RegionRoute: Regional Style Transfer with Diffusion Model

Bowen Chen[1] Jake Zuena[2] Alan C. Bovik[3] Divya Kothandaraman[2]
[1]The University of Texas at Austin [2]Dolby Laboratories, Inc. [3]University of Colorado Boulder
{bwchen}@utexas.edu,{Alan.Bovik}@colorado.edu,{Jake.Zuena, Divya.Kothandaraman}@dolby.com

Figure 1. When provided with region-specific editing instructions, our **RegionRoute** framework more precisely interprets localized style modification prompts and produces visually coherent results. Given prompts such as "Make the man in pixel-art style and keep other areas unchanged," the baseline image editing model tends to either apply the style globally or distort unrelated regions. Each column shows, from top to bottom: the input context image, the baseline, Flux.1-Kontext [29] output, and our RegionRoute output.

Abstract

Precise spatial control in diffusion-based style transfer remains challenging. This challenge arises because diffusion models treat style as a global feature and lack explicit spatial grounding of style representations, making it difficult to restrict style application to specific objects or regions. To our knowledge, existing diffusion models are unable to perform true localized style transfer, typically relying on handcrafted masks or multi-stage post-processing that introduce boundary artifacts and limit generalization. To address this, we propose an attention-supervised diffusion framework that explicitly teaches the model where to apply a given style by aligning the attention scores of style tokens with object masks during training. Two complementary objectives, a Focus loss based on KL divergence and a Cover loss using binary cross-entropy, jointly encourage accurate localization and dense coverage. A modular LoRA-MoE design further enables efficient and scalable multi-style adaptation. To evaluate localized stylization, we introduce the Regional Style Editing Score, which measures Regional Style Matching through CLIP-based similarity within the target region and Identity Preservation via masked LPIPS and pixel-level consistency on unedited areas. Experiments show that our method achieves mask-free, single-object style transfer at inference, producing regionally accurate and visually coherent results that outperform existing diffusion-based editing approaches.

1. Introduction

Recent advances in diffusion-based generative models have shown remarkable capabilities in producing high-quality and diverse visual content across various domains [4, 12, 13, 22, 36, 39, 41–43, 45]. Models from the Stable Diffusion [13, 39, 42] and Flux [4, 29] series achieve strong perceptual realism and semantic consistency, supporting a wide range of tasks such as image generation, editing, and style transfer. Building upon these foundations, some diffusion-based editing approaches [5, 6, 19, 47] have further enabled controllable modification of visual attributes guided by text or reference images. Similarly, diffusion-driven style transfer methods [9, 28, 30, 44, 46, 48] demonstrate strong ability in transferring artistic styles and visual aesthetics.

Despite these advances, precise spatial control over where a style is applied remains an open challenge. Existing style transfer models typically treat style as a global feature [9, 20, 23, 26–28, 30, 44, 46, 48], applying it uniformly across the entire image without considering spatial boundaries. As a result, they are unable to modify specific objects or regions. Consequently, the only existing way to achieve localized style effects is through a two-stage pipeline: first performing a global style transfer on the entire image, and then using handcrafted masks to splice the stylized regions with the original content. While this strategy can roughly localize style effects, it introduces several limitations, such as the need for precise mask preparation and visible seams at boundaries. These limitations hinder generalization and restrict the practicality of existing methods.

In principle, diffusion models inherently learn attention maps that capture spatial correspondences between textual concepts and image regions [7, 19, 35, 52, 53]. However, such attentions are not explicitly guided to associate style concepts with specific objects. Therefore, even though the model "sees" the correct regions, it often fails to apply the style precisely, leading to global style shifts.

To address these limitations, we propose a novel attention-supervised diffusion framework that explicitly teaches the model where to apply a style by binding the attention maps of style tokens to the binary masks of target objects during training. Specifically, we fine-tune the pretrained Flux.1-Kontext [29] backbone using a parameter-efficient LoRA-MoE (Mixture-of-Experts LoRA) [51] strategy, and enforce KL-divergence and binary cross-entropy losses to match attention distributions with corresponding object masks. This supervision directly establishes a correspondence between style tokens and semantic regions, enabling the model to internalize localized style grounding. As a result, it achieves mask-free, single-object style editing at inference without requiring explicit segmentation or external spatial controls. Some results are shown in Figure 1.

Moreover, existing evaluation protocols primarily measure global style similarity or overall perceptual quality, which do not reflect how well a model performs localized style transfer. To fill this gap, we design a new evaluation metric that quantifies localized style fidelity and unedited region preservation, providing a more comprehensive and objective assessment of spatially controllable style transfer.

In summary, our contributions are:

- We propose an attention-guided training paradigm that explicitly aligns style token attentions with object masks, enabling precise and mask-free localized style transfer.
- We introduce a LoRA-MoE strategy for parameter-efficient fine-tuning, allowing multiple experts to specialize in diverse styles while keeping the model lightweight and stable.
- We design a new evaluation metric to quantitatively measure localized style fidelity and unedited region preservation, filling a key gap in current evaluation standards.

2. Related Work

Diffusion-based Image Editing. Diffusion-based image editing has rapidly advanced by leveraging the strong generative priors of pretrained diffusion models. Early approaches such as SDEdit [33], Blended Diffusion [1], and Stable Diffusion Inpainting [42] perform guided denoising or masked inpainting to enable localized edits while preserving global structure. Later methods including ControlNet [53], T2I-Adapter [35], and BrushNet [25] incorporate structural priors (e.g., edges, depth, segmentation) for precise spatial control, but rely on external supervision such as masks or sketches. Another line of work achieves implicit localization via prompt manipulation or attention modulation. Prompt-to-Prompt [19] and InstructPix2Pix [5] enable text-driven editing, while AnyEdit [24], DiffEditor [34], ICEdit [57], and MGIE [16] generalize this paradigm to multimodal or instruction-based editing. Recent large frameworks such as Flux.1-Kontext [29] and Qwen-Image-Edit [50] unify these ideas through hierarchical attention and multimodal conditioning, yet still depend on cross-attention to localize edits. Attend-and-Excite [7] modifies attention activations to strengthen underrepresented regions while TokenCompose [49] supervises cross-attention maps to bind textual tokens with visual objects. Our work follows this direction by leveraging attention scheme within the diffusion backbone to achieve region-aware style transfer.

