TGFM: Text-Guided Frequency Modulation for Source-Data-Free Adaptation of Vision-Language Models in VQA

Anonymous ACL submission

Abstract

Pre-trained vision-language models (VLMs) have achieved remarkable success in general-purpose multimodal learning. However, adapting them to domain-specific visual question answering (VQA) scenarios remains challenging due to scarce annotations, substantial distribution shifts, and the practical impossibility of accessing source-domain data in real-world deployments. Meanwhile, many existing adaptation strategies rely on domain- or task-specific architectures, limiting their scalability and transferability. We propose **T**ext-**G**uided **F**requency **M**odulation (TGFM), a source-data-free, target-supervised framework for VQA adaptation that enables fine-grained cross-modal interaction directly in the image frequency domain. TGFM employs a text-guided spectral mask to jointly modulate amplitude and phase, where amplitude captures global structure and phase encodes detailed semantic variations, providing a complementary pathway to spatial-domain adaptation. To ensure robust learning, we design a frequency loss combining low-frequency preservation, text-conditioned band alignment, and spectral regularization for sparsity, smoothness, and semantic coherence. Extensive experiments across six domain-specific VQA benchmarks demonstrate that TGFM consistently outperforms both conventional fine-tuning and state-of-the-art source-data-free approaches, incurring only around 1 million additional parameters.

1 Introduction

Large-scale vision-language models such as CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) have revolutionized general-purpose multimodal learning through massive image-text pre-training. However, when transferred to specialized domains, such as medical visual question answering, remote sensing interpretation, and art analysis, their performance drops substantially (Lobry et al., 2020; Wang et al., 2022b). This degradation is primarily caused by two factors: (1) limited annotated data in target domains, making large-scale retraining impractical, and (2) substantial visual and semantic distribution shifts, which disrupt the alignment between visual and textual features. In safety-critical settings, even small misalignments can lead to unacceptable errors. Consequently, it remains an open problem to develop a *lightweight* and *robust* adaptation framework that can effectively leverage target supervision *without access to source-domain data*.

Existing solutions fall short on several fronts. Source-dependent adaptation requires access to pre-training data for distribution alignment, raising issues of scalability, privacy, and storage (Li et al., 2024). In contrast, source-free adaptation methods avoid source data but are often restricted to specific models or training settings, limiting their generality. Parameter-efficient fine-tuning (PEFT) methods, such as Adapter (Houlsby et al., 2019) and LoRA (Hu et al., 2022), reduce optimization costs by restricting trainable parameters, but do not explicitly address semantic misalignment under domain shift. More recently, FSA (Liu et al., 2025) has shown that frequency augmentation can enhance robustness; however, it operates on pixel- or output-level signals and lacks representation-level cross-modal interaction, which are essential for VQA due to its reliance on vision-language semantic grounding.

To address these challenges, we propose Text-Guided Frequency Modulation (TGFM), a source-data-free but target-supervised framework that introduces text-aware cross-modal interaction in the image frequency domain of intermediate representations. The image frequency domain provides a natural decomposition where amplitude encodes global structures and phase captures semantic variations (Yang and Soatto, 2020). Specifically, intermediate image features are transformed into the Fourier domain and decomposed into amplitude $A(\omega)$ and phase $\phi(\omega)$. A text-attended semantic vector t is used to condition a lightweight mask gen-

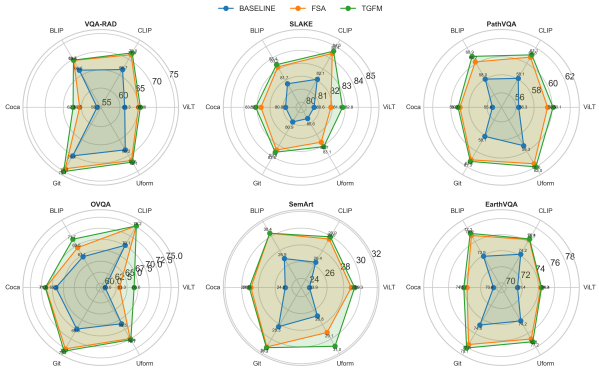

Figure 1: Overall performance comparison of direct fine-tuning (baseline), FSA, and the proposed TGFM across six VLMs and six domain-specific VQA datasets.

erator, which produces amplitude mask $M_a(\omega; t)$ and phase mask $M_\phi(\omega; t)$. The process can be succinctly illustrated as:

$$\widehat{X}'(\omega) = (1 + M_a(\omega; t))\, A(\omega)\, e^{i(\phi(\omega) + M_\phi(\omega; t))},$$
$$X' = \mathcal{F}^{-1}\left(\widehat{X}'(\omega)\right).$$

where amplitude modulation redistributes spectral energy toward text-relevant global structures, while phase modulation refines local and semantic details essential for accurate image-text grounding (Xu et al., 2021; Oppenheim and Lim, 2005). To ensure stable adaptation, we design a three-term frequency loss consisting of (i) low-frequency preservation to retain global structure, (ii) text-guided spectral-band alignment to inject semantic priors into frequency, and (iii) mask regularization promoting sparsity, smoothness, and semantic coherence.

Our setting aligns with practical VQA deployment, with labeled target data but no access to source data. We evaluate TGFM on six general-domain VLMs across six domain-specific VQA datasets. As shown in Figure 1, TGFM outperforms direct fine-tuning and source-data-free FSA with only about 1M extra parameters ($< 0.58\%$ of the model), and detailed results are reported in Tables 1 and 2. Additional analyses are in the **appendix**, covering related work (App. A), method details (App. B), theoretical analysis (App. C), extensive experiments (App. D), and algorithm pseudocode (App. E). Code will be released upon acceptance.

2 Method

2.1 Text-Guided Frequency Modulation

As shown in Figure 2, an input image and text are first processed by a pretrained VLM. TGFM is applied after the final block of the image and text encoder. The visual token sequence (excluding the `[CLS]` token, if present) is reshaped into a spatial feature map $\mathbf{x} \in \mathbb{R}^{B\times C\times H\times W}$, where B denotes the batch size, C the channel dimension, and $H = W$ the spatial resolution of the reshaped map. The text encoder provides a sequence of token embeddings: $\mathbf{t} \in \mathbb{R}^{B\times L\times D}$, where L is the number of text tokens and D the embedding dimension.

Global Text Semantic Representation. To summarize the token-level embeddings into a global semantic, we compute text attention features:

$$\boldsymbol{\alpha}_b = \mathrm{softmax}\left(\frac{1}{D}\sum_{d=1}^{D}\mathbf{t}_{b,:,d}\right), \quad \boldsymbol{\alpha}_b \in \mathbb{R}^{L},$$
$$\mathbf{t}_b^{\mathrm{att}} = \sum_{l=1}^{L}\boldsymbol{\alpha}_{b,l}\,\mathbf{t}_{b,l} \in \mathbb{R}^{D}.$$

This results in a global text vector $\mathbf{t}_b^{\mathrm{att}} \in \mathbb{R}^D$ that encodes weighted semantics for each text sample feature. This fine-grained token-wise attention provides a lightweight summary of global text intent while reducing the risk of overfitting.

Radial Frequency Encoding. For a frequency position (i, j) in the $H \times W$ grid, we define its normalized radial distance to the image center as:

$$\mathbf{r}_{i,j} = \sqrt{\left(\frac{2i}{H} - 1\right)^2 + \left(\frac{2j}{W} - 1\right)^2}, \mathbf{r}_{i,j} \in [0, 1].$$

This positional encoding helps distinguish between low-, mid-, and high-frequency regions. For each frequency location (i, j), we construct an augmented condition vector by concatenating expanded text and frequency for each position:

$$\mathbf{z}_{i,j} = \left[\mathbf{t}_b^{\mathrm{att}},\ \mathbf{r}_{i,j}\right] \in \mathbb{R}^{D+1},$$

expanding $\mathbf{t}_b^{\mathrm{att}}$ and collecting all positions yields $\mathbf{Z}_b \in \mathbb{R}^{(H\cdot W)\times(D+1)}$, where each position corresponds to a frequency position (i, j), combining text semantics with its radial frequency encoding.

Amplitude and Phase Modulation. The augmented vectors $\mathbf{Z}_b$ are fed into a dual-branch MLP consisting of a linear branch and a nonlinear branch. Each branch contains two linear layers that independently predict amplitude and phase signals by projecting $\mathbf{Z}_b \in \mathbb{R}^{(D+1)}$ to the channel dimension $\mathbb{R}^C$. The linear branch directly maps the $\mathbf{Z}_b$:

$$\Delta A_b^{\mathrm{lin}} = \mathbf{W}^{\mathrm{lin},A} \cdot \mathbf{Z}_b,$$
$$\Delta \phi_b^{\mathrm{lin}} = \mathbf{W}^{\mathrm{lin},\phi} \cdot \mathbf{Z}_b.$$

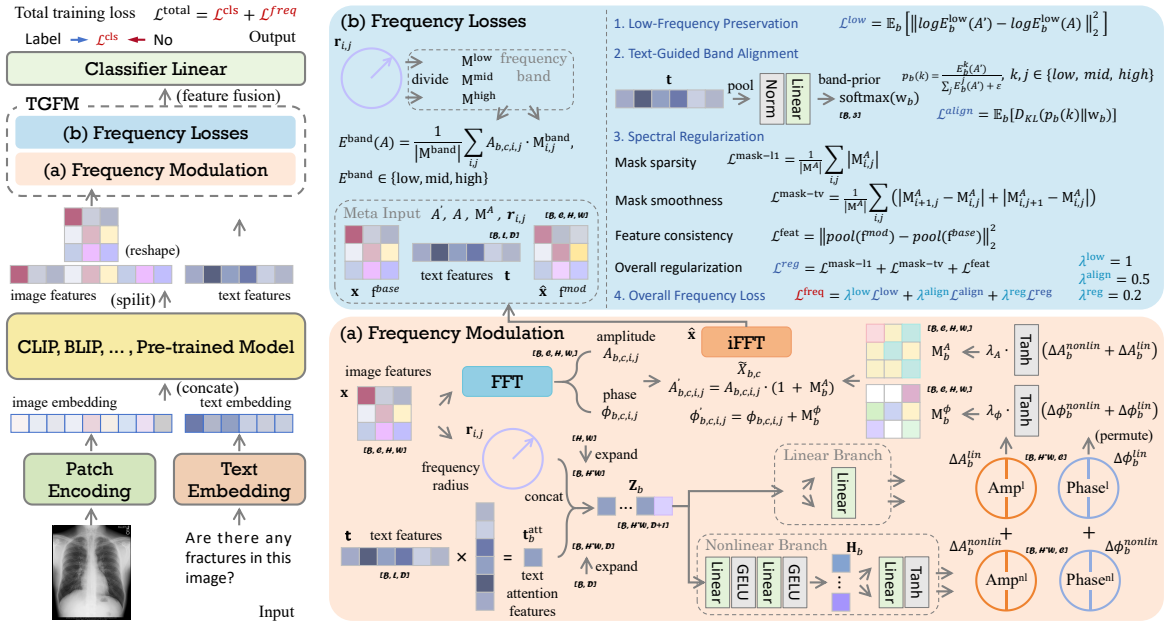

Figure 2: Text-Guided Frequency Modulation transfers general pre-trained VLMs to specific VQA tasks.

This branch provides a stable linear pathway that facilitates effective modulation under distribution shifts. In parallel, in the nonlinear branch, the vector $\mathbf{Z}_b$ is first projected through two feed-forward layers with GELU activation, producing an intermediate representation $\mathbf{H}_b$:

$$\mathbf{H}_b = \mathrm{GELU}\Big(\mathbf{W}^{(2)} \cdot \mathrm{GELU}(\mathbf{W}^{(1)} \cdot \mathbf{Z}_b)\Big),$$
$$\Delta A_b^{\text{nonlin}} = \tanh(\mathbf{W}^{\text{nonlin},A} \cdot \mathbf{H}_b),$$
$$\Delta \phi_b^{\text{nonlin}} = \tanh(\mathbf{W}^{\text{nonlin},\phi} \cdot \mathbf{H}_b).$$

For the representation $\mathbf{H}_b$, two separate linear layers generate the amplitude and phase modulation signals, followed by a $\tanh$ activation to bound their range. This branch allows the model to learn flexible, nonlinear modulations that adapt to complex text frequency interactions. This dual-branch design balances expressive nonlinear modulation with linear stability. The outputs from the nonlinear and linear branches are aggregated and permuted to the $C \times H \times W$ format to match the channel-first layout of feature maps. The resulting modulation masks are scaled by small hyperparameters λ^A and λ^ϕ to ensure stable frequency perturbations:

$$\mathbf{M}_b^A = \tanh\big(\Delta A_b^{\text{nonlin}} + \Delta A_b^{\text{lin}}\big) \cdot \lambda^A \in \mathbb{R}^{C\times H\times W},$$
$$\mathbf{M}_b^\phi = \tanh\big(\Delta \phi_b^{\text{nonlin}} + \Delta \phi_b^{\text{lin}}\big) \cdot \lambda^\phi \in \mathbb{R}^{C\times H\times W}.$$

These masks are reshaped and permuted to align with the spatial frequency grid, ensuring that each channel receives its own amplitude and phase modulation map. The scaling factors (e.g., $\lambda^A = 0.5$, $\lambda^\phi = 0.1$) act as regularizers that constrain the magnitude of perturbations and prevent excessive distortion (Yang and Soatto, 2020; Xu et al., 2021).

Frequency Feature Enhancement. With the modulation masks available, we operate in the Fourier domain to enhance the input features. Each image spatial feature map $\mathbf{x}_{b,c,:,:}$ is first transformed by a 2D FFT (Nussbaumer and Nussbaumer, 1982):

$$\mathcal{F}(\mathbf{x}_{b,c}) = \mathcal{F}_{2D}\left(\mathbf{x}_{b,c,:,:}\right) \in \mathbb{C}^{H\times W}.$$

We then decompose the complex-valued spectrum into its amplitude and phase:

$$A_{b,c,i,j} = \left|\mathcal{F}(\mathbf{x}_{b,c})_{i,j}\right|,$$
$$\phi_{b,c,i,j} = \angle\mathcal{F}(\mathbf{x}_{b,c})_{i,j},$$

where $A_{b,c,i,j} \geq 0$ represents the magnitude of the frequency response at location (i, j), and $\phi_{b,c,i,j}$ denotes its corresponding phase angle.

