

Text-Guided Multi-Scale Frequency Representation Adaptation

Weicai Yan¹, Xinhua Ma², Wang Lin¹, Tao Jin^{1*}
¹Zhejiang University, ²Nanyang Technological University
Correspondence: jint_zju@zju.edu.cn

Abstract

Parameter-efficient fine-tuning methods introduce a small number of training parameters, enabling pre-trained models to adapt rapidly to new data distributions. While these methods have shown promising results, they exhibit notable limitations. First, most existing methods operate in the signal space domain, which results in substantial information redundancy. Second, most existing methods utilize fixed prompts or adaptation layers, failing to fully account for the multi-scale characteristics of signals. To address these challenges, we propose the Multi-Scale **F**requency **A**dapter (FreqAdapter), which integrates textual information and performs multi-scale fine-tuning of signals in the frequency domain. Additionally, we introduce a multi-scale adaptation strategy to optimize receptive fields across different frequency ranges, further enhancing the model’s representational capacity. Extensive experiments on multimodal models, including CLIP and LLaVA, demonstrate that FreqAdapter significantly improves both performance and efficiency. FreqAdapter improves performance with minimal cost and fast convergence within one epoch. Code is available at <https://github.com/Kelvin-ywc/FreqAdapter>.

1 Introduction

The existing multimodal foundation models (Radford et al., 2021; Ilharco et al., 2021; Rombach et al., 2022; Li et al., 2024a; Bai et al., 2025; Guo et al., 2025) demonstrate powerful feature representation capabilities. Benefiting from the rapid advancements in large language models, recent research (Liu et al., 2023b,a, 2024) has shifted its focus to developing large vision-language models. These models encode images and align visual features with large language models, enabling efficient multimodal integration. For example,

*Corresponding author.

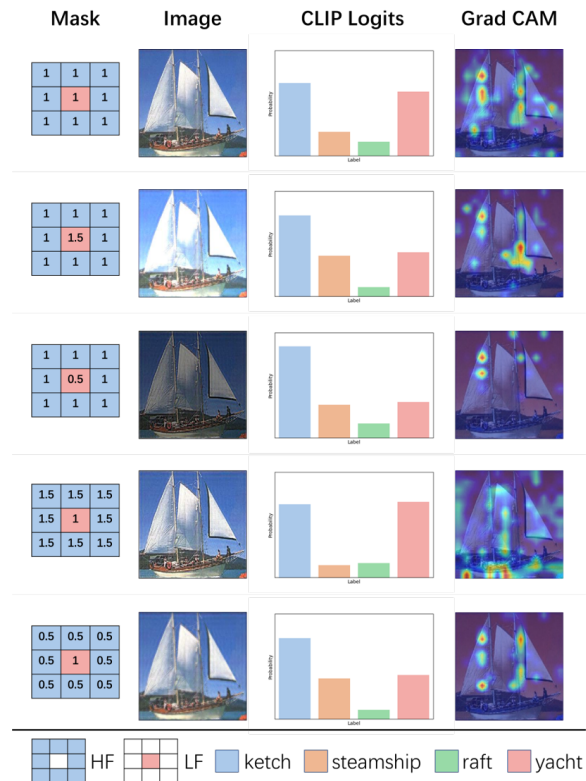


Figure 1: The effect of different frequency adaptations on CLIP predictions and attention. **Mask** represents the adjustments applied to the frequency information. **Image** shows the RGB image after these adjustments. **CLIP logits** indicate the prediction probabilities for four classes: ketch (the correct label), steamship, raft, and yacht. **Grad CAM** visualizes the attention regions of CLIP.

CLIP (Radford et al., 2021) is pre-trained on large-scale datasets and learns joint visual-textual representations. Building upon CLIP’s vision encoder as the image feature extractor, LLaVA (Liu et al., 2023b) learns a projection layer to align image features with the large language model Vicuna (Zheng et al., 2023).

Numerous studies focus on improving the performance of multimodal foundation models under constrained computational resources. Efficient fine-tuning techniques, such as prompt tuning (Zhou

et al., 2022b,a; Khattak et al., 2023; Qiu et al., 2024; Xiao et al., 2025), adapter tuning (Zhang et al., 2021; Gao et al., 2024; Yang et al., 2024; Zarei et al., 2025), and LoRA-based methods (Hu et al., 2021; Singhal et al., 2025), achieve this by optimizing only a small subset of parameters, enabling the models to effectively adapt to new data distributions. Additionally, some approaches (Yu et al., 2024; Yang et al., 2023a,b) leverage prompt construction or extract informative features from other pre-trained models as useful information. However, existing studies exhibit several limitations. First, most methods directly fine-tune features in the spatial domain, which contains a significant amount of redundant information. Second, most existing approaches adjust the entire feature set uniformly, failing to consider its inherent multi-scale characteristics.

To address the aforementioned issues, this paper aims to adjust image frequency domain information to fully harness model performance. We first conduct an experiment, as illustrated in the Fig. 1. We transform images into the frequency domain and apply different masks to the frequency components. The modified images are then converted back to the spatial domain and classified using the CLIP model. Additionally, Grad-CAM is employed to visualize the regions of interest attended to by CLIP. The results reveal that adjustments in the frequency domain significantly impact both CLIP logits and Grad-CAM visualizations.

Motivated by these findings, this paper proposes FreqAdapter, which leverages textual information to perform multi-scale adjustments on image frequency domain features. Specifically, we first apply a Discrete Cosine Transform (DCT) to convert the image from the spatial domain to the frequency domain. Then, FreqAdapter performs multi-scale adaptations conditioned on textual information to refine the frequency-domain features. The adjusted frequency domain features are subsequently transformed back to the spatial domain using an Inverse Discrete Cosine Transform (IDCT) and finally fed into the encoder for feature extraction. Subsequently, we conduct experiments on multiple multimodal models, including CLIP and LLaVA, and the results demonstrate the effectiveness of our proposed method. Our contributions can be summarized as follows: (1) We propose FreqAdapter, which adjusts visual information in the frequency domain according to the textual information. (2) We design a multi-scale adaptation strategy to per-

form multi-scale adjustments on frequency domain features. (3) We conduct experiments on multi-modal models, and the results demonstrate the effectiveness of our method.

2 Preliminary

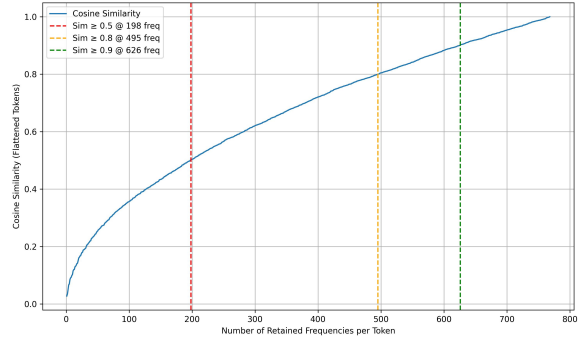


Figure 2: Information Concentration Illustration.

2.1 Frequency vs Spatial Adaptation

Proposition 1 (Information Concentration in the Frequency Domain). To analyze the information distribution, we transform a visual embedding E_i into its frequency representation X_i via DCT. We then create an approximation by retaining only the first k low-frequency components and reconstructing it via IDCT. Finally, we compute the cosine similarity between the original and the reconstructed embedding.

We empirically observe that semantic information is highly concentrated in the low-frequency components. As shown in Fig. 2, retaining just 25.8% of frequencies (198/768) achieves a cosine similarity of 0.5. This similarity exceeds 0.8 and 0.9 when retaining 64.5% (495) and 81.5% (626) of components, respectively. This evidence supports our proposition that DCT-transformed embeddings concentrate essential information in low-frequency bands, enabling compact and effective adaptation.

2.2 Aggregation of Frequency Coefficients over Spatial Neighborhoods

For a sequence of W spatially contiguous tokens, $\mathbf{E} = [E_0, \dots, E_{W-1}]$, each token E_j is first transformed into its frequency-domain representation X_j via the DCT.

