WaveletLoRA: Frequency-Aware Content-Style Decomposition for Personalized Image Generation

Peiyao Wang[1,2]* **Jiahui Sun**[1,2]* **Weining Wang**[1,2]† **Jing Liu**[1,2]

[1]Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences

Abstract

Content-style decomposition from a single image is a fundamental challenge in personalized image generation, aiming to separate subject identity from visual style for flexible recomposition. Existing approaches primarily operate in the spatial domain, without explicitly exploiting the distinct frequency characteristics naturally associated with content and style. In this paper, we propose WaveletLoRA, a novel frequency-aware framework for explicit content-style decomposition through wavelet-domain supervision. Specifically, we first apply a multi-level discrete wavelet transform (DWT) to decompose the reference image into low-frequency and high-frequency subbands. We then assign a dedicated LoRA branch to each subband, where the low-frequency branch models global appearance for style acquisition, while the high-frequency branches capture structural details for content preservation. To further improve disentanglement quality, we introduce a dense Mixture-of-Experts aggregation module together with a frequency-domain regularization objective, which reduces style leakage and enhance content fidelity. In addition, we establish WaveBench and develop a VLM-based evaluation protocol that explicitly measures both attribute similarity and attribute separation for disentanglement assessment. Extensive experiments demonstrate that WaveletLoRA consistently outperforms existing methods and produces results that better align with human preference. Code and the proposed dataset will be released in `https://github.com/PinkYearwww/WaveletLoRA`.

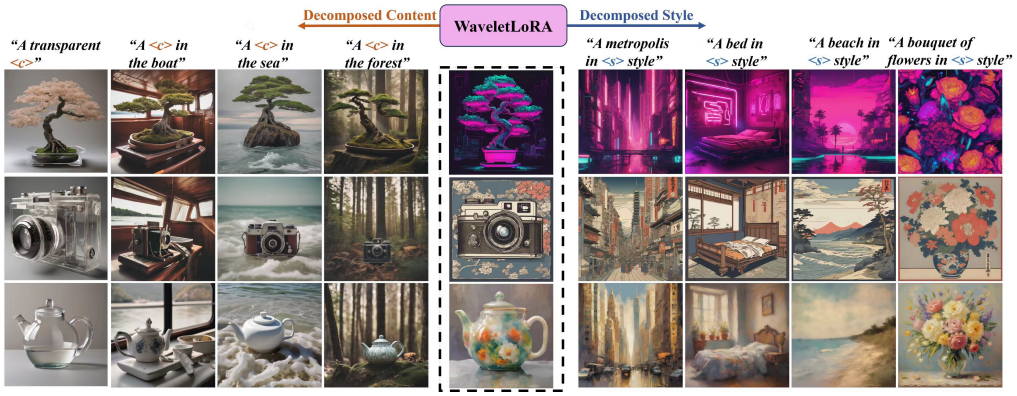

Figure 1: Given a single input image (middle), WaveletLoRA learns a disentangled content or subject LoRA (left) and style LoRA (right) that can be used to generate new images with the learned concepts.

*Equal Contribution.
†Corresponding Author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Personalized image generation has recently emerged as a major research direction in generative modeling, driven by the growing demand for controllable and customized visual content generation [19, 20, 22, 5, 32]. Although modern text-to-image diffusion models have demonstrated remarkable generation capability, textual prompts alone often fail to specify fine-grained instance-level attributes, such as the identity of a particular subject or the unique visual characteristics of a specific artistic style. To address this limitation, example-driven personalization methods, such as DreamBooth [19] and StyleDrop [22], fine-tune a pretrained model on a small set of reference images to inject subject or style concepts into the generation process. While effective, these methods typically specialize in either content or style in isolation. Disentangling both factors simultaneously from a single image, while preserving clean separation and recompositional flexibility, remains a challenging open problem.

To address single image content–style decomposition, several methods have recently been proposed [5, 20, 14]. However, existing methods predominantly operate in the spatial or feature domain and learn to separate content and style through architectural constraints and optimization heuristics, rather than explicitly modeling the intrinsic and fundamental disparities between structural content and visual appearance. As a result, content and style representations often remain entangled, leading to style leakage in content generation or subject residue in style transfer. We argue that a key reason for this limitation is the neglect of an important signal-processing prior: content and style naturally exhibit different frequency characteristics.

Classical studies in image processing have shown that high-frequency components primarily encode structural information, such as edges and fine-grained textures, whereas low-frequency components mainly capture global appearance attributes, including color distribution and artistic style [8, 27]. This observation suggests that frequency decomposition provides a natural inductive bias for disentangling content and style. Existing frequency-aware generative methods have largely relied on Fourier-domain analysis [25, 4, 29, 13], which captures global spectral statistics but lacks spatial locality. In contrast, the discrete wavelet transform (DWT) provides localized multi-scale frequency decomposition while preserving both spatial and spectral information, making it more suitable for separating structural content from stylistic appearance.

Motivated by these insights, we propose **WaveletLoRA**, a novel frequency-aware framework for personalized content–style decomposition. Specifically, we first decompose the reference image into low-frequency and high-frequency wavelet subbands, and assign dedicated LoRA branches to each subband for explicit frequency-domain supervision. The low-frequency branch is optimized to model global style characteristics, while the high-frequency branches focus on preserving multi-scale structural details. To effectively integrate structural information across different frequency bands, we further introduce a dense Mixture-of-Experts (MoE) module that dynamically aggregates high-frequency LoRA branches into a unified content representation. In addition, we design a frequency-domain disentanglement objective that combines weight-level frequency regularization with AdaIN-based style statistics alignment to further encourage clean separation between content and style. Another challenge in single-image content–style decomposition lies in evaluation. Existing metrics, such as DINOv2 [10] and CLIP-I [2], primarily measure holistic visual similarity and cannot reliably assess disentanglement quality. To address this issue, we construct WaveBench, a large-scale benchmark together with a VLM-based evaluation framework that explicitly measures both attribute similarity and attribute separation.

In summary, our main contributions are as follows:

- We observe that content and style exhibit distinct frequency characteristics, and propose wavelet-domain decomposition as a principled prior for single-image content–style disentanglement.
- We propose WaveletLoRA, a novel frequency-aware framework that learns disentangled content and style LoRAs from multi-scale wavelet subbands, together with dense MoE aggregation and frequency-domain regularization for improved structural preservation and reduced style leakage.
- We establish WaveBench, a large-scale benchmark for single-image content–style decomposition with a VLM-based evaluation protocol, and demonstrate state-of-the-art performance in content preservation and style fidelity.