The learned masks $\mathbf{M}^A_{b,c,i,j}$ ($\mathbf{M}^A_b$) and $\mathbf{M}^\phi_{b,c,i,j}$ ($\mathbf{M}^\phi_b$) are applied to perturb the amplitude and phase of each frequency component:

$$A'_{b,c,i,j} = A_{b,c,i,j} \cdot \left(1 + \mathbf{M}^A_{b,c,i,j}\right),$$
$$\phi'_{b,c,i,j} = \phi_{b,c,i,j} + \mathbf{M}^\phi_{b,c,i,j}.$$

This formulation reflects two complementary design choices. **Amplitude modulation** adopts a multiplicative form (Yang et al., 2022; Huang et al., 2021), as $A_{b,c,i,j}$ encodes spectral energy. Scaling by $(1 + \mathbf{M}^A_{b,c,i,j})$ adaptively enhances or attenuates

specific frequencies without changing their polarity. In contrast, **phase modulation** is additive, since phase captures angular shifts in the complex plane. Adding $\mathbf{M}^{\phi}_{b,c,i,j}$ produces controlled rotations that preserve the original amplitude distribution (Zhang et al., 2022; Fein-Ashley, 2025). The resulting amplitude and phase are then recombined to yield:

$$\tilde{X}_{b,c} = A'_{b,c,i,j} \cdot e^{i\phi'_{b,c,i,j}}, \quad \tilde{X}_{b,c} \in \mathbb{C}^{H\times W}.$$

Inverse Transform. Finally, the enhanced spatial-domain feature is reconstructed via inverse FFT:

$$\hat{x}_{b,c} = \Re\left(\mathcal{F}^{-1}_{2D}\left(\tilde{X}_{b,c}\right)\right) \in \mathbb{R}^{H\times W},$$

Stacking over channels and batch yields the enhanced feature tensor:

$$\hat{\mathbf{x}} \in \mathbb{R}^{B\times C\times H\times W}.$$

Meta for Frequency Loss. Beyond generating the enhanced features, we also preserve auxiliary variables in a dictionary `meta`. We record the original amplitude $A_{b,c,i,j}$ (A), the modulated amplitude $A'_{b,c,i,j}$ (A'), the amplitude mask $\mathbf{M}^{A}_{b,c,i,j}$ ($\mathbf{M}^{A}$), and the normalized radial frequency map $\mathbf{r}_{i,j}$. The detailed design of `meta` is provided in App. B.6.

2.2 Frequency-Modulation Loss

To ensure that the text-guided frequency modulation produces stable, interpretable, and semantically consistent transformations, we introduce a loss function with three complementary terms.

We partition the frequency spectrum into three radial bands using masks $\mathbf{M}^{\text{low}}, \mathbf{M}^{\text{mid}}, \mathbf{M}^{\text{high}}$. Each mask is defined over the normalized radial frequency grid $\mathbf{r}_{i,j} \in \mathbb{R}^{H\times W}$ and separates low-, mid-, and high-frequency components according to fixed thresholds (0.33, 0.66). For a given amplitude spectrum $A_{b,c,i,j}$, the energy of each band is computed as the average amplitude within the corresponding masked region:

$$E^{\text{band}}(A) = \frac{1}{|\mathbf{M}^{\text{band}}|}\sum_{i,j} A_{b,c,i,j} \cdot \mathbf{M}^{\text{band}}_{i,j},$$

where $|\mathbf{M}^{\text{band}}|$ is the number of pixels in the mask and E^{band} includes low-, mid-, and high-frequency.

Low-Frequency Preservation. The low-frequency components of the Fourier spectrum primarily capture global structure, which should remain stable under text-guided fine-grained edits. We therefore anchor the low-frequency energy between the original and enhanced spectra (Salmela et al., 2016). We define $E^{\text{band}}(A')$ for the enhanced amplitude spectrum A'. During training, we require the low-frequency energy of the enhanced spectrum A' to match that of the original spectrum A. We compare the energies in the logarithmic domain:

$$\mathcal{L}^{\text{low}} = \mathbb{E}_b\left[\left\|\log E^{\text{low}}_b(A') - \log E^{\text{low}}_b(A)\right\|_2^2\right].$$

This MSE loss penalizes deviations in energy ratios, ensuring that the enhanced representation does not arbitrarily amplify or suppress global structure. By anchoring only the low-frequency energy, our framework achieves a balance between stability and flexibility: it preserves global semantics for coherent structure while allowing the model to enrich fine-grained, text-guided details.

Text-Guided Band Alignment. To couple textual semantics with spectral energy redistribution, we aggregate the sequence of text embeddings $\mathbf{t}_b$ into a pooled representation $\bar{\mathbf{t}}_b \in \mathbb{R}^D$ that roughly captures the global semantic intent of the text features. This embedding is then normalized and linearly projected ($\mathbf{W}$) into a band-level prior distribution:

$$\mathbf{w}_b = \text{softmax}(\mathbf{W}\bar{\mathbf{t}}_b) \in \mathbb{R}^3,$$

where the softmax is taken over the band dimension for each sample, indicating how much emphasis should be given to $\{\text{low}, \text{mid}, \text{high}\}$ frequencies.

Given the enhanced amplitude tensor A', we compute per-sample band distributions:

$$p_b(k) = \frac{E^k_b(A')}{\sum_j E^j_b(A') + \varepsilon}, k, j \in \{\text{low, mid, high}\},$$

where $\varepsilon > 0$ (1e-6) ensures numerical stability. The text-induced prior $\mathbf{w}_b$ is softmax-normalized over bands. We then minimize the KL divergence:

$$\mathcal{L}^{\text{align}} = \mathbb{E}_b\left[D_{\text{KL}}(p_b(k) \parallel \mathbf{w}_b)\right],$$

pushing spectral energy toward text-relevant bands. The chosen direction $p||w$ emphasizes fitting the empirical distribution to the prior; symmetric KL or MSE in log-energies yields similar trends without affecting stability.

Spectral Regularization. To prevent training degeneration and ensure stable optimization dynamics, we incorporate a set of regularization terms that act on the modulation mask and semantic features.

Mask sparsity. The amplitude modulation mask $\mathbf{M}^{A}$ is expected to highlight only the frequencies

relevant to the textual guidance, rather than introducing indiscriminate modifications. We penalize the mean absolute value of the mask:

$$\mathcal{L}^{\text{mask-l1}} = \frac{1}{|\mathbf{M}^A|} \sum_{i,j} |\mathbf{M}^A_{i,j}|.$$

This sparsity constraint suppresses unnecessary frequency changes, promoting compact and localized spectral modulation.

Mask smoothness. Although sparsity reduces overmodulation, it may also result in scattered or noisy patterns in the mask. We encourage spatial smoothness in $\mathbf{M}^A$ by minimizing its total variation (TV):

$$\mathcal{L}^{\text{mask-tv}} = \frac{1}{|\mathbf{M}^A|} \sum_{i,j} \left(|\Delta_x \mathbf{M}^A_{i,j}| + |\Delta_y \mathbf{M}^A_{i,j}| \right).$$

This term suppresses local artifacts, resulting in coherent and structured modulation.

Feature consistency. To preserve global semantic consistency, we compare the pooled backbone features before and after frequency modulation. Let $\mathbf{f}^{\text{base}}(\mathbf{x})$ denote the unmodified backbone features and $\mathbf{f}^{\text{mod}}(\hat{\mathbf{x}})$ the features after text-guided image frequency modulation. We apply global average pooling to obtain compact representations and compute their mean squared error (MSE):

$$\mathcal{L}^{\text{feat}} = \left\| pool(\mathbf{f}^{\text{mod}}) - pool(\mathbf{f}^{\text{base}}) \right\|_2^2.$$

Overall Regularization. The total regularization loss is a weighted sum of the above terms:

$$\mathcal{L}^{\text{reg}} = \mathcal{L}^{\text{mask-l1}} + \mathcal{L}^{\text{mask-tv}} + \mathcal{L}^{\text{feat}}.$$

These three components are aggregated into a regularization block to balance spectral flexibility.

Joint Loss. To integrate the frequency-domain objectives, we define the joint frequency loss as:

$$\mathcal{L}^{\text{freq}} = \lambda^{\text{low}} \mathcal{L}^{\text{low}} + \lambda^{\text{align}} \mathcal{L}^{\text{align}} + \lambda^{\text{reg}} \mathcal{L}^{\text{reg}},$$

where $\mathcal{L}^{\text{low}}$ anchors global structure by preserving the low-frequency spectrum, $\mathcal{L}^{\text{align}}$ encourages semantic coupling between text and frequency bands, and $\mathcal{L}^{\text{reg}}$ stabilizes training via mask regularization and semantic consistency. The default configuration ($\lambda^{\text{low}} = 1.0, \lambda^{\text{align}} = 0.5, \lambda^{\text{reg}} = 0.2$) provides a balance between structural stability and semantic preservation. The overall task optimization:

$$\mathcal{L}^{\text{total}} = \mathcal{L}^{\text{cls}} + \mathcal{L}^{\text{freq}},$$

where $\mathcal{L}^{\text{cls}}$ is the cross-entropy classification loss. These two losses make the role of each component explicit: $\mathcal{L}^{\text{freq}}$ enforces spectral interpretability and semantic consistency, while $\mathcal{L}^{\text{total}}$ balances these loss objectives with task-level supervision.

3 Experiments and Analyses

We evaluate our proposed TGFM across six domain-specific VQA datasets, which cover medical imaging, remote sensing, and art. VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021b), and PathVQA (He et al., 2020) represent the radiology and pathology domains; OVQA (Huang et al., 2022) focuses on cross-domain orthopedic imaging VQA; EarthVQA (Wang et al., 2024a) is used for remote sensing analysis; and SemArt (Garcia et al., 2020) targets art interpretation, see App. D.1.

3.1 Baselines

(1) Domain-specific baselines. For medical VQA tasks, we include classical architectures such as MMQ (Do et al., 2021), MEVF (Nguyen et al., 2019), CR (Zhan et al., 2020), and CPRD (Liu et al., 2021a), as well as pretrained medical VLMs like MMBert (Khare et al., 2021), Med-CLIP (Eslami et al., 2021), M3AE (Chen et al., 2022), and PMC-CLIP (Lin et al., 2023), each leveraging large-scale clinical corpora. For remote sensing and art domains, we consider attention-based methods including SAN (Yang et al., 2016), BAN (Kim et al., 2018), and BUTD (Anderson et al., 2018), specialized models such as RSVQA (Zheng et al., 2021) and SOBA (Wang et al., 2024a) for EarthVQA, and VIKING (Garcia et al., 2020) for SemArt.

(2) General-domain VLMs. We further benchmark against general multimodal pretrained models, including ViLT (Kim et al., 2021), CLIP (Radford et al., 2021), BLIP (Li et al., 2022), CoCa (Yu et al., 2022), Git (Wang et al., 2022a), and uform (unum-cloud, 2023), see App. D.2. We also report fine-tuned results for FSA and our TGFM to highlight the contribution of text-guided modulation.

3.2 Implementation Details

The fine-tuning configurations and data augmentation strategies are summarized in App. D and the corresponding Table 8. All experiments are conducted with distributed training on NVIDIA V100 GPUs, and the batch size is selected based on the backbone capacity (16-128). We adopt the AdamW optimizer with ($\beta_1 = 0.9, \beta_2 = 0.999$) and a cosine annealing warm-restart scheduler (Loshchilov and Hutter, 2016) (initial learning rate 2×10^{-5}, $T_0 = 5$, $T_{\text{mult}} = 2$, $\eta_{\min} = 10^{-5}$). For fairness, we follow the training schedule of the FSA framework (Liu et al., 2025), ensuring a consistent comparison under identical optimization settings.

Methods	Radiology Domain						Pathology Domain			Cross Orthopedic		
	VQA-RAD			SLAKE			PathVQA			OVQA		
	Open	Closed	Overall	Open	Closed	Overall	Open	Closed	Overall	Open	Closed	Overall
MMQ	53.7	75.8	67.0	-	-	-	13.4	84.0	48.8	56.9	76.2	68.5
MEVF	49.2	77.2	66.1	77.8	79.8	78.6	8.1	81.4	44.8	36.3	76.3	60.4
CR	60.0	79.3	71.6	77.8	79.8	78.6	-	-	-	51.6	77.7	67.7
CPRD	52.5	77.9	67.8	79.5	83.4	81.1	-	-	-			
MMBert	63.1	77.9	72.0	-	-	-	-	-	-	37.9	80.2	63.3
Med-CLIP	60.1	80.0	72.1	78.4	82.5	80.1	-	-	-	-	-	-
M3AE	**67.2**	83.4	77.0	80.3	87.8	83.2	-	-	-	-	-	-
PMC-CLIP	67.0	**84.0**	**77.6**	**81.9**	88.0	84.3	-	-	-	52.6	82.3	70.5
ViLT (118.9M)	39.6	75.5	61.3	77.7	85.1	80.6	28.3	84.1	56.3	39.6	73.6	60.0
ViLT † (+1.0%)	49.1	79.5	65.1	78.5	87.2	81.9	33.7	85.2	59.5	42.7	77.0	63.3
ViLT ∗ (+0.6%)	53.0	73.9	65.6	79.7	87.5	82.8	32.9	87.2	60.1	46.1	80.2	66.6
CLIP (151.4M)	54.2	74.9	66.7	78.8	87.2	82.1	31.9	84.1	58.1	53.0	81.4	70.1
CLIP † (+0.9%)	59.8	78.9	71.4	**81.9**	88.3	84.4	36.5	85.0	60.8	**62.8**	83.4	75.2
CLIP ∗ (+0.6%)	60.8	79.4	72.0	81.7	**89.2**	**84.6**	35.0	87.0	61.1	60.6	84.8	75.2
BLIP (386.0M)	52.8	75.4	66.5	79.2	85.5	81.7	29.5	86.2	58.0	48.5	79.5	67.1
BLIP † (+0.4%)	58.1	77.0	69.6	79.8	88.3	83.2	33.8	86.5	60.2	52.3	81.0	69.5
BLIP ∗ (+0.3%)	57.6	77.6	69.7	79.9	88.7	83.4	34.6	87.2	60.9	52.9	84.1	71.7
Coca (253.7M)	32.2	71.0	55.7	79.4	83.0	80.8	25.5	85.0	55.4	52.3	80.4	69.2
Coca † (+0.9%)	45.3	70.5	60.5	80.2	86.6	82.7	30.7	87.2	59.0	55.2	82.3	71.5
Coca ∗ (+0.6%)	45.8	73.1	62.3	80.4	87.2	83.1	31.8	86.4	59.2	55.4	82.3	71.6
Git (153.8M)	59.2	77.3	70.1	77.7	86.0	80.9	31.9	84.1	58.1	52.8	81.2	69.9
Git † (+2.2%)	62.1	82.2	74.2	80.2	88.3	83.4	36.4	85.7	61.1	62.0	83.8	75.1
Git ∗ (+1.3%)	63.7	82.6	75.1	80.2	88.7	83.6	34.9	87.4	61.3	61.4	**85.2**	**75.7**
uform (1,500M)	55.8	76.2	68.2	78.7	83.6	80.6	32.0	86.4	59.3	51.3	79.9	68.5
uform † (+0.2%)	59.2	79.6	71.6	81.0	85.4	82.7	**37.4**	85.7	61.6	58.1	81.9	72.4
uform ∗ (+0.1%)	64.3	77.2	72.1	79.2	**89.2**	83.1	34.6	**89.0**	**62.0**	56.1	83.7	72.7

Table 1: Performance of different models on VQA tasks in the medical domain. † indicates models fine-tuned with FSA, and ∗ indicates models fine-tuned with our TGFM. +1.0% and +0.6% denote the relative increase in parameters compared with the original pretrained baseline for FSA and TGFM, respectively.

