Anchor-Guided Gradient Alignment for Incomplete Multimodal Learning

Supplementary Material

1. Dataset Overview and Baseline Details

1.1. Dataset Overview

In this subsection, we provide an overview of each dataset below, and detailed dataset statistics are provided in Table 1.

HateMemes [2]: Released by Facebook AI, this dataset is designed for multimodal hate content detection and contains 10,000 image-text pairs, where both modalities are required to infer the underlying intention. To avoid unimodal shortcut learning, the dataset is intentionally curated as a modality-dependent benchmark. The official split includes 8,500 training samples, 500 validation samples, and 1,500 test samples.

MM-IMDb [8]: This dataset is one of the largest publicly accessible benchmarks for multimodal movie genre prediction, containing 25,959 samples with both poster images and associated textual descriptions. As each movie may correspond to multiple categories, the task is formulated as multi-label classification. The official dataset split comprises 15,552 training examples, 2,608 validation examples, and 7,799 test examples.

Food-101 [10]: This large-scale benchmark contains 90,688 image–text pairs sourced from Google Image Search and covers 101 types of food, following the category taxonomy of the ETHZ Food-101 dataset. The dataset is divided into 67,972 samples for training and 22,716 for testing. Each instance includes a food image accompanied by a short textual description, which may contain noise due to web collection.

Table 1. Statistics of three benchmark datasets.

Dataset	Image	Text	Train	Val	Test
HateMemes	10,000	10,000	8,500	500	1,500
MM-IMDb	25,959	25,959	15,552	2,608	7,799
Food-101	90,688	90,688	67,972	–	22,716

1.2. Baseline Details

In this subsection, we summarize the compared SOTA baselines as follows.

IF-MMIN [16]: This method learns modality-invariant features by aligning the central moments of feature distributions across modalities and reconstructs missing modalities through an invariant imagination module.

ShaSpec [9]: This method learns both shared and modality-specific representations by combining distribution alignment, domain classification, and residual feature fusion, enabling robust handling of missing modalities across different downstream tasks.

DrFuse [15]: This method disentangles shared and modality-specific representations to handle missing clinical modalities and employs disease-wise attention to adaptively weight each modality for final prediction.

CorrKD [6]: This method performs correlation-driven knowledge distillation through contrastive semantic transfer and prototype-based feature alignment to handle uncertain missing modalities.

MoMKE [13]: This method adopts a two-stage mixture-of-experts framework that first trains unimodal experts and then jointly learns unimodal and multimodal representations through expert knowledge mixing with a soft routing mechanism for improved robustness.

ViLT [3]: This method simplifies vision-and-language pre-training by removing convolution-based visual encoders and directly processing image patches with a transformer for efficient multimodal representation learning.

MAPs [5]: This method introduces modality-missing-aware prompts that can be inserted into multimodal transformers to handle diverse missing-modality cases while requiring less than 1% learnable parameters.

MSPs [1]: This method designs modality-specific prompt tokens with orthogonality constraints to capture distinct modality characteristics, reducing prompt complexity while improving robustness under various missing-modality settings in practice.

IPer [7]: This method introduces invertible prompt learning that generates prompts for missing modalities using the available modality, while enforcing reconstruction to reduce semantic distortion for robust multimodal tracking.

AcMAE [12]: This method learns predictive coding for missing modalities by randomly dropping modality features during training and reconstructing them using the remaining modalities for achieving robust multimodal action recognition in practice.

DiCMoR [11]: This method recovers missing modalities by transferring class-conditional feature distributions using normalizing flow, ensuring distribution consistency between recovered and real modality features.

MMDS [14]: This method synthesizes complete multimodal samples via modal-mixup data imputation and employs a bilateral network with deep supervision to enhance mono-modal representations for disease classification.

RAGPT [4]: This method leverages retrieval-augmented dynamic prompt tuning by retrieving relevant instances, generating missing modality information, and producing context-aware prompts to enhance robustness under incomplete modality conditions.