Diffusion-based Style Transfer. Classical neural style transfer (NST) methods, beginning with Gatys *et al*. [17], optimize global content and style statistics extracted from CNN features. Feed-forward variants such as AdaIN [23], WCT [31], and SANet [37], and transformer-based StyTR^2 [11], accelerate stylization by matching global feature statistics. Diffusion-based approaches leverage strong generative priors for high-fidelity stylization. InST [55],

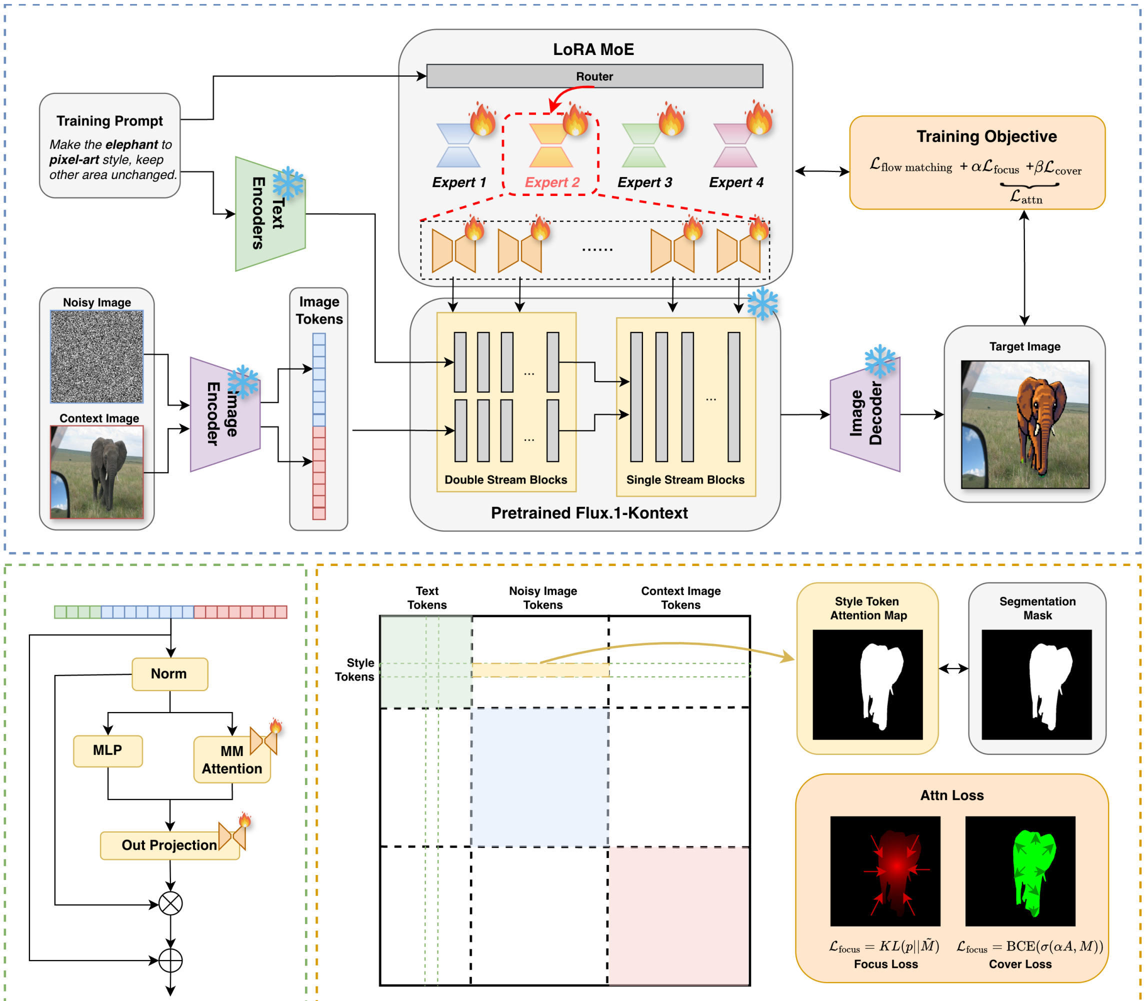

Figure 2. **Overview of the proposed framework.** The upper figure illustrates the overall pipeline based on the pretrained Flux.1-Kontext [29]. Given a context image, a noisy input, and a regional style prompt, image and text tokens are processed through Flux.1-Kontext, where LoRA-MoE modules [51] adapt attention and projection layers for style-specific learning. The model is optimized with flow matching, focus loss, and cover loss to reconstruct the target stylized image. Style-related attention maps are guided by binary masks, where focus loss concentrates attention within the target area and cover loss ensures spatial coverage for precise localized stylization.

StyleDiffusion [48], and FreeStyle [18] guide pretrained diffusion models using textual or latent style embeddings, while D^{2}Styler [46] improve stability or attempt localized control via attention injection. Style Injection in Diffusion [10] performs training-free key–value replacement to transfer global style. STAM [15] employs attention modulation for zero-shot transfer, U-StyDiT [56] utilizes diffusion transformers for high-resolution stylization, and StyleStudio [30] enables selective manipulation of stylistic elements. While promising, these methods model style as a global latent feature, offering limited region-level controllability. In contrast, our framework performs region-aware style transfer by aligning attention maps of style tokens and object binary mask, enabling localized style modulation without prompts, masks, or external supervision.