Method	Remote	Art
	EarthVQA	SemArt
SAN	75.7	-
BAN	76.7	22.4
BUTD	76.4	21.8
Instruct-BLIP	75.2	-
RSVQA	70.8	-
SOBA	**78.1**	-
VIKING	-	20.4
ViLT	71.4	23.9
ViLT †	74.3	28.9
ViLT ∗	74.4	29.3
CLIP	74.2	26.4
CLIP †	76.3	29.6
CLIP ∗	76.4	29.9
BLIP	73.9	26.9
BLIP †	76.9	30.4
BLIP ∗	77.2	30.4
Coca	70.4	24.8
Coca †	73.7	28.9
Coca ∗	74.1	29.1
Git	74.8	28.3
Git †	77.6	31.1
Git ∗	**78.1**	**31.2**
uform	74.2	26.8
uform †	76.8	29.1
uform ∗	77.2	31.0

Table 2: Performance of different models on VQA tasks in remote sensing and art domains.

3.3 Main Results and Analysis

Tables 1 and 2 summarize performance across six domain-specific VQA benchmarks. *1. Strong Gains Across Heterogeneous Domains.* Across all six domain-specific VQA datasets, TGFM surpasses both general-purpose backbones (ViLT, CLIP, etc.) and multiple specific-domain models. *2. Significant Improvements over Direct Fine-Tuning.* TGFM improves over naive fine-tuning by **+3.7%** average overall accuracy. *3. Outperforming the FSA Framework.* Compared with FSA, TGFM yields consistent gains (**+0.6%** on average) despite introducing fewer additional parameters. Whereas FSA perturbs raw pixel-level frequencies, TGFM operates on latent features, enabling semantically grounded and fine-grained frequency modulation that captures subtle cross-modal interactions essential for VQA. *4. Training Dynamics Reveal Synergistic Adaptation.* Figure 3 illustrates the learning behavior of Git equipped with TGFM on OVQA. (a) Training and testing accuracy rise steadily, showing stable convergence. (b) Classification and frequency losses decrease jointly. (c) Individual frequency-loss components converge early and remain stable. (d) Both open- and closed-set accuracies improve continuously. Together, these behaviors validate the stability, coordination, and effectiveness of frequency multimodal adaptation.

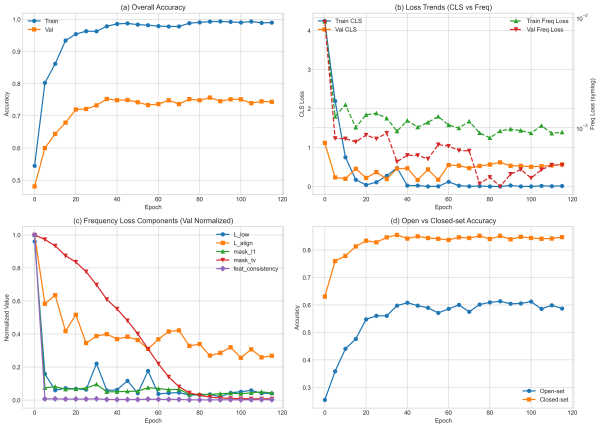

Figure 3: Dynamic behavior during training of the TGFM-based Git model on the OVQA dataset.

3.4 LoRA vs. LoRA+TGFM

To evaluate TGFM under parameter-efficient fine-tuning, we study four representative VLMs (CLIP, BLIP, CoCa, and Git) using LoRA and LoRA+TGFM. Results are reported in Table 3, with full fine-tuning baselines in Tables 1 and 2. Across most datasets, LoRA underperforms full fine-tuning. A notable exception occurs on SemArt, where LoRA occasionally surpasses full fine-tuning. This is likely due to SemArt's slower convergence: full fine-tuning tends to overfit early, whereas LoRA's constrained updates encourage more gradual adaptation. Integrating TGFM with LoRA yields consistent improvements, achieving an average gain of **+2.4%** overall accuracy.

3.5 Ablation Validation

Effectiveness Without Spectral Losses. We isolate the effect of the TGFM modulation module

Methods	Radiology Domain						Pathology Domain			Cross Orthopedic			Remote	Art
	VQA-RAD			SLAKE			PathVQA			OVQA			EarthVQA	SemArt
	Open	Closed	Overall	Open	Closed	Overall	Open	Closed	Overall	Open	Closed	Overall	Overall	Overall
CLIP	43.1	70.3	59.5	78.8	81.7	80.0	30.8	83.6	57.3	43.2	77.9	64.1	72.9	28.5
CLIP ∗	50.2	72.1	63.4	79.3	85.0	81.6	31.2	85.1	58.2	47.1	79.6	66.9	75.6	29.6
BLIP	45.9	75.4	63.7	78.0	82.5	79.8	28.6	85.8	57.3	44.0	77.8	64.4	73.9	30.6
BLIP ∗	50.3	75.4	65.5	79.1	**87.3**	**82.3**	30.4	**86.7**	58.6	47.6	**81.3**	67.8	76.1	**31.1**
Coca	39.1	71.7	58.7	77.6	82.5	79.5	26.0	83.3	54.7	39.2	76.4	61.6	71.1	27.1
Coca ∗	50.3	72.1	63.5	80.2	83.2	81.4	29.8	83.8	56.9	45.7	78.5	65.4	72.3	29.2
Git	57.6	75.0	68.1	77.7	83.2	79.9	30.7	84.8	57.8	42.6	80.9	65.6	74.8	27.3
Git ∗	**59.8**	**77.6**	**70.6**	**80.5**	84.7	82.1	**32.5**	86.2	**59.4**	**49.5**	80.6	**68.2**	**77.6**	30.4

Table 3: Comparison of LoRA (r=16, α=32) and LoRA+TGFM across four multimodal VLMs on six domain-specific VQA benchmarks. TGFM ($*$) consistently improves PEFT across all backbones.

Methods	Radiology Domain						Pathology Domain			Cross Orthopedic			Remote	Art
	VQA-RAD			SLAKE			PathVQA			OVQA			EarthVQA	SemArt
	Open	Closed	Overall	Open	Closed	Overall	Open	Closed	Overall	Open	Closed	Overall	Overall	Overall
CLIP	54.2	74.9	66.7	78.8	87.2	82.1	31.9	84.1	58.1	53.0	81.4	70.1	74.2	26.4
CLIP ∗	57.6	75.4	68.3	**80.2**	**87.5**	**83.0**	33.7	**87.2**	60.5	55.8	82.7	72.0	75.7	28.7
BLIP	52.8	75.4	66.5	79.2	85.5	81.7	29.5	86.2	58.0	48.5	79.5	67.1	73.9	26.9
BLIP ∗	58.1	75.7	68.7	**80.2**	85.6	82.3	**34.6**	86.9	**60.8**	53.5	81.6	70.4	74.9	29.3
Coca	32.2	71.0	55.7	79.4	83.0	80.8	25.5	85.0	55.4	52.3	80.4	69.2	70.4	24.8
Coca ∗	39.1	73.6	59.9	79.4	86.3	82.1	30.6	86.4	58.6	53.5	80.6	69.8	73.7	28.2
Git	59.2	77.3	70.1	77.7	86.0	80.9	31.9	84.1	58.1	52.8	81.2	69.9	74.8	28.3
Git ∗	**61.4**	**79.4**	**72.2**	80.0	86.1	82.4	34.3	87.0	60.7	**57.6**	**84.2**	**73.6**	**76.8**	**29.7**

Table 4: Performance of frequency modulation-only models (marked with $*$) across six VQA domains.

by removing all frequency-domain loss terms. Results are reported in Table 4. To ensure fair comparison, all experimental settings are kept identical. Frequency modulation ($*$) alone yields consistent gains (average **+2.2%**) across all six VQA benchmarks, demonstrating that text-guided amplitude and phase modulation already strengthen cross-modal alignment and facilitate domain adaptation. However, without frequency-domain losses, the modulation becomes under-constrained.

Components of Frequency Modulation. To assess the role of each component in our modulation design, we individually remove the *linear mapping* (-Linear), *nonlinear mapping* (-Nonlin), *phase modulation* (-Phase), and *amplitude modulation* (-Amplitude). Figure 4 reports overall accuracies of CLIP, BLIP, and Git on VQA-RAD, OVQA, and SemArt. Removing any single module consistently degrades performance across backbones and datasets. Notably, removing phase modulation leads to substantial degradation, especially for CLIP and BLIP on VQA-RAD, even underperforming the no-modulation setting (NoFreqMo), highlighting the critical role of phase modulation.

Complementary Effects of Frequency Loss. To analyze the contribution of each frequency-domain loss, we ablate Low-Frequency Preservation (*low*), Text-Guided Band Alignment (*align*), and Spectral Regularization (*reg*). In the main text, we report results for CLIP, while complete experiments on CLIP-, BLIP-, and Git-based TGFM across

Freq_loss	low	align	reg	VQA-RAD	OVQA	SemArt
CLIP				68.3	72.0	28.7
CLIP	✓			67.8	70.6	29.2
CLIP		✓		66.7	72.2	29.4
CLIP			✓	68.2	72.1	29.5
CLIP	✓	✓		71.5	74.6	29.7
CLIP		✓	✓	69.1	72.4	29.5
CLIP	✓		✓	70.1	74.1	29.6
CLIP	✓	✓	✓	**72.0**	**75.2**	**29.9**

Table 5: Ablation of the three frequency-domain losses.

VQA-RAD, OVQA, and SemArt are provided in App. D.4. As shown in Table 5, using any single loss in isolation even underperforms the no-frequency-loss baseline. Applying only *low* or *align* leads to clear degradation on CLIP. Figure 5 further illustrates this effect: *align* alone induces strong optimization oscillations (a), while removing *reg* results in unstable modulation patterns (b). Together, these results show that the three losses are mutually complementary, yielding stable and consistent improvements when used jointly.

Supplementary Experiments. Due to space limitations, detailed ablations and comparisons are provided in **App. D**. We analyze TGFM's sensitivity to amplitude/phase modulation scales (λ^A, λ^ϕ, App. D.5), frequency-loss weights (λ^{low}, λ^{align}, λ^{reg}, App. D.6), and frequency band partitioning and thresholds (App. D.7), where a three-band design with (0.33, 0.66) performs best across datasets, indicating stable behavior. Inserting TGFM at later backbone blocks yields consistently stronger gains (App. D.8), as higher-level features capture more complete semantics. Multi-seed mean $\pm$ std results show that TGFM outperforms the baseline FSA with lower variance (App. D.9), confirming robustness. Finally, TGFM is effective across multiple PEFT strategies, such as FourierFT (App. D.10), and outperforms spatial-domain FiLM modulation (App. D.12) and recent multimodal adaptation methods (App. D.13), supporting its generality.

4 Discussion and Analysis

Spectral Energy Variation Induced by TGFM. We analyze frequency-band energy evolution during CLIP training on VQA-RAD. At each epoch, we compute the energy difference between TGFM-modulated features and the original CLIP features ($\text{energy}_{\text{modulated}}$ - $\text{energy}_{\text{orig}}$), aggregated over low-, mid-, and high-frequency bands. As shown in Figure 6, TGFM induces a structured, band-dependent redistribution of spectral energy: high-frequency components decay most, mid-frequency components are moderately refined, and low-frequency energy remains largely preserved. This behav-

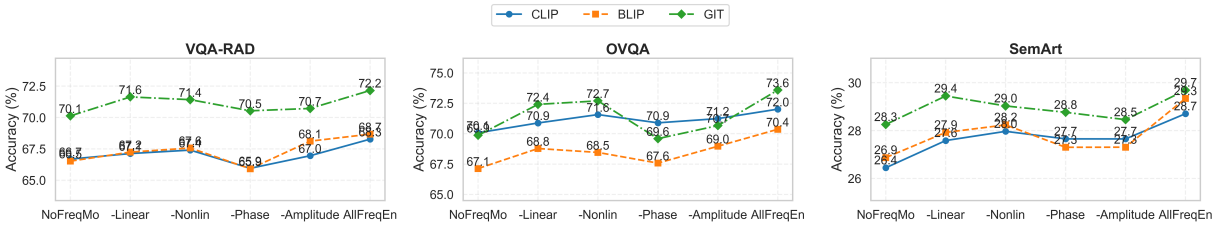

Figure 4: Ablation of individual components in the frequency modulation module.

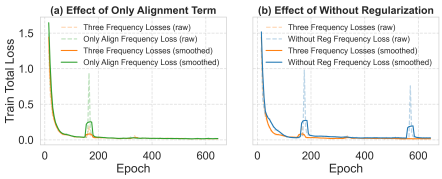

Figure 5: Training instability induced by imbalanced frequency-domain loss components in CLIP (VQA-RAD). Severe oscillation occurs with band alignment alone or without frequency regularization.

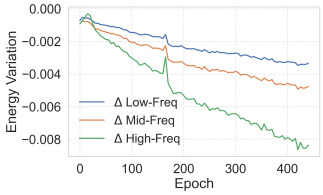

Figure 6: TGFM spectral energy redistribution by CLIP on VQA-RAD, showing variation across low-, mid-, and high-frequency bands.