We then compute the aggregated frequency feature \bar{X} by averaging the coefficients across the

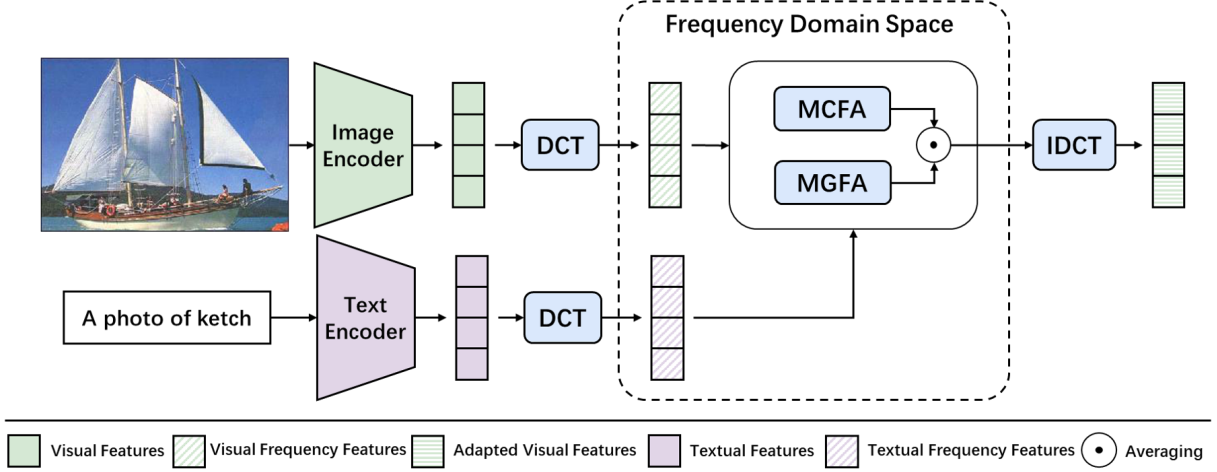


Figure 3: The framework of Multi-Scale Frequency Adapter. First, the CLIP encoder is used to encode images into visual embeddings and text into textual embeddings. At the same time, both the visual and textual embeddings are converted from the spatial domain to the frequency domain using the DCT. In the frequency domain, with the guidance of textual signals, MGFA and MCFA modules are employed to fine-tune the visual signals. Subsequently, the visual signals in the frequency domain are transformed back into the spatial domain via the IDCT as the adapted visual features.

spatial neighborhood:

$$\bar{X} = \frac{1}{W} \sum_{j=0}^{W-1} X_j. \quad (1)$$

This operation aggregates coefficients for each frequency channel independently, producing for every frequency k an averaged coefficient $\bar{X}[k] = \frac{1}{W} \sum_{j=0}^{W-1} X_j[k]$. Such aggregation effectively smooths the features in the frequency domain, yielding a stable representation that captures the overall spectral characteristics of the local region while suppressing token-level noise. A detailed mathematical interpretation and stability analysis are provided in Appendix A.

3 Multi-Scale Frequency Adapter

We introduce **FreqAdapter**, an efficient framework for text-guided fine-tuning of visual embeddings in the frequency domain. An image–text pair is first encoded by CLIP into a shared space, yielding visual embeddings $E_v \in \mathbb{R}^{S_v \times D_v}$ and textual embeddings $E_t \in \mathbb{R}^{S_t \times D_t}$. Both are transformed to the frequency domain via the DCT, producing X_v and X_t . Operating in the frequency domain offers compact, band-wise control and supports stable multi-scale aggregation.

FreqAdapter refines X_v under the guidance of X_t , producing adapted representations \tilde{X}_v . It consists of two lightweight components: the **Multi-Scale Global Frequency Adapter (MGFA)** for

global calibration and the **Multi-Scale Cross-Modal Frequency Adapter (MCFA)** for text-guided refinement. The refined features are mapped back to the spatial domain through the Inverse DCT to obtain \tilde{E}_v :

$$\tilde{X}_v = \text{FreqAdapter}(\text{DCT}(E_v), \text{DCT}(E_t)), \quad (2)$$

$$\tilde{E}_v = \text{IDCT}(\tilde{X}_v). \quad (3)$$

3.1 Multi-Scale Adaptation Strategy

To capture diverse receptive fields, we adopt a **multi-scale adaptation strategy** in the frequency domain. Given $X_v \in \mathbb{R}^{S_v \times D_v}$ and $X_t \in \mathbb{R}^{S_t \times D_t}$, the goal is to refine X_v by aggregating contextual information across scales.

We define N scales with two adapter sets: $\mathbf{G} = \{\mathcal{G}_1, \dots, \mathcal{G}_N\}$ (MGFA) and $\mathbf{C} = \{\mathcal{C}_1, \dots, \mathcal{C}_N\}$ (MCFA). At each scale n , features are reshaped to a grid $X_v^{(0)} \in \mathbb{R}^{H \times W \times D_v}$ and downsampled by $\text{Down}(\cdot, 2^{n-1})$:

$$X_v^{(n)} = \text{Down}(X_v^{(0)}, 2^{n-1}). \quad (4)$$

Let $X_{v,n} = \text{Seq}(X_v^{(n)})$ denote the flattened sequence. We then refine and fuse:

$$G_n = \mathcal{G}_n(X_{v,n}), \quad C_n = \mathcal{C}_n(X_{v,n}, X_t), \quad (5)$$

$$\tilde{X}_{v,n} = G_n + w C_n, \quad (6)$$

where w balances global and cross-modal adaptation. An interleave-repeat upsampling restores the

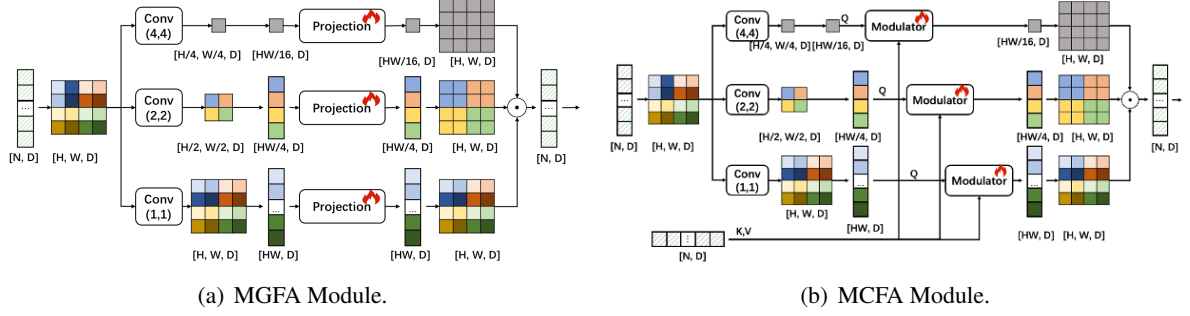


Figure 4: Overview of the proposed FreqAdapter. The CLIP vision encoder divides an image into patches and encodes them into a flattened representation of shape $[N, D]$, where $N = H \times W$ and D is the embedding dimension. For multi-scale aggregation, the sequence is reshaped to $[H, W, D]$. Spatial downsampling over (H, W) (e.g., strided average pooling or strided convolution) aggregates information under different receptive fields. The aggregated features are then refined by MGFA and MCFA and finally restored to the original $[N, D]$ via a repeat-interleave operation.

original spatial size:

$$\hat{X}_{v,n} = \text{IR}(\tilde{X}_{v,n}, 2^{n-1}). \quad (7)$$

Finally, outputs from all scales are averaged:

$$\tilde{X}_v = \frac{1}{N} \sum_{n=1}^N \hat{X}_{v,n}. \quad (8)$$

3.2 Multi-Scale Global Frequency Adapter

The MGFA module globally calibrates visual frequency features to ensure stable and consistent adaptation across scales. Because frequency coefficients are inherently coupled, local modulation alone may cause scale misalignment. MGFA applies a lightweight bottleneck transformation that adjusts each frequency channel globally:

$$G_n = f(X_{v,n}), \quad (9)$$

where $f(\cdot)$ is a two-layer projection with a ReLU activation.

3.3 Multi-Scale Cross-Modal Frequency Adapter

The MCFA module injects textual guidance into visual frequency representations for fine-grained cross-modal alignment. Working in the frequency domain aggregates semantic cues across bands while reducing spatial redundancy. At each scale n , MCFA predicts modulation parameters from X_t and applies them to the visual features:

$$\gamma, \beta = \text{Modulator}(X_t), \quad (10)$$

$$C_n = \gamma \odot X_{v,n} + \beta, \quad (11)$$

where $\text{Modulator}(\cdot)$ is a lightweight two-layer MLP. Together with MGFA, MCFA provides complementary text-aware adaptation, enhancing overall representational consistency.

3.4 Training and Inference

We optimize the adapter with the CLIP contrastive loss on paired image–text data. Inference follows the same forward procedure. The trained FreqAdapter is plug-and-play for CLIP-based vision–language models (e.g., LLaVA), enabling richer, text-conditioned visual features without modifying the backbone.

4 Experiments

In Section 4.1, we first train FreqAdapter on the CLIP model and evaluate its performance on the image–text retrieval task. Subsequently, in Section 4.2, we integrate FreqAdapter into LLaVA and assess its performance on Visual Question Answering (VQA) tasks.

4.1 Quantitative Analysis on Retrieval Tasks

Foundation Model. We conduct experiments on the CLIP-B/16, CLIP-L/14, and CLIP-L/14-336 as foundation models. The input image size for CLIP-B/16 and CLIP-L/14 is 224×224 , while for CLIP-L/14-336, it is 336×336 . It is worth noting that the vision encoder in CLIP-L/14-336 is the same as the one used in LLaVA 1.5.

Dataset. FreqAdapter is first trained on the MS COCO 2017 (Lin et al., 2014) training set, and then evaluated on the COCO 2017 validation set. Additionally, to assess the zero-shot capability, we evaluate the finetuned models on Flickr30K (Plummer et al., 2015) validation and test set.