3. Method

3.1. Overview

Our goal is to enable a diffusion model to automatically determine where a visual style should be applied within an image, e.g., applying a pixel-art style only to a cat without explicit segmentation masks during inference. We achieve this through attention-guided training, which supervises the model's internal attention maps during fine-tuning to establish spatial correspondence between style tokens and object regions.

Built upon Flux.1-Kontext [29], a DiT-based diffusion model with joint text–image self-attention, our framework introduces attention supervision that aligns the attention maps of style tokens with target object masks. This correspondence is enforced via dedicated loss terms, allowing the model to ground style concepts to visual regions and perform mask-free localized style transfer at inference.

To efficiently handle multiple styles, we integrate a LoRA-MoE mechanism [51], where each style is represented by a specialized LoRA expert attached to a shared diffusion backbone. The backbone learns *where* to apply the style, while each expert defines *how* the style is rendered, enabling modular, plug-and-play control without retraining or cross-style interference. The overview of the proposed framework are shown in Figure 2.

3.2. Attention Map Extraction

Flux.1-Kontext [29] employs a transformer-based diffusion backbone, where each DiT block performs multi-head self-attention jointly over both image and text tokens. Given a text prompt containing a style phrase (e.g., "pixel-art style"), we extract the *text-to-image attention slice* associated with the style token s, which measures how each image token attends to the textual concept representing the target style (Figure 2).

For each attention layer ℓ with multi-head attention $A^{(\ell)} \in \mathbb{R}^{H\times N\times N}$, we first isolate the attention from the image queries Q_{img} to the style tokens K_s. We then average over heads, layers, and style tokens to obtain the aggregated style-conditioned attention map:

$$\hat{M}_s = \frac{1}{L}\sum_{\ell\in\mathcal{L}}\frac{1}{H}\sum_{h=1}^{H}\frac{1}{|K_s|}\sum_{k\in K_s} A_h^{(\ell)}[Q_{\text{img}}, k], \tag{1}$$

where L is the set of layers used for supervision, and H is the number of attention heads. The resulting map $\hat{M}_s \in \mathbb{R}^{h\times w}$ captures how strongly each spatial token attends to the style token s. For supervision, the ground-truth mask $M_s \in [0,1]^{h\times w}$ is obtained by downsampling the object segmentation map to match the attention map.

3.3. Attention Supervision Losses

To teach the model to attend accurately and comprehensively to the style-relevant region, we introduce two complementary losses:

- **Focus Loss** — aligns the overall spatial distribution of attention mass with the target object.
- **Cover Loss** — enforces uniform coverage within the object region, discouraging sparse or partial attention.

Focus Loss. The focus loss aligns the global shape of the predicted attention with the ground-truth mask. We interpret both the predicted attention and mask as normalized probability distributions and minimize their Kullback–Leibler divergence:

$$\mathcal{L}_{\text{focus}} = \sum_{s=1}^{S}\mathrm{KL}\Big(\mathrm{softmax}(\hat{M}_s/\tau)\,\Big\|\,\mathrm{norm}(M_s)\Big), \tag{2}$$

where $\mathrm{norm}(Z) = \frac{Z}{\sum Z}$ and τ controls the sharpness of the attention distribution. This encourages the attention map of each style token to concentrate its mass in the same spatial region as the corresponding object.

Cover Loss. While the focus loss ensures global alignment, it does not prevent the model from collapsing attention to a small part of the object. To encourage spatially dense and uniform coverage, we introduce a binary cross-entropy loss that operates at the token level:

$$\mathcal{L}_{\text{cover}} = \sum_{s=1}^{S}\mathrm{BCE\_logits}\big(\alpha\,\hat{M}_s,\; M_s\big), \tag{3}$$

where $\mathrm{BCE\_logits}$ is the numerically stable binary cross-entropy operating on logits, and α is a contrast factor that amplifies attention magnitude for stronger gradients. This term penalizes attention outside the object region ($M_s = 0$) and rewards attention inside ($M_s = 1$), producing smooth and coherent attention over the object.

Together, these two objectives ensure that the model learns not only to localize the correct style-relevant region (via $\mathcal{L}_{\text{focus}}$) but also to distribute attention densely within it (via $\mathcal{L}_{\text{cover}}$), resulting in coherent and spatially consistent style application.

3.4. LoRA-MoE Adaptation

To efficiently support multiple visual styles, we introduce a modular Low-Rank Adaptation with LoRA-MoE [51] scheme. Instead of fine-tuning a single LoRA across all styles, which often causes interference and degraded style fidelity, we assign each style a lightweight LoRA expert trained independently on the same shared diffusion backbone. During training, only the expert corresponding to the current style is activated, while the backbone remains frozen to preserve the attention-grounded spatial reasoning learned earlier. At inference, the appropriate expert is selected based on the target style token, enabling plug-and-play style control. This design provides three key benefits: (i) parameter efficiency — new styles are added without retraining the backbone, (ii) specialization — each expert learns distinct style patterns, and (iii) stability — the shared backbone ensures consistent spatial alignment across experts.

3.5. Training Objective

The overall objective combines the diffusion reconstruction loss with attention supervision:

$$\mathcal{L} = \mathcal{L}_{\epsilon} + \lambda_f\mathcal{L}_{\text{focus}} + \lambda_c\mathcal{L}_{\text{cover}}, \tag{4}$$

where $\mathcal{L}_\epsilon = \|\hat{\epsilon} - \epsilon\|_2^2$ is the standard noise prediction loss, and λ_f, λ_c balance the relative strengths of attention alignment and coverage.