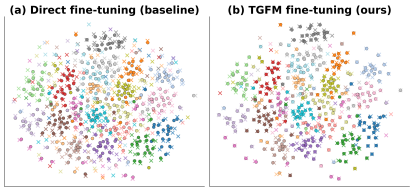

Figure 7: t-SNE visualization of CLIP multimodal embeddings on OVQA. TGFM yields more compact and aligned source–target clusters compared to direct fine-tuning.

ior aligns with TGFM's objective of suppressing domain-specific artifacts while maintaining global semantic structure. Overall, TGFM performs controlled, structure-aware frequency modulation.

t-SNE Visualization of Cross-Domain Embedding Geometry. We visualize learned image-text representations and compare them with original CLIP features. Figure 7 presents t-SNE visualizations of the first twenty answer categories from the OVQA dataset. For each category, twenty paired samples are drawn, and their fused embeddings, extracted from the multimodal projection head, are first compressed via PCA and subsequently mapped to 2D using t-SNE. Direct fine-tuning results in substantial geometric drift and inter-class overlap, whereas TGFM produces more compact intra-class clusters aligned with the original CLIP manifold. This suggests TGFM acts as an effective spectral regularizer, preserving global low-frequency structure while selectively refining discriminative bands under source-free adaptation.

Efficiency and Practicality. Table 6 reports the training and inference efficiency of TGFM under both full fine-tuning and parameter-efficient settings on VQA-RAD using CLIP, with a batch size of 64 on a single NVIDIA V100 GPU. TGFM introduces only modest training overhead (+8.44%) per epoch, substantially lower than that of FSA (+15.70%). At the same time, it maintains comparable inference latency. When combined with LoRA, TGFM preserves parameter efficiency, increasing

Method	Trainable Params (M)	Training Time (s / epoch)	Training Time Overhead(%)	Inference (s / per)
Fine-tuning	151.4M	58.98	-	0.102
FSA	152.9M	68.24	+15.70%	0.107
TGFM	152.3M	63.96	+8.44%	0.104
LoRA	2.08M	53.82	-	0.110
LoRA+TGFM	2.17M	55.79	+3.66%	0.113

Table 6: Training time, overhead, inference latency, and trainable parameters on CLIP for VQA-RAD.

trainable parameters by only 0.09M and incurring a marginal training overhead of 3.66%. These results demonstrate that TGFM is a lightweight and practical module without sacrificing runtime efficiency.

5 Conclusion

We proposed Text-Guided Frequency Modulation (TGFM), a source-data-free, target-supervised framework for cross-domain vision-language adaptation. TGFM performs text-conditioned amplitude and phase modulation at the representation level. This enables structurally coherent and semantically selective frequency adaptation under domain shift. Across six specialized VQA benchmarks spanning radiology, pathology, remote sensing, and art, TGFM consistently outperforms direct fine-tuning and recent source-data-free methods while introducing only ∼1M additional parameters. TGFM also provides a principled view of frequency-aware multimodal adaptation, showing that structured, text-guided spectral control improves cross-domain alignment without source data.

Limitations

TGFM is not merely a vision-side frequency regularizer, but a language-conditioned representation alignment mechanism, where text acts as an explicit semantic controller over representation geometry. Despite its effectiveness, TGFM has several limitations that point to directions for future work. First, TGFM operates on the Fourier spectrum of intermediate visual feature maps and therefore assumes that these representations preserve grid-aligned spatial structure, as is common in CNN- and ViT-based VQA encoders. Architectures with non-spatial embeddings or irregular tokenization may require customized frequency parameterizations. Second, while TGFM demonstrates robust performance across diverse VQA benchmarks, its applicability to other multimodal tasks, such as image-text captioning or retrieval, remains to be explored.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Nenad Tomasev, Jovana Mitrović, Patricia Strachan, and 1 others. 2023. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6):756–779.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in neural information processing systems*, 35:32897–32912.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, and 1 others. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Qi Bi, Jingjun Yi, Hao Zheng, Haolan Zhan, Yawen Huang, Wei Ji, Yuexiang Li, and Yefeng Zheng. 2024. Learning frequency-adapted vision foundation model for domain generalized semantic segmentation. *Advances in Neural Information Processing Systems*, 37:94047–94072.

Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731.

Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. 2022. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pages 679–689. Springer.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703.

Tuong Do, Binh X Nguyen, Erman Tjiputra, Minh Tran, Quang D Tran, and Anh Nguyen. 2021. Multiple meta-model quantifying for medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 64–74. Springer.

Zhekai Du, Jingjing Li, Hongzu Su, Lei Zhu, and Ke Lu. 2021. Cross-domain gradient discrepancy minimization for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3937–3946.

Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. 2021. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*.

Jacob Fein-Ashley. 2025. The fft strikes back: An efficient alternative to self-attention. *arXiv e-prints*, pages arXiv–2502.

Chris Finlay, Jeff Calder, Bilal Abbasi, and Adam Oberman. 2018. Lipschitz regularized deep neural networks generalize and are adversarially robust. *arXiv preprint arXiv:1808.09540*.

Ziqi Gao, Qichao Wang, Aochuan Chen, Zijing Liu, Bingzhe Wu, Liang Chen, and Jia Li. 2024. Parameter-efficient fine-tuning with discrete fourier transform. *arXiv preprint arXiv:2405.03003*.

Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. 2020. A dataset and baselines for visual question answering on art. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 92–108. Springer.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Xuefeng Hu, Ke Zhang, Lu Xia, Albert Chen, Jiajia Luo, Yuyin Sun, Ken Wang, Nan Qiao, Xiao Zeng, Min Sun, and 1 others. 2024. Reclip: Refine contrastive language image pre-training with source free domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2994–3003.

Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. 2021. Rda: Robust domain adaptation via fourier adversarial attacking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8988–8999.

Yefan Huang, Xiaoli Wang, Feiyan Liu, and Guofeng Huang. 2022. Ovqa: A clinically generated visual question answering dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2924–2938.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer.

Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. 2021. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1033–1036. IEEE.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. *Advances in Neural Information Processing Systems*, 31.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex Kot. 2020. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in neural information processing systems*, 33:3118–3129.

Jingjing Li, Zhiqi Yu, Zhekai Du, Lei Zhu, and Heng Tao Shen. 2024. A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5743–5762.

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

Xinjin Li, Yulie Lu, Jinghan Cao, Yu Ma, Zhenglin Li, and Yeyang Zhou. 2025. Catch: A modular cross-domain adaptive template with hook. *arXiv preprint arXiv:2510.26582*.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer.

Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. 2021a. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 210–220. Springer.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021b. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

Lei Liu, Xiangdong Su, and Guanglai Gao. 2024a. Leveraging convolutional models as backbone for medical visual question answering. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2177–2182. IEEE.

Lei Liu, Xiangdong Su, and Guanglai Gao. 2024b. Optimizing transformer and mlp with hidden states perturbation for medical visual question answering. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 5515–5522. IEEE.

Lei Liu, Xiangdong Su, and Guanglai Gao. 2025. Fourier self-adaptation for transferring general pre-trained models to specific domains. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3605–3614.

Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. 2021c. Feddg: Federated domain generalization on medical image segmentation via episodic

learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023.

Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. 2020. Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE transactions on medical imaging*, 39(9):2713–2724.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. 2020. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. 2019. Overcoming data limitation in medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 522–530. Springer.

Henri J Nussbaumer and Henri J Nussbaumer. 1982. *The fast Fourier transform*. Springer.

Alan V Oppenheim and Jae S Lim. 2005. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Viljami R Salmela, Linda Henriksson, and Simo Vanni. 2016. Radial frequency analysis of contour shapes in the visual cortex. *PLoS Computational Biology*, 12(2):e1004719.

Ruoqi Shen, Sébastien Bubeck, and Suriya Gunasekar. 2022. Data augmentation as feature manipulation. In *International conference on machine learning*, pages 19773–19808. PMLR.

Ryan Soklaski, Michael Yee, and Theodoros Tsiligkaridis. 2022. Fourier-based augmentations for improved robustness and uncertainty calibration. *arXiv preprint arXiv:2202.12412*.

Kaito Tanaka, Benjamin Tan, and Brian Wong. 2024. Optimizing vision-language interactions through decoder-only models. *arXiv preprint arXiv:2412.10758*.

unum-cloud. 2023. Pocket-sized multimodal ai for content understanding and generation. https://github.com/unum-cloud/uform.

Aladin Virmaux and Kevin Scaman. 2018a. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31.

Aladin Virmaux and Kevin Scaman. 2018b. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022a. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Junjue Wang, Zhuo Zheng, Zihang Chen, Ailong Ma, and Yanfei Zhong. 2024a. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5481–5489.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, and 1 others. 2024c. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022b. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876.

Zhiquan Wen, Yaowei Wang, Mingkui Tan, Qingyao Wu, and Qi Wu. 2023. Digging out discrimination information from generated samples for robust visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6910–6928.

Weixi Weng, Rui Zhang, Xiaojun Meng, Jieming Zhu, Qun Liu, and Chun Yuan. 2025. Unsupervised domain adaptive visual question answering in the era of multi-modal large language models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6248–6258. IEEE.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. 2021. A fourier-based framework for domain generalization. In *Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition, pages 14383–14392.

Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. 2020a. Open-ended visual question answering by multi-modal domain adaptation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 367–376.

Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. 2020b. Open-ended visual question answering by multi-modal domain adaptation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 367–376.

Chen Yang, Xiaoqing Guo, Zhen Chen, and Yixuan Yuan. 2022. Source free domain adaptation for medical image segmentation with fourier style mining. *Medical Image Analysis*, 79:102457.

Yanchao Yang and Stefano Soatto. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4085–4095.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29.

Maxwell J Yin, Boyu Wang, Yue Dong, and Charles Ling. 2024. Source-free domain adaptation for question answering with masked self-training. *Transactions of the Association for Computational Linguistics*, 12:721–737.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. 2020. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2345–2354.

Jingyi Zhang, Jiaxing Huang, Zichen Tian, and Shijian Lu. 2022. Spectral unsupervised domain adaptation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9829–9840.

Xinxin Zhang, Jun Sun, Simin Hong, and Taihao Li. 2024. Amanda: Adaptively modality-balanced domain adaptation for multimodal emotion recognition. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14448–14458.

Xiangtao Zheng, Binqiang Wang, Xingqian Du, and Xiaoqiang Lu. 2021. Mutual attention inception network for remote sensing visual question answering. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14.

Wei Zhu, Le Lu, Jing Xiao, Mei Han, Jiebo Luo, and Adam P Harrison. 2022. Localized adversarial domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7108–7118.

A Related Work

A.1 Multimodal Domain Adaptation

Multimodal domain adaptation (MMDA) (Xu et al., 2020a; Zhang et al., 2024; Wen et al., 2023) aims to enhance the generalization of vision-language models (VLMs) under cross-domain distribution shifts. Prior work can be broadly grouped into data-level, architecture-level, and representation-level approaches. Data-level methods reduce domain gaps by manipulating input distributions, such as frequency augmentation applied directly to image pixels (Xu et al., 2021; Bi et al., 2024), or adversarial sample generation (Bousmalis et al., 2017). These approaches operate at the input level and do not modify intermediate representations within models. Architecture- or training-level approaches modify model structures or learning objectives, including ensemble learning (Liu et al., 2020), meta-learning (Liu et al., 2021c), self-supervised objectives (Azizi et al., 2023), and more recently expert-based or mixture-of-modality designs in unified VLMs (Bao et al., 2022; Wang et al., 2024c; Tanaka et al., 2024). While effective, these methods often require architectural changes, large-scale training, or substantial computational resources.

Feature- and representation-level techniques instead focus on aligning latent visual and textual features across domains, through cross-modal alignment (Li et al., 2020), adversarial domain learning (Zhu et al., 2022), or feature-wise modulation. Among them, FiLM (Perez et al., 2018) conditions intermediate visual features on external signals via affine transformations in the spatial domain, enabling flexible cross-modal interaction without modifying backbone parameters.

Most MMDA methods rely on access to source-domain data for distribution matching or feature alignment, which raises concerns regarding privacy, storage, and real-world deployment (Yin et al., 2024). To address these limitations, *source-free domain adaptation* has emerged, where adaptation is performed using only a pre-trained source model and labeled or unlabeled target-domain data, without revisiting source samples or statistics.

Within this source-free setting, frequency-based adaptation has recently attracted attention. Fourier Self-Adaptation (FSA) (Liu et al., 2025) is a representative source-data-free, target-supervised approach that introduces frequency-domain regularization to encourage spectral consistency during adaptation. However, existing frequency-based methods (Xu et al., 2021), including FSA, primarily manipulate image-level spectra or constrain output-level frequency statistics and treat frequency modulation as a unimodal visual operation. As a result, they do not explicitly regulate frequency behavior at the level of intermediate visual representations, nor do they incorporate linguistic signals to guide which frequency components are semantically relevant. This limits their effectiveness for multimodal reasoning tasks such as VQA, where fine-grained, text-dependent alignment between visual representations and language is critical.

In contrast, our Text-Guided Frequency Modulation (TGFM) performs *representation-level frequency adaptation* by directly regulating the spectral behavior of intermediate visual features under text guidance. TGFM explicitly models amplitude and phase modulation conditioned on linguistic context, together with frequency-aware regularization that preserves global structure while selectively refining text-relevant bands. Unlike architecture-level expert models or parameter-efficient fine-tuning (PEFT) methods that modify or constrain model parameters, TGFM targets global representation adaptation without introducing task-specific parameter blocks, remaining lightweight and complementary to PEFT-style techniques.

Recent large-scale transfer-oriented VLMs, such as PaliGemma (Beyer et al., 2024) and Qwen2-VL (Wang et al., 2024b), achieve strong generalization through billion-scale pre-training. Due to hardware and training budget constraints, our experiments are limited to models up to 1.5B parameters. Importantly, TGFM is backbone-agnostic and can, in principle, be applied to larger VLMs as long as intermediate visual representations preserve grid-aligned structure; we leave large-scale validation to future work.

A.2 Domain-Specific VQA Tasks

Although large-scale Internet datasets support general-domain VQA training, many specialized domains remain severely resource-constrained, largely due to the high cost of expert annotation and data acquisition. In the medical domain, benchmarks such as VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021b), PathVQA (He et al., 2020), and OVQA (Huang et al., 2022) focus on radiology and pathology images. In remote sensing, EarthVQA (Wang et al., 2024a) focuses on geospatial understanding, whereas art-related reasoning is benchmarked on datasets such as SemArt

(Garcia et al., 2020). These datasets collectively highlight the unique challenges of domain-specific VQA, where visual styles, linguistic patterns, and semantic priors differ substantially from general-domain settings.