2 Related Works

Personalized image generation. With the rise of diffusion models [11, 18, 23], researchers have developed powerful fine-tuning strategies tailored to personalized image generation. DreamBooth [19] and Textual Inversion [6] enabled faithful subject or style learning from a handful of reference images. However, these methods [22, 6, 19] are predominantly designed to capture a single concept at a time. To address the more challenging scenario of extracting multiple concepts from one image, Inspiration Tree [26] and CusConcept [34] learn hierarchical or customized text embeddings via textual inversion, but the resulting embeddings offer limited generation control compared to full weight-space adaptations. In parallel, B-LoRA [5], ZipLoRA [20], and UnZipLoRA [14] tackle content–style separation more directly by jointly or independently training LoRA modules under block-wise, column-wise, and prompt-level constraints, achieving compatible content and style representations suitable for recomposition. Despite these advances, existing methods operate purely in the spatial or feature domain and overlook a key signal-processing insight, namely that high frequencies carry structural edges and fine detail, while low frequencies encode global color and texture. This gap motivates our WaveletLoRA framework, which leverages multi-scale wavelet decomposition to cleanly separate low-frequency style from high-frequency content.

Frequency perspective in image generation. Exploiting frequency representations has benefited diverse vision tasks. Early works incorporated the Discrete Cosine Transform (DCT) into CNNs for faster learning [9] or designed frequency-adaptive networks for super-resolution [33], demonstrating the promise of spectral processing beyond spatial analysis. In diffusion-based generation, frequency cues have been increasingly adopted to improve synthesis quality and editing control. For instance, FDDiff [28] and FDG-Diff [36] enhance high-frequency components to reduce blurriness, while WaveDiff [16] decomposes latent features into wavelet subbands for richer textures. FreeU [21] and ConsisID [35] modulate frequency responses to boost generation quality and identity preservation, and FCDiffusion [7] employs frequency-specific pathways for multi-modal translation. Despite these advances, existing frequency-based methods either rely on hand-crafted architectural changes or auxiliary losses, and none tackle the problem of disentangling content and style from a single image. More importantly, the use of wavelet decomposition for personalized image generation remains unexplored. WaveletLoRA fills this gap by directly conditioning LoRA branches with wavelet subbands from the reference image, thereby naturally assigning low-frequency supervision to style and high-frequency details to content.

3 Method

3.1 Overview of Our Method

Given a reference image containing both subject content and visual style, our goal is to disentangle structural content from stylistic appearance for flexible personalized image generation. Existing methods typically perform disentanglement in the spatial or feature domain, often leading to residual entanglement between content and style. To address this issue, we propose WaveletLoRA, a wavelet-guided framework for explicit content–style disentanglement. As illustrated in Fig. 2, we first decompose the reference image into low-frequency and high-frequency subbands through a multi-level discrete wavelet transform (DWT). The low-frequency branch models globally coherent appearance information for style representation, while the high-frequency branches preserve structural details for content representation. The resulting frequency-specific representations are further integrated through adaptive aggregation and optimized with frequency-domain disentanglement supervision for cleaner separation between content and style.

3.2 Wavelet-based Content–Style Decomposition

Structural content and visual appearance often exhibit different frequency tendencies. In general, high-frequency components tend to preserve structural details such as edges and contours, whereas low-frequency components are more strongly associated with global appearance attributes including color distribution and artistic style. Motivated by this observation, we argue that frequency decomposition provides a natural inductive bias for content–style disentanglement. We employ a two-level DWT to decompose the reference image into localized frequency subbands. For a one-level decomposition, DWT transforms the input image into four subbands,

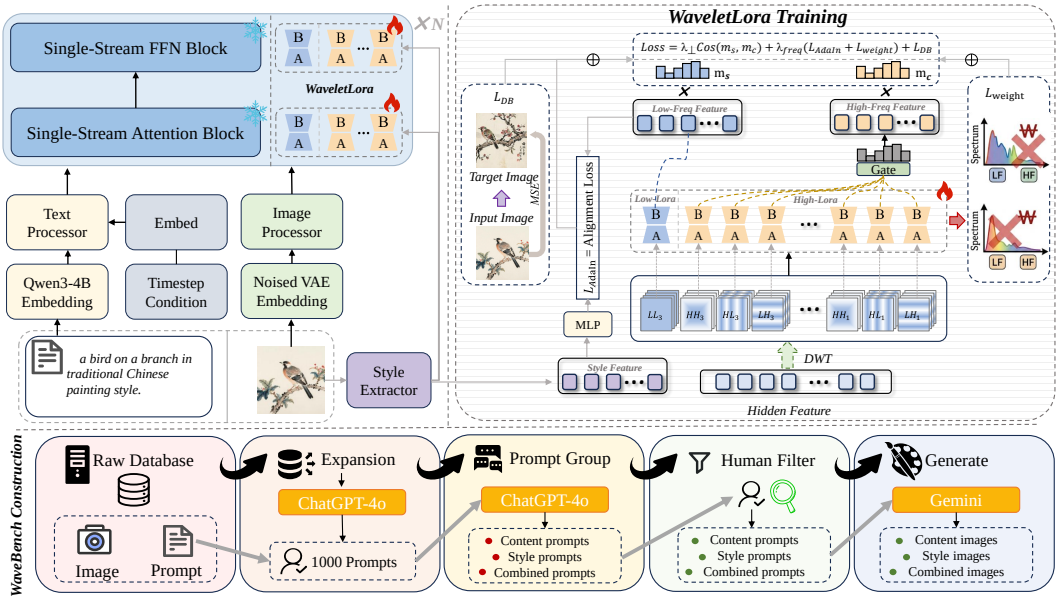

Figure 2: **Overview of WaveletLoRA and WaveBench.** (Top) Given a reference image, a two-level DWT decomposes the image into one low-frequency subband and multiple high-frequency subbands. The low-frequency branch learns global appearance attributes for style modeling, while the high-frequency branches preserve multi-scale structural details for content representation. An adaptive aggregation MoE module fuses high-frequency representations into a unified content LoRA, followed by frequency-domain disentanglement supervision and AdaIN-based style alignment. (Bottom) Construction pipeline of WaveBench, including prompt generation, manual verification, and reference image synthesis.

namely $\mathbf{F} \xrightarrow{\text{DWT}} \{\mathbf{LL}_1, \mathbf{LH}_1, \mathbf{HL}_1, \mathbf{HH}_1\}$, where $\mathbf{LL}_1$ is the low-frequency approximation, and $\mathbf{LH}_1, \mathbf{HL}_1, \mathbf{HH}_1$ capture horizontal, vertical, and diagonal high-frequency details, respectively. As illustrated in Fig. 3, we further adopt a two-level DWT to decompose the image into seven subbands, the low-frequency $\mathbf{LL}_2$ and six high-frequency components $(\mathbf{LH}_1, \mathbf{HL}_1, \mathbf{HH}_1, \mathbf{LH}_2, \mathbf{HL}_2, \mathbf{HH}_2)$. Among these subbands, the $\mathbf{LL}_2$ subband discards fine structural details but preserves global color and shading, characterizing artistic style as a supervision target for the style LoRA. In contrast, the high-frequency subbands retain multi-scale structural information, including contours, textures, and identity-related details, which are important for preserving subject structure and content consistency. In practice, we adopt a two-level decomposition, which provides a favorable balance between global appearance modeling and structural detail preservation. As validated in Sec. 4.4, shallower decomposition retains excessive structural information in the low-frequency branch, while deeper decomposition removes informative mid-frequency structures and degrades content preservation.