4. Regional Style Editing Score

Existing metrics such as FID [21] or CLIP-similarity [40] mainly capture global appearance, lacking sensitivity to whether the style is accurately localized and whether unedited regions are preserved. To address this gap, we introduce the Regional Style Editing Score (RSE-Score), a metric specifically designed for evaluating *single-object regional style transfer*. RSE-Score decomposes the evaluation into two complementary aspects: (1) *Regional Style Matching (RSM)*, assessing whether the target region successfully reflects the intended style, and (2) *Identity Preservation* components, evaluating the perceptual and pixel-level fidelity of unedited areas. Together, these measures provide a comprehensive and interpretable assessment of both local style accuracy and spatial controllability.

Let x denote the original image, $\hat{x}$ the edited image, and $M \in \{0,1\}^{H\times W}$ the binary target mask, where $M_p = 1$ if pixel p belongs to the object region and $M_p = 0$ otherwise. The complementary background region is denoted by $(1 - M)$.

4.1. Regional Style Matching (RSM)

RSM quantifies how well the style within the target region matches the desired textual description. Instead of relying on patch-level features, we crop the edited image to the minimal bounding box enclosing the target mask (with a small padding margin) and compute its similarity to the style text using CLIP [40] encoders:

$$\mathrm{RSM} = \frac{1}{2}\big(1 + \cos\big(f_{\text{img}}(\hat{x}_{\text{crop}}), f_{\text{text}}(s)\big)\big), \tag{5}$$

where $\hat{x}_{\text{crop}}$ is the cropped region around the object, and $f_{\text{img}}(\cdot)$, $f_{\text{text}}(\cdot)$ denote CLIP's image and text feature extractors. The cosine similarity is linearly mapped to $[0, 1]$ for interpretability. This formulation focuses the style evaluation strictly within the edited region, mitigating interference from unrelated background areas.

4.2. Identity Preservation

To separately assess fidelity in unedited regions, we compute two complementary metrics on the background:

Perceptual Consistency (LPIPS). We use the spatial version of LPIPS [54] to measure perceptual distance between the edited and original images within the unedited region:

$$\mathrm{LPIPS}_{\text{bg}} = \frac{\sum_p (1 - M_p)\,\mathrm{LPIPS}_p(\hat{x}, x)}{\sum_p (1 - M_p)}, \tag{6}$$

where LPIPS_p denotes the per-pixel perceptual difference. Lower values indicate better background preservation.

Pixel Consistency (MSE). In addition, we compute a masked mean squared error over the same background region:

$$\mathrm{MSE}_{\text{bg}} = \frac{\sum_p (1 - M_p)\,\|\hat{x}_p - x_p\|_2^2}{\sum_p (1 - M_p)}. \tag{7}$$

This term captures fine-grained pixel alignment and complements the perceptual similarity measure.

Unlike the previous unified identity score, we now report $\mathrm{LPIPS}_{\text{bg}}$ and MSE_{bg} independently, offering a clearer diagnostic view of both perceptual and structural preservation. In summary, RSM evaluates style correctness within the edited region, while $\mathrm{LPIPS}_{\text{bg}}$ and MSE_{bg} quantify background fidelity—together forming a comprehensive benchmark for regional style transfer quality.

5. Experiments

5.1. Dataset and Pseudo-GT Generation

We use a subset of the Grounded COCO dataset introduced in TokenCompose [49], built upon MS-COCO [32] image–caption pairs. We randomly sample 150 image–caption pairs for fine-tuning and analysis. For each image, one target object (with its binary mask) is selected, and a pseudo ground-truth (pseudo-GT) image is generated by applying a diffusion-based style transfer model and compositing it with the original image. Benefiting from Flux.1-Kontext [29]'s strong instance recognition, precise mask alignment is not required, as the model can naturally learn smooth region boundaries. To ensure stylistic diversity, we generate pseudo-GT images in four representative styles—pixel art, cyberpunk, expressionism, and line art—resulting in 600 training samples (150 per style) for fine-tuning. More details of the pseudo-GT generation are explained in Supplementary Materials.

5.2. Experimental Setup

We fine-tuned the Flux.1-Kontext [29] model using LoRA-MoE [51] on a single NVIDIA GH200 GPU (120 GB VRAM). Training is performed at 1024×1024 resolution using `bf16` mixed precision and 8-bit Adam optimizer. The LoRA rank is set to 4, with a learning rate of 1×10^{-4}, batch size of 2, and gradient accumulation of 4. We train for 5000 steps under a constant learning rate schedule without warmup. Focus and cover losses are weighted by 0.1 and 0.2, respectively.

5.3. Experiment Results

Baselines. We compare our model with several representative instruction-based image editing approaches spanning different paradigms. Flux.1-Kontext [29] is a recent text-guided editor supporting soft regional conditioning, while