Evaluation Metrics. We evaluate the model’s capabilities in both image-to-text retrieval and text-to-image retrieval, and report the results for R@1, R@5, and R@10. R@k represents the percentage of cases where the correct label appears within

Foundation Model	Method	COCO 2017						Flickr30K Val					
		I2T			T2I			I2T			T2I		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP-B/16	-	51.82	76.80	84.34	32.65	57.60	68.04	85.30	97.00	98.60	62.28	85.64	92.14
	CoOp	54.76	78.64	86.78	38.53	63.19	75.27	84.60	97.50	99.10	<u>72.04</u>	93.22	<u>96.26</u>
	MaPLe	52.72	76.18	85.10	37.26	62.13	73.21	82.40	97.20	99.00	70.36	92.04	95.04
	CLIP-Adapter	<u>56.30</u>	<u>80.32</u>	<u>87.88</u>	<u>41.60</u>	<u>68.57</u>	<u>78.32</u>	83.90	97.60	<u>99.20</u>	71.26	92.64	96.08
	MMA	55.80	80.04	87.66	39.41	65.53	76.09	<u>86.00</u>	<u>97.80</u>	98.90	70.56	91.56	95.88
	FreqAdapter	57.96	82.78	89.66	43.30	70.58	80.04	86.80	98.50	99.30	73.42	<u>92.88</u>	96.56
CLIP-L/14	-	56.14	79.64	86.94	35.51	59.90	70.26	86.60	97.90	<u>99.40</u>	66.36	88.50	93.62
	CoOp	59.02	82.30	88.74	43.26	69.84	78.29	87.70	<u>98.50</u>	99.30	74.96	92.12	96.14
	MaPLe	57.26	81.76	88.02	44.21	71.26	77.26	86.10	98.00	99.20	71.02	91.28	94.16
	CLIP-Adapter	<u>60.38</u>	<u>83.12</u>	89.58	43.18	69.92	<u>79.28</u>	87.30	98.20	99.30	75.76	<u>93.24</u>	<u>96.44</u>
	MMA	60.20	<u>83.12</u>	89.76	42.58	68.32	78.22	87.20	98.10	99.30	73.14	92.38	96.10
	FreqAdapter	61.02	83.30	<u>89.68</u>	<u>44.18</u>	<u>70.04</u>	79.55	<u>87.50</u>	98.70	99.60	<u>75.72</u>	93.72	96.86
CLIP-L/14-336	-	57.34	80.38	87.64	36.08	60.70	70.66	89.80	99.10	99.80	69.12	90.20	94.94
	CoOp	59.64	82.70	88.96	<u>44.74</u>	71.63	79.14	90.20	<u>98.70</u>	99.60	75.62	94.22	<u>97.28</u>
	MaPLe	57.32	82.34	89.02	44.45	70.16	<u>79.62</u>	88.30	97.50	99.10	71.24	92.04	96.38
	CLIP-Adapter	60.42	82.24	90.76	44.62	70.20	79.58	90.00	98.40	99.60	<u>77.28</u>	<u>94.72</u>	97.24
	MMA	<u>60.90</u>	82.58	<u>90.48</u>	43.55	69.46	78.89	89.90	99.10	99.80	75.68	93.66	96.60
	LoR-VP	59.62	83.72	88.20	42.35	68.17	77.46	<u>90.30</u>	98.30	99.50	73.46	93.86	96.40
FreqAdapter	61.42	<u>83.64</u>	90.10	45.23	<u>70.92</u>	80.02	90.90	<u>98.70</u>	<u>99.70</u>	77.60	95.06	97.44	

Table 1: Evaluation on Image-Text Retrieval. **Bold: best results**, Underline: second best results.

the top-k predicted scores. For the COCO 2017 and Flickr30k datasets, each image corresponds to multiple texts. In the retrieval task, the retrieval is considered successful if any of the associated texts are retrieved.

Baseline. We choose several parameter-efficient fine-tuning methods as baselines, including CoOp (Zhou et al., 2022b), MaPLe (Khattak et al., 2023), CLIP-Adapter (Gao et al., 2024), and MMA (Yang et al., 2024), and LoR-VP (Jin et al., 2025). Further experimental details are provided in Appendix C.

Experiment Detail. During training, the batch size is set to 128, the epoch is set to 1 and the AdamW optimizer is employed with a learning rate of 0.001. The multimodal weight parameter w is set to 0.01 for retrieval tasks and 1.0 for VQA tasks. All experiments were conducted on a single A100-40G GPU.

Result. As shown in Tab. 1, our FreqAdapter consistently achieves state-of-the-art performance among parameter-efficient fine-tuning methods on both the COCO 2017 and Flickr30K benchmarks. Results for Flickr30K test subset are provided in Appendix G. On COCO, FreqAdapter notably surpasses all prompt- and adapter-based baselines, improving the R@1 score by up to **4.9%** for the text-to-image retrieval task under the CLIP-L/14 setting. On Flickr30K, it further demonstrates superior generalization, outperforming other approaches across

all retrieval directions and achieving the highest scores on both I2T and T2I tasks.

We attribute these gains to the advantages of frequency-domain adaptation. Unlike the spatial domain—which contains high redundancy and often suffers from overfitting under limited fine-tuning budgets—the frequency domain provides compact and disentangled representations that facilitate stable optimization. Moreover, FreqAdapter’s multi-scale design effectively integrates fine- and coarse-grained frequency information, enabling efficient and robust adaptation without sacrificing generalization—a common weakness of traditional spatial adapters.

4.2 Quantitative Analysis on VQA Tasks

Base Model and Dataset. We select LLaVA1.5 (Liu et al., 2023a) as the base model and validate the effectiveness of the method on both the 7B and 13B versions. Experiments are conducted on two datasets, MM-Vet (Yu et al., 2023) and LLaVA-Bench (Liu et al., 2023c). Both datasets use GPT-based evaluation scores to assess model performance.

Baseline. We follow the work API (Yu et al., 2024) and directly use the results reported in the paper as our baseline for comparison. We conduct additional experiments on the LLaVA 1.5 7B model. The evaluation methods included having the model answer questions directly (referred to as “w/o

Foundation Model	Method	Dataset	
		MM-Vet	LLaVA-Bench
LLaVA 1.5-7B	w/o prompt	30.9	64.3
	CLIP-Adapter	27.1	61.8
	FreqAdapter	31.8	64.8
LLaVA 1.5-13B	w/o prompt	32.8	71.9
	+Step-by-Step	33.7	73.5
	FGVP(Mask)	31.0	57.4
	FGVP(RBM)	25.0	57.4
	SoM	26.4	56.1
	API (CLIP)	35.3	74.1
	API (LLaVA)	36.6	74.8
	CLIP-Adapter	32.9	64.9
FreqAdapter	37.4	72.4	

Table 2: Performance Evaluation on VQA.

prompt”) and using a chain-of-thought approach (named “Step-by-step”). The compared approaches include visual prompting methods FGVP (Yang et al., 2023b) and SoM (Yang et al., 2023a), as well as CLIP-Adapter and FreqAdapter.

Result. The experimental results are shown in the Tab. 2. Detailed result is provided in Appendix D. From the table, we can conclude that the FreqAdapter trained on CLIP can be directly integrated into LLaVA. On the LLaVA-Bench dataset, FreqAdapter shows a notable improvement over the "w/o prompt" baseline, while on the MM-Vet dataset, it achieves impressive performance on both the 7B and 13B models. In contrast, CLIP-Adapter exhibits a performance degradation trend, likely because it is tailored to a specific data distribution, which, to some extent, compromises the model’s generalization capability. Meanwhile, FreqAdapter facilitates cross-modal interaction in the frequency domain, enabling the integration of visual features extracted from textual information, thereby providing a stronger expressive capability.

4.3 Qualitative Analysis

This section provides a qualitative analysis of FreqAdapter’s performance on both CLIP and LLaVA. Additional results can be found in the Appendix F, and qualitative analysis for CLIP model is in Appendix E.

Qualitative Analysis for LLaVA. In this section, we qualitatively analyze the effect of integrating the adapter tuning method into the LLaVA approach. Additional results can be found in the supplementary material, and the outcomes are illustrated in the Fig. 5. In the image, various gasoline prices are provided, and the task is to calculate how many gallons can be purchased with \$50. LLaVA’s re-


Question:	
How many gallons of supreme gasoline can I get with \$50?	
LLaVA: To determine the number of gallons of supreme gasoline you can get with \$50, we need to know the price per gallon. Unfortunately, the image does not provide the price per gallon. However, we can still calculate the number of gallons you can get with \$50 by dividing the amount by the price per gallon.	
CLIP-Adapter: To determine how many gallons of supreme gasoline you can get with \$50, we need to know the current price per gallon. In the image, the price of supreme gasoline is displayed on a sign, but the exact price per gallon is not provided. If you could provide the price per gallon, we could calculate the number of gallons you can purchase with \$50.	
FreqAdapter: To determine how many gallons of supreme gasoline you can get with \$50, we need to know the price per gallon. In the image, the price of supreme gasoline is \$3.66 per gallon. If you have \$50, you can calculate the number of gallons you can purchase by dividing the amount of money by the price per gallon:	
$\$50 / \$3.66 = 13.72$ gallons	
So, with \$50, you can purchase approximately 13.72 gallons of supreme gasoline.	
Answer: 13.6<OR>13.7	

Figure 5: Qualitative Analysis for LLaVA.

sponse indicates that it fails to fully comprehend the content of the image. Although CLIP-Adapter exhibits a stronger understanding of the image, it only offers a description without identifying the specific numerical values. In contrast, FreqAdapter accurately identifies the price of supreme gasoline, performs the arithmetic calculation, and ultimately produces the correct result.