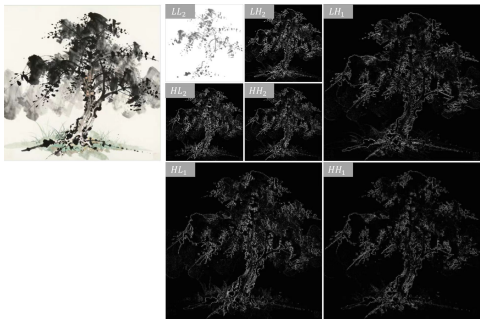

Figure 3: **Illustration of two-level DWT.** Two-level DWT decomposes an image into one low-frequency and six high-frequency subbands at two scales.

3.3 Frequency-specific LoRA Parameterization

Guided by the wavelet decomposition introduced in Sec. 3.2, we decompose the input image I via a two-level DWT into seven subbands: $\mathcal{W}(I) = \{\text{LL}_2, \text{LH}_2, \text{HL}_2, \text{HH}_2, \text{LH}_1, \text{HL}_1, \text{HH}_1\}$, where LL_2 represents the low-frequency approximation, while the remaining six capture high-frequency details

at different scales and orientations. Each subband S_k is reshaped to match the latent space dimension of the diffusion model and passed through a small learnable projector to obtain a conditioning vector x_k (of the same dimension as the text embedding). For a diffusion backbone parameterized by weight matrix $W^{(l)}$, LoRA introduces a low-rank update $W^{(l)\prime} = W_0^{(l)} + \Delta W^{(l)}$, where $\Delta W^{(l)} = B^{(l)} A^{(l)}$, and $A^{(l)} \in \mathbb{R}^{r\times d}$, $B^{(l)} \in \mathbb{R}^{d\times r}$ are low-rank matrices with rank $r \ll d$.

Instead of learning a shared LoRA representation, we assign independent LoRA branches to different wavelet subbands. Specifically, the low-frequency branch is optimized using the LL_2 subband to model global appearance attributes for style representation, while the high-frequency branches are supervised by the remaining subbands to preserve structural details for content representation. As a result, for each layer l of the DiT, we instantiate seven independent LoRA branches: $\Delta W_{\mathrm{LL}_2}^{(l)}$, $\Delta W_{\mathrm{LH}_2}^{(l)}$, $\Delta W_{\mathrm{HL}_2}^{(l)}$, $\Delta W_{\mathrm{HH}_2}^{(l)}$, $\Delta W_{\mathrm{LH}_1}^{(l)}$, $\Delta W_{\mathrm{HL}_1}^{(l)}$, $\Delta W_{\mathrm{HH}_1}^{(l)}$.

During training, we aim for each branch to specialize in reconstructing its corresponding subband. For a linear layer in the DiT backbone, the adapted forward pass is

$$h_{\text{out}} = W_0^{(l)} h_{\text{in}} + \sum_{k=1}^{7} \Delta W_k^{(l)}\, c_k, \tag{1}$$

where h_{in} is the input hidden state and the sum runs over all K wavelet subbands, with c_k denoting the frequency-specific conditioning vector derived from the k-th subband. Each update $\Delta W_k^{(l)} = B_k^{(l)} A_k^{(l)}$ follows the standard low-rank factorization. The low-frequency branch $\Delta W_{\mathrm{LL}_2}^{(l)}$ is designated as the **style LoRA**:

$$\Delta W_s^{(l)} = \Delta W_{\mathrm{LL}_2}^{(l)}. \tag{2}$$

The six high-frequency branches capture fine details and textures essential for content representation.

3.4 Dense MoE for Adaptive Frequency Aggregation

To integrate the six high-frequency LoRAs into a coherent content representation, we employ a dense MoE architecture with softmax gating. For each layer l, we define six experts $\{E_j^{(l)}\}_{j=1}^{6}$ corresponding to the high-frequency LoRAs:

$$E_j^{(l)} = \Delta W_{\mathrm{HF}_j}^{(l)}, \quad \mathrm{HF}_j \in \{\mathrm{LH}_2, \mathrm{HL}_2, \mathrm{HH}_2, \mathrm{LH}_1, \mathrm{HL}_1, \mathrm{HH}_1\}. \tag{3}$$

A gating network $G^{(l)}(\cdot)$ computes a probability distribution over experts based on the layer input hidden states h:

$$G^{(l)}(h) = \mathrm{softmax}\left(W_g^{(l)} h\right), \tag{4}$$

where $W_g^{(l)}$ is a learnable matrix. The content LoRA output is then the weighted sum of expert outputs:

$$\Delta W_c^{(l)}(h) = \sum_{j=1}^{6} G^{(l)}(h)_j \cdot E_j^{(l)}(h). \tag{5}$$

The gating networks are trained jointly with all LoRA branches, allowing the model to dynamically weight different frequency experts depending on the content characteristics. This formulation enables, for example, emphasizing horizontal edges via LH experts or textures via HH experts.

3.5 Training Objective

Our full training objective combines a reconstruction loss, an orthogonality prior, and a composite frequency-domain regularization that jointly enforces weight-level frequency separation and global style-statistics alignment.

Frequency-domain regularization loss. To explicitly steer the style and content branches toward low- and high-frequency specialization, we introduce a composite regularization $\mathcal{L}_{\text{freq}}$ consisting of two complementary terms.

Weight-level frequency separation. For each layer l, let $\Delta W_s^{(l)}$ and $\Delta W_c^{(l)}$ denote the style and content LoRA updates. Applying a 2D DWT to these weight matrices, we denote by $\mathcal{H}(\cdot)$ the extraction of

all high-frequency subbands (LH, HL, HH) and by $\mathcal{L}(\cdot)$ the low-frequency approximation (LL). We penalize high-frequency energy in the style weights and low-frequency energy in the content weights:

$$\mathcal{L}_{\text{weight}} = \frac{1}{N}\sum_{l=1}^{N}\Big(\|\mathcal{H}(\Delta W_s^{(l)})\|_F^2 + \|\mathcal{L}(\Delta W_c^{(l)})\|_F^2\Big), \tag{6}$$

with $\|\cdot\|_F$ the Frobenius norm. This loss encourages the style LoRA to be smooth and the content LoRA to retain fine detail.

