EFFICIENT REMOTE SENSING WITH HARMONIZED TRANSFER LEARNING AND MODALITY ALIGNMENT

Tengjun Huang
Shandong University
htj@mail.sdu.edu.cn

ABSTRACT

With the rise of Visual and Language Pretraining (VLP), an increasing number of downstream tasks are adopting the paradigm of pretraining followed by fine-tuning. Although this paradigm has demonstrated potential in various multimodal downstream tasks, its implementation in the remote sensing domain encounters some obstacles. Specifically, the tendency for same-modality embeddings to cluster together impedes efficient transfer learning. To tackle this issue, we review the aim of multimodal transfer learning for downstream tasks from a unified perspective, and rethink the optimization process based on three distinct objectives. We propose "Harmonized Transfer Learning and Modality Alignment (HarMA)", a method that simultaneously satisfies task constraints, modality alignment, and single-modality uniform alignment, while minimizing training overhead through parameter-efficient fine-tuning. Remarkably, without the need for external data for training, HarMA achieves state-of-the-art performance in two popular multimodal retrieval tasks in the remote sensing field. Experiments demonstrate that, even with minimal adjustable parameters, HarMA can still achieve performance comparable to or even better than that of a fully fine-tuned model. Due to its simplicity, HarMA can be integrated into almost all existing multimodal pretraining models. We hope this can help to better apply large models to various downstream tasks in the future with minimal resource consumption. Code is available on https://github.com/seekerhuang/HarMA.

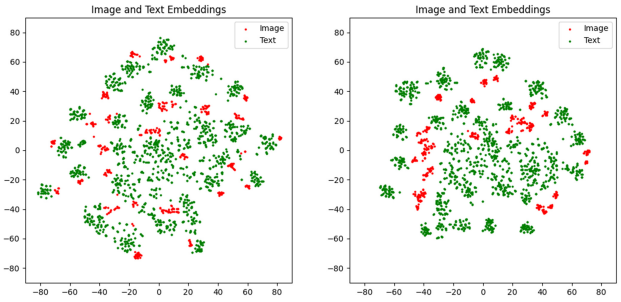
1 INTRODUCTION

The advent of Visual and Language Pretraining (VLP) has spurred a surge in studies employing large-scale pretraining and subsequent fine-tuning for diverse multimodal tasks (Tan & Bansal, 2019; Li et al., 2020; 2021; 2022; Liu et al., 2023). When conducting transfer learning for downstream tasks in the multimodal domain, the common practice is to first perform large-scale pre-training and then fully fine-tune on a specific domain dataset (Hu & Singh, 2021; Akbari et al., 2021; Zhang et al., 2024b), which is also the case in the field of remote sensing image-text retrieval (Cheng et al., 2021; Pan et al., 2023b). However, this method has at least two notable limitations. Firstly, fully fine-tuning a large model is extremely expensive and not scalable (Zhang et al., 2024a). Secondly, the pre-trained model has already been trained on a large dataset for a long time, and fully fine-tuning on a small dataset may lead to reduced generalization ability or overfitting.

Recently, several works have attempted to use Parameter-Efficient Fine-Tuning (PEFT) to address this issue, aiming to freeze most of the model parameters and fine-tune only a few (Houlsby et al., 2019; Mao et al., 2021; Zhang et al., 2022). This strategy seeks to incorporate domain-specific knowledge into the model while preserving the bulk of its original learned information. For example, Houlsby et al. (2019) attempted to fine-tune the pre-trained model by simply introducing a single-modality MLP layer. In contrast, Yuan et al. (2023) designed a cross-modal interaction adapter, aiming to enhance the model's ability to integrate multimodal knowledge. Although the above works have achieved promising results, they either concentrate on single-modality features or overlook potential semantic mismatches when modeling the visual-language joint space.

We have observed that poorly performing models sometimes exhibit a clustering phenomenon within the same modality embedding. Figure 1 illustrates the visualization of the last layer embeddings for

two models with differing performance in the filed of remote sensing image-text retrieval; the clustering phenomenon is noticeably more pronounced in the right image than in the left. We hypothesize that this may be attributed to the high intra-class and inter-class similarity of remote sensing images, leading to semantic confusion when modeling a low-rank visual-language joint space. This raises a critical question: *"How can we model a highly aligned visual-language joint space while ensuring efficient transfer learning?"*



(A) Retrieving a better performing model (B) Retrieving a worse performing model

Figure 1: In remote sensing image-text retrieval, excessive clustering of the same modality sometimes leads to a decrease in performance. The experiment was conducted on the RSITMD dataset.

In the brains of congenitally blind individuals, parts of the visual cortex can take on the function of language processing (Bedny et al., 2011). Concurrently, in the typical human cortex, several small regions—such as the Angular Gyrus and the Visual Word Form Area (VWFA)—serve as hubs for integrated visual-language processing (Houk & Wise, 1995). These areas hierarchically manage both low-level and high-level stimuli information (Chen et al., 2019). Inspired by this natural phenomenon, we propose "Efficient Remote Sensing with **Har**monized Transfer Learning and **M**odality **A**lignment (HarMA)". Specifically, similar to the information processing methods of the human brain, we designed a hierarchical multimodal adapter with mini-adapters. This framework emulates the human brain's strategy of utilizing shared mini-regions to process neural impulses originating from both visual and linguistic stimuli. It models the visual-language semantic space from low to high levels by hierarchically sharing multiple mini-adapters. Finally, we introduced a new objective function to alleviate the severe clustering of features within the same modality. Thanks to its simplicity, the method can be easily integrated into almost all existing multimodal frameworks.