2. Additional Experiments

In this section, we present additional experimental results of the proposed ANGA. We also provide the training algorithm of ANGA in Algorithm 1.

Algorithm 1: Incomplete Multimodal Learning with ANGA.

Input : Training set $\mathcal{D} = \{\mathcal{D}^c, \mathcal{D}^m\}$.
Output: Learned parameters θ of ANGA.
INIT : Initialize parameters θ^0, memory bank $\mathcal{M}$, retrieved size K, curriculum stage Z_{grow}, sample-ratio range $[\lambda_{\min}, \lambda_{\max}]$, cosine threshold τ, prompt layer b, and maximum epochs T.

/* Calculate sample reliability. */
for each sample $\tilde{\boldsymbol{x}}_i$ in $\mathcal{D}^m$ **do**
  Obtain prediction entropy $H(\tilde{\boldsymbol{x}}_i)$ based on Eq. (5).
  Add $\tilde{\boldsymbol{x}}_i$ to $\tilde{\mathcal{D}}^m_{\text{rank}}$ in ascending order of $H(\tilde{\boldsymbol{x}}_i)$.
end
/* Train ANGA. */
for $t = 0$ **to** $T-1$ **do**
  /* Construct optimization anchor. */
  Calculate the proportion of reliable reconstructed samples $\lambda(z)$ with Eq. (6).
  Obtain current epoch anchor set $\mathcal{A}_z$ with Eq. (7).
  Compute batch-wise anchor gradient $\mathbf{g}_A$ with Eq. (4) and Eq. (8).
  /* Gradient alignment. */
  Compute reconstructed gradient $\mathbf{g}_{M'}$ with $\mathbf{g}_A$ via cosine formulation $\texttt{sim}(\mathbf{g}_{M'}, \mathbf{g}_A)$.
  if $\texttt{sim}(\mathbf{g}_{M'}, \mathbf{g}_A) \geq \tau$ **then**
    Obtain modulated gradient with Eq. (9).
  else if $0 < \texttt{sim}(\mathbf{g}_{M'}, \mathbf{g}_A) < \tau$ **then**
    Obtain modulated gradient with Eq. (10)-(13).
  else if $\texttt{sim}(\mathbf{g}_{M'}, \mathbf{g}_A) < 0$ **then**
    Obtain modulated gradient with Eq. (14).
  end
  /* Semantic-enhanced adapter. */
  Generate dynamic prompts $\mathbf{P}^t_i$ and $\mathbf{P}^v_i$ via Eq. (16), and insert them into the b-th MSA layer.
  Train ANGA with batch data and update parameters θ with Eq. (15).
  Update $t = t + 1$.
end

2.1. Plug-and-Play Evaluation

To validate the flexibility and compatibility of our method, we perform a plug-and-play evaluation by directly integrating ANGA (using only the gradient alignment module) into several representative baselines [3–5, 12, 16] without modifying their original architectures or loss functions. As reported in Table 2, ANGA leads to consistent performance improvements across all baselines on the HateMemes dataset. These results indicate that ANGA can serve as a lightweight and generalizable enhancement module that is easily applicable to different multimodal frameworks.

Table 2. Plug-and-play performance gain of ANGA when applied to different baselines under a 70% missing rate across various missing-modality scenarios.

Metric →	AUROC (%)	AUROC (%)	AUROC (%)
Method ↓	Text	Image	Both
IF-MMIN [16]	57.62	53.44	55.19
IF-MMIN + ANGA	**61.37**(+3.75)	**57.18**(+3.74)	**57.76**(+2.57)
ViLT [3]	56.61	57.15	56.42
ViLT + ANGA	**57.82**(+1.21)	**60.62**(+3.47)	**59.54**(+3.12)
MAPs [5]	58.62	60.16	58.89
MAPs + ANGA	**59.42**(+0.80)	**62.10**(+1.94)	**60.18**(+1.29)
AcMAE [12]	55.74	59.66	57.25
AcMAE + ANGA	**58.72**(+2.98)	**61.82**(+2.16)	**60.54**(+3.29)
RAGPT [4]	64.10	62.57	63.47
RAGPT + ANGA	**67.23**(+3.13)	**63.26**(+0.69)	**64.86**(+1.39)