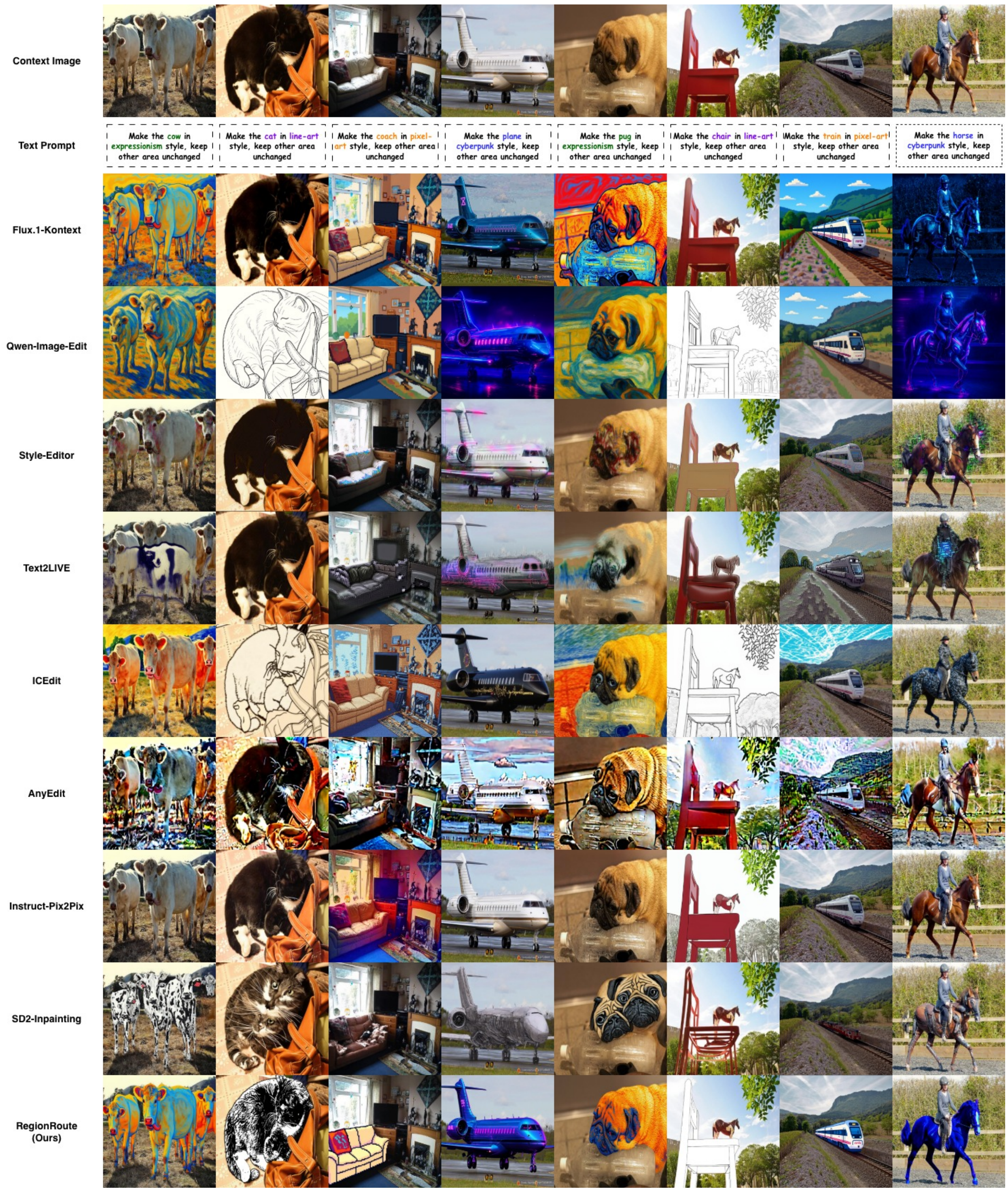

Figure 3. Qualitative comparison of state-of-the-art instruction-based image editing methods.

Qwen-Image-Edit [50] is an MLLM-based system that interprets natural language instructions to directly produce edited images. Style-Editor [38] is a text-driven object-centric style editing model. Text2LIVE [3], ICEdit [57], and AnyEdit [24] are diffusion models for localized edits aligned with text. We also include Instruct-Pix2Pix [5], and SD2-Inpainting [42], the Stable Diffusion v2 inpainting variant for mask-based regional editing.

Table 1. Quantitative comparison with baseline methods across three datasets. We report mean_{std} for RSM (↑), LPIPS_{bg} (↓), and MSE_{bg} (↓). Darker cells indicate better ranks (1–3) within each dataset and metric.

Model	COCO			Pascal VOC			BIG		
	RSM ↑	LPIPS_{bg} ↓	MSE_{bg} ↓	RSM ↑	LPIPS_{bg} ↓	MSE_{bg} ↓	RSM ↑	LPIPS_{bg} ↓	MSE_{bg} ↓
Flux.1-Kontext [29]	$0.6126_{0.0180}$	$0.4546_{0.2960}$	$0.1699_{0.2254}$	$0.6135_{0.0170}$	$0.4326_{0.3135}$	$0.1455_{0.2258}$	$0.6169_{0.0228}$	$0.4815_{0.3125}$	$0.1981_{0.2408}$
Qwen-Image-Edit [50]	$0.6235_{0.0164}$	$0.7530_{0.1605}$	$0.4398_{0.3876}$	$0.6333_{0.0104}$	$0.8078_{0.1550}$	$0.4477_{0.3767}$	$0.6329_{0.0103}$	$0.7185_{0.1687}$	$0.4284_{0.3731}$
Style-Editor [38]	$0.6071_{0.0157}$	$0.2235_{0.0697}$	$0.0093_{0.0086}$	$0.6091_{0.0152}$	$02077_{0.0633}$	$0.0087_{0.0069}$	$0.6047_{0.0156}$	$0.3061_{0.1137}$	$0.0108_{0.0084}$
Text2LIVE [3]	$0.6041_{0.0197}$	$0.2505_{0.0899}$	$0.0614_{0.0501}$	$0.6090_{0.0195}$	$0.2293_{0.0880}$	$0.0514_{0.0451}$	$0.6102_{0.0189}$	$0.3198_{0.1243}$	$0.0473_{0.0385}$
ICEdit [57]	$0.6086_{0.0152}$	$0.3512_{0.1979}$	$0.1568_{0.2800}$	$0.6102_{0.0147}$	$0.3506_{0.1963}$	$0.1284_{0.2304}$	$0.6115_{0.0164}$	$0.4282_{0.2315}$	$0.1570_{0.2697}$
AnyEdit [24]	$0.6085_{0.0104}$	$0.6895_{0.1564}$	$0.2633_{0.2913}$	$0.6081_{0.0107}$	$0.6834_{0.1715}$	$0.2345_{0.2454}$	$0.6078_{0.0097}$	$0.7109_{0.1679}$	$0.2446_{0.2813}$
Instruct-Pix2Pix [5]	$0.5978_{0.0115}$	$0.1867_{0.1995}$	$0.0516_{0.1587}$	$0.5969_{0.0111}$	$0.1554_{0.1472}$	$0.0299_{0.1078}$	$0.5939_{0.0111}$	$0.1466_{0.1377}$	$0.0347_{0.1403}$
SD2-Inpainting [42]	$0.6028_{0.0135}$	$0.0859_{0.0566}$	$0.0039_{0.0047}$	$0.6094_{0.0137}$	$0.0960_{0.0410}$	$0.0043_{0.0049}$	$0.6016_{0.0139}$	$0.1019_{0.0492}$	$0.0063_{0.0073}$
RegionRoute (Ours)	$\mathbf{0.6128}_{0.0162}$	$\mathbf{0.2103}_{0.1933}$	$\mathbf{0.0729}_{0.1281}$	$\mathbf{0.6147}_{0.0151}$	$\mathbf{0.1331}_{0.1429}$	$\mathbf{0.0269}_{0.0571}$	$\mathbf{0.6159}_{0.0193}$	$\mathbf{0.1593}_{0.1681}$	$\mathbf{0.0474}_{0.0967}$