2 METHOD

2.1 OVERALL FRAMEWORK

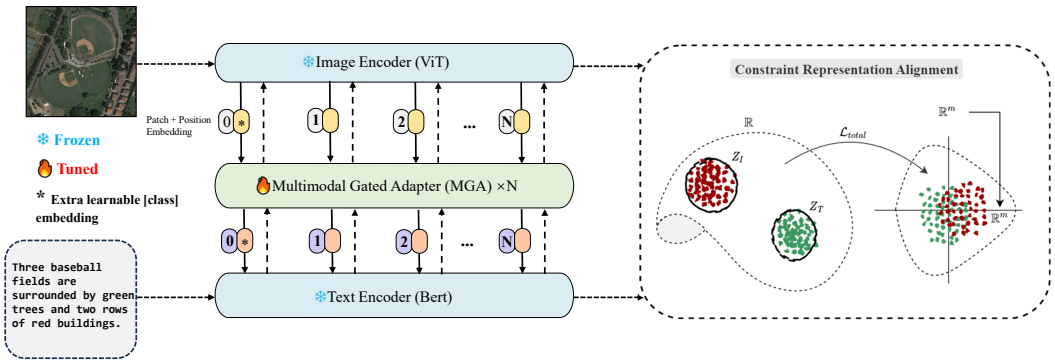


Figure 2: The overall framework of the proposed method.

Figure 2 illustrates our proposed HarMA framework. It initiates the process with the extraction of representations using image and text encoders, similar to CLIP (Radford et al., 2021b). These features are then processed by our unique multimodal gated adapter to obtain refined feature representations. Unlike the simple linear layer interaction used in (Yuan et al., 2023), we employ a

shared mini adapter as our interaction layer within the entire adapter. After that, we optimize using a contrastive learning objective and our adaptive triplet loss.

2.2 MULTIMODAL GATED ADAPTER

Previous parameter-efficient fine-tuning methods in the multimodal domain (Jiang et al., 2022b; Yuan et al., 2023) use a simple shared-weight method for modal interaction, potentially causing semantic matching confusion in the inherent modal embedding space. To address this, we designed a cross-modal adapter with an adaptive gating mechanism (Figure 3).

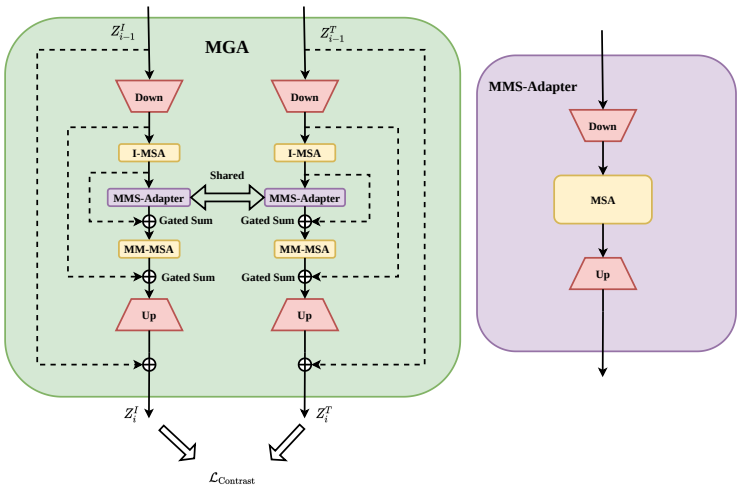


Figure 3: The specific structure of the multimodal gated adapter.The overall structure is shown on the left, while the structure of the shared multimodal sub-adapter is displayed on the right.

In this module, the extracted features z_I and z_T are first projected into low-dimensional embeddings. Different features z_I (z_T) are further enhanced in feature expression after non-linear activation and subsequent processing by I-MSA. I-MSA and the subsequent MM-MSA share parameters. The features are then fed into our designed Multimodal Sub-Adapter (MMS-Adapter) for further interaction, the structure of this module is shown on the right side of Figure 3.

The MMS-Adapter, akin to standard adapters, aligns multimodal context representations via shared-weight self-attention. However, direct post-projection output of these aligned representations negatively impacts image-text retrieval performance, likely due to off-diagonal semantic key matches in the feature's low-dimensional manifold space. This contradicts contrastive learning objectives.

To tackle this, the already aligned representations are further processed in the I-MSA with shared weights, thereby reducing model parameters and leveraging prior modality knowledge. And to ensure a finer-grained semantic match between image and text, we introduce early image-text matching supervision in the MGA output, significantly mitigating the occurrence of above issue.

Ultimately, features are projected back to their original dimensions before adding the skip connection. The final layer is initialized to zero to safeguard the performance of the pre-trained model during the initial stages of training. Algorithm 1 summarizes the proposed method.

2.3 OBJECTIVE FUNCTION

In the realm of multimodal learning, when engaging in transfer learning for downstream tasks, it is typically necessary to devise objective functions tailored to distinct tasks and to align different modalities embeddings. we can initially define the objective of multimodal learning applied to all downstream tasks as follows:

$$\min_{\theta^*} \left(\sum_i \mathbb{E}_{x_i \sim \mathcal{D}^i} [L^i_{\text{task}}(f(x_i; \theta^*))] + \sum_{j \neq k} \mathbb{E}_{(x_j, x_k) \sim \mathcal{D}^j \times \mathcal{D}^k} [L^{jk}_{\text{align}}(f(x_j; \theta^*), f(x_k; \theta^*))] \right). \quad (1)$$

Here, L^i_{task} represents the task loss for the i-th task, and L^{jk}_{align} denotes the alignment loss between different pairs of modalities (j, k). The expectation is taken over the data distribution $\mathcal{D}$ for each task. θ^* represents the target parameters for transfer learning.