2.2. Sensitivity to Hyperparameters

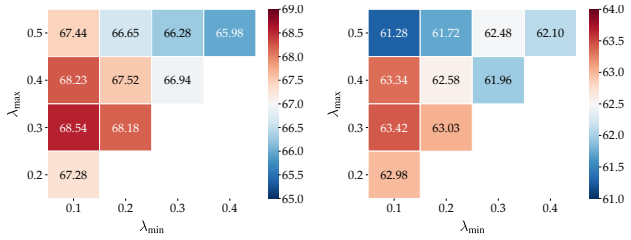
We further analyze the sensitivity of ANGA to the hyperparameters $\lambda_{\min}$ and $\lambda_{\max}$ on HateMemes, which control the lower and upper bounds of the sample ratio during curriculum learning. For simplicity, we fix the curriculum stage Z_{grow} to 5 and vary $\lambda_{\min}$ over $\{0.1, 0.2, 0.3, 0.4\}$ and $\lambda_{\max}$ over $\{0.2, 0.3, 0.4, 0.5\}$. The results are shown in Figure 1.

Overall, the performance of ANGA remains stable across different choices of $\lambda_{\min}$ and $\lambda_{\max}$, indicating low sensitivity to these hyperparameters. The best result is obtained at $(\lambda_{\min}, \lambda_{\max}) = (0.1, 0.3)$, while other moderate combinations such as $(0.1, 0.4)$ and $(0.2, 0.3)$ also achieve competitive performance. Extreme settings, however, result in noticeable performance degradation for different reasons. When only a very small portion of reconstructed samples is introduced, the optimization anchor formed from mixed samples becomes insufficiently representative and fails to provide a stable direction for alignment. Conversely, when a large portion of reconstructed samples is used, the optimization is dominated by noisy gradients from imperfect reconstructions, weakening the influence of complete-modal samples. These results indicate that ANGA benefits from maintaining a reasonable proportion of reconstructed samples, allowing the anchor to remain reliable while preventing excessive noise during training.

3. Limitations and Future Works

Although the proposed ANGA demonstrates strong robustness under various missing-modality settings, several directions can further improve our work in the future. The main limitations and possible extensions are listed as follows:

- **Limitation in modality scale.** The current framework is mainly designed for dual-modal tasks, while real multi-



(a) Text-missing Scenario (b) Image-missing Scenario

Figure 1. Sensitivity to hyperparameters λ_{min} and λ_{max}.

modal systems often involve more than two modalities. Extending ANGA to richer modality combinations remains an important direction for future exploration.

- **Limitation in gradient alignment.** The gradient modulation component depends on a fixed similarity threshold τ to determine the alignment strategy. This static rule may not fully reflect the evolving optimization dynamics during training. Developing adaptive or learnable alignment mechanisms may produce more stable and generalizable optimization behavior across diverse missing-modality conditions.
- **Limitation in real-world validation.** The current evaluation is conducted on widely used benchmark datasets. These datasets are curated and may not fully reflect the complexity of real-world multimodal scenarios. Applying ANGA to practical settings such as remote sensing analysis can provide stronger evidence of its generalization ability and practical value.

References

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

[2] 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. 1

[3] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, pages 5583–5594, 2021. 1, 2

[4] 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

[5] 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

[6] 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. 1

[7] 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. 1

[8] John Edison Arevalo Ovalle, Thamar Solorio, Manuel Montes-y-Gómez, and Fabio A. González. Gated multimodal units for information fusion. In *ICLR Workshop*, 2017. 1

[9] 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

[10] 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. 1

[11] Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *ICCV*, pages 21968–21977, 2023. 1

[12] 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

[13] Wenxin Xu, Hexin Jiang, and Xuefeng Liang. Leveraging knowledge of modality experts for incomplete multimodal learning. In *MM*, pages 438–446, 2024. 1

[14] 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. 1

[15] 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. 1

[16] 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