Table 2. Controllability and semantic reliability evaluation via Vision-Language Model (VLM). Q1: "Is the *object* in the *target style*?"; Q2: "Is the *background* in the *target style*?; Q3: "Is the *object* in the *negative style*?"; Q4: "Is the *background* in the *negative style*?" Each cell shows the probability (%) of a "Yes" response to the corresponding binary question.

Model	COCO				Pascal VOC				BIG			
	Q1↑	Q2↓	Q3↓	Q4↓	Q1↑	Q2↓	Q3↓	Q4↓	Q1↑	Q2↓	Q3↓	Q4↓
Flux.1-Kontext [29]	0.63	0.44	0.08	0.06	0.65	0.43	0.08	0.05	0.63	0.39	0.11	0.10
Qwen-Image-Edit [50]	0.98	0.86	0.01	0.00	0.97	0.78	0.02	0.01	0.98	0.90	0.05	0.04
Style-Editor [38]	0.63	0.42	0.14	0.10	0.64	0.42	0.23	0.15	0.61	0.37	0.23	0.12
Text2LIVE [3]	0.65	0.24	0.20	0.07	0.68	0.26	0.44	0.12	0.70	0.23	0.43	0.07
ICEdit [57]	0.72	0.40	0.11	0.05	0.50	0.32	0.06	0.04	0.63	0.36	0.15	0.06
AnyEdit [24]	0.50	0.41	0.57	0.47	0.46	0.28	0.47	0.28	0.58	0.50	0.58	0.47
Instruct-Pix2Pix [5]	0.06	0.08	0.03	0.03	0.10	0.05	0.06	0.03	0.10	0.09	0.06	0.07
SD2-Inpainting [42]	0.36	0.01	0.15	0.01	0.24	0.04	0.07	0.02	0.54	0.04	0.32	0.03
RegionRoute (Ours)	**0.73**	**0.07**	**0.12**	**0.00**	**0.74**	**0.09**	**0.10**	**0.02**	**0.76**	**0.07**	**0.23**	**0.04**

Datasets. We evaluate our method on three benchmark datasets: COCO, Pascal VOC, and BIG. We intentionally select segmentation-based datasets that provide pixel-level object masks, which are essential for evaluating regional style transfer. The availability of accurate masks allows us to precisely define the edited region and its complement for computing the proposed metrics, ensuring consistent and spatially aligned evaluation across all methods. For the COCO dataset, we use the Grounded COCO dataset introduced in [49] while excluding all images used for training in our setup; only the remaining images are employed for evaluation. The Pascal VOC dataset [14] follows the relabeled version introduced in [8], which provides more accurate and consistent annotations than the original Pascal VOC labels. For the BIG dataset [8], we use both the official evaluation and test subsets for quantitative and qualitative comparisons.

Results and Analysis. Table 1 presents quantitative comparisons across three datasets, while Figure 3 illustrates qualitative examples. RSM reflects regional style accuracy, and LPIPS_{bg}/MSE_{bg} measure background preservation, where higher RSM and lower LPIPS_{bg}, MSE_{bg} indicate better localized editing performance.

Across all datasets, consistent patterns can be observed. Flux.1-Kontext [29] and Qwen-Image-Edit [50] achieve high RSM values, showing strong style generation ability, but their elevated background distortion indicates a tendency toward global style transfer.While Style-Editor [38] and Text2LIVE [3] are capable of achieving a certain level of localization, they exhibit less accurate style control and are prone to style leakage beyond the target region. ICEdit [57] and AnyEdit [24] obtain moderate RSM with less stable regional control, while Instruct-Pix2Pix [5] and SD2-Inpainting [42] preserve unedited regions effectively but provide limited stylization strength. Overall, existing methods tend to emphasize either stylistic fidelity or background consistency, but rarely achieve both simultaneously. In contrast, our proposed **RegionRoute** attains a favorable balance between regional style fidelity and background preservation. It maintains competitive RSM while substantially lowering LPIPS_{bg} and MSE_{bg}, indicating that edits are well localized and semantically coherent. Qualitative results in Figure 3 further confirm that **RegionRoute** applies the target style precisely within the intended area, preserves the structure of unedited regions, and maintains visual harmony across the entire image.