Style-statistics alignment. To reinforce style consistency, we introduce an AdaIN [12]-based alignment mechanism. We employ a pretrained encoder to extract style features $\mathbf{F}_{\text{ref}}$ from the reference image and align them with the projected LL2 subband features $\mathbf{F}_{\text{LL2}}$ in the latent space. The AdaIN alignment loss is then the mean squared error between the corresponding statistics:

$$\mathcal{L}_{\text{AdaIN}} = \frac{1}{C}\sum_{c=1}^{C}\Big((\mu_{\text{LL}_2,c} - \mu_{\text{style},c})^2 + (\sigma_{\text{LL}_2,c} - \sigma_{\text{style},c})^2\Big), \tag{7}$$

where C is the number of feature channels, and $\mu_{\text{LL}_2,c}$, $\sigma_{\text{LL}_2,c}$ (resp. $\mu_{\text{style},c}$, $\sigma_{\text{style},c}$) denote the per-channel mean and standard deviation of the projected LL_2 features (resp. the style features extracted from the reference image). This term directly enforces that the global color palettes, texture granularity, and contrast of the generated style match those of the reference.

The full frequency-domain regularization is then

$$\mathcal{L}_{\text{freq}} = \mathcal{L}_{\text{weight}} + \lambda_{\text{AdaIN}}\,\mathcal{L}_{\text{AdaIN}}. \tag{8}$$

Orthogonality regularization. We promote disentanglement by encouraging column-wise orthogonality between the content and style LoRA masks:

$$\mathcal{L}_{\perp} = \sum\nolimits_l |m_c^{(l)} \cdot m_s^{(l)}|, \tag{9}$$

where $m_c^{(l)}$ and $m_s^{(l)}$ are binary column masks for content and style LoRAs at layer l.

Overall objective. The complete training loss combines a frequency-domain regularization term, a diffusion denoising loss $\mathcal{L}_{\text{DB}}$ (the standard DreamBooth reconstruction objective [19]), and an orthogonality regularization loss:

$$\mathcal{L} = \lambda_{\text{freq}}\mathcal{L}_{\text{freq}} + \lambda_{\perp}\mathcal{L}_{\perp} + \mathcal{L}_{\text{DB}}, \tag{10}$$

with hyperparameters λ_{freq}, λ_{AdaIN} and $\lambda_{\perp}$ controlling the orthogonality, frequency separation, and style-statistics alignment strengths, respectively. All LoRA branches, gating networks and subband projectors are optimized jointly.

After training, the style LoRA $L_s = \{\Delta W_{\text{LL}_2}^{(l)}\}$ and the content LoRA $L_c = \{\sum_{j=1}^{6} G^{(l)}(\cdot)_j \cdot \Delta W_{\text{HF}_j}^{(l)}\}$ are extracted. They can be applied individually for style or subject transfer, or composed via direct addition $L_s \oplus L_c$ for recontextualization without further fine-tuning.

4 Experiments

4.1 Experimental Settings

Experimental setup. We use Z-Image [2] as our base model. All LoRA branches are trained with rank $r = 64$ using the Adam optimizer with learning rate 5×10^{-5} for 600 steps and batch size 1. The base model weights and text encoders remain frozen. For the two-level DWT, we use the Haar wavelet. Each subband is reshaped to 64×64 and projected via a small learnable MLP to match the cross-attention dimension. For column separation, we set the column cap $N = 30\%$, recalibration interval $t = 200$, and orthogonality loss weight $\lambda_{\perp} = 0.02$. The style-statistic alignment loss weight is set to $\lambda_{\text{AdaIn}} = 0.2$. During training, we use the descriptive prompts provided by WaveBench to reconstruct the reference image, whereas at test time we adopt the content-only and style-only prompts provided by WaveBench for decomposition evaluation. The frequency-domain loss weight is set to $\lambda_{\text{freq}} = 0.01$. All experiments are conducted on a single NVIDIA A800 GPU.

Evaluation metrics. Following prior work, we measure content alignment with DINOv2 [10] features and style alignment with CLIP-I [17] embeddings. For each image in our WaveBench dataset (1,000 examples), we generate 4 samples per method and compute the average cosine similarity between the generated outputs and the reference image.

4.2 WaveBench

To rigorously evaluate single-image decomposition, we introduce WaveBench, a large-scale benchmark which comprises 1,000 diverse examples spanning a wide range of subjects (animals, objects, human figures, scenes, etc.) and styles (painting, cartoon, art, futuristic, etc.), as shown in Fig. 4. Each entry includes a reference image paired with a suite of auxiliary prompts designed to test content/style transfer and fusion, offering superior scale and diversity over existing datasets. Beyond traditional feature-space metrics, we propose a VLM-based evaluation protocol using Qwen2.5-VL-72B to assess semantic disentanglement. The VLM rates two key dimensions on a $[0, 1]$ scale, namely 1) separation, which measures the removal of complementary attributes (e.g., style residue in content-only tasks), and 2) similarity, which reflects fidelity to the intended attribute. The final Content and Style Scores are computed as a weighted average of separation score and similarity score, providing a human-aligned assessment of disentanglement quality across the benchmark. Details of WaveBench and VLM-based evaluation are provided in the Appendix.

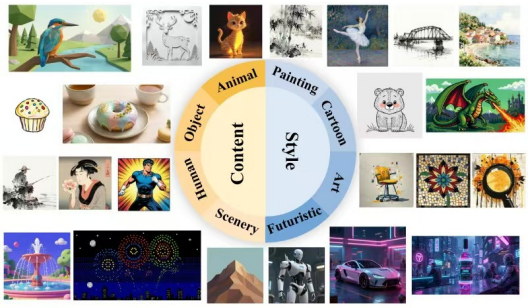

Figure 4: **Diversity of WaveBench.**

4.3 Frequency Retention across Artistic Styles

The frequency assignment in WaveletLoRA is a task-level inductive bias rather than a strict semantic partition. Low-frequency components tend to emphasize global appearance, while high-frequency components retain structural details. To test this tendency across diverse styles, we selected eight commonly occurring style categories from WaveBench (cartoon, Chinese painting, futuristic, oil painting, pixel art, pointillism, sketch, and water color), randomly sampled 20 images from each (160 total). For each image, we applied a two-level DWT and reconstructed an LL_2-only variant by zeroing the high-frequency subbands, as well as a high-only variant by zeroing LL_2. We then used Qwen2.5-VL-72B, a third-party VLM, to assess style and content contributions via four-choice questions. For each images, the VLM selected between one correct and three distractor options for both artistic style and subject content. Four accuracy rates were calculated: style and subject accuracy for LL-only images, and style and subject accuracy for High-only images. The results are presented in Table 1.

Table 1: VLM-based recognition accuracy on frequency-retained reconstructions. LL_2 retains more style information, while high-frequency bands retain more subject information across all 8 categories.