4.4 Ablation Study

In this section, we validate the effectiveness of different modules. In appendix H, we explore the impact of three hyperparameters, top-k, multimodal weight, and downsampling factor on the experimental results. We observe that FreqAdapter, when applied in the frequency domain, can be effectively integrated with spatial-domain fine-tuning method clip CLIP-Adapter, leading to further performance improvements. The experimental setup is consistent with the comparative experiments.

MGFA	MCFA	MSCOCO I2T			MSCOCO T2I		
		R@1	R@5	R@10	R@1	R@5	R@10
-	-	57.34	80.38	87.64	36.08	60.70	70.66
-	✓	58.16	81.90	88.94	42.81	67.86	77.43
✓	-	58.70	82.28	89.54	43.47	68.66	78.04
✓	✓	61.42	83.64	90.10	45.23	70.92	80.02

Table 3: Effectiveness of FreqAdapter Modules.

Effectiveness of Frequency Adapter. We evaluate the proposed modules on CLIP-L/14-336, and the results in Tab. 3 show that both MCFA and MGFA improve retrieval performance. MCFA brings a larger gain, highlighting its importance for multimodal alignment. MGFA provides additional benefits by capturing frequency-specific information. Combining both modules achieves the best overall results, confirming their complementarity in

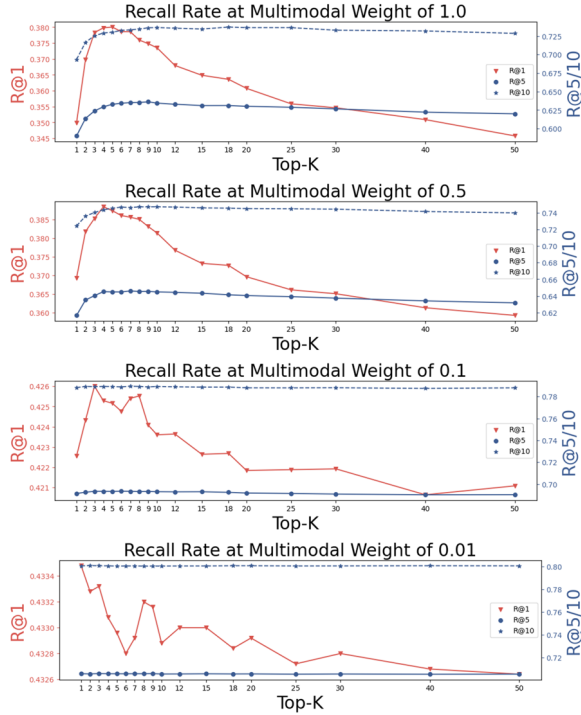


Figure 6: Effectiveness of Top-K and Multimodal Weight.

enhancing cross-modal understanding and adaptability.

5 Hyperparameter Analysis

Top-K and Multimodal Weight. In this section, we conduct ablation experiments on top-k and multimodal weight, visualizing the relationship between R@K and top-k under different multi-modal weights, where the multi-modal weight takes values of 1.0, 0.5, 0.1, and 0.01. As can be seen from the Fig. 6, relatively smaller multi-modal weights result in higher R@k values, indicating that excessive cross-modal interaction reduces accuracy. This is because excessive cross-modal information interferes with the feature extraction process of the current modality. At the same time, when the multi-modal weight is larger, R@1 is more affected by top-k. As the top-k value increases, R@1 first increases and then decreases, reaching its maximum around top-k = 5. For R@5 and R@10, the values first increase and then slowly decrease, eventually stabilizing. One reason is that for the COCO 2017, each image has five associated captions. Selecting the top five relevant captions allows for more comprehensive fine-tuning of visual features. Selecting fewer captions leads to information shift, while selecting too many captions introduces excessive irrelevant information. By comparison, R@5

and R@10 emphasize the model’s recall ability and are therefore more tolerant of precision. When the multi-modal weight is smaller, the impact of top-k on R@k is relatively minor.

6 In-depth Analysis

In this section, we provide a comprehensive analysis of the advantages of frequency-domain fine-tuning over spatial-domain adaptation, the role of multi-scale mechanisms, and the computational complexity of the proposed model. Furthermore, in Appendix I, we present an extended discussion on the integration of FreqAdapter and CLIPAdapter, along with additional comparative results between frequency-domain and spatial-domain adaptations.

Frequency vs. Spatial Adaptation. To verify the effectiveness of frequency-domain fine-tuning, we conduct a controlled comparison in which we remove the frequency transform and fine-tune directly in the spatial domain, denoted *SpatialAdapter*. *SpatialAdapter* is strictly matched to *FreqAdapter* in backbone, parameter count, optimization settings, and training schedule. We train on COCO 2017 and report curves for training loss, validation loss, and image-to-text accuracy on COCO and Flickr30k. The COCO 2017 dataset contains a total of 591,753 image-text pairs, with 4,623 steps per epoch. The results are shown in Fig. 7. Additional metrics, including T2I and further diagnostics are provided in the Appendix. I.2.

From the plots, FreqAdapter trains more stably and converges within a single epoch, with consistent gains in fine-tuning accuracy on COCO validation and in generalization on Flickr30k validation. In contrast, *SpatialAdapter* exhibits overfitting and collapses in the latter half of the epoch. We attribute this to the advantages of frequency-domain adaptation: representations in the spectral domain are less redundant, with low- and high-frequency structure disentangled and noise separated, yielding cleaner parameter updates and a smoother optimization trajectory.

Effectiveness of Multi-Scale. Fig. 8 illustrates the visual embedding refinement effect of FreqAdapter. The first column shows the original image, while the following three columns correspond to the aggregation results using 1×1, 2×2, and 4×4 windows respectively. For each given caption, we visualize the feature modifications under different scales, where darker colors indicate stronger adjustments made by FreqAdapter. As observed, FreqAdapter

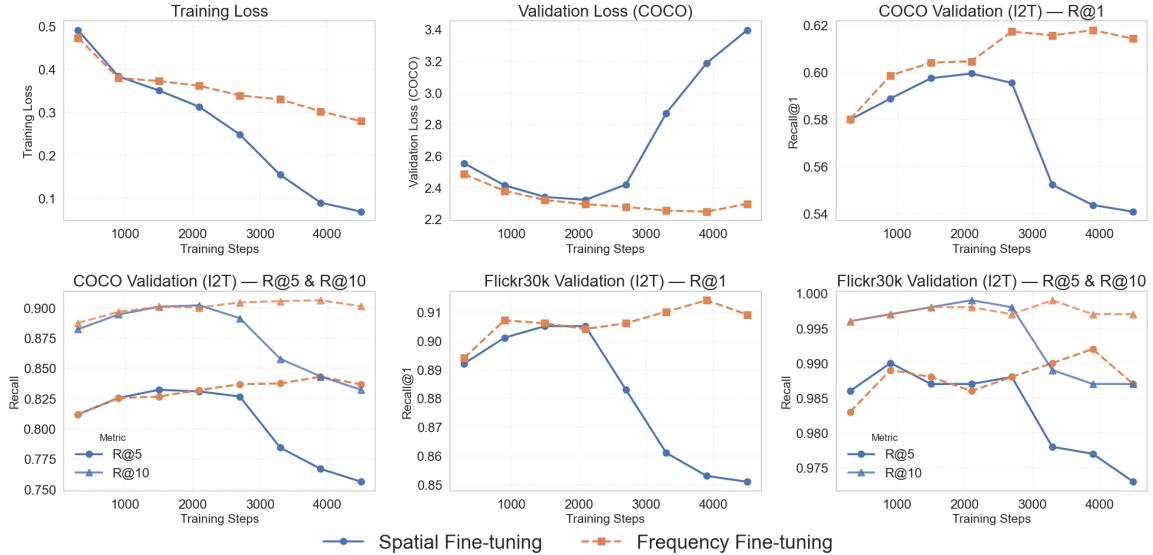


Figure 7: Frequency vs. Spatial Adaptation.

adaptively enhances the semantically relevant regions of the visual representation. For example, in the first row, the caption mentions a clock tower, and the model consistently focuses on the tower region across multiple scales — finer scales emphasize detailed components, while coarser scales capture the global structure. This demonstrates that FreqAdapter effectively integrates semantic cues to perform multi-scale frequency adaptation.

Method	Param	Param%	GFLOPS
CLIP	-	-	362.5
CoOp	16.4k	0.003	370.8
MaPLE	798.7k	0.19	362.9
CLIP-Adapter	524.3k	0.12	362.5
MMA	118.7k	0.03	362.7
FreqAdapter	476.4k	0.11	362.6

Table 4: Computational Complexity.

Computational Complexity. The detailed network architecture is provided in the Appendix. B. As shown in Table 4, the proposed FreqAdapter introduces a highly competitive number of additional parameters. Compared with other multimodal interaction methods such as MaPLE, FreqAdapter requires substantially fewer parameters—less than that of a single linear layer—while maintaining comparable performance. In terms of computational cost, the GFLOPs of FreqAdapter remain almost identical to other efficient fine-tuning approaches such as CLIP-Adapter and MMA. Considering the total inference cost of the backbone CLIP model (362.5

GFLOPs), the additional computational overhead introduced by FreqAdapter is negligible.