However, in the remote sensing field, we observe that underperforming models sometimes exhibit a phenomenon where same-modality embeddings cluster together, as shown in Figure 1. Wang & Isola (2020) highlighted that the low uniformity of alignment of modalities from the same distribution may limit the transferability of embeddings. To ensure that embeddings from the same modality are uniformly aligned without excessive clustering, the unified objective for multimodal learning applied to remote sensing downstream tasks can be defined as:

$$\min_{\theta^*} \left(L_{\text{ini}} + \lambda_1 \sum_i \mathbb{E}_{x_i \sim \mathcal{D}^i}[L^i_{\text{uniform}}(f(x_i; \theta^*))] \right) \quad \text{s.t. } D(\theta, \theta^*) \leq \delta. \tag{2}$$

In this equation, L_{ini} represents the initial optimization objective (Equation 1), which is composed of the task loss and alignment loss. L^i_{uniform} denotes the singe-modality uniformity loss for the i-th modality, and $D(\theta, \theta^*)$ is a cost measure between the original and updated model parameters, constrained to be less than δ. δ is the minimum parameter update cost in the ideal state.

We observe that existing works often only explore one or two objectives, with most focusing either on how to efficiently fine-tune parameters for downstream tasks (Jiang et al., 2022b; Jie & Deng, 2022; Yuan et al., 2023) or on modality alignment (Chen et al., 2020; Ma et al., 2023; Pan et al., 2023a). Few can simultaneously satisfy the three requirements outlined in the above formula. We have satisfied the need for efficient transfer learning by introducing adapters that mimic the human brain. This prompts us to ask: how can we fulfill the latter two objectives—high alignment of embeddings across different modalities while preventing excessive clustering of embeddings within the same modality?

2.3.1 Adaptive Triplet Loss

In image-text retrieval tasks, the bidirectional triplet loss established by Faghri et al. (2017) has become a mainstream loss function. However, if we choose to accumulate the losses of all samples in image-text matching, the model may struggle to optimize due to the high intra-class similarity and inter-class similarity prevalent in most regions of the images within the field of remote sensing (Yuan et al., 2022a). Therefore, We propose an Adaptive Triplet Loss that automatically mines and optimizes hard samples:

$$\mathcal{L}_{\text{ada-triplet}} = \sum_{i=1}^{N} w_i [m + s_{ij} - s_{ii}]_+ + \sum_{j=1}^{N} w_j [m + s_{ji} - s_{ii}]_+, \tag{3}$$

where s_{ij} is the dot product between image feature i and text feature j, w_i and w_j are the weights of sample i and j, determined by the loss size of different samples:

$$w_i = (1 - \exp(-[m + s_{ij} - s_{ii}]_+))^\gamma, w_j = (1 - \exp(-[m + s_{ji} - s_{ii}]_+))^\gamma, \tag{4}$$

where γ is a hyperparameter adjusting the size of the weights. This loss function aims to bring the features of positive samples closer together, while distancing those between positive and negative samples. By dynamically adjusting the focus between hard and easy samples, our approach effectively satisfies the other two objectives proposed above. It not only aligns different modality samples at a fine-grained level but also prevents over-aggregation among samples of the same modality, thereby enhancing the model's matching capability. Also, following the approach of (Radford et al., 2021b), we utilize a contrastive learning objective to align image and text semantic features. Consequently, the total objective is defined as:

$$\mathcal{L}_{\text{total}} = (\lambda_1 \mathcal{L}_{\text{ada-triplet}} + \lambda_2 \mathcal{L}_{\text{contrastive}}). \tag{5}$$

The λ_1 and λ_2 are parameters for balancing the loss. We offer detailed information on contrastive learning in Appendix A.2.

3 EXPERIMENTS

We evaluate our proposed HarMA framework on two widely used remote sensing (RS) image-text datasets: RSICD (Lu et al., 2017) and RSITMD (Yuan et al., 2022a). We use standard recall at TOP-K (R@K, K = 1, 5, 10) and mean recall (mR) to assess our model.

3.1 COMPARATIVE EXPERIMENTS

In this section, we compare the proposed method with state-of-the-art retrieval techniques on two remote sensing multimodal retrieval benchmarks. The backbone networks employed in our experiments are the CLIP (ViT-B-32) (Radford et al., 2021a) and the GeoRSCLIP (Zhang et al., 2023).

Table 1: Retrieval Performance Test. ‡ : RSICD Test Set; ⁕ : RSITMD Test Set; † : The parameter amount of a single adapter module. Red: Our method; Blue: Full fine-tuned CLIP.