Style category	LL_2-only style	LL_2-only subject	High-only style	High-only subject
Cartoon	0.85	0.30	0.15	0.85
Chinese painting	0.80	0.25	0.10	0.80
Futuristic	0.90	0.40	0.25	0.85
Oil painting	0.70	0.30	0.05	0.65
Pixel art	0.85	0.45	0.20	0.70
Pointillism	0.75	0.30	0.05	0.65
Sketch	0.85	0.50	0.20	0.65
Watercolor	0.70	0.35	0.05	0.75
Average	0.80	0.36	0.13	0.74

LL_2-only reconstructions achieve 80% style-recognition accuracy, compared with 13% for high-only reconstructions. In the reverse direction, high-only reconstructions achieve 74% subject-recognition

accuracy, compared with 36% for LL_2-only images. The same complementary tendency appears in each category. These results support our frequency mapping as a dominant task-level tendency.

4.4 Experimental Results

Table 2: **Quantitative comparison with state-of-the-art methods.** We report feature space alignment metrics (DINOv2 for content consistency and CLIP-I for style consistency), VLM evaluation scores (content-decom., style-decom.), and user study scores (content, style, combined). Best results are highlighted in bold and second-best results are underlined.

Method	WaveBench		VLM Evaluation		User Study Scores		
	Content-align.	Style-align.	Content-decom.	Style-decom.	Content.	Style.	Combined.
Qwen-Image-Edit	-	-	0.6404	0.5805	0.5925	0.4025	**0.8575**
OmniGen2	-	-	0.5333	0.4252	0.3325	0.2425	0.7425
B-LoRA	0.5486	0.6115	0.7541	0.8396	0.5725	0.6250	0.6250
UnZipLoRA	0.5529	<u>0.6286</u>	<u>0.8373</u>	0.8693	<u>0.7250</u>	0.7800	0.7075
USO	**0.8031**	0.5157	0.6503	<u>0.8707</u>	0.3900	<u>0.8325</u>	0.5825
WaveletLoRA	<u>0.5866</u>	**0.6330**	**0.8563**	**0.8730**	**0.7925**	**0.8400**	<u>0.7750</u>

Quantitative results. As shown in Table 2, we compare WaveletLoRA against five representative and competitive methods, including two general-purpose instruction-tuned text-to-image models, Qwen-Image-Edit [30] and OmniGen2 [31], and three dedicated content–style decomposition approaches, B-LoRA [5], UnZipLoRA [14], and USO [32]. For fair comparison, we use the same prompt templates for all methods. Specifically, content generations are guided by the content-only prompts from WaveBench, and style generations are guided by the style-only prompts. The two general-purpose models are evaluated via the VLM protocol and user study, while feature-space metrics (DINOv2 content alignment and CLIP-I style alignment) are reported for the dedicated methods.

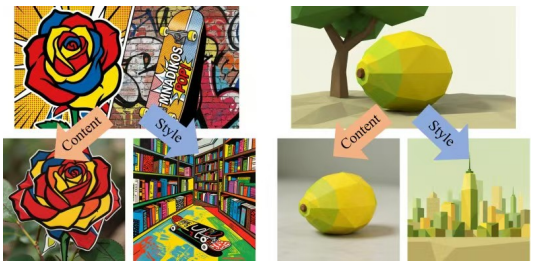

Figure 5: **Style leakage in USO.** USO's content-decoupled outputs often retain strong style artifacts.

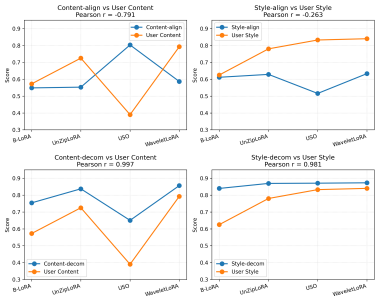

Figure 6: **Correlation between VLM content-decomposition scores and user study.**

Table 2 summarizes the quantitative results. Under feature space evaluation, USO achieves the highest DINOv2 content score (0.8031) but exhibits relatively weak style consistency (0.5157), indicating that high feature similarity does not necessarily correspond to clean disentanglement. In contrast, WaveletLoRA achieves the best CLIP-I style consistency (0.6330) together with competitive content consistency (0.5866), demonstrating a more balanced separation between structural content and visual appearance. Under VLM-based evaluation, WaveletLoRA achieves the best performance on both content preservation (0.8563) and style preservation (0.8730), outperforming all competing methods. This confirms that wavelet supervision enables superior disentanglement with minimal leakage. Overall, WaveletLoRA achieves the best combined result (0.7750) among specialized methods, significantly surpassing USO and generic editing baselines.

We further conducted a user study to evaluate the perceptual quality of the decoupled content and style representations produced by different methods. Specifically, we randomly sampled 200 images from the test set and split them into 10 non-overlapping groups of 20 images each without replacement. Each group was assigned to a different participant (10 participants in total). For each image, participants were asked to rate the decoupling results of six baseline methods along three

dimensions: *content*, *style*, and *combined*. As shown in Table 2, all ratings utilize a 5-point Likert scale and were normalized after data collection to ensure a fair comparison.

VLM-based evaluation. VLM-based evaluation aligns more closely with human judgment than conventional feature-space metrics. While DINOv2 and CLIP-I measure holistic similarity, they fail to explicitly disentangle content fidelity from style residue. This limitation is exemplified by USO, whose generations often retain significant stylistic traces (Fig. 5). These residues misleadingly inflate its DINOv2 content score to a top-tier 0.8031. In contrast, our VLM protocol directly assesses the purity of attribute separation and remains robust against such style leakage. The alignment between USO's VLM scores and our user study further validates the reliability of this metric. The alignment between VLM scores and human preferences is corroborated by the correlation curves in Fig. 6, where VLM-based metrics track human ratings more closely than DINOv2 across both content and style dimensions. These observations, supported by higher Pearson correlation coefficients, validate that our VLM protocol provides a more faithful and semantically meaningful evaluation of disentanglement than raw feature-space similarity.

Analysis of dense MoE. To verify that the MoE module adaptively utilizes frequency information, we visualize its gating weights across diverse samples (Fig. 7). The distribution varies markedly by content type: for the pencil-sketch wolf, the MoE prioritizes LH_1 and HL_1 experts to capture fine-grained contours. In contrast, the pixel-art cactus triggers stronger HH_1 activation, targeting its blocky diagonal structures. For the watercolor boat and artwork camera, the weighting shifts to second-level subbands (LH_2, HL_2, HH_2) to encode mid-level textures rather than sharp edges. This sample-dependent variation confirms that our MoE avoids representation collapse, learning to selectively fuse multi-scale structural cues while preventing low-frequency style leakage.