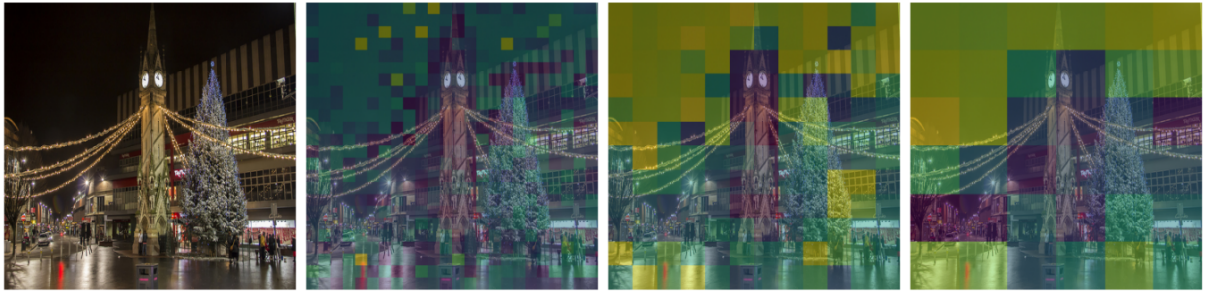
7 Related Work

7.1 Multimodal Foundation Model

The advent of the Transformer (Vaswani, 2017) architecture marked a significant breakthrough in natural language processing and has since been extended to various multimodal tasks. Works like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) are pre-trained using the contrastive learning to align images and text into a shared embedding space. GIT (Wang et al., 2022) and BLIP2 (Li et al., 2023) use image-text pairs for training, further scaling up the model size. Recent works (Bai et al., 2025; Li et al., 2024a; Guo et al., 2025; Liu et al., 2023c; Cheng et al., 2025) on multimodal large language models employ a pre-trained vision encoder to extract image embeddings, which are aligned with the language model via a vision projection layer, bridging the modality gap and enabling joint visual–language reasoning.

7.2 Parameter-Efficient Fine-Tuning

Common efficient fine-tuning techniques include prompt tuning (Zhou et al., 2022b,a; Khattak et al., 2023; Qiu et al., 2024; Fu et al., 2024), adapter tuning (Zhang et al., 2021; Gao et al., 2024; Yang et al., 2024; Zarei et al., 2025; Xie et al., 2025), and LoRA (Hu et al., 2021; Dettmers et al., 2023; Zhang et al., 2023). In prompt tuning, learnable tokens are introduced at the input layer of the model,



Caption: Series of light decorations at a city center with a tree and a clock tower.



Caption: A pair of people ride a small, horse drawn buggy near a large puddle of water.

Figure 8: Effectiveness of Multi-Scale Strategy.

allowing the model to adapt to new tasks. Adapter tuning introduces learnable networks between layers, typically connected in a residual manner, enabling the model to adjust with minimal changes to its architecture. LoRA learns low-rank vectors and injects them into the original model to adapt the model without significant overhead. For ensemble methods (Yu et al., 2024; Yang et al., 2023a,b), common strategies include generating prompts with new models or extracting attention maps from the model itself for self-reflection. Recent work (Li et al., 2025) explores cross-modal fusion during the fine-tuning stage. Most existing approaches, however, perform such fine-tuning directly in the spatial domain.

7.3 Fourier Transform

Fourier-based transformations, including Fourier Transform, Discrete Fourier Transform, Discrete Cosine Transform, and Wavelet Transform, have played pivotal roles in signal and image processing. Recent advances (Xu et al., 2020; Li et al., 2024b; Qian et al., 2024; Tan et al., 2024) have seen these transformations combined with modern deep learning techniques to enhance model performance. SpectFormer (Patro et al., 2023) integrates wavelet transform with transformer architecture. VFPT (Zeng et al., 2024) integrates the Fast Fourier Transform into prompt embeddings, effectively combining spatial and frequency domain

information. SFMFusion (Sun et al., 2025) introduces frequency-domain processing for visual features, where representations are transformed back to the spatial domain for fusion. DAFF-Net (Zhou and Hu, 2024), on the other hand, models interactions between high- and low-frequency components within a single modality. Building on prior work, our method extends frequency-domain modeling to cross-modal settings by fusing latent embeddings directly in the frequency domain, preserving an end-to-end frequency-domain formulation.

8 Conclusion

In this paper, we propose FreqAdapter, which performs multi-scale adaptation and enables modality interactions in the frequency domain. Compared to methods adapting in the spatial domain, our approach achieves superior fitting performance, stronger generalization capability, and faster adaptation speed, which provides a unique perspective for parameter-efficient fine-tuning methods.

9 Acknowledge

This work was supported by the ‘‘Pioneer’’ and ‘‘Leading Goose’’ R&D Program of Zhejiang under (Grant No. 2025C02110), Public Welfare Research Program of Ningbo under (Grant No. 2024S062), and Yongjiang Talent Project of Ningbo under (Grant No. 2024A-161-G).

Limitations

Although FreqAdapter achieves strong empirical results, its theoretical depth can be further explored to better explain the effectiveness of frequency-domain adaptation. Moreover, our current study is limited to CLIP-based settings with relatively small parameter scales, which may constrain the upper bound of performance. Future work could extend this approach to larger multimodal language models, and investigate how scaling parameters balances performance gains with computational efficiency.

References

- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Xize Cheng, Dongjie Fu, Chenyuhao Wen, Shannon Yu, Zehan Wang, Shengpeng Ji, Siddhant Arora, Tao Jin, Shinji Watanabe, and Zhou Zhao. 2025. **AHAbench: Benchmarking audio hallucinations in large audio-language models**. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.
- Dongjie Fu, Xize Cheng, Xiaoda Yang, Wang Hanting, Zhou Zhao, and Tao Jin. 2024. Boosting speech recognition robustness to modality-distortion with contrast-augmented prompts. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3838–3847.
- Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595.
- Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, and 1 others. 2025. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. **Openclip**. If you use this software, please cite it as below.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.
- Can Jin, Ying Li, Mingyu Zhao, Shiyu Zhao, Zhenting Wang, Xiaoxiao He, Ligong Han, Tong Che, and Dimitris N. Metaxas. 2025. **Lor-VP: Low-rank visual prompting for efficient vision model adaptation**. In *The Thirteenth International Conference on Learning Representations*.
- Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Ke Li, Di Wang, Zhangyuan Hu, Wenxuan Zhu, Shaofeng Li, and Quan Wang. 2024b. Unleashing channel potential: Space-frequency selection convolution for sar object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17323–17332.
- Qilei Li, Wenhao Song, Mingliang Gao, Wenzhe Zhai, Qiang Zhou, and Zhao Huang. 2025. Towards text-refereed multi-modal image fusion by cross-modality interaction. *Signal Processing*, 237:110073.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13*, pages 740–755. Springer.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. **Llava-next: Improved reasoning, ocr, and world knowledge**.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Badri N Patro, Vinay P Namboodiri, and Vijay Srinivas Agneeswaran. 2023. Spectformer: Frequency and attention is what you need in a vision transformer. *arXiv preprint arXiv:2304.06446*.
- Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.
- Yurui Qian, Qi Cai, Yingwei Pan, Yehao Li, Ting Yao, Qibin Sun, and Tao Mei. 2024. Boosting diffusion models with moving average sampling in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8911–8920.
- Chen Qiu, Xingyu Li, Chaithanya Kumar Mummadi, Madan Ravi Ganesh, Zhenzhen Li, Lu Peng, and Wan-Yi Lin. 2024. Federated text-driven prompt generation for vision-language models. In *The Twelfth International Conference on Learning Representations*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.
- Raghav Singhal, Kaustubh Ponkshe, and Praneeth Vepakomma. 2025. Fedex-lora: Exact aggregation for federated and efficient fine-tuning of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1316–1336.
- Hui Sun, Long Lv, Pingping Zhang, Tongdan Tang, Feng Tian, Weibing Sun, and Huchuan Lu. 2025. Spatial-frequency enhanced mamba for multi-modal image fusion. *IEEE Transactions on Image Processing*, 34:7684–7696.
- Jiangtong Tan, Jie Huang, Naishan Zheng, Man Zhou, Keyu Yan, Danfeng Hong, and Feng Zhao. 2024. Revisiting spatial-frequency information integration from a hierarchical perspective for panchromatic and multi-spectral image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25922–25931.
- A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.
- Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.
- Xi Xiao, Yunbei Zhang, Xingjian Li, Tianyang Wang, Xiao Wang, Yuxiang Wei, Jihun Hamm, and Min Xu. 2025. Visual instance-aware prompt tuning. *arXiv preprint arXiv:2507.07796*.
- Zeun Xie, Chuxin Wang, Yeqiang Wang, Sihang Cai, Shulei Wang, and Tao Jin. 2025. Chat-driven text generation and interaction for person retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5259–5270.
- Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. 2020. Learning in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023a. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.
- Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. 2023b. Fine-grained visual prompting. *Advances in Neural Information Processing Systems*, 36:24993–25006.
- Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. 2024. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23826–23837.
- Runpeng Yu, Weihao Yu, and Xinchao Wang. 2024. Attention prompting on image for large vision-language models. In *European Conference on Computer Vision*, pages 251–268. Springer.
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.
- Mohammad Reza Zarei, Abbas Akkasi, and Majid Komeili. 2025. Dual adapter tuning of vision-language models using large language models. *International Journal of Computational Intelligence Systems*, 18(1):109.
- Runjia Zeng, Cheng Han, Qifan Wang, Chunshu Wu, Tong Geng, Lifu Huang, Ying Nian Wu, and Dongfang Liu. 2024. Visual fourier prompt tuning. *Advances in Neural Information Processing Systems*, 37:5552–5585.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *Preprint*, arXiv:2306.05685.