Methods	Backbone (image/text)	Trainable Params	Text-to-image			Image-to-text			mR
			R@1	R@5	R@10	R@1	R@5	R@10	
Traditional methods									
GaLR with MR (Yuan et al., 2022b)	ResNet18, biGRU ‡	46.89M	6.59	19.85	31.04	4.69	19.48	32.13	18.96
PIR (Pan et al., 2023a)	Swin Transformer, Bert ‡	-	9.88	27.26	39.16	6.97	24.56	38.92	24.46
CLIP-based methods									
Zero-shot CLIP (Radford et al., 2021a)	CLIP(ViT-B-32) ‡	0.00M	6.77	15.37	23.15	5.01	15.75	24.21	15.04
Full-FT CLIP (Radford et al., 2021a)	CLIP(ViT-B-32) ‡	151M	**13.54**	**30.83**	**43.46**	**11.55**	**33.14**	**49.83**	**30.39**
Full-FT GeoRSCLIP (Zhang et al., 2023)	GeoRSCLIP(ViT-B-32-RET-2) ‡	151M	**18.85**	**38.15**	**53.16**	**14.27**	**39.71**	**57.49**	**36.94**
Full-FT GeoRSCLIP (w/ Extra Data)	GeoRSCLIP(ViT-B-32-RET-2) ‡	151M	**21.13**	**41.72**	**55.63**	**15.59**	**41.19**	**57.99**	**38.87**
Adapter (Houlsby et al., 2019)	CLIP(ViT-B-32) ‡	0.17M †	8.73	24.73	37.81	8.43	26.02	43.33	24.84
CLIP-Adapter (Gao et al., 2021)	CLIP(ViT-B-32) ‡	0.52M †	7.11	19.48	31.01	7.67	24.87	39.73	21.65
AdaptFormer (Chen et al., 2022)	CLIP(ViT-B-32) ‡	0.17M †	12.46	28.49	41.86	9.09	29.89	46.81	28.10
Cross-Modal Adapter (Jiang et al., 2022a)	CLIP(ViT-B-32) ‡	0.16M †	11.18	27.31	40.62	9.57	30.74	48.36	27.96
UniAdapter (Lu et al., 2023)	CLIP(ViT-B-32) ‡	0.55M †	12.65	30.81	42.74	9.61	30.06	47.16	28.84
PE-RSITR (Yuan et al., 2023)	CLIP(ViT-B-32) ‡	0.16M †	14.13	31.51	44.78	11.63	33.92	50.73	31.12
Ours (HarMA w/o Extra Data)	CLIP(ViT-B-32) ‡	0.50M †	**15.21**	**33.46**	**47.42**	**11.67**	**35.49**	**51.71**	**32.49**
Ours (HarMA w/o Extra Data)	GeoRSCLIP(ViT-B-32-RET-2) ‡	0.50M †	**20.52**	**41.37**	**54.66**	**15.84**	**41.92**	**59.39**	**38.95**
Zero-shot CLIP (Radford et al., 2021a)	CLIP(ViT-B-32) ⁕	0.00M	9.29	26.33	37.39	7.79	23.67	38.89	23.89
Full-FT CLIP (Radford et al., 2021a)	CLIP(ViT-B-32) ⁕	151M	**26.99**	**46.9**	**58.85**	**20.53**	**52.35**	**71.15**	**46.13**
Full-FT GeoRSCLIP (Zhang et al., 2023)	GeoRSCLIP(ViT-B-32-RET-2) ⁕	151M	**30.53**	**49.78**	**63.05**	**24.91**	**57.21**	**75.35**	**50.14**
Full-FT GeoRSCLIP (w/ Extra Data)	GeoRSCLIP(ViT-B-32-RET-2) ⁕	151M	**32.30**	**53.32**	**67.92**	**25.04**	**57.88**	**74.38**	**51.81**
UniAdapter (Lu et al., 2023)	CLIP(ViT-B-32) ⁕	0.55M †	19.86	36.32	51.28	17.54	44.89	56.46	39.23
PE-RSITR (Yuan et al., 2023)	CLIP(ViT-B-32) ⁕	0.16M †	23.67	44.07	60.36	20.10	50.63	67.97	44.47
Ours (HarMA w/o Extra Data)	CLIP(ViT-B-32) ⁕	0.50M †	**25.81**	**48.37**	**60.61**	**19.92**	**53.27**	**71.21**	**46.53**
Ours (HarMA w/o Extra Data)	GeoRSCLIP(ViT-B-32-RET-2) ⁕	0.50M †	**32.74**	**53.76**	**69.25**	**25.62**	**57.65**	**74.60**	**52.27**

Table 1 presents the retrieval performance on RSICD and RSITMD. Firstly, as indicated in the first column, our method surpasses traditional state-of-the-art approaches while requiring significantly fewer tuned parameters. Secondly, when using CLIP (ViT-B-32) (Radford et al., 2021a) as the backbone, our approach achieves competitive and even superior performance compared to fully fine-tuned methods. Specifically, when matched with methods that have a similar number of tunable parameters, our method's Mean Recall (MR) sees an approximate increase of 50% over CLIP-Adapter (Gao et al., 2021) and 12.7% over UniAdapter (Lu et al., 2023) on RSICD, and an 18.6% improvement over UniAdapter on RSITMD. Remarkably, by utilizing the pretrained weights of GeoRSCLIP, HarMA establishes a new benchmark in the remote sensing field for two popular multimodal retrieval tasks. It only modifies less than 4% of the total model parameters, outperforming all current parameter-efficient fine-tuning methods and even surpassing the image-text retrieval performance of fully fine-tuned GeoRSCLIP on RSICD and RSITMD.

4 CONCLUSION

In this paper, we have revisited the learning objectives of multimodal downstream tasks from a unified perspective and proposed HarMA, an efficient framework that addresses the suboptimal multimodal alignment in remote sensing. HarMA uniquely enhances uniform alignment while preserving pretrained knowledge. Through the use of lightweight adapters and adaptive losses, HarMA achieves state-of-the-art retrieval performance with minimal parameter updates, surpassing even full fine-tuning. Despite all the benefits, one potential limitation is that designing pairwise objective functions may not provide more robust distribution constraints. Our future work will focus on extending this approach to more multimodal tasks.

REFERENCES

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021.

Marina Bedny, Alvaro Pascual-Leone, David Dodell-Feder, Evelina Fedorenko, and Rebecca Saxe. Language processing in the occipital cortex of congenitally blind adults. *Proceedings of the National Academy of Sciences*, 108(11):4429–4434, 2011.

Lang Chen, Demian Wassermann, Daniel A Abrams, John Kochalka, Guillermo Gallardo-Diez, and Vinod Menon. The visual word form area (vwfa) is part of both language and attention circuitry. *Nature communications*, 10(1):5601, 2019.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Adv. Neural Inf. Process. Syst.*, 35:16664–16678, 2022.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020.

Qimin Cheng, Yuzhuo Zhou, Peng Fu, Yuan Xu, and Liang Zhang. A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 14:4284–4297, 2021.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv:2110.04544*, 2021.