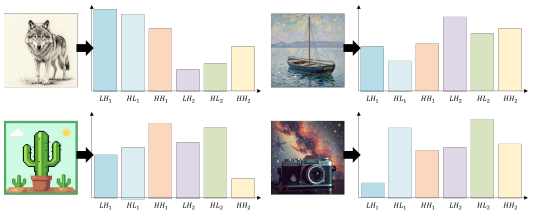

Figure 7: **Visualization of dense MoE gating weights.**



Figure 8: **Qualitative comparison.** WaveletLoRA produces cleaner content–style separation than other methods. Content outputs (top rows) retain subject identity without style residues, and style outputs (bottom rows) capture artistic textures without subject leakage.

Qualitative results. As shown in Fig. 8, existing methods often leave style residues in content outputs or lose subject identity. For example, in the first row, UnZipLoRA, USO and other methods still retain obvious watercolor textures and stylized color patterns in the generated woman image,

leading to unrealistic facial appearance and blurred structural details. In contrast, WaveletLoRA reconstructs a more realistic face with cleaner contours and more faithful identity-related structures while effectively removing the original artistic style. Similar observations can also be found in the second row. Moreover, the style generation results further demonstrate the effectiveness of our frequency-aware disentanglement. These results demonstrate that wavelet-domain supervision enables cleaner disentanglement between structural content and visual style.

4.5 Ablation Study

We evaluate the contribution of each component by assessing content and style alignment on WaveBench, utilizing DINOv2 and CLIP-I respectively.

Decomposition depth. Replacing the two-level DWT with a single-level transform raises content alignment to 0.5904 but severely degrades style alignment to 0.6112. The drop stems from residual structural contours in the LL_1 subband, which contaminate the style LoRA and produce an inflated content score through style leakage. A three-level DWT excessively filters mid-frequency details, lowering content alignment to 0.5788 without style gains. Our two-level design thus strikes the optimal balance.

High-frequency fusion. Equal averaging causes the largest content drop (0.5752), as it ignores spatial and temporal relevance. Learnable per-band scalars improve content to 0.5769, but remain globally fixed. Cross-attention reaches 0.5831 by dynamically conditioning on layer features, yet it incurs extra projections and lacks expert-specialised gating. Our dense MoE achieves the best content alignment (0.5866) while preserving style, confirming the advantage of input-dependent, expert-wise weighting.

Table 3: **Ablation on decomposition depth and fusion strategy.**

Configuration	Content-align.	Style-align.
WaveletLoRA	**0.5866**	**0.6330**
Decomposition Depth		
1-level DWT	0.5904	0.6112
3-level DWT	0.5788	0.6317
High-frequency Fusion		
Equal averaging	0.5752	0.6329
Learnable scalars α_j	0.5769	0.6324
Cross-attention	0.5831	0.6326
Style Feature Extractor		
w/o AdaIN	0.5863	0.6321

Style regularization. Removing the AdaIN loss marginally reduces style alignment (0.6321), validating its role as a lightweight style refinement. In particular, for styles whose defining characteristics may be concentrated in high-frequency components, our AdaIN-based style alignment provides complementary supervision for global style statistics that may not be fully captured by the LL_2-based objective. To validate this, we conducted an additional ablation on the pointillism subset of WaveBench using Qwen2.5-VL-72B style scores. As shown in Table 4, removing AdaIN lowers the style score from 0.703 to 0.655, a drop of 0.048. This result confirms that AdaIN contributes meaningfully to style fidelity on this highly textured subset.

Table 4: **AdaIN ablation on the pointillism subset of WaveBench.**

Full model	w/o AdaIN	Drop
0.703	0.655	0.048

Moreover, we further analyze the sensitivity of the framework to the two key loss weights, λ_{freq} and λ_{AdaIN}, as reported in appendix.

5 Conclusion

We present WaveletLoRA, a frequency-aware framework for content–style disentanglement in personalized image generation. By introducing wavelet-domain supervision, our method explicitly separates low-frequency appearance information from high-frequency structural details through frequency-specific LoRA branches. We further establish WaveBench and a VLM-based evaluation protocol for disentanglement assessment. Extensive experiments demonstrate that WaveletLoRA achieves cleaner content-style separation and more realistic generation results than existing methods.

Acknowledgments and Disclosure of Funding

This research is supported by Artificial Intelligence-National Science and Technology Major Project (2023ZD0121200), the National Natural Science Foundation of China (62531026, 62437001, 62436001), the Natural Science Foundation of Jiangsu Province under Grant BK20243051, and the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDB1350103.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[2] Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

[3] Google DeepMind. Gemini 3 pro: Advancing multimodal intelligence. `https://deepmind.google/gemini/`, 2025.

[4] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7890–7899, 2020.

[5] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora. In *European Conference on Computer Vision*, pages 181–198. Springer, 2024.

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[7] Xiang Gao, Zhengbo Xu, Junhan Zhao, and Jiaying Liu. Frequency-controlled diffusion model for versatile text-guided image-to-image translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1824–1832, 2024.

[8] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[9] Arthita Ghosh and Rama Chellappa. Deep feature extraction in the dct domain. In *2016 23rd International conference on pattern recognition (ICPR)*, pages 3536–3541. IEEE, 2016.

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[12] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[13] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24174–24184, 2024.

[14] Chang Liu, Viraj Shah, Aiyu Cui, and Svetlana Lazebnik. Unziplora: Separating content and style from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16776–16785, 2025.

[15] Quang-Binh Nguyen, Minh Luu, Quang Nguyen, Anh Tran, and Khoi Nguyen. Csd-var: Content-style decomposition in visual autoregressive models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17013–17023, 2025.

[16] Hao Phung, Quan Dao, and Anh Tran. Wavelet diffusion models are fast and scalable image generators. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10199–10208, 2023.

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

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[19] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

[20] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. In *European Conference on Computer Vision*, pages 422–438. Springer, 2024.

[21] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4733–4743, 2024.

[22] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023.

[23] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[24] Xinyang Song, Libin Wang, Weining Wang, Zhiwei Li, Jianxin Sun, Dandan Zheng, Jingdong Chen, Qi Li, and Zhenan Sun. 3sgen: Unified subject, style, and structure-driven image generation with adaptive task-specific memory. *arXiv preprint arXiv:2512.19271*, 2025.

[25] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

[26] Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. Concept decomposition for visual exploration and inspiration. *ACM Transactions on Graphics (TOG)*, 42(6):1–13, 2023.

[27] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[28] Xingjian Wang, Li Chai, and Jiming Chen. Frequency-domain refinement with multiscale diffusion for super resolution. *arXiv preprint arXiv:2405.10014*, 2024.

[29] Zhe Wang, Ziqiu Chi, Yanbing Zhang, et al. Fregan: Exploiting frequency components for training gans under limited data. *Advances in Neural Information Processing Systems*, 35:33387–33399, 2022.

