# UniCSG: Unified High-Fidelity Content-Constrained Style-Driven Generation via Staged Semantic and Frequency Disentanglement

Jingwei Yang[1], Ruoxi Wu[2⋆], Wei Shen[2], Meng Li[2], Yulong Liu[2], Huimin She[2], and Lunxi Yuan[2]

[1] China University of Mining and Technology, Beijing, China
[2] OPPO Artificial Intelligence Center, Beijing, China

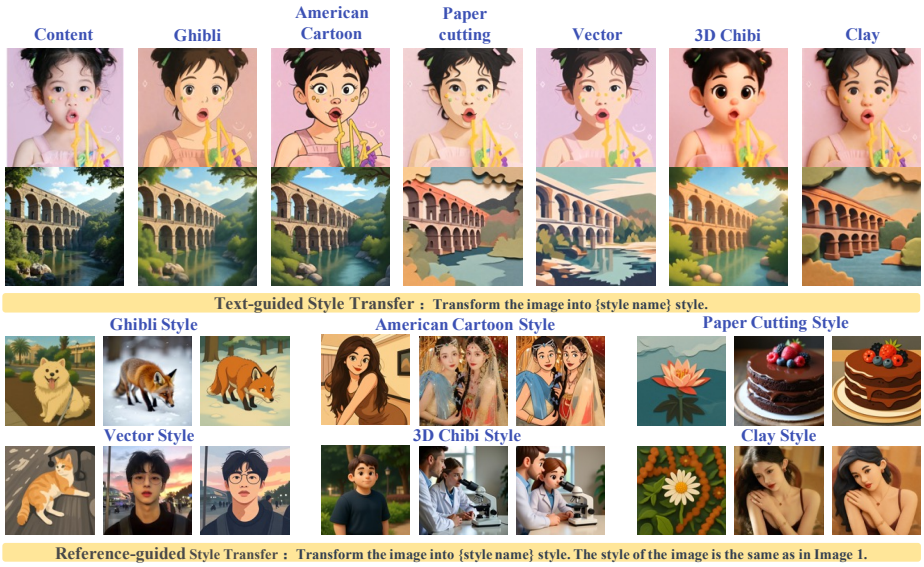


**Fig. 1:** Teaser of UniCSG on text-guided and reference-guided style transfer. Given a content image, UniCSG transforms it into user-specified styles under both text prompts and reference exemplars, showing faithful content preservation and style alignment. In the reference-guided rows, each triplet shows (left to right) the reference image, the content image, and the generated result.

**Abstract.** Style transfer must match a target style while preserving content semantics. DiT-based diffusion models often suffer from content–style entanglement, leading to reference-content leakage and unstable generation. We present UniCSG, a unified framework for content-constrained, style-driven generation in both text-guided and reference-guided settings. UniCSG employs staged training: (i) a latent-space semantic disentanglement stage that combines low-frequency preprocessing with conditioning corruption to encourage content–style separation,

⋆ Corresponding author, Email: wuruoxi@oppo.com

and (ii) a latent-space frequency-aware detail reconstruction stage that refines details via multi-scale frequency supervision. We further incorporate pixel-space reward learning to align latent objectives with perceptual quality after decoding. Experiments demonstrate improved content faithfulness, style alignment, and robustness in both settings.

**Keywords:** Style transfer · Diffusion models · Content-style disentanglement

## 1 Introduction

Style transfer [11] is a core controllable generation task. Text-guided style transfer applies a language-specified style to a content image, whereas reference-guided style transfer transfers the visual appearance of a reference image to the target content. In applications such as art creation, advertising, and game development, the goal is to preserve content semantics while achieving high-fidelity style alignment [51].

Most diffusion-based style transfer systems rely on either SDXL-style UNets [26] or DiT backbones [8, 14, 25]. UNets benefit from local inductive biases [49, 52] and often yield better empirical content–style separation [36, 50], yet their scalability and global modeling capacity can limit performance at high resolution and in complex scenes. DiT architectures scale better, but they typically lack fine-grained disentanglement mechanisms [51]. As a result, they can suffer from style overfitting [18], reference-content leakage [2, 39, 42], and instability such as checkerboard artifacts [18, 19, 53].

We attribute these issues to three factors. First, latent diffusion models are often trained with a single reconstruction objective, which biases learning toward low-level textures and under-emphasizes high-level semantics such as contours and layout [43]. Second, frequency coupling further entangles content and style, reducing controllability and fidelity. DiT backbones tend to model low-frequency structure more effectively than high-frequency textures [22]. Third, optimization occurs in latent space whereas evaluation is performed in pixel space. Latent-optimal solutions can diverge from perceptual optimality after decoding.

To address these challenges, we propose UniCSG, a unified framework for high-fidelity, content-constrained, style-driven generation. UniCSG follows a staged, progressive paradigm. In latent space, two sequential stages with decoupled objectives handle semantic disentanglement and frequency-aware detail reconstruction, respectively. A pixel-space perceptual alignment module further bridges the latent–pixel objective gap.

Latent-space Stage 1 (Semantic Disentanglement): we combine low-frequency preprocessing with conditioning corruption to encourage the model to rely on high-level cues, improving content–style disentanglement and reducing content leakage.

