering driven by consensus semantic learning. In *CVPR*, pages 5071–5081, 2025. 2

[3] Chenzhuang Du, Tingle Li, Yichen Liu, Zixin Wen, Tianyu Hua, Yue Wang, and Hang Zhao. Improving multi-modal learning with uni-modal teachers. *CoRR*, abs/2106.11059, 2021. 2

[4] Siyuan Duan, Yuan Sun, Dezhong Peng, Zheng Liu, Xiaomin Song, and Peng Hu. Fuzzy multimodal learning for trusted cross-modal retrieval. In *CVPR*, pages 20747–20756, 2025. 1

[5] Xiyuan Gao, Bing Cao, Pengfei Zhu, Nannan Wang, and Qinghua Hu. Asymmetric reinforcing against multi-modal representation bias. In *AAAI*, pages 16754–16762, 2025. 2

[6] Xiyuan Gao, Bing Cao, Pengfei Zhu, Nannan Wang, and Qinghua Hu. Synergistic prompting for robust visual recognition with missing modalities. In *ICCV*, pages 1881–1890, 2025. 1, 4

[7] Zhi-Hao Guan, Qing-Yuan Jiang, and Yang Yang. Balance-aware sequence sampling makes multi-modal learning better. In *IJCAI*, pages 2838–2846, 2025. 2

[8] Zhihao Guan, Jia-Qi Yang, Yang Yang, Hengshu Zhu, Wenjie Li, and Hui Xiong. Jobformer: Skill-aware job recommendation with semantic-enhanced transformer. *TKDD*, 19(1): 18:1–18:20, 2025. 1

[9] Lianyu Hu, Tongkai Shi, Wei Feng, Fanhua Shang, and Liang Wan. Deep correlated prompting for visual recognition with missing modalities. In *NeurIPS*, 2024. 1

[10] Yuchen Hu, Ruizhe Li, Chen Chen, Heqing Zou, Qiushi Zhu, and Eng Siong Chng. Cross-modal global interaction and local alignment for audio-visual speech recognition. In *IJCAI*, pages 5076–5084, 2023. 1

[11] Jaehyuk Jang, Yooseung Wang, and Changick Kim. Towards robust multimodal prompting with missing modalities. In *ICASSP*, pages 8070–8074, 2024. 1, 2, 3, 5, 6

[12] Guanzhou Ke, Shengfeng He, Xiaoli Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Yi Xie, and Hexing Su. Knowledge bridger: Towards training-free missing modality completion. In *CVPR*, pages 25864–25873, 2025. 2

[13] 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 *NeurIPS*, 2020. 2, 5

[14] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, pages 5583–5594, 2021. 3, 5, 6

[15] Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning. In *AAAI*, pages 18035–18043, 2025. 1, 2, 4, 5, 6, 7, 8

[16] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *CVPR*, pages 14943–14952, 2023. 1, 2, 3, 5, 6

[17] Mingcheng Li, Dingkang Yang, Yang Liu, Shunli Wang, Jiawei Chen, Shuaibing Wang, Jinjie Wei, Yue Jiang, Qingyao Xu, Xiaolu Hou, Mingyang Sun, Ziyun Qian, Dongliang Kou, and Lihua Zhang. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. In *NeurIPS*, 2024. 2

[18] Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *CVPR*, pages 12458–12468, 2024. 2, 5, 6

[19] Quanjiang Li, Tingjin Luo, and Jiahui Liao. Theory-inspired deep multi-view multi-label learning with incomplete views and noisy labels. In *CVPR*, pages 20706–20715, 2025. 1

[20] 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 *NeurIPS*, 2022. 8

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[22] Andong Lu, Chenglong Li, Jiacong Zhao, Jin Tang, and Bin Luo. Modality-missing RGBT tracking: Invertible prompt learning and high-quality benchmarks. *Int. J. Comput. Vis.*, 133(5):2599–2619, 2025. 5, 6

[23] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605, 2008. 8

[24] John Edison Arevalo Ovalle, Thamar Solorio, Manuel Montes-y-Gómez, and Fabio A. González. Gated multimodal units for information fusion. In *ICLR Workshop*, 2017. 5

[25] Bikang Pan, Qun Li, Xiaoying Tang, Wei Huang, Zhen Fang, Feng Liu, Jingya Wang, Jingyi Yu, and Ye Shi. Nlprompt: Noise-label prompt learning for vision-language models. In *CVPR*, pages 19963–19973, 2025. 3

[26] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *CVPR*, pages 8228–8237, 2022. 2

[27] 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 *ICML*, pages 8748–8763, 2021. 4

[28] Merey Ramazanova, Alejandro Pardo, Bernard Ghanem, and Motasem Alfarra. Test-time adaptation for combating missing modalities in egocentric videos. In *ICLR*, 2025. 2