[30] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

[31] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025.

[32] Shaojin Wu, Mengqi Huang, Yufeng Cheng, Wenxu Wu, Jiahe Tian, Yiming Luo, Fei Ding, and Qian He. Uso: Unified style and subject-driven generation via disentangled and reward learning. *arXiv preprint arXiv:2508.18966*, 2025.

[33] Wenbin Xie, Dehua Song, Chang Xu, Chunjing Xu, Hui Zhang, and Yunhe Wang. Learning frequency-aware dynamic network for efficient super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4308–4317, 2021.

[34] Zhi Xu, Shaozhe Hao, and Kai Han. Cusconcept: Customized visual concept decomposition with diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3678–3687. IEEE, 2025.

[35] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyang Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *Proceedings of the computer vision and pattern recognition conference*, pages 12978–12988, 2025.

[36] Ruicheng Zhang, Kanghui Tian, Zeyu Zhang, Qixiang Liu, and Zhi Jin. Fdg-diff: Frequency-domain-guided diffusion framework for compressed hazy image restoration, 2025.

A WaveBench

WaveBench Construction. To rigorously benchmark single-image content–style decomposition, we construct WaveBench, a large-scale dataset of 1,000 diverse and carefully curated examples. As illustrated in Fig. 2, the construction follows a multi-stage pipeline:

- **Raw Database & Expansion:** The initial prompt pool is aggregated from CSD-100 [15] (100 prompts) and 3SGen [24] (135 prompts). To ensure scale and diversity, we employ **ChatGPT-4o** to expand the collection with 765 additional unique prompts, reaching a total of 1,000 candidate entries.
- **Prompt Grouping:** For each original prompt P_i^{orig}, ChatGPT-4o acts as a semantic engine to generate a structured prompt group. This includes: (1) two **content-only** prompts $\{P_{i,j}^{\text{cont}}\}_{j=1}^{2}$ that port the original subject into new contexts; (2) two **style-only** prompts $\{P_{i,k}^{\text{sty}}\}_{k=1}^{2}$ that apply the style to disparate subjects; and (3) two **combined** prompts $\{P_{i,\ell}^{\text{comb}}\}_{\ell=1}^{2}$ for holistic reconstruction.
- **Human Filter:** Every prompt group undergoes rigorous manual inspection. Human annotators verify semantic correctness, eliminate duplicates, and ensure a broad coverage of artistic styles and complex subjects, reducing potential evaluation noise.
- **Reference Generation:** Finally, the verified P_i^{orig} is fed into **Gemini-3 Pro** [3] to synthesize high-fidelity reference images I_i (1024×1024).

Each WaveBench entry is formally defined as an **8-tuple**:

$$\mathcal{D}_i = \big(P_i^{\text{orig}}, I_i, \{P_{i,j}^{\text{cont}}\}_{j=1}^{2}, \{P_{i,k}^{\text{sty}}\}_{k=1}^{2}, \{P_{i,\ell}^{\text{comb}}\}_{\ell=1}^{2}\big). \tag{11}$$

WaveBench significantly advances previous content-style decomposition benchmarks in scale, annotation quality, and diversity. As visualized in Figure. 4, with 1,000 curated examples, it is significantly larger than existing datasets, providing a more reliable foundation for statistical comparison. Manual screening at every stage guarantees clean, unambiguous annotations, reducing evaluation noise and providing a faithful testbed for disentanglement methods.

WaveBench Dataset Statistics WaveBench contains 1,000 examples. Table 5 summarizes the top-level style and content labels.

Table 5: Top-level style and content category distributions in WaveBench.

Style category	Count
Painting	426
Cartoon	97
Futuristic	171
Art	128
Other	178
Total	1,000

Content category	Count
Animals	195
Human figures	157
Landscapes	202
Objects	222
Other	224
Total	1,000

VLM-based evaluation. Leveraging WaveBench, we design an automatic evaluation protocol powered by Qwen2.5-VL-72B [1], a large vision-language model, to assess disentanglement in a human-aligned, semantic manner.

For a given method and each sample $\mathcal{D}_i$, we generate two content-decoupled images: one for each $P_{i,1}^{\text{cont}}$ and $P_{i,2}^{\text{cont}}$, denoted $\{I_{i,1}^{\text{cont}}, I_{i,2}^{\text{cont}}\}$. Similarly, we obtain two style-decoupled images $\{I_{i,1}^{\text{sty}}, I_{i,2}^{\text{sty}}\}$ from the two style-only prompts. The VLM is queried with the reference I_i paired with each generated image to rate two aspects on a $[0, 1]$ scale:

- **Separation**: how well the complementary attribute has been removed (style residue for a content-decoupled image, subject residue for a style-decoupled image);
- **Similarity**: how similar the two images are in terms of the intended attribute.

For a content-decoupled image I^{cont}, let $s_{\text{sep}}(I_i, I^{\text{cont}})$ be the style-separation score and $s_{\text{subj}}(I_i, I^{\text{cont}})$ the subject-similarity score. The per-image content score is

$$s_{\text{cont}}(I_i, I^{\text{cont}}) = 0.6 \cdot s_{\text{sep}}(I_i, I^{\text{cont}}) + 0.4 \cdot s_{\text{subj}}(I_i, I^{\text{cont}}). \tag{12}$$

The sample-level content score is the average over the two generated content images:

$$s_{\text{cont}}(\mathcal{D}_i) = \frac{1}{2} \sum_{m=1}^{2} s_{\text{cont}}(I_i, I_{i,m}^{\text{cont}}). \tag{13}$$

Analogously, for a style-decoupled image I^{sty} we obtain a subject-separation score s'_{sep} and a style-similarity score s'_{sty}. We define the per-image style score as

$$s_{\text{sty}}(I_i, I^{\text{sty}}) = 0.6 \cdot s'_{\text{sep}}(I_i, I^{\text{sty}}) + 0.4 \cdot s'_{\text{sty}}(I_i, I^{\text{sty}}), \tag{14}$$

and average across the two style generations:

$$s_{\text{sty}}(\mathcal{D}_i) = \frac{1}{2} \sum_{n=1}^{2} s_{\text{sty}}(I_i, I_{i,n}^{\text{sty}}). \tag{15}$$

The final **Content Score** and **Style Score** reported for each method are the averages of $s_{\text{cont}}(\mathcal{D}_i)$ and $s_{\text{sty}}(\mathcal{D}_i)$ over all 1,000 WaveBench samples, respectively. This VLM-based metric directly evaluates semantic disentanglement, complementing traditional feature-space measures.

B Additional Evaluation Analyses

Cross-VLM Evaluation To test sensitivity to the choice of evaluator, we evaluated all methods on the same samples with the same protocol using Gemini-3.5-Flash and GPT-5.1. WaveletLoRA obtains the highest score on both dimensions under all three VLMs (Table 6), indicating that the main comparison is not specific to one evaluator.