James C. Houk and Steven P. Wise. Feature Article: Distributed Modular Architectures Linking Basal Ganglia, Cerebellum, and Cerebral Cortex: Their Role in Planning and Controlling Action. *Cerebral Cortex*, 5(2):95–110, 03 1995. ISSN 1047-3211. doi: 10.1093/cercor/5.2.95. URL https://doi.org/10.1093/cercor/5.2.95.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proc. 36th Int. Conf. Mach. Learn.*, volume 97, pp. 2790–2799, 2019.

Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1439–1449, 2021.

Haojun Jiang, Jianke Zhang, Rui Huang, Chunjiang Ge, Zanlin Ni, Jiwen Lu, Jie Zhou, Shiji Song, and Gao Huang. Cross-modal adapter for text-video retrieval. *arxiv.2211.09623*, 2022a.

Haojun Jiang, Jianke Zhang, Rui Huang, Chunjiang Ge, Zanlin Ni, Jiwen Lu, Jie Zhou, Shiji Song, and Gao Huang. Cross-modal adapter for text-video retrieval. *arXiv preprint arXiv:2211.09623*, 2022b.

Shibo Jie and Zhi-Hong Deng. Convolutional bypasses are better vision transformer adapters. *arXiv preprint arXiv:2207.07039*, 2022.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Adv. Neural Inf. Process. Syst.*, 34:9694–9705, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pp. 121–137. Springer, 2020.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Haoyu Lu, Mingyu Ding, Yuqi Huo, Guoxing Yang, Zhiwu Lu, Masayoshi Tomizuka, and Wei Zhan. Uniadapter: Unified parameter-efficient transfer learning for cross-modal modeling. *arxiv.2302.06605*, 2023.

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017.

Qing Ma, Jiancheng Pan, and Cong Bai. Direction-oriented visual-semantic embedding model for remote sensing image-text retrieval. *arXiv preprint arXiv:2310.08276*, 2023.

Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Wen-tau Yih, and Madian Khabsa. Unipelt: A unified framework for parameter-efficient language model tuning. *arXiv preprint arXiv:2110.07577*, 2021.

Jiancheng Pan, Qing Ma, and Cong Bai. A prior instruction representation framework for remote sensing image-text retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 611–620, 2023a.

Jiancheng Pan, Qing Ma, and Cong Bai. Reducing semantic confusion: Scene-aware aggregation network for remote sensing cross-modal retrieval. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pp. 398–406, 2023b.

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 *Proc. Int. Conf. Mach. Learn.*, volume 139, pp. 8748–8763, 2021a.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021b.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Yuan Yuan, Yang Zhan, and Zhitong Xiong. Parameter-efficient transfer learning for remote sensing image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *arXiv preprint arXiv:2204.09868*, 2022a.

Zhiqiang Yuan, Wenkai Zhang, Changyuan Tian, Xuee Rong, Zhengyuan Zhang, Hongqi Wang, Kun Fu, and Xian Sun. Remote sensing cross-modal text-image retrieval based on global and local information. *IEEE Trans. Geosci. Remote Sens.*, 60:1–16, 2022b.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models, 2024a.

Xinxin Zhang, Yuan Yuan, and Xuelong Li. Reweighted low-rank and joint-sparse unmixing with library pruning. *IEEE Trans. Geosci. Remote Sens.*, 60:1–16, 2022.

Yan Zhang, Zhong Ji, Di Wang, Yanwei Pang, and Xuelong Li. User: Unified semantic enhancement with momentum contrast for image-text retrieval. *IEEE Transactions on Image Processing*, 2024b.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m: A large scale vision-language dataset for remote sensing vision-language foundation model. *arXiv preprint arXiv:2306.11300*, 2023.

APPENDIX

A PRELIMINARIES

A.1 TRIPLET LOSS

In the domain of image-text retrieval tasks, the bidirectional triplet loss introduced by Faghri et al. (2017) has been widely adopted as a standard loss function. The formulation of the triplet loss is given by:

$$\mathcal{L}_{\text{triplet}} = \sum_{i=1}^{N} \max(0, m + s_{ij} - s_{ii}) + \sum_{j=1}^{N} \max(0, m + s_{ji} - s_{ii}). \tag{6}$$

The m denotes the margin, a hyperparameter that defines the minimum distance between the non-matching pairs. The function s_{ij} represents the similarity score between the i-th image and the j-th text, and s_{ii} is the similarity score between matching image-text pairs. The objective of the triplet loss is to ensure that the distance between non-matching pairs is greater than the distance between matching pairs by at least the margin m.

A.2 CONTRASTIVE LEARNING

Given a mini-batch of N positive image-text pairs $\{(I_i, T_i)\}_{i\in\{1,\ldots,N\}}$, we obtain the final image and text embeddings $\{(v_i, t_i)\}_{i\in\{1,\ldots,N\}}$, and compute the cosine similarity $\text{sim}(v_i, t_j) = v_i^{\mathrm{T}} t_j$. The vision-to-text and text-to-vision contrastive losses are defined as:

$$\begin{aligned} \mathcal{L}_{v2t} &= -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(v_i, t_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(v_i, t_j)/\tau)}, \\ \mathcal{L}_{t2v} &= -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(t_i, v_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(t_i, v_j)/\tau)}, \end{aligned} \tag{7}$$

where τ is the temperature parameter that scales the distribution of similarities. The overall contrastive loss is computed as the average of the vision-to-text and text-to-vision losses:

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{2}\sum_{i=1}^{N} \left(\mathcal{L}_{v2t} + \mathcal{L}_{t2v}\right). \tag{8}$$

This loss function encourages the model to distinguish between positive pairs (where images and texts correspond to each other) and negative pairs (where they do not), effectively learning a joint embedding space where similar concepts are closer together.