Hongchao Zhou and Shunbo Hu. 2024. Dual-attention frequency fusion at multi-scale for joint segmentation and deformable medical image registration. *arXiv preprint arXiv:2409.19658*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022a. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

A Preliminary

A.1 Multimodal Foundation Model

CLIP consists of a vision encoder \mathcal{V} and a text encoder \mathcal{L} , which encode image and text information, respectively, into a unified embedding space. Both the vision and text encoder have L transformer layers. The image I is divided into N_v patches, which is then projected into patch embeddings $E_v^0 \in \mathbb{R}^{N_v \times D_v}$. The vision encoder processes the patch embeddings to embeddings E_v^L , and then projected into the shared space F_v . The text T is tokenized into a sequence of tokens. Each token is embedded into token embeddings E_t^0 . Similar to that in vision encoding, the token embeddings is processed by the text encoder, resulted in E_t^L , and projected into textual feature F_t .

LLaVA utilizes the CLIP vision encoder to process the input image I , encoding it into visual embeddings. A trainable projection layer is then employed to align these visual embeddings with the latent space of the large language model (LLM), resulting in the visual embeddings E_v . Simultaneously, the input text undergoes tokenization and

embedding, producing the textual embeddings E_t . These embeddings, E_v and E_t , are then integrated and fed into the large model \mathcal{M} , which generates the final output.

A.2 Discrete Cosine Transform

The modal information is encoded as a one-dimensional tensor. The Discrete Cosine Transform (DCT) transforms this signal from the spatial domain to the frequency domain. For a signal $E \in \mathbb{R}^D$, the DCT is applied as follows:

$$X[k] = \alpha(k) \sum_{n=0}^{D-1} E[n] \cos\left(\frac{\pi(2n+1)k}{2D}\right),$$

$$k = 0, 1, \dots, D-1,$$

where X is the frequency domain data, and $\alpha(k)$ are normalization factors given by:

$$\alpha(k) = \begin{cases} \sqrt{\frac{1}{D}}, & k = 0, \\ \sqrt{\frac{2}{D}}, & k \neq 0. \end{cases}$$

The Inverse Discrete Cosine Transform converts frequency domain data back to the spatial domain:

$$\tilde{E}[n] = \sum_{k=0}^{D-1} \alpha(k) X[k] \cos\left(\frac{\pi(2n+1)k}{2D}\right),$$

$$n = 0, 1, \dots, D-1.$$

A.3 More Proposition

Proposition 2. *Averaging the frequency-domain representations of W spatially adjacent tokens is equivalent to independently averaging the coefficients of each cosine basis across the spatial dimension.*

Derivation. Let $\{E_j\}_{j=0}^{W-1}$ denote W spatially contiguous tokens, and $\mathbf{X}_j = \text{DCT}(E_j)$ their frequency-domain representations. The aggregation is defined as:

$$\bar{\mathbf{X}} = \frac{1}{W} \sum_{j=0}^{W-1} \mathbf{X}_j.$$

For the k -th frequency component:

$$\begin{aligned}\bar{X}[k] &= \frac{1}{W} \sum_{j=0}^{W-1} X_j \\ &= \frac{1}{W} \sum_{j=0}^{W-1} \alpha(k) \sum_{n=0}^{D-1} E_j[n] \cos\left(\frac{\pi(2n+1)k}{2D}\right) \\ &= \frac{\alpha(k)}{W} \sum_{n=0}^{D-1} \sum_{j=0}^{W-1} E_j[n] \cos\left(\frac{\pi(2n+1)k}{2D}\right)\end{aligned}$$

Hence, frequency-domain averaging equals spatial aggregation of $\sum_{j=0}^{W-1} E_j[n]$ over the same cosine basis. Each frequency channel k is averaged independently, acting as a spatial low-pass filter that yields a smoother and more stable spectral representation.

Proposition 3. (*Structured and Stable Optimization in the Frequency Domain*) Let $T \in \mathbb{R}^{D \times D}$ denote the DCT matrix. For a loss function \mathcal{L} ,

$$\nabla_{\mathbf{e}_i} \mathcal{L} = T^\top \nabla_{\hat{\mathbf{e}}_i} \mathcal{L}, \quad \|\nabla_{\mathbf{e}_i} \mathcal{L}\|_2 = \|\nabla_{\hat{\mathbf{e}}_i} \mathcal{L}\|_2.$$

This orthogonality ensures identical gradient magnitudes in both domains, preserving optimization stability. Moreover, each dimension of $\hat{\mathbf{e}}_i$ corresponds to a distinct frequency band, enabling interpretable and selective adaptation—e.g., tuning low frequencies for global semantics and high frequencies for local details.

B Module Parameter Analysis

Let D_v , D_t , and H denote the visual, textual, and hidden dimensions, respectively. The number of trainable parameters in each module is:

$$P_{\text{MGFA}} = 2D_v H + (H + D_v),$$

$$P_{\text{MCFA}} = D_t H + 2D_v H + (H + 2D_v).$$

For N scales without parameter sharing, the total number of trainable parameters is:

$$P_{\text{Total}} = N \times (P_{\text{MGFA}} + P_{\text{MCFA}}). \quad (12)$$

For CLIP-L/14-336, $D_v = 1024$, $D_t = 768$, $H = 32$, and $N = 3$, the parameter counts are 475,776.

C Experiment Detail

For the comparative methods, CoOp introduces prompts only for the visual input, while MaPLe introduces prompts in the vision-language modality

and uses a linear layer to establish connections between the modalities. CLIP-Adapter, on the other hand, adjusts the outputs of both the vision encoder and text encoder using adapters. MMA establish the cross-modal using shared and modality-specific adapter within the last few transformer layers. LoR-VP introduces low rank prompts which enables shared and patch-specific information across image rows and columns.

In the retrieval task, where N_t captions are used to retrieve N_v images, directly interacting each image-text pair results in a time complexity of $O(N_t N_v)$, which is impractical. Therefore, in our experiment, for each image, we select the K most relevant text features and average them for cross-modal interaction. This reduces the time complexity to $O(N_v)$, meaning each image only needs to extract its features once.

FreqAdapter fine-tunes the penultimate layer of the CLIP vision encoder while simultaneously extracting features from the penultimate layer of the language encoder to facilitate interaction. For retrieval tasks, the newly obtained visual embeddings are fed into the final transformer layer, and the [CLS] token is extracted as the visual feature. For VQA tasks, these new visual embeddings are directly passed into the LLaVA multimodal projector.

For CoOp and MaPLe, we follow the original implementations by inserting prompts at the encoder input. Regarding CLIP-Adapter, in comparative experiments, we add it at the output of the vision encoder; in contrast, in the ablation studies described in Sec. I.1, we concatenate CLIP-Adapter features to the penultimate layer of the encoder.

For all experiments and analyses involving CLIP models, the text is encoded into a fixed sequence of 77 tokens. Unless otherwise specified (such as in, Appendix. 5, and Appendix. I.1, where CLIP-B/16 is used), we adopt CLIP-L/14-336 for the rest of the experiments.

Since LLaVA employs a different tokenizer and text encoder from CLIP, we apply an additional text processing step using CLIP’s tokenizer and text encoder to extract textual embeddings. These embeddings are used exclusively in the FreqAdapter module to enable consistent cross-modal interaction in the frequency domain.

D Detail Result For LLaVA

We provide detailed experimental results on the MM-Vet dataset, as shown in the Fig. 9. Com-

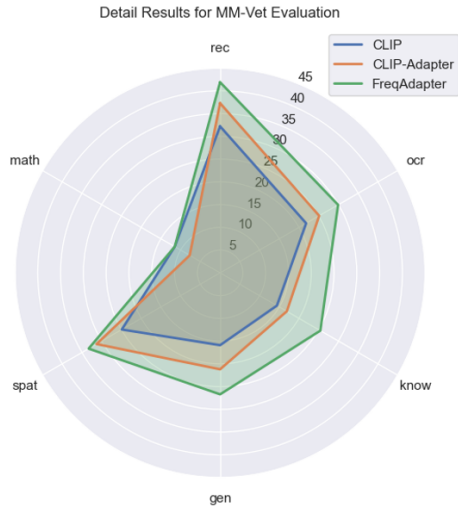


Figure 9: Detail Result For LLaVA.

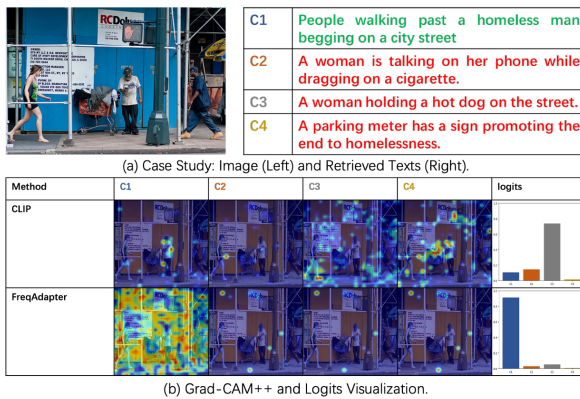


Figure 10: Caption

pared to CLIP and CLIP-Adapter, FreqAdapter demonstrates consistent performance improvements across five question-answering categories, except for math.