Adapting general pre-trained VLMs to such heterogeneous domains is therefore highly challenging. Conventional fine-tuning tends to overfit small datasets and fails to correct the frequency and semantic misalignment introduced by domain shift. TGFM explicitly aligns frequency-domain representations under text guidance, enabling efficient, stable, and source-data-free adaptation of general-purpose VLMs to diverse specialized VQA tasks.

B Detailed Method

B.1 FFT Formulation and Real-Valued Reconstruction

Our frequency modulation operates on real-valued intermediate visual feature maps $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$. We apply a 2D complex Fast Fourier Transform (FFT) independently on each channel:

$$\mathcal{F}(\mathbf{X}) = \texttt{FFT2}(\mathbf{X}), \quad \mathcal{F}(\mathbf{X}) \in \mathbb{C}^{B\times C\times H\times W},$$

using an orthonormal normalization scheme. We adopt the full complex FFT rather than real FFT (rFFT), as this allows explicit and decoupled modulation of amplitude and phase. For real-valued spatial signals, the FFT spectrum naturally satisfies conjugate symmetry. Our modulation is defined element-wise and symmetric across the spectrum, which maintains approximate conjugate symmetry. After modulation, we reconstruct spatial features via inverse FFT and retain the real component:

$$\mathbf{X}' = \Re\left(\texttt{IFFT2}(\mathcal{F}'(\mathbf{X}))\right).$$

Empirically (Liu et al., 2025; Xu et al., 2021; Yang et al., 2022), discarding the negligible imaginary part does not affect downstream performance, consistent with common practices in frequency-domain feature manipulation. The imaginary part remains small due to the smoothness of the learned modulation and the regularization imposed by the frequency losses. We emphasize that enforcing strict conjugate symmetry is not required in our setting, as the inverse FFT followed by real-valued projection yields stable and consistent representations in practice. This results in stable training and consistent performance without explicit symmetry constraints; enforcing strict conjugate symmetry did not yield measurable improvements.

B.2 Token-to-Grid Reshaping and Spatial Assumptions

For both CNN-based and Transformer-based visual encoders, the intermediate visual representations naturally preserve a block-wise spatial structure. CNN backbones directly produce feature maps of shape $H \times W$, while Transformer encoders operate on patch embeddings that correspond to non-overlapping spatial blocks of the input image.

Given a token sequence of length N, we reshape it into a 2D feature map of size $H \times W$ such that $N = H \cdot W$. If a special classification token (e.g., `[CLS]`) is present, it is removed prior to reshaping. This reshaping is a deterministic and invertible operation that does not alter the underlying spatial correspondence of visual features. In our implementation, we operate on square feature maps with $H = W$, which is standard for commonly used pre-trained vision-language encoders (e.g., ViT-based models with 14×14 patch grids).

B.3 Band Energy Aggregation

For each frequency band $b \in \{low, mid, high\}$, we compute the band-wise spectral magnitude by first averaging the Fourier amplitude within the band for each channel, and then averaging across channels:

$$e_b = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{|\Omega_b|}\sum_{(i,j)\in\Omega_b}\left|\mathcal{F}_c(i,j)\right|,$$

where Ω_b denotes the set of frequency indices belonging to band b, and $\mathcal{F}_c$ denotes the Fourier transform of the c-th feature channel.

B.4 Motivation of Three-Band Frequency Decomposition

We adopt a three-band frequency decomposition (low / mid / high) as a deliberate bias–variance tradeoff rather than an arbitrary heuristic.

Low-frequency components primarily capture global structure and semantics, mid-frequency components encode object-level patterns, while high-frequency components correspond to fine details and noise. Empirically (Yang and Soatto, 2020; Salmela et al., 2016), finer band partitioning increases estimation variance, especially in target-supervised or low-data adaptation settings, leading to unstable gradients and degraded generalization. The three-band design provides sufficient expressive granularity while maintaining robustness across domains.

B.5 Text-Guided Band Alignment Objective

Rather than enforcing hard frequency constraints, we align the relative energy distribution across frequency bands with text-derived soft targets.

Given band-wise energies $\mathbf{e} \in \mathbb{R}^3$, we normalize them into a probability distribution and minimize the KL divergence to text-predicted band weights. This formulation encourages semantic-consistent frequency emphasis without over-constraining individual frequency coefficients. In practice, different question semantics induce distinct band priors, providing interpretable text–frequency coupling.

B.6 Meta Information for Frequency Loss

Beyond generating the enhanced features, we also preserve auxiliary variables in a dictionary `meta`. We record the original amplitude $A_{b,c,i,j}$ (A), the modulated amplitude $A'_{b,c,i,j}$ (A'), the amplitude mask $\mathbf{M}^A_{b,c,i,j}$ ($\mathbf{M}^A$), and the normalized radial frequency map $\mathbf{r}_{i,j}$.

This design serves two purposes. First, retaining (A, A') and $\mathbf{M}^A$ enables frequency-aware learning objectives, which both regularize the spectral energy distribution and control the magnitude of modulation. Second, keeping both unmodified and modulated amplitudes allows frequency losses that complement spatial-domain training. This ensures that the perturbed frequencies remain physically meaningful rather than arbitrary modifications. We explicitly regularize only amplitude, as cross-domain shifts predominantly manifest as changes in spectral energy (Xu et al., 2021; Yang et al., 2022; Yang and Soatto, 2020). Phase receives no explicit supervision. Instead, it is modulated via a small scaling factor ($\lambda^\phi = 0.1$), encouraging semantic diversity without destabilizing gradients or compromising spatial coherence. This treatment yields a balanced form of controllable frequency adaptation that preserves structural integrity while improving text-guided spectral alignment. The phase $\phi'_{b,c,i,j}$ is perturbed slightly to introduce spectral diversity, but we do not apply phase supervision, as such perturbations primarily serve as stochastic regularization.

Unless otherwise specified, TGFM is applied on top of standard fine-tuning setups: for full fine-tuning, both the vision and text encoders are updated; for PEFT settings, only the corresponding PEFT modules and TGFM parameters are trainable.

C Theoretical Analysis

C.1 Text-Guided Frequency Augmentation

C.1.1 Training Risk under Spectral Perturbation with Text Guidance

To theoretically analyze the effect of our text-guided frequency modulation, we adopt a risk-based perspective under feature-level augmentation (Liu et al., 2024b,a; Shen et al., 2022). Unlike standard augmentations that only affect the input image $\mathbf{x}$, our module jointly conditions on image-text pairs $(\mathbf{x}, \mathbf{t})$ to produce enhanced features.

Let $\mathbf{h}(\cdot,\cdot)$ be a deep encoder that extracts visual-textual features from both image and text. Let $f(\cdot) = \mathbf{W}^\top \mathbf{h}(\cdot,\cdot)$ be a linear classifier. Denote the dataset triplet as $(\mathbf{x}, \mathbf{t}, \mathbf{y})$ with image $\mathbf{x}$, associated text $\mathbf{t}$, and label $\mathbf{y}$.

Let $\hat{\mathbf{x}}$ be the enhanced feature obtained from $\mathbf{x}$ using the modulation masks generated from both $\mathbf{t}$ and $\mathbf{x}$. The empirical risk under our augmentation becomes:

$$\hat{\mathcal{R}}^{\text{fa}} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}_{\Delta A, \Delta\phi}\left[\ell\left(\mathbf{W}^\top \mathbf{h}(\hat{\mathbf{x}}_n, \mathbf{t}_n), \mathbf{y}_n\right)\right]. \tag{1}$$

Key difference: the enhanced $(\hat{\mathbf{x}}, \mathbf{t})$ reflects *text-aware, frequency-domain modulations*, unlike classical augmentations.

Applying a second-order Taylor expansion (Virmaux and Scaman, 2018a) of the loss ℓ around the expectation $\bar{\mathbf{h}} = \mathbb{E}[\mathbf{h}(\hat{\mathbf{x}}, \mathbf{t})]$ gives:

$$\mathbb{E}\left[\ell(\mathbf{W}^\top \mathbf{h}(\hat{\mathbf{x}}, \mathbf{t}), \mathbf{y})\right] \approx \ell(\mathbf{W}^\top \bar{\mathbf{h}}, \mathbf{y}) \tag{2}$$

$$+\frac{1}{2}\mathbb{E}\left[\Delta^\top \mathbf{H} \Delta\right], \tag{3}$$

where:

$$\begin{aligned} \bar{\mathbf{h}} &= \mathbb{E}[\mathbf{h}(\hat{\mathbf{x}}, \mathbf{t})], \\ \Delta &= \mathbf{W}^\top\left(\bar{\mathbf{h}} - \mathbf{h}(\hat{\mathbf{x}}, \mathbf{t})\right), \\ \mathbf{H} &= \nabla^2 \ell(\mathbf{W}^\top \bar{\mathbf{h}}, \mathbf{y}). \end{aligned} \tag{4}$$

Here, $\mathbf{H}$ is the Hessian of the loss (e.g., cross-entropy), and the second term quantifies the variance-induced penalty from modulated features.

C.1.2 Feature Decomposition and Frequency Variance

Assume that the feature extractor $\mathbf{h}(\cdot,\cdot)$ operates over both the amplitude and phase components of $\hat{\mathbf{x}}$:

$$\mathbf{h}(\hat{\mathbf{x}}, \mathbf{t}) = \mathbf{h}^a(\hat{A}, \mathbf{t}) + \mathbf{h}^\phi(\hat{\phi}, \mathbf{t}). \tag{5}$$

If modulation-induced perturbations to amplitude or phase cause large variance, the model may suppress reliance on that component:

$$\begin{aligned} &\mathrm{Var}[\mathbf{h}^a] \gg \mathrm{Var}[\mathbf{h}^\phi] \Rightarrow w_i^a \to 0 \\ &\qquad\qquad \text{(amplitude suppressed)}, \\ &\mathrm{Var}[\mathbf{h}^\phi] \gg \mathrm{Var}[\mathbf{h}^a] \Rightarrow w_i^\phi \to 0 \\ &\qquad\qquad \text{(phase suppressed)}. \end{aligned}$$

Baseline methods that apply random or global frequency perturbations tend to increase variance uniformly (Soklaski et al., 2022), which often destabilizes learning or discards important cues.

C.1.3 Text-Guided Semantic Variance Control

Let $\Delta A(\mathbf{x}, \mathbf{t}, \mathbf{r})$ and $\Delta\phi(\mathbf{x}, \mathbf{t}, \mathbf{r})$ be the learned text-conditioned perturbation masks over radial frequency coordinate $\mathbf{r}$. Suppose two texts $\mathbf{t}_1$ and $\mathbf{t}_2$ cause significantly different modulations:

$$\begin{aligned} &\|\Delta A(\mathbf{x}, \mathbf{t}_1, \mathbf{r}) - \Delta A(\mathbf{x}, \mathbf{t}_2, \mathbf{r})\| > \epsilon_1, \\ &\|\Delta\phi(\mathbf{x}, \mathbf{t}_1, \mathbf{r}) - \Delta\phi(\mathbf{x}, \mathbf{t}_2, \mathbf{r})\| > \epsilon_2. \end{aligned} \tag{6}$$

Then the variance of the extracted features becomes text-sensitive: $\mathrm{Var}[\mathbf{h}^a]$, $\mathrm{Var}[\mathbf{h}^\phi]$ increase selectively along semantic dimensions.

Thus, instead of inducing global high variance, our modulation selectively increases variance in semantically relevant frequency regions, allowing the model to focus on discriminative frequency cues while ignoring irrelevant ones.

C.2 Loss-Regularized Risk Control

The previous subsections show that the second-order Taylor term $\mathbb{E}[\Delta^\top \mathbf{H} \Delta]$ governs how augmentation-induced variance affects empirical risk. We now make this connection explicit (Du et al., 2021) and show how each loss term in $\mathcal{L}^{\text{total}} = \lambda^{\text{low}}\mathcal{L}^{\text{low}} + \lambda^{\text{align}}\mathcal{L}^{\text{align}} + \lambda^{\text{reg}}\mathcal{L}^{\text{reg}}$ reduces an upper bound on that term.

C.2.1 Upper-bound of the Taylor variance term

Let $\lambda^{\max}(\mathbf{H})$ be the largest eigenvalue of the Hessian $\mathbf{H}$. By the Rayleigh quotient (or operator norm) bound, we have

$$\Delta^\top \mathbf{H}\, \Delta \le \lambda^{\max}(\mathbf{H}) \|\Delta\|_2^2, \tag{7}$$

and therefore

$$\mathbb{E}\big[\Delta^\top \mathbf{H}\, \Delta\big] \le \lambda^{\max}(\mathbf{H})\, \mathbb{E}\big[\|\Delta\|_2^2\big]. \tag{8}$$

By definition $\Delta = \mathbf{W}^\top(\bar{\mathbf{h}} - \mathbf{h}(\hat{\mathbf{x}}, \mathbf{t}))$, so using the operator norm $\|\mathbf{W}\|_{\text{op}}$ we get

$$\|\Delta\|_2 \le \|\mathbf{W}\|_{\text{op}} \big\|\mathbf{h}(\hat{\mathbf{x}}, \mathbf{t}) - \mathbf{h}(\mathbf{x}, \mathbf{t})\big\|_2, \tag{9}$$

and thus

$$\begin{aligned} \mathbb{E}\big[\Delta^\top \mathbf{H}\, \Delta\big] \le \lambda^{\max}(\mathbf{H})\, \|\mathbf{W}\|_{\text{op}}^2\, \mathbb{E}\big[\\ \big\|\mathbf{h}(\hat{\mathbf{x}}, \mathbf{t}) - \mathbf{h}(\mathbf{x}, \mathbf{t})\big\|_2^2\big]. \end{aligned} \tag{10}$$

Hence, it suffices to bound the expected squared change of the encoder output under spectral modulation, $\mathbb{E}\|\mathbf{h}(\hat{\mathbf{x}}, \mathbf{t}) - \mathbf{h}(\mathbf{x}, \mathbf{t})\|_2^2$.

C.2.2 Decompose feature sensitivity to amplitude/phase.