A.3 RETRIEVAL EVALUATION METRICS

The retrieval evaluation metrics are defined as the mean of the recall rates at different cutoff points for both text-to-image and image-to-text retrieval tasks:

$$\text{mR} = \left(\underbrace{R@1 + R@5 + R@10}_{\text{Text-to-image}} + \underbrace{R@1 + R@5 + R@10}_{\text{Image-to-text}} \right) /6, \tag{9}$$

where $R@k$ is the recall at rank k, indicating the percentage of queries for which the correct item is found among the top k retrieved results. The mean recall mR averages these recall values to provide a single performance metric that captures retrieval effectiveness at various depths of the result list.

B ALGORITHM

Algorithm 1 MultiModal Gated Adapter (MGA) for cross-modal interaction.

Input: Feature tensors Z_I and Z_T from image and text encoders, respectively.
Parameters: Weight matrices $\mathbf{W_1}, \mathbf{W_2}, \mathbf{W_i}$, bias vectors $\mathbf{b_1}, \mathbf{b_2}, \mathbf{b_i}$, and learnable gate parameters λ_1, λ_2.
Output: Enhanced feature tensors f_{end}^I and f_{end}^T for image and text.
function $\sigma(\cdot)$ **is** non-linear activation function (e.g., GELU)
function $MSA(\cdot)$ **is** Multi-Head Self-Attention mechanism
function $MMSA(x)$ **is** Multi-Modal Sub-Adapter mechanism defined as:
$MMSA(x) = \mathbf{W_i}^{\mathbf{Up}}(MSA(\sigma(\mathbf{W_i}^{\mathbf{Down}}x + \mathbf{b_i}^{\mathbf{Down}}))) + \mathbf{b_i}^{\mathbf{Up}}$
for each feature tensor Z in $\{Z_I, Z_T\}$ **do**
$f_1 = \sigma(\mathbf{W_1}Z + \mathbf{b_1})$ # Process image and text feature tensors
$f_2 = MSA(f_1)$
$f_3 = \lambda_1 MMSA(f_2) + (1 - \lambda_1)f_2$ # Apply multi-modal sub-adapter with gating
$f_4 = \lambda_2 MSA(f_3) + (1 - \lambda_2)f_1$
$f_{end} = (\mathbf{W_2}f_4 + \mathbf{b_2}) + Z$
end for
return f_{end}^I, f_{end}^T

C EXPERIMENT DETAILS

C.1 DATASETS

RSICD. The Remote Sensing Image Captioning Dataset (RSICD) serves as a benchmark for the task of captioning remote sensing images (Lu et al., 2017). It encompasses over ten thousand remote sensing images sourced from Google Earth, Baidu Maps, MapABC, and Tianditu. The resolution of these images varies, with each being resized to a fixed dimension of 224x224 pixels. The dataset comprises a total of 10,921 images, with each image accompanied by five descriptive sentences.

RSITMD. The Remote Sensing Image-Text Matching Dataset (RSITMD) is a fine-grained and challenging dataset for remote sensing multimodal retrieval tasks, introduced by Yuan et al. (2022a). Unlike other remote sensing image-text pairing datasets, it features detailed descriptions of the relationships between objects. Additionally, the dataset includes keyword attributes (ranging from one to five keywords per image), facilitating keyword-based remote sensing text retrieval tasks. It comprises 4,743 images spanning 32 scenes, with a total of 23,715 annotations, of which 21,829 are unique.

C.2 IMPLEMENTATION DETAILS

Consistent with the experimental details established by Yuan et al. (2022a) and Pan et al. (2023a), we partitioned the dataset into distinct sets for training, validation, and testing. For CLIP or GeoRSCLIP, we configured the output dimensions of both visual and textual encoders to 768, which were then linearly projected to a 512-dimensional space. We set the temperature coefficient for the contrastive loss at 0.07 and the margin for the adaptive triplet loss at 0.2. Training was executed on either four A40 GPUs (48GB × 4) or eight RTX 4090 GPUs (24GB × 8) with a batch size of 1024. We utilized an AdamW optimizer (Loshchilov & Hutter, 2017) with a learning rate of 8e-5, a weight decay of 0.04, and implemented a linear decay strategy for the learning rate.

C.3 ABLATION STUDY

To demonstrate the effectiveness of our proposed method, we compared it against three baselines: 1) a CLIP model fine-tuned with downstream domain data, 2) a CLIP model fully fine-tuned with downstream domain data, and 3) a CLIP model with MGA but without introducing a new objective function. We evaluate the experimental performance on the RSITMD dataset. Table 2 presents the results of the ablation study.

Table 2: Comparison of different methods on RSITMD dataset. The first row represents the results of the fully fine-tuned CLIP (Full-ft CLIP). The subsequent rows, from top to bottom, represent the non-fine-tuned CLIP (CLIP w/o fine-tuning), CLIP with the Multimodal Gating Adapter (MGA) where only the MGA is fine-tuned (CLIP w/ MGA), and CLIP with both the MGA and the adaptive triplet loss, where only the MGA is fine-tuned (CLIP w/ MGA + Adaptive Triplet Loss).

Method	Backbone	Module		Text-to-image			Image-to-text			MR
		MGA	Adaptive Triplet Loss	R@1	R@5	R@10	R@1	R@5	R@10	
Full-FT CLIP	ViT-B-32	✗	✗	**26.99**	**46.9**	**58.85**	**20.53**	**52.35**	**71.15**	**46.13**
Zero-shot CLIP	ViT-B-32	✗	✗	9.29	26.33	37.39	7.79	23.67	38.89	23.89
CLIP	ViT-B-32	✓	✗	**25.33**	**47.96**	**60.26**	**18.71**	**53.13**	**70.52**	**45.98**
CLIP (HarMA)	ViT-B-32	✓	✓	**25.81**	**48.37**	**60.61**	**19.92**	**53.27**	**71.21**	**46.53**

Table 3: Model trainable parameters comparison.