To further assess controllability and semantic reliability, Table 2 introduces four binary questions evaluated by Qwen2.5-VL-7B-Instruct [2]: (Q1) "Is the *object* in the *target style*?" (higher is better); (Q2) "Is the *background* in the *target style*?" (lower is better, indicating less style leakage); (Q3) "Is the *object* in the *negative style*?"; and (Q4) "Is the *background* in the *negative style*?" Q3–Q4 serve as sanity checks to ensure the model does not produce false positives or semantically inconsistent outputs. **RegionRoute** achieves high Q1 probabilities with minimal Q2 leakage and low Q3–Q4 values across all datasets, reflecting accurate regional stylization, strong semantic reliability, and minimal background contamination. Most baseline methods show consistent behavior, while AnyEdit exhibits elevated Q3–Q4 due to its semantically chaotic outputs, which sometimes confuse the vision-language evaluator.

Table 3. Ablation of key loss components, network streams, and LoRA ranks on three datasets. We report mean_{std} for RSM (↑), LPIPS_{bg} (↓), and MSE_{bg} (↓). ✓ denotes the component is enabled.

Variant	Enabled Components / Blocks					COCO			Pascal VOC			BIG		
	$\mathcal{L}_{\text{cover}}$	$\mathcal{L}_{\text{focus}}$	*Double*	*Single*	*LoRA*	**RSM** ↑	LPIPS_{bg} ↓	MSE_{bg} ↓	**RSM** ↑	LPIPS_{bg} ↓	MSE_{bg} ↓	**RSM** ↑	LPIPS_{bg} ↓	MSE_{bg} ↓
Ours	✓	✓	✓	✓	4	$\mathbf{0.6128}_{0.0162}$	$\mathbf{0.2103}_{0.1933}$	$\mathbf{0.0729}_{0.1281}$	$\mathbf{0.6147}_{0.0151}$	$\mathbf{0.1331}_{0.1429}$	$\mathbf{0.0269}_{0.0571}$	$\mathbf{0.6159}_{0.0193}$	$\mathbf{0.1593}_{0.1681}$	$\mathbf{0.0474}_{0.0681}$
w/o $\mathcal{L}_{\text{cover}}$	✗	✓	✓	✓	4	$0.6120_{0.0163}$	$0.2174_{0.1889}$	$0.0730_{0.1069}$	$0.6142_{0.0155}$	$0.1383_{0.1530}$	$0.0359_{0.0580}$	$0.6152_{0.0193}$	$0.1605_{0.1680}$	$0.0488_{0.0681}$
w/o $\mathcal{L}_{\text{focus}}$	✓	✗	✓	✓	4	$0.6127_{0.0159}$	$0.2132_{0.1852}$	$0.0740_{0.1117}$	$0.6147_{0.0153}$	$0.1359_{0.1393}$	$0.0325_{0.0497}$	$0.6158_{0.0190}$	$0.1612_{0.1620}$	$0.0542_{0.0917}$
w/o Double	✓	✓	✗	✓	4	$0.6168_{0.0168}$	$0.4225_{0.2838}$	$0.1409_{0.1831}$	$0.6175_{0.0155}$	$0.3887_{0.3064}$	$0.0980_{0.1435}$	$0.6195_{0.0198}$	$0.3843_{0.3086}$	$0.1275_{0.1458}$
w/o Single	✓	✓	✓	✗	4	$0.6190_{0.0163}$	$0.5203_{0.3060}$	$0.2284_{0.2216}$	$0.6185_{0.0158}$	$0.4252_{0.3303}$	$0.1390_{0.1884}$	$0.6192_{0.0205}$	$0.3933_{0.3110}$	$0.1402_{0.1595}$
Rank = 8	✓	✓	✓	✓	8	$0.6137_{0.0154}$	$0.2007_{0.1739}$	$0.0752_{0.1491}$	$0.6158_{0.0144}$	$0.1289_{0.1365}$	$0.0305_{0.0811}$	$0.6169_{0.0183}$	$0.1378_{0.1280}$	$0.0387_{0.0726}$
Rank = 16	✓	✓	✓	✓	16	$0.6126_{0.0150}$	$0.1876_{0.1583}$	$0.0671_{0.1229}$	$0.6158_{0.0140}$	$0.1182_{0.1099}$	$0.0212_{0.0480}$	$0.6152_{0.0179}$	$0.1177_{0.1059}$	$0.0265_{0.0491}$

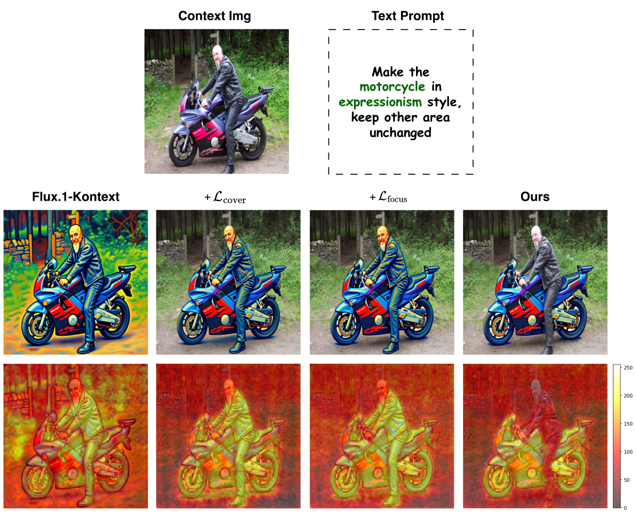

Figure 4. Visualization of attention maps under different loss configurations. Using only $\mathcal{L}_{\text{cover}}$ or $\mathcal{L}_{\text{focus}}$ causes attention spillover to nearby area, whereas our full objective focuses on the *motorcycle* without leakage to surrounding regions, demonstrating its ability to maintain precise and consistent attention.