Assume the encoder decomposes contributions from amplitude and phase:

$$\mathbf{h}(\hat{\mathbf{x}}, \mathbf{t}) = \mathbf{h}^a(\hat{A}, \mathbf{t}) + \mathbf{h}^\phi(\hat{\phi}, \mathbf{t}). \tag{11}$$

We make the (standard) Lipschitz-type assumptions (Finlay et al., 2018; Virmaux and Scaman, 2018b): there exist constants $L^a, L^\phi \ge 0$ such that for any two spectra (A_1', ϕ_1') and (A_2', ϕ_2'),

$$\|\mathbf{h}^a(A_1', \mathbf{t}) - \mathbf{h}^a(A_2', \mathbf{t})\|_2 \le L^a \|A_1' - A_2'\|_F, \tag{12}$$

$$\|\mathbf{h}^\phi(\phi_1', \mathbf{t}) - \mathbf{h}^\phi(\phi_2', \mathbf{t})\|_2 \le L^\phi \|\phi_1' - \phi_2'\|_F. \tag{13}$$

Combining these,

$$\begin{aligned} \big\|\mathbf{h}(\hat{\mathbf{x}}, \mathbf{t}) - \mathbf{h}(\mathbf{x}, \mathbf{t})\big\|_2 \le L^a \|A' - A\|_F \\ + L^\phi \|\phi' - \phi\|_F. \end{aligned} \tag{14}$$

C.2.3 Relate amplitude difference to the amplitude mask.

By design $A' = A \odot (1 + \mathbf{M}^A)$, hence

$$A' - A = A \odot \mathbf{M}^A. \tag{15}$$

Using standard norm inequalities (elementwise product and sup-norm),

$$\|A' - A\|_F = \|A \odot \mathbf{M}^A\|_F \le \|A\|_\infty \|\mathbf{M}^A\|_F, \tag{16}$$

where $\|A\|_\infty = \max_{b,c,i,j} |A_{b,c,i,j}|$. In practice, intermediate feature maps are normalized, which bounds $\|A\|_\infty$. Substituting (16) into (14) yields

$$\begin{aligned} \big\|\mathbf{h}(\hat{\mathbf{x}}, \mathbf{t}) - \mathbf{h}(\mathbf{x}, \mathbf{t})\big\|_2 \le L^a \|A\|_\infty \|\mathbf{M}^A\|_F \\ + L^\phi \|\phi' - \phi\|_F. \end{aligned} \tag{17}$$

C.2.4 Combine to bound the Hessian quadratic term

Plug (17) into (10) to obtain

$$\mathbb{E}\big[\Delta^\top \mathbf{H}\,\Delta\big] \le \lambda^{\max}(\mathbf{H})\,\|\mathbf{W}\|_{\mathrm{op}}^2\,\mathbb{E}\Big[\qquad (18)$$
$$(L^a)^2\|A\|_\infty^2\|\mathbf{M}^A\|_F^2 \quad + (L^\phi)^2\|\phi' - \phi\|_F^2\Big].$$

This inequality highlights two key control mechanisms: (i) reduce the mask norm $\mathbb{E}\|\mathbf{M}^A\|_F^2$; (ii) limit uncontrolled phase perturbations $\mathbb{E}\|\phi' - \phi\|_F^2$ or mitigate them via feature-level constraints.

C.2.5 How each loss term contributes.

The implemented loss terms map directly to components in (19):

- **Low-frequency preservation** ($\mathcal{L}^{\text{low}}$). If the model's sensitivity to the low-band energy can be locally bounded by a constant L^{low} (i.e., the low-band contribution to $\mathbf{h}$ is Lipschitz w.r.t. the band energy), then controlling the log-energy difference $\big|\log E^{\text{low}}(A') - \log E^{\text{low}}(A)\big|$ directly limits the portion of $\|\mathbf{h}(\hat{\mathbf{x}}) - \mathbf{h}(\mathbf{x})\|_2$ coming from low frequencies. Since low frequencies often dominate classifier-relevant features (large L^a there), $\mathcal{L}^{\text{low}}$ is crucial to reduce the most harmful part of the bound.

- **Text-guided band alignment** ($\mathcal{L}^{\text{align}}$). The KL divergence between the normalized band-energy vector of A' and the text-predicted prior $w(\mathbf{t})$ constrains how energy is redistributed across bands. Under a band-wise Lipschitz model for the model, limiting discrepancies in band energies controls band-specific feature deviations, i.e., it reduces variance in bands that are important for classification according to $\mathbf{t}$.

- **Mask regularization** ($\mathcal{L}^{\text{mask-l1}}$ and $\mathcal{L}^{\text{mask-tv}}$). $\mathcal{L}^{\text{mask-l1}} = \|\mathbf{M}^A\|_1$ and the tv penalty encourage small magnitude and spatial smoothness. Since $\|v\|_2 \le \|v\|_1$, $\|\mathbf{M}^A\|_F^2 \le \|\mathbf{M}^A\|_1^2$, so minimizing the $l1$ term reduces the dominant amplitude-contribution in (19). tv regularization further prevents high-frequency oscillations in the mask that could enlarge $\|\mathbf{M}^A\|_F$ or amplify local effects.

- **Feature consistency regularization** ($\mathcal{L}^{\text{feat}}$). This term directly constrains a low-dimensional projection of the encoder outputs (pooled and optionally projected features). By minimizing $\mathcal{L}^{\text{feat}}$ we directly shrink an empirical estimate of $\mathbb{E}\|\mathbf{h}(\hat{\mathbf{x}}) - \mathbf{h}(\mathbf{x})\|_2^2$, which, via (10), immediately reduces the risk upper bound regardless of amplitude/phase decomposition.

C.2.6 Why we omit explicit phase supervision.

There are two practical and theoretical reasons:

1. **Circularity and instability.** Phase is angular, direct supervision requires circular distance metrics, and is numerically delicate. Small numeric errors in phase can lead to large spatial shifts after inverse FFT, complicating stable gradient-based training.

2. **Indirect control via feature consistency.** Although $\|\phi' - \phi\|_F$ appears in (19), the feature consistency term $\mathcal{L}^{\text{feat}}$ constrains the combined amplitude+phase effect on encoder outputs. If the encoder is less sensitive to phase (i.e., $L^\phi \ll L^a$), amplitude regularization plus feature MSE suffice to keep classifier-relevant deviations small while allowing phase to affect fine-grained geometric adjustments.

C.2.7 Summary bound and design intuition.

Collecting the inequalities gives the qualitative bound

$$\mathbb{E}\big[\Delta^\top \mathbf{H}\,\Delta\big] \lesssim \lambda^{\max}(\mathbf{H})\,\|\mathbf{W}\|_{\mathrm{op}}^2 \qquad (19)$$
$$\Big((L^a)^2\|A\|_\infty^2\,\mathbb{E}\|\mathbf{M}^A\|_F^2 + (L^\phi)^2\,\mathbb{E}\|\phi' - \phi\|_F^2\Big).$$

Each implemented loss term targets one component of this bound: $\mathcal{L}^{\text{low}}$ controls the low-band contribution, $\mathcal{L}^{\text{align}}$ concentrates permissible variance into text-relevant bands; mask norms and tv control $\mathbb{E}\|\mathbf{M}^A\|_F^2$, and $\mathcal{L}^{\text{feat}}$ reduces the encoder-level deviation. Together, they provide interpretive insight into how frequency perturbations affect optimization stability.

D Experiments

In this work, we consider a source-data-free, target-supervised adaptation setting, where a pre-trained VLM is adapted using labeled target-domain data without any access to source-domain samples or statistics. This differs from classical unsupervised SFDA, which assumes unlabeled target data only; we focus on a practical deployment scenario common in domain-specific VQA, where limited labeled target data is available, but source data cannot be accessed due to privacy or storage constraints.

Datasets		Radiology			Pathology	Remote sensing	Art
		VQA-RAD	SLAKE	OVQA	PathVQA	EarthVQA	SemArt
train	images	315	450	2,000	2,499	2,522	21,384
	questions	3,064	4,919	15,216	17,325	88,166	69,812
test	images	315	96	1,223	1,000	1,809	21,384
	questions	451	1,061	1,902	6,012	63,225	4,912

Table 7: Details of the domain-specific VQA datasets.

D.1 Datasets

As summarized in Table 7, VQA-RAD is a relatively small radiology dataset and thus relies heavily on pre-trained medical priors for effective reasoning. OVQA focuses on orthopedic imaging and serves as a challenging benchmark for fine-grained visual-textual alignment. SemArt involves abstract artistic interpretation, where many questions depend on external cultural or stylistic knowledge. This diversity in visual modalities, domain priors, and reasoning complexity makes these benchmarks well-suited for analyzing the robustness and interpretability of cross-domain adaptation.

D.2 Baselines

- **ViLT** (Kim et al., 2021) is a lightweight transformer pre-trained with joint vision-language objectives, including image-text matching and masked language modeling, on 4M image-text pairs.
- **CLIP** (Radford et al., 2021) adopts a dual-encoder contrastive learning framework trained on 400M image-text pairs to align vision and language representations.
- **BLIP** (Li et al., 2022) combines contrastive and generative objectives, pre-trained on 14M image-text pairs for flexible multimodal reasoning.
- **CoCa** (Yu et al., 2022) integrates contrastive and captioning losses under a unified encoder-decoder framework, pre-trained on JFT-3B and ALIGN datasets.
- **Git** (Wang et al., 2022a) is a transformer decoder conditioned on CLIP embeddings, trained with teacher forcing on 0.8B image-text pairs.
- **uform** (unum-cloud, 2023) is a generative multimodal model designed for captioning and VQA, pre-trained on MSCOCO, Visual Genome, and related datasets.

configuration	value
batch size	16, 32, 64, 128
optimizer	AdamW
weight decay	$1e^{-6}$
gradient clip val	0.5
base learning rate	$2e^{-5}$
optimizer momentum	$\beta 1, \beta 2 = 0.9, 0.999$
learning rate schedule	cosine decay
warmup epochs	200
RandomResizedCrop	(0.8,1.0)
RandomAugment (Cubuk et al., 2020)	(2, 12)
ColorJitter	(0.2,0.2,0.2)
RandomHorizontalFlip	0.5
RandomErasing	0.2

Table 8: Experimental configurations for fine-tuning different models.

D.3 Implementation Details

Evaluation strictly follows prior protocols: open-ended, closed-ended, and overall accuracy for medical VQA datasets, and overall accuracy for remote sensing and art datasets. Detailed experimental information is provided in Table 8.

D.4 Ablation on the Complementary Roles of the Three Frequency-Domain Losses

To dissect the functional role of each frequency-domain loss, we ablate the three components, Low-Frequency Preservation (*low*), Text-Guided Band Alignment (*align*), and Spectral Regularization (*reg*), both individually and in combination. Experiments are performed on CLIP-, BLIP-, and Git-based TGFM across VQA-RAD, OVQA, and SemArt. For consistency, we report overall accuracy.

As summarized in Table 9, using any single loss in isolation often performs worse than removing frequency losses entirely. This is most evident in CLIP, where applying only *low* or *align* results in clear degradation. The reason is structural: each loss controls only one axis of spectral behavior, and removing the others breaks the balance between semantic alignment, frequency stability, and cross-modal robustness. In the CLIP model fine-tuning on the VQA-RAD dataset, the *align* term alone induces strong optimization oscillations, as visualized in Figure 5(a), revealing its tendency to aggressively reshape spectral bands without stabilizing constraints. Pairwise combinations alleviate part of this imbalance but remain insufficient. In particular, removing the *reg* loss leads to unstable training dynamics, as shown in Figure 5(b). This

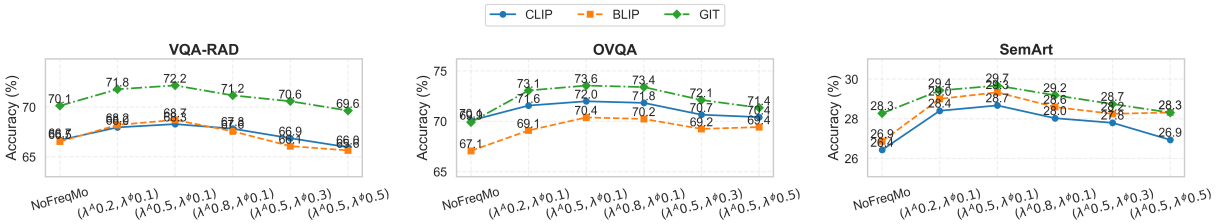

Figure 8: Sensitivity analysis of amplitude-phase modulation ratios. The results reveal the stability boundary of phase perturbation and highlight the optimal balance at $\lambda^A = 0.5$ and $\lambda^\phi = 0.1$ for cross-domain VQA adaptation.

Freq_loss	low	align	reg	VQA-RAD	OVQA	SemArt
CLIP				68.3	72.0	28.7
CLIP	✓			67.8	70.6	29.2
CLIP		✓		66.7	72.2	29.4
CLIP			✓	68.2	72.1	29.5
CLIP	✓	✓		71.5	74.6	29.7
CLIP		✓	✓	69.1	72.4	29.5
CLIP	✓		✓	70.1	74.1	29.6
CLIP	✓	✓	✓	**72.0**	**75.2**	**29.9**
BLIP				68.7	70.4	29.3
BLIP	✓			69.0	70.6	29.6
BLIP		✓		67.9	70.4	29.4
BLIP			✓	68.8	70.8	29.9
BLIP	✓	✓		69.1	71.0	29.9
BLIP		✓	✓	69.3	71.6	30.2
BLIP	✓		✓	69.3	71.5	30.3
BLIP	✓	✓	✓	**69.7**	**71.7**	**30.4**
Git				72.2	73.6	29.7
Git	✓			73.1	73.7	30.0
Git		✓		73.3	74.1	30.2
Git			✓	72.4	74.0	30.2
Git	✓	✓		73.2	75.2	31.0
Git		✓	✓	73.8	74.2	30.4
Git	✓		✓	73.4	73.7	30.2
Git	✓	✓	✓	**75.1**	**75.7**	**31.2**

Table 9: Ablation of the three frequency-domain losses, Low-Frequency Preservation (low), Text-Guided Band Alignment (align), and Spectral Regularization (reg). Results on VQA transfer tasks indicate that a single loss function is insufficient to achieve optimal performance.

demonstrates that *reg* acts as an essential smoothness prior, preventing the model from overfitting to high-frequency distortions introduced during modulation. Overall, the three frequency-domain losses are mutually complementary, which establishes a principled trade-off among spectral coherence, structural stability, and text-conditioned alignment, yielding the most reliable and consistent improvements across all evaluated domains.