Method	Trainable Params (%)
Full-FT CLIP	100.00
Zero-shot CLIP	0.00
CLIP w/ MGA (Ours)	3.82
HarMA (Ours)	3.82

In our experiments, for a fair comparison, we selected CLIP (ViT-B-32) (Radford et al., 2021b) as the backbone model to conduct ablation studies. As shown in Table 2, the first row demonstrates the performance of CLIP after full fine-tuning on downstream data, yielding a substantial improvement compared to the non-fine-tuned CLIP in the second row (46.13 vs 23.89). However, this approach necessitates fine-tuning the entire model, which is computationally expensive and difficult to scale. By merely introducing the Multimodal Gated Adapter (MGA) in the third row, the retrieval performance significantly surpasses that of the non-fine-tuned CLIP ($45.98 > 23.89$) and achieves comparable results to full fine-tuning (45.98 vs 46.13). The fourth row shows that incorporating our proposed adaptive triplet loss further boosts performance, slightly exceeding full fine-tuning ($46.53 > 46.13$). This improvement can be attributed to the semantic ambiguity caused by unimodal embedding aggregation when modeling the joint visual-language space, as mentioned in Section 1. Our experiments validate the effectiveness of the proposed modules, offering a promising solution to address this challenge. Table 3 presents the percentage of parameters that require fine-tuning for each method, relative to the total number of parameters in the model.

D Qualitative Analysis

D.1 Image-to-Text Results Analysis

Figure 4 presents the qualitative results for image-to-text retrieval, comparing the top-5 retrieved captions from HarMA (Ours) and the fully fine-tuned CLIP (Full-ft CLIP). Overall, HarMA demonstrates superior retrieval performance compared to Full-ft CLIP.

Notably, in the first case, even though our method's fifth-ranked retrieval result is incorrect, it effectively captures the overall semantics rather than irrelevant details. For instance, we identify the "two tennis courts" as the core semantic element of the image and retrieve a highly similar text (see the incorrect example in the rightmost column). In contrast, Full-ft CLIP mistakenly treats the "slight shadows" of the tennis courts and the surrounding "trees and lawn" as the primary semantics, as evidenced by the second (shadow of green farmland) and third (some trees and pentagonal squares of lawns) retrieved texts. This observation may provide two insights from an interpretable perspective: **1) HarMA is better at recognizing overall semantics, and 2) Full-ft CLIP may be prone to overfitting, focusing excessively on unimportant details in the image.**

In the second scenario, we focus on a forest image, which poses a greater challenge. Unlike the first scenario, here trees typically seen in the background are brought to the foreground, testing the model's ability to generalize. Here, although HarMA retrieves two incorrect texts, each corresponding to a specific image. Interestingly, as shown on the far right, the retrieved image is also a forest. From a human evaluation perspective, the incorrectly retrieved results may provide more detailed

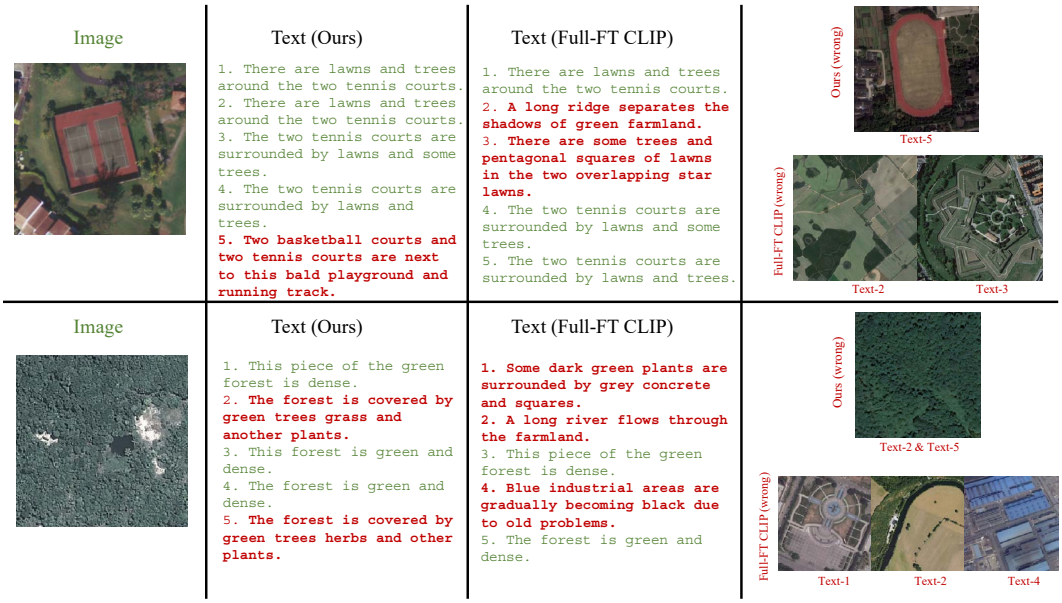


Figure 4: **Image-to-Text Retrieval Visual Results.** We compare the top-5 retrieved captions from HarMA (Ours) and the fully fine-tuned CLIP (Full-FT CLIP). Overall, HarMA demonstrates superior retrieval performance compared to Full-FT CLIP. Our method accurately captures the semantic elements, such as the tennis court in the first image, and associates them with the overall context. In contrast, Full-FT CLIP tends to overemphasize irrelevant details, mistakenly treating partial shadows and surrounding trees as the main subject matter.

descriptions than the correct text, while the original dataset annotation appears rather simplistic. The results retrieved by Full-ft CLIP, on the other hand, deviate significantly from the actual image semantics, with the primary semantics being squares (TOP-1), farmland (TOP-2), and industrial areas (TOP-4). We hypothesize that HarMA, by introducing downstream domain knowledge through adapters while retaining some of the original large-scale pre-training priors, is better equipped to handle the "noise problem" in small datasets compared to full fine-tuning.