5.4. Ablation Studies

We conduct a comprehensive ablation study on three datasets. Table 3 summarizes the quantitative results, while Figure 4 provides qualitative comparisons on each component of our training loss.

As shown in Table 3, removing either $\mathcal{L}_{\text{cover}}$ or $\mathcal{L}_{\text{focus}}$ leads to a consistent degradation in all metrics across datasets. The $\mathcal{L}_{\text{cover}}$ term encourages the model to preserve context coverage, and $\mathcal{L}_{\text{focus}}$ sharpens object-specific adaptation. Figure 4 visualizes the attention distributions under different settings. The full training objective aligns attention exclusively on the *motorcycle* while cleanly suppressing activations around surrounding regions. This confirms that the proposed joint loss formulation effectively enforces precise spatial localization and prevents attention leakage. More visualizations and analysis about the attention loss are in Supplementary Materials.

The Single and Double stream blocks originate from the Flux.1-Kontext [29] architecture, each modeling complementary aspects of spatial context. As shown in Table 3, disabling LoRA on either stream (*w/o Double* or *w/o Single*) slightly increases RSM, but substantially worsens LPIPS_{bg} and MSE_{bg}. This indicates that while the target region may look more "on-style," the model loses control over the consistency of the rest of the regions.

We further vary the LoRA rank ($r = \{4, 8, 16\}$) to analyze the trade-off between efficiency and representation capacity. As shown in Table 3, increasing the rank consistently improves all metrics, with higher ranks yielding better background consistency and lower reconstruction errors. Nevertheless, the model already performs competitively with a low rank ($r = 4$), demonstrating that **RegionRoute** can achieve strong adaptation and generalization under highly compact low-rank constraints.

6. Conclusion

In this work, we presented an attention-supervised diffusion framework for precise and mask-free localized style transfer. By explicitly aligning the attention maps of style tokens with object masks during training, our method learns spatially grounded style associations without relying on handcrafted segmentation at inference. The proposed LoRA-MoE adaptation further enhances parameter efficiency and stylistic diversity, while our newly designed evaluation metric offers a more objective measure of localized style fidelity and identity preservation. Extensive experiments across multiple datasets demonstrate that our approach achieves controllable, semantically consistent, and high-quality regional stylization, advancing the practicality of diffusion-based visual editing. While our approach achieves promising results, challenges remain for small, occluded, or ambiguous objects, highlighting the need for stronger spatial alignment. Extending the framework from text-driven editing to example-based style transfer is another promising direction.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18208–18218, 2022. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 7

[3] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pages 707–723. Springer, 2022. 6, 7

[4] BlackForest. Black forest labs; frontier ai lab, 2024. 2

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 2, 6, 7

[6] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22560–22570, 2023. 2

[7] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2

[8] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8890–8899, 2020. 7

[9] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8795–8805, 2024. 2

[10] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8795–8805, 2024. 3

[11] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11326–11336, 2022. 2

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 2

[14] Mark Everingham, Luc Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 7

[15] Masud An Nur Islam Fahim, Nazmus Saqib, and Jani Boutellier. Stam: Zero-shot style transfer using diffusion model via attention modulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6333–6343, 2025. 3

[16] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023. 2

[17] 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. 2

[18] Feihong He, Gang Li, Mengyuan Zhang, Leilei Yan, Lingyu Si, Fanzhang Li, and Li Shen. Freestyle: Free lunch for text-guided style transfer using diffusion models. *arXiv preprint arXiv:2401.15636*, 2024. 3

[19] 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. 2

[20] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024. 2

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[23] 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. 2

[24] Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Guojun Ma, Mingyang Wan, Xiang Wang, Xiangnan He, and Tat-seng Chua. Anyedit: Edit any knowledge encoded in language models. *arXiv preprint arXiv:2502.05628*, 2025. 2, 6, 7

[25] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, pages 150–168. Springer, 2024. 2

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[28] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022. 2

[29] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 1, 2, 3, 4, 5, 7, 8

[30] Mingkun Lei, Xue Song, Beier Zhu, Hao Wang, and Chi Zhang. Stylestudio: Text-driven style transfer with selective control of style elements. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23443–23452, 2025. 2, 3

[31] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017. 2

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[33] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[34] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8488–8497, 2024. 2

[35] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4296–4304, 2024. 2

[36] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 2

[37] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5880–5888, 2019. 2

[38] Jihun Park, Jongmin Gim, Kyoungmin Lee, Seunghun Lee, and Sunghoon Im. Style-editor: Text-driven object-centric style editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18281–18291, 2025. 6, 7

[39] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 2

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. 5

[41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[42] 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. 2, 6, 7

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[44] 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. 2

[45] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[46] Onkar Susladkar, Gayatri Deshmukh, Sparsh Mittal, and Parth Shastri. D2styler: Advancing arbitrary style transfer with discrete diffusion methods. In *International Conference on Pattern Recognition*, pages 63–82. Springer, 2025. 2, 3

[47] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1921–1930, 2023. 2

[48] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7677–7689, 2023. 2, 3

[49] Zirui Wang, Zhizhou Sha, Zheng Ding, Yilin Wang, and Zhuowen Tu. Tokencompose: Text-to-image diffusion with token-level supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8553–8564, 2024. 2, 5, 7

[50] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei

Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 2, 6, 7

[51] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of lora experts. *arXiv preprint arXiv:2404.13628*, 2024. 2, 3, 4, 5

[52] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18381–18391, 2023. 2

[53] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2

[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[55] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10146–10156, 2023. 2

[56] Zhanjie Zhang, Ao Ma, Ke Cao, Jing Wang, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, and Yuhui Yin. U-stydit: Ultra-high quality artistic style transfer using diffusion transformers. *arXiv preprint arXiv:2503.08157*, 2025. 3

[57] Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*, 2025. 2, 6, 7