D.5 Ablation on the Balance Between Amplitude and Phase Modulation

To understand how different frequency perturbation magnitudes affect multimodal adaptation, we systematically vary the amplitude and phase modulation ratios λ^A and λ^ϕ. As illustrated in Figure 8, we evaluate CLIP-, BLIP-, and Git-based TGFM models on VQA-RAD, OVQA, and SemArt under a range of ratio configurations.

A consistent trend emerges across all architectures and datasets: increasing the phase ratio degrades performance, in some cases falling below the non-modulated baseline. The phase spectrum encodes essential spatial and structural information (Yang and Soatto, 2020); overly strong phase perturbations distort these spatial priors, resulting in unstable optimization and weakened multimodal alignment. Conversely, moderate phase modulation introduces controlled spectral diversity, enhancing cross-modal correspondence while preserving the structural fidelity required for reliable semantic grounding. Amplitude modulation behaves more gently. Since amplitude primarily influences the magnitude distribution of frequency responses (e.g., contrast and global intensity), increasing λ^A leads to progressive and less disruptive changes in adaptation behavior. This makes amplitude modulation more tolerant to scaling, although excessively large values still diminish marginal gains. Overall, the optimal configuration is achieved at $\lambda^A = 0.5$ and $\lambda^\phi = 0.1$. This setting provides the best trade-off between representational diversity (from amplitude) and spatial consistency (from phase), defining a stable operating region for frequency-domain modulation. These findings further validate TGFM's design: phase modulation must remain lightweight to preserve structure, while amplitude modulation can serve as the primary lever for spectral adaptation.

D.6 Ablation on the Relative Weighting of Frequency-Domain Losses

To assess the sensitivity of TGFM to the relative contributions of its frequency-domain losses, we systematically vary the weights of the Low-Frequency Preservation (λ^{low}), Text-Guided Band Alignment (λ^{align}), and Spectral Regularization (λ^{reg}) terms. Table 10 reports results for CLIP-, BLIP-, and Git-based models across VQA-RAD,

Freq_loss	λ^{low}	λ^{align}	λ^{reg}	VQA-RAD	OVQA	SemArt
CLIP				68.3	72.0	28.7
CLIP	0.5	0.5	0.2	71.2	73.7	29.4
CLIP	1.0	0.5	0.2	**72.0**	**75.2**	**29.9**
CLIP	1.5	0.5	0.2	70.8	74.0	29.5
CLIP	1.0	0.2	0.2	68.9	72.9	29.2
CLIP	1.0	1.0	0.2	69.1	74.4	29.5
CLIP	1.0	0.5	0.1	71.2	73.9	29.3
CLIP	1.0	0.5	0.3	70.2	73.0	29.2
BLIP				68.7	70.4	29.3
BLIP	0.5	0.5	0.2	69.4	70.8	29.7
BLIP	1.0	0.5	0.2	**69.7**	**71.7**	**30.4**
BLIP	1.5	0.5	0.2	69.0	**71.7**	30.0
BLIP	1.0	0.2	0.2	68.7	70.6	29.7
BLIP	1.0	1.0	0.2	67.7	71.6	29.5
BLIP	1.0	0.5	0.1	68.8	71.1	29.5
BLIP	1.0	0.5	0.3	69.4	71.2	30.3
Git				72.2	73.6	29.7
Git	0.5	0.5	0.2	72.6	74.1	30.7
Git	1.0	0.5	0.2	**75.1**	**75.7**	**31.2**
Git	1.5	0.5	0.2	73.8	73.8	30.3
Git	1.0	0.2	0.2	73.7	74.4	30.0
Git	1.0	1.0	0.2	72.4	73.2	30.2
Git	1.0	0.5	0.1	74.4	74.6	30.7
Git	1.0	0.5	0.3	74.3	74.5	**31.2**

Table 10: Effect of varying the relative weights of the three frequency-domain objectives, Low-Frequency Preservation (λ^{low}), Text-Guided Band Alignment (λ^{align}), and Spectral Regularization (λ^{reg}), across VQA transfer benchmarks.

Model	Band Partition	VQA-RAD	OVQA	SemArt
CLIP	2-band (0.5)	71.6	74.8	29.6
CLIP	3-band (0.2, 0.8)	71.5	74.6	29.6
CLIP	3-band (0.33, 0.66)	**72.0**	**75.2**	**29.9**
CLIP	3-band (0.4, 0.6)	71.8	74.8	**29.9**
CLIP	4-band (0.25, 0.5, 0.75)	71.2	74.4	29.1
BLIP	2-band (0.5)	69.3	71.5	30.1
BLIP	3-band (0.2, 0.8)	69.2	71.2	30.0
BLIP	3-band (0.33, 0.66)	**69.7**	**71.7**	**30.4**
BLIP	3-band (0.4, 0.6)	69.6	71.4	30.2
BLIP	4-band (0.25, 0.5, 0.75)	68.8	71.1	29.9
Git	2-band (0.5)	74.6	75.2	30.8
Git	3-band (0.2, 0.8)	74.2	74.9	30.7
Git	3-band (0.33, 0.66)	**75.1**	**75.7**	**31.2**
Git	3-band (0.4, 0.6)	74.6	75.4	31.0
Git	4-band (0.25, 0.5, 0.75)	74.0	74.3	30.4

Table 11: Ablation study on frequency band partitioning and threshold selection. We compare different numbers of bands (2–4) and varying threshold positions to evaluate their impact on domain-specific VQA performance across CLIP, BLIP, and Git backbones.

OVQA, and SemArt.

A consistent trend emerges across all backbones and domains: the configuration ($\lambda^{low} = 1.0$, $\lambda^{align} = 0.5$, $\lambda^{reg} = 0.2$) yields the best overall performance. This balance reflects an effective trade-off among spectral stability, semantic alignment, and optimization smoothness. When λ^{low} is too small (e.g., 0.5), the model lacks sufficient low-frequency anchoring, resulting in reduced structural coherence and weakened cross-modal alignment. Conversely, overly large λ^{low} (e.g., 1.5) over-constrains the latent spectrum, restricting the flexibility needed for domain adaptation. A similar trend holds for λ^{align}: higher values aggressively reshape band-level responses and may lead to optimization instability. The role of λ^{reg} is complementary; too small a value weakens the smoothness prior and increases training oscillations, whereas excessive regularization suppresses beneficial spectral diversity. Taken together, these findings demonstrate that the three frequency-domain objectives interact as a tightly coupled multi-objective system rather than independent components.

D.7 Frequency Band Partition and Threshold Analysis

To better understand the effect of frequency band decomposition in our Text-Guided Frequency Modulation (TGFM) framework, we conduct ablation studies from two perspectives: (i) the number of frequency bands and (ii) the thresholds used for band separation.

We first vary the number of frequency bands from 2 to 4 while keeping the thresholds either fixed or equally spaced. As shown in Table 11, the 3-band setting with thresholds (0.33, 0.66) achieves the best performance across all datasets and backbones. Using only 2 bands tends to underrepresent mid-frequency content, which is important for object-level patterns, whereas 4 bands introduce more variance in low-data target supervision, slightly reducing performance.

Next, we investigate different threshold positions for the 3-band configuration. Thresholds (0.2, 0.8) or (0.4, 0.6) are slightly worse than (0.33, 0.66), suggesting that extremely narrow or wide mid-frequency bands either fail to capture sufficient object-level information or lead to unstable gradient estimation. This result supports our choice of the 3-band, evenly balanced threshold as a bias-variance tradeoff between capturing informative frequency content and maintaining stable adaptation.

Overall, these results validate our design choice of the 3-band frequency partition with thresholds (0.33, 0.66). It balances low-, mid-, and high-frequency components effectively, maximizing semantic alignment and adaptation performance across heterogeneous VQA datasets.

Model	Block	VQA-RAD	OVQA	SemArt
CLIP	1/4	68.5	71.6	26.9
CLIP	1/2	70.3	73.0	28.5
CLIP	3/4	70.7	73.7	29.1
CLIP	Final	72.0	75.2	29.9
BLIP	1/4	67.2	68.6	27.8
BLIP	1/2	68.6	70.0	29.1
BLIP	3/4	69.0	70.7	29.6
BLIP	Final	69.7	71.7	30.4
Git	1/4	70.9	70.6	29.1
Git	1/2	72.8	73.5	30.2
Git	3/4	73.7	74.2	30.5
Git	Final	75.1	75.7	31.2

Table 12: Effect of TGFM insertion depth across different backbone models. TGFM is inserted at different transformer block depths (1/4, 1/2, 3/4, and final layer), and evaluated on three representative VQA benchmarks.

D.8 Effect of TGFM Insertion Depth

We investigate the sensitivity of TGFM to the depth at which it is inserted into the vision-language backbone. Specifically, we apply TGFM at four representative positions corresponding to the first quarter (1/4), middle (1/2), third quarter (3/4), and final transformer blocks, and evaluate performance on three representative datasets across different models, including CLIP, BLIP, and Git.

As shown in Table 12, TGFM exhibits a consistent performance improvement as it is inserted at deeper layers across all backbones and datasets. Early-layer insertion yields limited gains, while mid-to-late layers provide progressively greater improvements, with the final block achieving the best overall performance. This trend suggests that TGFM benefits from operating on semantically mature representations, where global structure and cross-modal alignment cues are more explicitly encoded. In contrast, early-layer features lack sufficient semantic abstraction, limiting the effectiveness of frequency-based modulation. The observed pattern is consistent across diverse architectures, indicating that TGFM is not sensitive to a specific backbone design. These results support our design choice of inserting TGFM at deeper layers, where it can effectively refine discriminative frequency components while preserving high-level semantic structure.

D.9 Cross-Domain Adaptation Robustness Across Seeds

Since our framework is evaluated on six adaptation scenarios with heterogeneous VQA datasets and diverse model architectures, performing full multi-seed experiments for every model–dataset pair would incur substantial computational overhead. The cost is particularly prohibitive for large-scale models such as uform and dataset-heavy tasks such as EarthVQA. To ensure statistical reliability while maintaining feasible resource usage, we therefore select the four strongest models and conduct 3-seed validation on four representative datasets.

Model	Method	VQA-RAD	PathVQA	OVQA	SemArt
CLIP	FSA	71.4 ± 0.6	60.8 ± 0.5	75.2 ± 0.7	29.6 ± 0.3
CLIP	TGFM	72.0 ± 0.5	61.1 ± 0.3	75.2 ± 0.5	29.9 ± 0.1
BLIP	FSA	69.6 ± 0.4	60.2 ± 0.5	69.5 ± 0.4	30.4 ± 0.3
BLIP	TGFM	69.7 ± 0.4	60.9 ± 0.3	71.7 ± 0.3	30.4 ± 0.2
Coca	FSA	60.5 ± 0.5	59.0 ± 0.3	71.5 ± 0.4	28.9 ± 0.4
Coca	TGFM	62.3 ± 0.3	59.2 ± 0.1	71.6 ± 0.4	29.1 ± 0.3
Git	FSA	74.2 ± 0.6	61.1 ± 0.3	75.1 ± 0.6	31.1 ± 0.4
Git	TGFM	75.1 ± 0.5	61.3 ± 0.3	75.7 ± 0.5	31.2 ± 0.4

Table 13: Multi-seed cross-domain adaptation performance (mean $\pm$ std over three seeds) of FSA and TGFM across four VQA benchmarks.

To contextualize the stability of TGFM, we additionally report multi-seed results for the strongest source-data-free baseline FSA under the same experimental settings. As shown in Table 13, TGFM consistently achieves higher mean performance with comparable or lower variance across seeds, while FSA exhibits slightly larger fluctuations on several benchmarks. Overall, the averaged results closely align with the best single-seed scores reported in Tables 1 and 2, indicating that the performance gains of TGFM are not attributable to favorable random initialization. These results confirm that TGFM provides more stable and reproducible cross-domain adaptation compared to the existing frequency-based source-data-free FSA method under the same training conditions.

D.10 Comparison with Alternative Parameter-Efficient Tuning Methods

To further contextualize the effectiveness of the proposed TGFM, we compare it with alternative parameter-efficient fine-tuning (PEFT) strategies under the same CLIP backbone, including Visual Prompt Tuning (VPT) (Jia et al., 2022), Adapter-based tuning (Houlsby et al., 2019), LoRA, and a representative frequency-domain PEFT baseline, FourierFT (Gao et al., 2024). FourierFT performs parameter-efficient adaptation by inserting learnable filters in the Fourier domain of intermediate visual features, allowing for frequency-selective tuning without requiring explicit spatial-domain modules.

As shown in Table 14, VPT exhibits substan-

Method	VQA-RAD	PathVQA	OVQA	SemArt
VPT	47.6	44.6	46.2	21.7
VPT + TGFM	49.2	46.2	48.4	23.1
Adapter	57.8	55.4	61.1	26.9
Adapter + TGFM	59.3	56.9	63.0	28.3
LoRA	59.5	57.3	64.1	28.5
LoRA + TGFM	63.4	58.2	66.9	29.6
FourierFT	60.5	57.7	64.9	28.9
FourierFT + TGFM	63.5	58.1	66.8	29.4

Table 14: Performance comparison of VPT, Adapter, LoRA, and FourierFT on CLIP across four cross-domain VQA datasets. Results with TGFM highlight its consistent improvements under different PEFT fine-tuning strategies.

Model	Method	VQA-RAD	PathVQA	OVQA	SemArt
CLIP	-	66.7	58.1	70.1	26.4
CLIP	FiLM	66.9	58.2	70.5	26.7
CLIP	TGFM	68.3	60.5	72.0	28.7
BLIP	-	66.5	58.0	67.1	26.9
BLIP	FiLM	66.6	58.2	67.5	27.4
BLIP	TGFM	68.7	60.8	70.4	29.3
Coca	-	55.7	55.4	69.2	24.8
Coca	FiLM	56.3	56.0	69.5	25.9
Coca	TGFM	59.9	58.6	69.8	28.2
Git	-	70.1	58.1	69.9	28.3
Git	FiLM	70.2	58.6	70.1	28.5
Git	TGFM	72.2	60.7	73.6	29.7

Table 15: Comparison between FiLM-style spatial feature modulation and TGFM without frequency-domain losses across different vision-language backbones. All modulation modules are inserted at the last layer of the backbone, ensuring a fair comparison of modulation mechanisms.

tially lower performance across all datasets. This is expected, as VPT only optimizes a small set of prompt tokens and lacks the capacity to adapt intermediate visual representations under severe domain shifts. Adapter-based tuning achieves stronger results by introducing lightweight trainable modules within transformer blocks, enabling moderate feature adaptation. Incorporating TGFM further enhances adapter performance across all datasets, indicating that frequency-domain modulation provides complementary benefits beyond spatial-domain adaptation.