D.2 Text-to-Image Results Analysis

Figure 5 illustrates the top-5 text-to-image retrieval results on the RSITMD test set for our HarMA and the fully fine-tuned CLIP models. Similar to the image-to-text case, our HarMA outperforms the fully fine-tuned CLIP overall.

Let us examine the first case. In the first row, the query image depicts "a city with a lot of green plants." Our HarmA successfully retrieves a matching image in the top-1 result. Encouragingly, the top-2 to top-4 results are primarily semantically associated with "city" and "green plants." Although the final top-5 result is semantically unrelated, the distorted river channel bears striking resemblance to the winding road in the query image. The second row shows the retrieval results from the fully fine-tuned CLIP, which, similar to the image-to-text case, focuses on irrelevant details (e.g., the green lake in top-4 and the green playground in top-5).

In the second case, both HarMA and the fully fine-tuned CLIP exhibit varying degrees of "hallucination" by associating the winding river with distorted roads. However, HarMA still outperforms the fully fine-tuned CLIP in terms of retrieval confidence and overall semantic relevance. For instance, HarMA still captures the overall semantics of "A river with dark green water."

In conclusion, the image-to-text and text-to-image retrieval results demonstrate HarMA's effectiveness in mitigating hallucinations and resisting noise. In the future, we aim to extend it to LLMs.

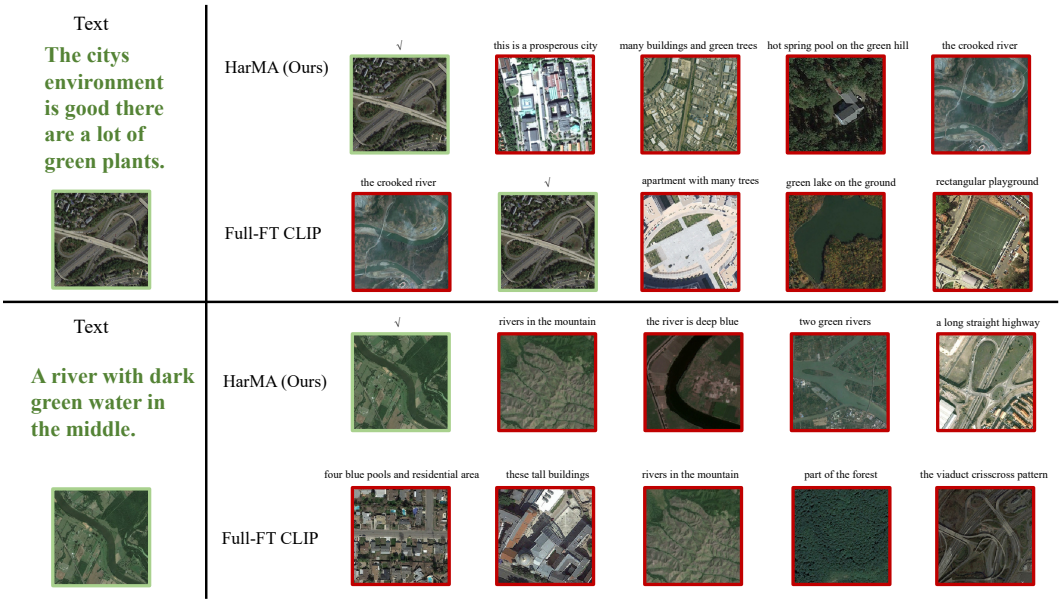


Figure 5: **Text-to-Image Retrieval Visual Results.** On the left side, we show the query text, and the ground-truth image is displayed below. The top-5 retrieved images based on the query text are presented on the right side, where green boxes indicate matches and red boxes indicate mismatches. For the mismatched retrieved images, we identify the main semantics of their associated text above the images. It is worth noting that in the rsitmd dataset, the relationship between images and texts is one-to-many, meaning that Image-to-Text can retrieve multiple results, while Text-to-Image has only one correct result.

D.3 Visual Results of Embedding Space

In this subsection, we present the t-SNE (Van der Maaten & Hinton, 2008) visualizations of the image and text embeddings. To capture the highest-level semantics, we select the embeddings from the final transformer layer for t-SNE processing. We employ a CLIP model with a ViT-B-32 backbone and conduct experiments on the RSITMD test set.

(1) Initial State (2) CLIP w/ MGA (3) HarMA

Figure 6: The leftmost figure illustrates the visualization results of the CLIP model outputs without fine-tuning on downstream data. The central figure depicts the visualization of the CLIP outputs with MGA. The rightmost figure showcases the results of our HarMA framework.

As shown in Figure 6, the visualization of the CLIP model outputs without fine-tuning on downstream data reveals a significant distance between the embeddings of different modalities, indicating

(1) Overfitted CLIP (2) Full-FT CLIP (3) HarMA

Figure 7: **Visualization of the embedding space.** The first image illustrates the output of an overfitted CLIP model, the second image depicts the results from a fully fine-tuned CLIP model, and the third image presents the outputs after fine-tuning with the HarMA framework.

suboptimal modal alignment. The CLIP output results with MGA demonstrate excellent modal alignment; however, the distances within the same modality are excessively small (exhibiting partial clustering), potentially hindering the capture of fine-grained semantic differences.

Figure 7 presents the visualization results of the embedding space under various conditions. The retrieval performance is ranked as follows: HarMA $>$ Fully fine-tuned CLIP $>$ Overfitted CLIP. The overfitted CLIP model shows almost no alignment between visual and textual embeddings, resulting in the poorest retrieval performance. The fully fine-tuned CLIP model achieves basic modality alignment but still exhibits some clustering phenomena mentioned in Section 1. Conversely, regardless of whether in Figure 6 or Figure 7, the output results of the complete HarMA framework not only exhibit robust modal alignment but also mitigate the aforementioned intra-modal clustering phenomenon, thereby validating the effectiveness of our proposed method.