E Further Qualitative Analysis

Qualitative Analysis for CLIP. We employ Grad-CAM++ to visualize text-aware regions within an image, which is shown in Fig.10. Specifically, given an image and four captions [C1, C2, C3, C4], only C1 is semantically relevant to the image. Visualization results reveal that the original CLIP model struggles to correctly interpret the semantics of prompt C1, assigning the highest confidence erroneously to the irrelevant prompt C3. In contrast, FreqAdapter demonstrates a comprehensive understanding of key concepts such as “walking”, “homeless man”, and “city street”, accurately capturing global semantic information across the entire image. Activations corresponding to the irrelevant

captions appear only sparsely scattered throughout the image, clearly indicating the model’s ability to discriminate between relevant and irrelevant textual information.

F More Visualization Results

Fig.11 and 12 demonstrate additional VQA results of FreqAdapter based on LLaVA 1.5-13B, reinforcing conclusions from the qualitative analysis: FreqAdapter exhibits superior image comprehension capabilities. Specifically, as shown in Fig. 11, FreqAdapter accurately identifies all textual content in the image. In Fig. 12, it effectively makes reasonable judgments based on the visual information.

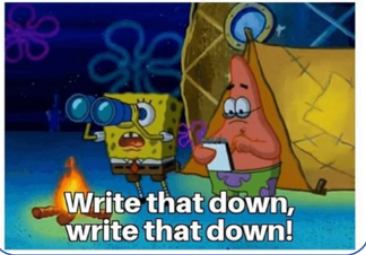
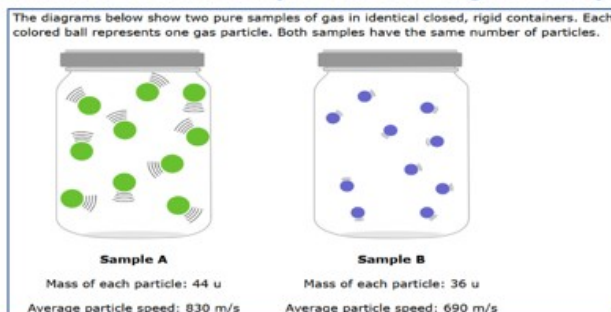
<p>Question: Can you explain this meme?</p>	<p>Professor: "Good morni..."</p> <p>Freshmen:</p> 
<p>LLaVA: The image is a meme featuring two cartoon characters, one of which is a professor and the other is a student. The professor is holding a clipboard, while the student is holding a telescope. The meme is a play on words, with the professor saying \"Good morning, freshmen\" and the student responding with \"Write tha' down, write tha' down.\" The scene is set in front of a tent, adding to the humorous and lighthearted nature of the meme.</p>	<p>CLIP-Adapter: The image is a meme featuring two popular characters from the animated TV show "Spongebob Squarepants." Spongebob is standing next to a tent, and Patrick Star is holding a paper. The meme is a play on the famous "Good morning" meme, which has been altered to include the two characters.</p> <p>The meme is set in a dark environment, possibly a cave, and the characters are positioned in a way that they appear to be looking at the camera. The combination of the two characters and the "Good morning" meme creates a humorous and lighthearted scene.</p>
<p>FreqAdapter: The image is a meme featuring two popular cartoon characters, Spongebob and Patrick, in a humorous scene. Spongebob is holding a telescope, while Patrick is holding a clipboard. The meme is captioned with a playful message, \"Professor: Good morning... Freshmen: Write that down, write that down!\" This suggests that the professor is giving instructions to the freshmen, and they are taking notes on their clipboards. The meme is a lighthearted take on the typical classroom setting, using the well-known characters to create a fun and relatable scene.</p>	<p>Answer: This meme is a joke about how new college students (freshmen) are often eager to take notes and learn everything they can in class. The image is from the cartoon SpongeBob SquarePants, where SpongeBob and Patrick are frantically writing down notes. The caption implies that the freshmen are so eager to learn that they are even writing down the professor's greeting.</p>

Figure 11: Sample 1.

Question: Compare the average kinetic energies of the particles in each sample. The diagrams below show two pure samples of gas in identical closed, rigid containers. Each colored ball represents one gas particle. Both samples have the same number of particles. Which sample has the higher temperature? Sample A or sample B?



LLaVA: To compare the average kinetic energies of the particles in each sample, we need to consider the temperature of the samples. The temperature of a gas is directly related to the average kinetic energy of its particles. In the image, we can see that Sample A has a higher number of colored balls compared to Sample B. **However, we cannot determine the temperature of the samples just by looking at the image.** To determine the temperature, we would need to know the specific heat capacity of the gas particles and the temperature at which the samples were prepared. Without this information, we cannot accurately compare the average kinetic energies of the particles in each sample.

CLIP-Adapter: To compare the average kinetic energy of the particles in each sample, we need to determine the temperature of each sample. The temperature of a gas is directly proportional to the average kinetic energy of its particles. In the image, there are two identical containers filled with colored balls, representing gas particles.

In Sample A, the balls are green, and in Sample B, the balls are blue. Since both samples have the same number of particles, we can assume that the containers have the same volume. The temperature of a gas can be determined by measuring the average kinetic energy of its particles. The average kinetic energy of the particles in Sample A is represented by the green balls, and the average kinetic energy of the particles in Sample B is represented by the blue balls.

Since the balls are moving around in the containers, we can infer that the particles in both samples have some kinetic energy. **However, without knowing the exact speed and distribution of the balls, we cannot determine the exact temperature of each sample. To determine the temperature, we would need to know the average kinetic energy of the particles, which is related to their speed and distribution.**

FreqAdapter: To compare the average kinetic energies of the particles in each sample, we need to consider the temperature of the samples. The temperature of a gas is directly related to the average kinetic energy of its particles. **In the image, we can see that Sample A has a higher temperature than Sample B. This is evident from the fact that the balls in Sample A are moving faster than those in Sample B. The balls in Sample A are more spread out, indicating higher kinetic energy, while the balls in Sample B are more clustered, suggesting lower kinetic energy. Therefore, Sample A has a higher average kinetic energy than Sample B.**

Answer: sample A

Figure 12: Sample 2.

G Zero-Shot Evaluation on Flickr30K Test

Table 5: Zero-Shot Evaluation on Flickr30K Test.

Foundation Model	Method	Flickr30K Test					
		I2T			T2I		
		R@1	R@5	R@10	R@1	R@5	R@10
CLIP-B/16	-	85.20	97.30	99.10	64.98	87.84	92.78
	CLIP-Adapter	85.50	98.40	99.10	74.00	92.84	96.00
	MMA	86.80	98.00	99.30	72.58	92.28	95.84
	FreqAdapter	87.80	97.50	98.80	75.14	92.76	96.34
CLIP-L/14	-	86.80	98.30	99.80	67.90	89.70	94.28
	CLIP-Adapter	88.40	98.50	99.60	77.36	94.22	97.10
	MMA	90.30	98.50	99.80	75.24	93.42	96.78
	FreqAdapter	90.00	98.20	99.50	77.28	94.58	97.36
CLIP-L/14-336	-	88.10	98.20	99.60	71.40	91.64	95.46
	CLIP-Adapter	89.60	99.20	99.80	78.74	95.10	97.42
	MMA	90.00	99.30	99.90	77.30	94.38	97.14
	FreqAdapter	90.50	98.90	99.70	78.58	95.16	97.60

The zero-shot evaluation on Flickr30K test results are shown in the Tab. 5, and the results are consistent with the conclusions of the Flickr30k Val set.

H More Ablation Study

H.1 Effectiveness of Multiple Scale

Table 6: Ablation Study for Downsampling Factor.

N	MSCOCO I2T			MSCOCO T2I		
	R@1	R@5	R@10	R@1	R@5	R@10
1	60.18	81.78	87.50	44.08	69.80	77.28
2	61.32	82.76	88.42	44.81	70.20	79.47
3	61.42	83.64	90.10	45.23	70.92	80.02
4	60.36	82.58	88.12	45.06	70.70	79.36

The Tab. 6 demonstrates the impact of the downsampling factor N on the experimental results. As N increases, the FreqAdapter learns more multi-scale visual features. The results indicate that incorporating additional scales promotes performance improvements, since fine-tuning at different scales enables the model to perceive image regions with various receptive fields. This allows it to extract richer visual features and gain a better understanding of the image content. However, we also observe that overly large receptive fields are less effective than moderate ones, as aggregating information over excessively large regions can lead to significant information loss and result in erroneous parameter learning.