Training and Inference Cost Table 7 reports the measured training time for 600 optimization steps, inference time per generated image, and additional trainable parameters. WaveletLoRA incurs additional computation from its frequency-specific branches and AdaIN component.

Table 6: Evaluation scores from three VLMs under the same protocol.

Method	Qwen2.5-VL-72B		Gemini-3.5-Flash		GPT-5.1	
	Content	Style	Content	Style	Content	Style
Qwen-Image-Edit	0.6404	0.5805	0.6103	0.5724	0.6152	0.5920
OmniGen2	0.5333	0.4252	0.5169	0.3985	0.5087	0.4003
B-LoRA	0.7541	0.8396	0.7346	0.8450	0.7326	0.8433
UnZipLoRA	0.8373	0.8693	0.8113	0.8584	0.7945	0.8608
USO	0.6503	0.8707	0.5752	0.8601	0.6032	0.8607
WaveletLoRA	0.8563	0.8730	0.8379	0.8667	0.7995	0.8642

Table 7: Computational cost comparison.

Component	Training time	Inference time	Additional parameters
Base Z-Image (without LoRA)	–	∼5 s	0
Base + style and content LoRA	∼230 min	∼6 s	$2 \times (2 \times 64 \times d)$
WaveletLoRA (full)	∼320 min	∼10 s	$7 \times (2 \times 64 \times d) + 6 \times d + \text{AdaIN}$

Parameter Sensitivity Analysis We analyze the sensitivity of the framework to the two key loss weights, λ_{freq} and λ_{AdaIN}, as reported in Table 8 and Table 9. For the frequency disentanglement loss, $\lambda_{\text{freq}} = 0.01$ attains the highest content alignment (0.5866) and the peak style alignment (0.6330). Smaller weights weaken the regularization that penalizes high-frequency energy in the style LoRA, whereas larger weights overly suppress the style branch's capacity, degrading content alignment without further style gains. For the AdaIN loss, $\lambda_{\text{AdaIN}} = 0.2$ yields the optimal style alignment (0.6330) while preserving content alignment (0.5866). Excessively large λ_{AdaIN} (e.g., 0.6) forces the style branch to mimic the reference statistics too aggressively, leading to a decline in style alignment (0.6317) and no content advantage. Consequently, we adopt $\lambda_{\text{freq}} = 0.01$ and $\lambda_{\text{AdaIN}} = 0.2$ throughout our experiments, which strikes a favorable balance between content preservation and style fidelity.

C More Qualitative Results

As shown in Figure 9, we extend our qualitative evaluation to a broader suite of content–style combinations beyond those presented in the main text. Across this spectrum, WaveletLoRA exhibits consistent and robust decomposition behavior. Content-decoupled generations faithfully retain the original subject's identity, pose, and fine structural details, while successfully discarding the color palettes, textures, and brushwork of the reference style. Style-decoupled outputs capture the defining artistic characteristics—stroke patterns, shading conventions, and palette distributions—and transfer them cleanly to new textual prompts without carrying over silhouettes or object boundaries from the source image. Baselines such as USO and UnZipLoRA, by contrast, show occasional style bleeding in content outputs or subject residuals in style outputs, particularly under strong stylization (e.g., heavy brush textures or low-contrast line drawings). These qualitative observations align with the

Table 8: Sensitivity Analysis of λ_{freq}

λ_{freq}	Content-align.	Style-align.
0.002	0.5860	0.6298
0.006	0.5863	0.6304
0.01	0.5866	0.6330
0.02	0.5862	0.6324
0.04	0.5862	0.6317

Table 9: Sensitivity Analysis of λ_{AdaIn}

λ_{AdaIn}	Content-align.	Style-align.
0.05	0.5865	0.6325
0.1	0.5866	0.6327
0.2	0.5866	0.6330
0.3	0.5865	0.6320
0.4	0.5867	0.6326
0.5	0.5866	0.6322
0.6	0.5865	0.6317

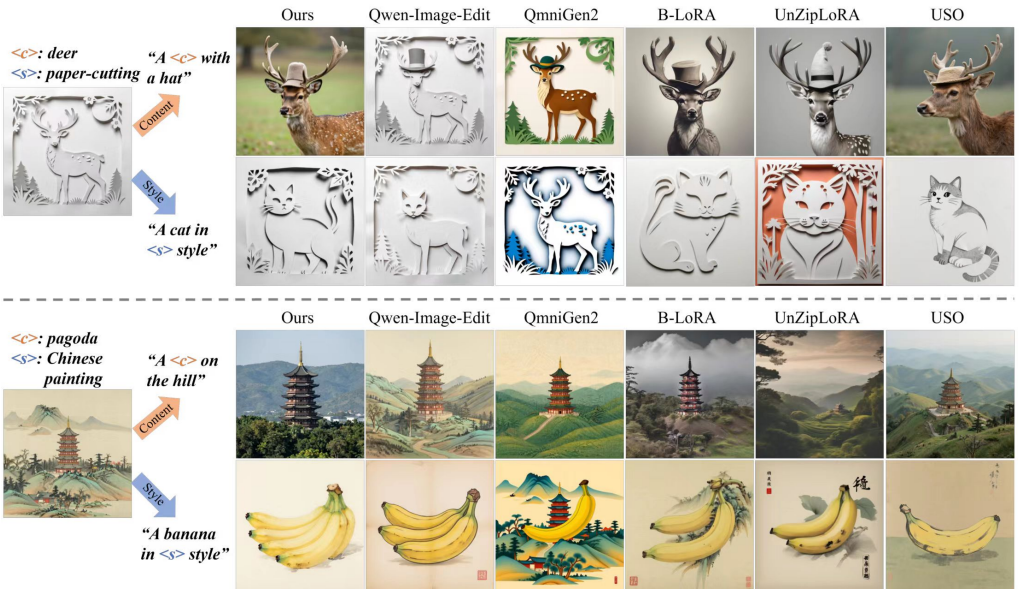

Figure 9: **Additional qualitative comparisons across diverse content–style pairs.** We show decomposition results on other set of subjects and artistic styles. For each example, the top row displays content-decoupled outputs and the bottom row shows style-decoupled outputs. WaveletLoRA consistently preserves subject identity without ghosting from the original style and extracts coherent artistic textures without silhouette leakage from the source content.

quantitative metrics reported in the main paper and further confirm that the frequency-supervised design of WaveletLoRA enables more thorough and generalizable content–style separation across a wide range of visual domains.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Abstract and Section 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 4.4.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: See Section 3.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Section 4 and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: See Supplementary Material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 4 and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: See all sections of the paper.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Section 1 and Section 2.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See all sections of the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: See Section 4 and Supplementary Material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: See Appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.