[29] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *IJCV*, 130(6):1526–1565, 2022. 2, 4

[30] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, pages 14398–14409, 2024. 2

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 3, 5

[32] 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 *CVPR*, pages 15878–15887, 2023. 1, 2, 5, 6

[33] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, pages 12692–12702, 2020. 2

[34] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frédéric Precioso. Recipe recognition with large multimodal food dataset. In *ICME Workshop*, pages 1–6, 2015. 5

[35] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *TPAMI*, 44(9):4555–4576, 2022. 2, 4

[36] Xiaoli Wang, Anqi Huang, Yongli Wang, Guanzhou Ke, Xiaobin Hong, and Jun Liu. Global-semantic alignment distillation for partial multi-view classification. In *AAAI*, pages 21287–21295, 2025. 2

[37] Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *ICCV*, pages 21968–21977, 2023. 2, 5, 6

[38] Yake Wei and Di Hu. Mmpareto: Boosting multimodal learning with innocent unimodal assistance. In *ICML*, pages 1–14, 2024. 2

[39] Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *CVPR*, pages 27328–27337, 2024. 2

[40] Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards good practices for missing modality robust action recognition. In *AAAI*, pages 2776–2784, 2023. 1, 2, 5, 6

[41] Haixiang Wu. Revisiting attention for multivariate time series forecasting. In *AAAI*, pages 21528–21535, 2025. 1

[42] Renjie Wu, Hu Wang, and Hsiang-Ting Chen. A comprehensive survey on deep multimodal learning with missing modality. *CoRR*, abs/2409.07825, 2024. 1, 2

[43] Peng Xu, Xiatian Zhu, and David A. Clifton. Multimodal learning with transformers: A survey. *TPAMI*, 45(10): 12113–12132, 2023. 1

[44] Wenxin Xu, Hexin Jiang, and Xuefeng Liang. Leveraging knowledge of modality experts for incomplete multimodal learning. In *MM*, pages 438–446, 2024. 5, 6

[45] Yingxue Xu, Fengtao Zhou, Chenyu Zhao, Yihui Wang, Can Yang, and Hao Chen. Distilled prompt learning for incomplete multimodal survival prediction. In *CVPR*, pages 5102–5111, 2025. 2

[46] Yang Yang, De-Chuan Zhan, Yi-Feng Wu, Zhi-Bin Liu, Hui Xiong, and Yuan Jiang. Semi-supervised multi-modal clustering and classification with incomplete modalities. *TKDE*, 33(2):682–695, 2021. 1

[47] Yanwu Yang, Hairui Chen, Zhikai Chang, Yang Xiang, Chenfei Ye, and Heather Ting Ma. Incomplete learning of multimodal connectome for brain disorder diagnosis via modal-mixup and deep supervision. In *MIDL*, pages 1006–1018, 2023. 2, 5, 6

[48] Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. Facilitating multimodal classification via dynamically learning modality gap. In *NeurIPS*, pages 62108–62122, 2024. 2

[49] Yang Yang, Wenjuan Xi, Luping Zhou, and Jinhui Tang. Rebalanced vision-language retrieval considering structure-aware distillation. *TIP*, 33:6881–6892, 2024. 1

[50] Yang Yang, Hongpeng Pan, Qing-Yuan Jiang, Yi Xu, and Jinhui Tang. Learning to rebalance multi-modal optimization by adaptively masking subnetworks. *TPAMI*, 47(6):4553–4566, 2025. 2

[51] Wenfang Yao, Kejing Yin, William K. Cheung, Jia Liu, and Jing Qin. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *AAAI*, pages 16416–16424, 2024. 5, 6

[52] Yiqun Yao and Rada Mihalcea. Modality-specific learning rates for effective multimodal additive late-fusion. In *ACL*, pages 1824–1834, 2022. 2

[53] Ziyi Yin, Ruijin Liu, Zhiliang Xiong, and Zejian Yuan. Multimodal transformer networks for pedestrian trajectory prediction. In *IJCAI*, pages 1259–1265, 2021. 1

[54] Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, and Changqing Zhang. Multimodal fusion on low-quality data: A comprehensive survey. *CoRR*, abs/2404.18947, 2024. 1, 2

[55] Shu-Peng Zhong, Zhihao Guan, Yu-Xuan Zhang, Xiao-Cong Lian, and Yang Yang. U^2dp: Unlocking unlabeled data potential for semi-supervised remote sensing image captioning. *TGRS*, 63:1–14, 2025. 1

[56] Haolin Zuo, Rui Liu, Jinming Zhao, Guanglai Gao, and Haizhou Li. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. In *ICASSP*, pages 1–5, 2023. 1, 2, 5, 6