Among all strategies, LoRA achieves the strongest baseline performance, and combining LoRA with TGFM consistently yields the largest absolute gains. Notably, TGFM also brings consistent improvements when paired with FourierFT, despite both operating in the frequency domain. This suggests that TGFM contributes additional benefits beyond generic frequency filtering, which are attributable to its explicit modeling of amplitude and phase, as well as its text-conditioned modulation mechanism. Overall, these results demonstrate that TGFM is complementary to both spatial-domain and frequency-domain PEFT methods and can be seamlessly integrated to further improve cross-domain VQA performance.

D.11 Runtime and Computational Overhead Analysis

We evaluate the runtime efficiency and computational overhead of TGFM in comparison with existing adaptation strategies under both full fine-tuning and parameter-efficient settings. All experiments are conducted using the CLIP backbone on the VQA-RAD dataset, with a batch size of 64 on a single NVIDIA V100 GPU.

Training time is measured as the average wall-clock time per epoch, while inference latency reports the average time required to process a single sample. As shown in Table 6, TGFM introduces only a modest training-time overhead compared to standard fine-tuning. Specifically, TGFM increases the per-epoch training time by 8.44%, which is substantially lower than the overhead introduced by FSA (+15.70%), while maintaining comparable inference latency. This demonstrates that TGFM achieves frequency-domain modulation with minimal additional computational cost.

When combined with LoRA, TGFM preserves the parameter-efficiency advantages of low-rank adaptation. LoRA+TGFM increases the number of trainable parameters by only 0.09M and incurs a marginal per-epoch training-time overhead of 3.66%. Overall, these results demonstrate that TGFM is a lightweight and efficient module that can be seamlessly integrated into both full fine-tuning and parameter-efficient adaptation pipelines without sacrificing runtime efficiency.

D.12 Comparison with Feature-wise Modulation (FiLM)

To distinguish the effect of frequency-domain representation modulation from conventional spatial feature conditioning, we compare TGFM with FiLM (Perez et al., 2018), a widely used feature-wise linear modulation method. For a fair comparison, both FiLM and TGFM are implemented as lightweight modulation modules inserted at the last layer of the backbone, and are trained using identical target-domain supervision. Notably, TGFM in this experiment excludes all frequency-domain losses, isolating the contribution of the modulation

mechanism itself.

As shown in Table 15, FiLM yields only marginal or negligible improvements over direct fine-tuning across all datasets and architectures, suggesting that spatial-domain affine modulation alone is insufficient to effectively address domain shift in source-data-free multimodal adaptation. In contrast, TGFM consistently achieves substantial performance gains across all backbones and benchmarks, even without explicit frequency loss regularization. These results indicate that the advantage of TGFM does not stem from text-conditioned feature modulation, but rather from its frequency-domain design, which enables structured and global adaptation of visual representations.

D.13 Comparison with Recent Domain Adaptation Methods

We also compare TGFM with four recent representative domain adaptation methods that can be applied or adapted to multimodal VQA: (1) **UDAM** (Weng et al., 2025) introduces semantic context and query feature alignment with a pairwise domain-aware prompt strategy for domain adaptive VQA with multi-modal large models. (2) **ReCLIP** (Hu et al., 2024) is a source-free domain adaptation method for vision-language models that refines visual and text encoders via cross-modality self-training with pseudo labels. (3) **CATCH**(Li et al., 2025) uses a plug-and-play domain classifier and dual adapter mechanism (prompt and visual adapters) to improve cross-domain VQA generalization with minimal changes to the backbone. (4) **Open-ended VQA Domain Adaptation** (Xu et al., 2020b) aligns joint embeddings across source and target using supervised multi-modal domain alignment for VQA tasks with limited target labels. For methods with publicly available implementations, we directly use the released code for evaluation; otherwise, we re-implement the core techniques following the original papers. We also perform source-data-free training, enabling these methods to use target domain labels to facilitate effective transfer of VQA tasks.

Although these baselines address domain shifts in multimodal settings, TGFM consistently outperforms them across CLIP, BLIP, and Git backbones on VQA-RAD, PathVQA, OVQA, and SemArt, as summarized in Table 16. This performance advantage stems from three key factors. First, TGFM operates directly on output representations by performing text-conditioned amplitude and phase modulation in the frequency domain, enabling structured and selective spectral adaptation that more effectively mitigates domain discrepancies than feature-level alignment or prompt tuning alone. Second, the proposed frequency-domain loss explicitly enforces spectral stability and interpretability across domains, whereas most prior methods focus primarily on cross-modal alignment without fine-grained control over frequency behavior. Third, TGFM is inherently plug-and-play: it can be seamlessly integrated into different vision–language models without introducing task- or backbone-specific interfaces. In contrast, many existing methods require customized adapters, prompt designs, or alignment modules tailored to particular architectures or feature representations. Together, these properties allow TGFM to deliver consistent and robust improvements across diverse datasets, tasks, and model families.

Model	Method	VQA-RAD	PathVQA	OVQA	SemArt
CLIP	-	66.7	58.1	70.1	26.4
	UDAM	70.2	60.1	73.5	29.2
	ReCLIP	70.2	60.7	74.0	29.4
	CATCH	69.8	60.2	72.6	28.7
	Open-Adaptation	67.7	58.4	69.8	26.5
	TGFM	**72.0**	**61.1**	**75.2**	**29.9**
BLIP	-	66.5	58.0	67.1	26.9
	UDAM	68.3	59.8	70.1	28.8
	ReCLIP	67.1	58.8	68.4	27.4
	CATCH	68.6	60.0	70.4	28.5
	Open-Adaptation	66.8	58.2	67.1	27.1
	TGFM	**69.7**	**60.9**	**71.7**	**30.4**
Git	-	70.1	58.1	69.9	28.3
	UDAM	73.2	60.1	73.5	30.4
	ReCLIP	71.3	59.0	71.2	29.1
	CATCH	73.0	60.4	73.2	30.1
	Open-Adaptation	70.6	58.3	70.3	28.3
	TGFM	**75.1**	**61.3**	**75.7**	**31.2**

Table 16: Performance comparison of TGFM with representative source-data-free adaptation methods across different VLM backbones and VQA datasets.

E Pseudocode

For clarity and reproducibility, we present the pseudocode of TGFM in this section, which closely follows our actual implementation. Algorithm 1 details the text-guided frequency modulation module, while Algorithm 2 specifies the frequency-domain loss function. The full implementation code and trained models will be publicly released upon acceptance to support exact replication of our results.

Algorithm 1 TGFM: Text-Guided Frequency Modulation

Require: image feature $\mathbf{x} \in \mathbb{R}^{B\times C\times H\times W}$, text tokens $\mathbf{t} \in \mathbb{R}^{B\times L\times D}$, parameters $C, H = W, \lambda^A, \lambda^\phi$ (defaults: λ^A=0.5, λ^ϕ=0.1)

Ensure: enhanced feature $\hat{\mathbf{x}} \in \mathbb{R}^{B\times C\times H\times W}$, meta dictionary

Text attention pooling:

  $\boldsymbol{\alpha}_{b,l} \leftarrow \mathrm{softmax}(\mathrm{mean}_d\, \mathbf{t}_{b,l,d})$ $[B, L]$

  $\mathbf{t}_b^{\mathrm{att}} \leftarrow \sum_l \boldsymbol{\alpha}_{b,l}\mathbf{t}_{b,l}$ $[B, D]$

Radial freq. map: compute $r_{i,j} = \sqrt{(\frac{2i}{H}-1)^2 + (\frac{2j}{W}-1)^2}$, normalize to $[0, 1]$ $[H, W]$

Build per-frequency conditions:

  For each sample b and position (i, j) form $\mathbf{z}_{b,(i,j)} = [\,\mathbf{t}_b^{\mathrm{att}};\ r_{i,j}\,] \in \mathbb{R}^{D+1}$

  Stack $\mathbf{Z}_b \in \mathbb{R}^{(H\cdot W)\times(D+1)}$

Dual-branch MLP (shared weights):

  $\mathbf{H}_b \leftarrow \mathrm{GELU}(W^{(2)}\,\mathrm{GELU}(W^{(1)}\mathbf{Z}_b)) \in \mathbb{R}^{HW\times H_{hid}}$

  $\Delta A^{\mathrm{nonlin}}, \Delta\phi^{\mathrm{nonlin}} \leftarrow \tanh(W^{\mathrm{nonlin},A}\mathbf{H}_b), \tanh(W^{\mathrm{nonlin},\phi}\mathbf{H}_b) \in \mathbb{R}^{HW\times C}$

  $\Delta A^{\mathrm{lin}},\ \Delta\phi^{\mathrm{lin}} \leftarrow W^{\mathrm{lin},A}\mathbf{Z}_b,\ W^{\mathrm{lin},\phi}\mathbf{Z}_b \in \mathbb{R}^{HW\times C}$

Combine & reshape to masks:

  $\mathbf{M}_b^A \leftarrow \tanh(\Delta A^{\mathrm{nonlin}} + \Delta A^{\mathrm{lin}})\cdot\lambda^A$, reshape to $[C, H, W]$ and align with channel-wise spectra ▷ final: $[B, C, H, W]$

  $\mathbf{M}_b^\phi \leftarrow \tanh(\Delta\phi^{\mathrm{nonlin}} + \Delta\phi^{\mathrm{lin}})\cdot\lambda^\phi$

FFT-domain modulation (per channel):

for each b, c **do**

  $X_{b,c} \leftarrow \mathrm{FFT2}(\mathbf{x}_{b,c,:,:})$ ▷ complex $[H, W]$

  $A_{b,c} \leftarrow |X_{b,c}|,\ \phi_{b,c} \leftarrow \angle X_{b,c}$

  $A'_{b,c} \leftarrow A_{b,c}\cdot(1 + \mathbf{M}^A_{b,c})$

  $\phi'_{b,c} \leftarrow \phi_{b,c} + \mathbf{M}^\phi_{b,c}$

  $\widetilde{X}_{b,c} \leftarrow \mathrm{polar}(A'_{b,c}, \phi'_{b,c})$ ▷ recompose complex spectrum

  $\hat{x}_{b,c} \leftarrow \Re(\mathrm{IFFT2}(\widetilde{X}_{b,c}))$ ▷ discard numerical imaginary residuals

end for

Stack $\hat{x}_{b,c}$ to produce $\hat{\mathbf{x}}_b$ and aggregate batch $\hat{\mathbf{x}}$

Save meta: $\texttt{meta}[b] = \{A,\ A',\ \mathbf{M}^A,\ r\}$, **return** $\hat{\mathbf{x}}$, `meta`

Algorithm 2 TGFM: Frequency-Modulation Losses

Require: original features $\mathbf{x} \in \mathbb{R}^{B\times C\times H\times W}$, enhanced features $\hat{\mathbf{x}}$, text tokens $\mathbf{t} \in \mathbb{R}^{B\times L\times D}$, `meta` (contains $A, A', \mathbf{M}^A, r$), band thresholds (t_1, t_2) $(0.33, 0.66)$, weights $\lambda_{\mathrm{low}}, \lambda_{\mathrm{align}}, \lambda_{\mathrm{reg}}$

Ensure: total loss $\mathcal{L}^{\mathrm{total}}$ and individual loss components

Pool features: $f_b^{\mathrm{orig}} \leftarrow \mathrm{GAP}(\mathbf{x}_b)$, $f_b^{\mathrm{new}} \leftarrow \mathrm{GAP}(\hat{\mathbf{x}}_b)$ $[B, C]$

Text pool: $\bar{\mathbf{t}}_b \leftarrow \mathrm{mean}_l(\mathbf{t}_{b,l})$ $[B, D]$

Build band masks:

  $m_{\mathrm{low}} = (r \le t_1),\ m_{\mathrm{mid}} = (t_1 < r \le t_2),\ m_{\mathrm{high}} = (r > t_2)$ each $[H, W]$

Band energy (per sample): compute for A and A':

  For each band $k \in \{\mathrm{low,mid,high}\}$:

    expand m_k to $[B, C, H, W]$, compute spatial average per channel, then average over channels to yield $E_b^k(A)$ and $E_b^k(A')$ $\in \mathbb{R}^B$

1) Low-frequency preservation:

  $L_{\mathrm{low}} \leftarrow \mathrm{MSE}(\log(E^{\mathrm{low}}(A')),\ \log(E^{\mathrm{low}}(A)))$

2) Text-guided band alignment:

  $\mathbf{w}_b \leftarrow \mathrm{softmax}(\mathrm{LayerNorm}(\bar{\mathbf{t}}_b)W_{\mathrm{text2band}})$ $[B, 3]$

  $p_b(k) \leftarrow \dfrac{E_b^k(A')}{\sum_j E_b^j(A')}$ empirical band distribution

  $L_{\mathrm{align}} \leftarrow \mathbb{E}_b\big[D_{\mathrm{KL}}(p_b \,\|\, \mathbf{w}_b)\big]$ ▷ match spectral energy to text-conditioned prior

3) Spectral Regularization:

  $L_{\mathrm{mask\_l1}} \leftarrow \mathrm{mean}(|\mathbf{M}^A|)$

  $L_{\mathrm{mask\_tv}} \leftarrow \mathrm{TV}(\mathbf{M}^A)$ ▷ sum of horiz/vert diffs averaged

  $L_{\mathrm{feat}} \leftarrow \mathrm{MSE}(\ \mathrm{proj}(f^{\mathrm{new}}),\ \mathrm{proj}(f^{\mathrm{orig}})\ )$

  $L_{\mathrm{reg}} \leftarrow L_{\mathrm{mask\_l1}} + L_{\mathrm{mask\_tv}} + L_{\mathrm{feat}}$

Aggregate:

  $\mathcal{L}^{\mathrm{freq}} \leftarrow \lambda_{\mathrm{low}}L_{\mathrm{low}} + \lambda_{\mathrm{align}}L_{\mathrm{align}} + \lambda_{\mathrm{reg}}L_{\mathrm{reg}}$

  $\mathcal{L}^{\mathrm{total}} \leftarrow \mathcal{L}^{\mathrm{cls}} + \mathcal{L}^{\mathrm{freq}}$, **return** $\mathcal{L}^{\mathrm{total}}$ and component dictionary