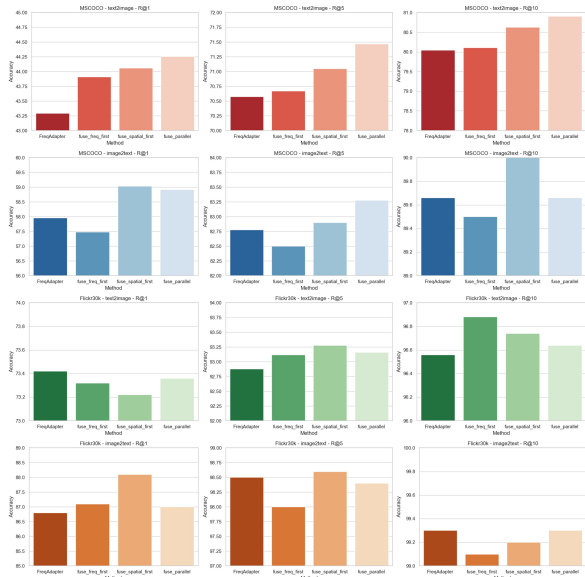


Figure 13: Ablation on FreqAdapter and CLIP-Adapter Ensemble.

I Further In-depth Analysis

I.1 Ensemble with CLIP-Adapter

In this section, we investigate the performance of integrating FreqAdapter and CLIP-Adapter. FreqAdapter performs adaptation in the frequency domain, while CLIP-Adapter operates in the spatial domain. We explore three integration methods: (1) adaptation in the frequency domain followed by spatial adaptation, denoted as fuse_freq_first; (2) spatial adaptation followed by frequency adaptation, denoted as fuse_spatial_first; and (3) simultaneous adaptation in both domains, denoted as fuse_parallel. The experimental setup is consistent with the comparative experiments described previously. The figure demonstrates the retrieval performance on the COCO 2017 and Flickr30k datasets for I2T and T2I tasks, evaluated using R@1, R@5, and R@10 metrics. Compared with the standalone FreqAdapter, the integrated methods generally exhibit improved performance. The fuse_spatial_first method typically outperforms fuse_freq_first across most scenarios. The fuse_parallel method achieves the best results on all metrics for the T2I retrieval task when fitting specific data distributions, while fuse_spatial_first and fuse_parallel show comparable performance on other metrics. Given that fuse_parallel can perform adaptations simultaneously, it offers greater advantages in practical applications.

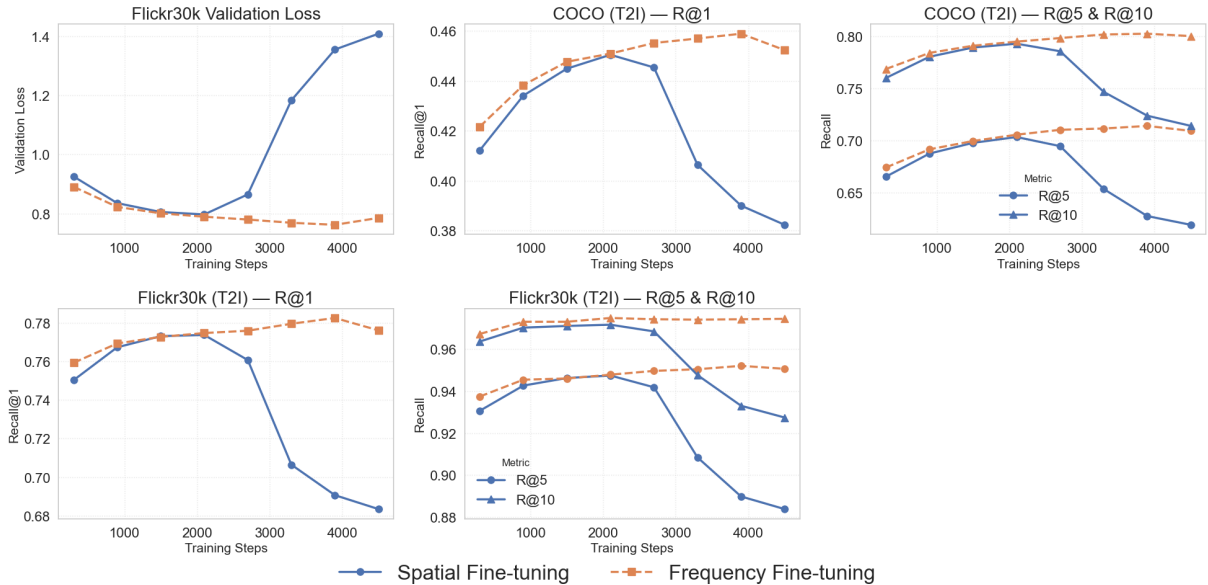


Figure 14: Frequency vs. Spatial Adaptation.

I.2 Frequency vs. Spatial Adaptation.

Additional Results (Flickr ValLoss and T2I). Fig. 14 consolidates five supplementary plots—Flickr30k validation loss and T2I retrieval on COCO and Flickr30k (R@1/5/10)—that corroborate the main-text findings. On Flickr30k validation loss, *FreqAdapter* continues to decrease or stabilize smoothly in late training, while *SpatialAdapter* rebounds and drifts away from its optimum, a clear sign of overfitting consistent with its behavior on COCO. For COCO T2I, *FreqAdapter* maintains a lead throughout training and remains stable in the latter half of the epoch, whereas *SpatialAdapter* oscillates and degrades after roughly half an epoch; the same pattern holds on Flickr30k T2I. Notably, these trends are consistent across R@1, R@5, and R@10, indicating that the gains are not confined to a single operating point but extend across recall thresholds. We attribute these effects to the advantages of frequency-domain adaptation—less redundant representations with disentangled low- and high-frequency structure and partial noise separation—yielding cleaner parameter updates, steadier gradients, and more reliable generalization.

I.3 Comparison Between Spatial and Frequency Domain Adaptation

To investigate the effect of adaptation space, we compare the proposed frequency-domain adapter *FreqAdapter* with its spatial-domain counterpart

SpatialAdapter under different learning rates and training steps on the COCO retrieval benchmark.

Across both learning rates and training stages, *FreqAdapter* consistently achieves lower evaluation loss and improved retrieval performance, particularly on T2I R@1. The advantage becomes more pronounced as training progresses, indicating more stable optimization behavior.

One possible explanation is that frequency-domain representations provide a more compact and structured view of embeddings. By leveraging the Discrete Cosine Transform (DCT), *FreqAdapter* operates on decorrelated components, which may reduce redundancy and allow the model to focus updates on more informative directions.

Moreover, since DCT is an orthogonal transform, it introduces negligible computational overhead while preserving information content. This property enables frequency-domain adaptation to maintain efficiency while improving optimization dynamics.

These results suggest that performing adaptation in the frequency domain can offer practical benefits over spatial-domain alternatives. While the improvements are moderate, they are consistent across different settings, indicating that frequency-domain adaptation provides a stable inductive bias for cross-modal representation learning. We further conjecture that such advantages may become more significant in settings with larger numbers of trainable parameters, where selectively updating informative components could lead to more efficient

Step	lr	Method	Train Loss	Eval Loss	I2T R@1	I2T R@5	T2I R@1	T2I R@5
1000	1e-5	Spatial	0.52	2.83	58.06	81.10	38.31	63.31
2000	1e-5	Spatial	0.46	2.65	57.94	81.50	39.87	65.21
3000	1e-5	Spatial	0.44	2.56	58.24	81.36	41.09	66.44
4000	1e-5	Spatial	0.43	2.51	57.76	81.60	41.80	67.01
1000	1e-5	Freq	0.50	2.74	58.32	81.46	39.31	64.44
2000	1e-5	Freq	0.44	2.56	58.08	81.52	41.16	66.40
3000	1e-5	Freq	0.42	2.49	58.72	81.42	42.21	67.34
4000	1e-5	Freq	0.42	2.47	58.44	81.70	42.53	67.50
1000	5e-5	Spatial	0.49	2.60	57.84	81.48	40.61	65.98
2000	5e-5	Spatial	0.38	2.40	58.94	82.12	43.56	68.81
3000	5e-5	Spatial	0.36	2.36	59.56	82.88	44.16	69.31
4000	5e-5	Spatial	0.35	2.34	59.94	83.26	44.53	69.79
1000	5e-5	Freq	0.39	2.38	58.82	82.40	43.73	68.88
2000	5e-5	Freq	0.37	2.34	60.16	82.66	44.42	69.71
3000	5e-5	Freq	0.36	2.31	60.54	83.22	44.80	70.26
4000	5e-5	Freq	0.35	2.28	61.18	83.46	45.25	70.69

Table 7: Comparison between SpatialAdapter and FreqAdapter under different learning rates and training steps on the COCO retrieval benchmark.

learning.

J Future Work

To the best of our knowledge, this work is the first attempt to fine-tune a pretrained model in the frequency domain and directly apply the fine-tuned model to large-scale multimodal models, demonstrating its effectiveness. Nonetheless, there remain several areas for improvement. First, while the introduction of cross-modal interaction modules enhances the model’s expressive capability, it inevitably increases the number of trainable parameters and computational complexity. In real-world scenarios, one can leverage existing network compression methods to prune or factorize model parameters, thereby striking a better balance between accuracy and resource consumption. In the future, we plan to further explore efficient data-compression strategies in the frequency domain, investigate deeper integrations of frequency- and spatial-domain approaches, and examine how to more effectively incorporate the fine-tuned modules into cutting-edge large models.

K LLM Usage

We use a large language model only for writing refinement, including grammar correction and expression polishing. The model does not contribute to the core research process, such as method design, and experimentation.