Latent-space Stage 2 (Frequency-aware Detail Reconstruction): we introduce multi-scale frequency decomposition and a frequency-weighted objective to refine high-frequency textures.

Pixel-space Reward Learning (Reward-guided Perceptual Alignment): we incorporate pixel-space reward learning to improve perceptual quality after decoding.

Through this disentangle–reconstruct–align pipeline, UniCSG jointly improves controllability, detail fidelity, and perceptual quality.

Our contributions can be summarized as follows:

- We introduce a **semantic disentanglement** strategy based on low-frequency preprocessing and conditioning corruption, which effectively separates style from content and prevents reference-content leakage.
- We propose a **frequency-aware detail reconstruction** mechanism that leverages multi-scale frequency decomposition and frequency-weighted supervision to enhance high-frequency texture recovery upon the semantic scaffold.
- We introduce a **reward-guided learning** mechanism in pixel space that bridges the gap between latent-space optimization objectives and perceptual quality after decoding.

## 2 Related Work

### 2.1 Content-Constrained Style-Driven Generation

Early CNN-based style transfer defines separate content and style losses to constrain structural and textural similarity, respectively [11]. Later methods add attention [6, 28] and multi-scale fusion [13] for finer local control. With diffusion models, the focus shifts to imposing precise content constraints on powerful backbones. IP-Adapter [44] injects reference features via cross-attention for disentangled control. ControlNet [47] conditions on spatial cues (edges, depth) to preserve structure. However, most approaches still operate at the pixel or shallow-feature level without explicit high-level semantic constraints, causing structural distortion and content-style entanglement in complex scenes. Frequency coupling further limits fidelity and visual quality. These limitations motivate stronger semantic constraints and frequency-aware optimization for robust, high-fidelity style transfer.

### 2.2 Semantic Disentanglement

In image generation, pixel-level reconstruction alone often drives latents to overfit textures and noise while under-capturing high-level semantics [43]. Yet latent semantic capacity correlates strongly with generation quality. Prior work addresses this gap by strengthening encoders (e.g., VAE–ViT hybrids or unified encoders [43]), aligning latents with semantic features via DINOv2 [23], or employing asynchronous denoising such as SFD [24]. In contrast, we retain the standard encoder and decouple objectives via staged training. Stage 1 learns semantic disentanglement under degraded inputs. Stage 2 refines details guided by the learned scaffold.

### 2.3 Frequency Disentanglement

Diffusion Transformers have popularized high/low-frequency separation. DeCo [22], DiP [3], PixelDiT [46], and JiT [20] assign different components to low-frequency semantics versus high-frequency textures, aligning with Transformers' strengths in low-frequency modeling. Common strategies employ heavy downsampling [22,46] or embedding bottlenecks [20] to constrain the backbone to low frequencies. Lightweight decoders [22] or auxiliary paths [46] then restore high-frequency textures. Another line decouples frequencies in latent space [33, 41], exploiting early denoising for low frequencies and later steps for high frequencies [45]. We deepen this perspective with a two-stage paradigm. Stage 1 concentrates on low-frequency structure and content–style integration. Stage 2 targets frequency-aware detail reconstruction, yielding stronger disentanglement and higher fidelity.

## 3 Methodology

We build UniCSG on top of a pretrained image-editing model (Qwen-Image-Edit-2509 [37]). We unify text-to-image and image-to-image settings with a four-tuple input $\langle text, ref\_img, source\_img, gt\_img \rangle$. Both text-guided and reference-guided stylization follow the same content-constrained generation framework, differing only in the source of style information: text-guided transfer uses the text prompt to specify the desired style, whereas reference-guided transfer uses the reference image as the style exemplar. We structure the methodology description into three core sections. Semantic Disentanglement (Stage 1) is covered in Sec. 3.1. Sec. 3.2 addresses Frequency-aware Detail Reconstruction (Stage 2). Pixel-space Reward Learning is then presented in Sec. 3.3. An overview is shown in Fig. 2.

### 3.1 Semantic Disentanglement

This stage targets semantic entanglement between content and style. We combine low-frequency preprocessing with conditioning corruption to increase the difficulty of the conditioning inputs. This encourages the model to learn disentangled content–style representations and a stable semantic scaffold that reduces content leakage.

Diffusion denoising naturally follows a coarse-to-fine trajectory: early steps primarily recover global contours and semantic layout [45]. Accordingly, we apply low-pass filtering [16, 17] to the input images before VAE encoding, retaining only frequencies below a threshold $\tau$ (default $\tau$=0.2) for the target $\mathbf{I}_{\text{target}}$ and the conditioning images $\mathbf{I}_{\text{content}}$ and $\mathbf{I}_{\text{style}}$:

$$\begin{aligned} \mathbf{I}^{\text{low}}_{\text{target}} &= \mathcal{F}_{\text{low\_pass}}(\mathbf{I}_{\text{target}}), \\ \mathbf{I}^{\text{low}}_{\text{content}} &= \mathcal{F}_{\text{low\_pass}}(\mathbf{I}_{\text{content}}), \\ \mathbf{I}^{\text{low}}_{\text{style}} &= \mathcal{F}_{\text{low\_pass}}(\mathbf{I}_{\text{style}}). \end{aligned} \tag{1}$$

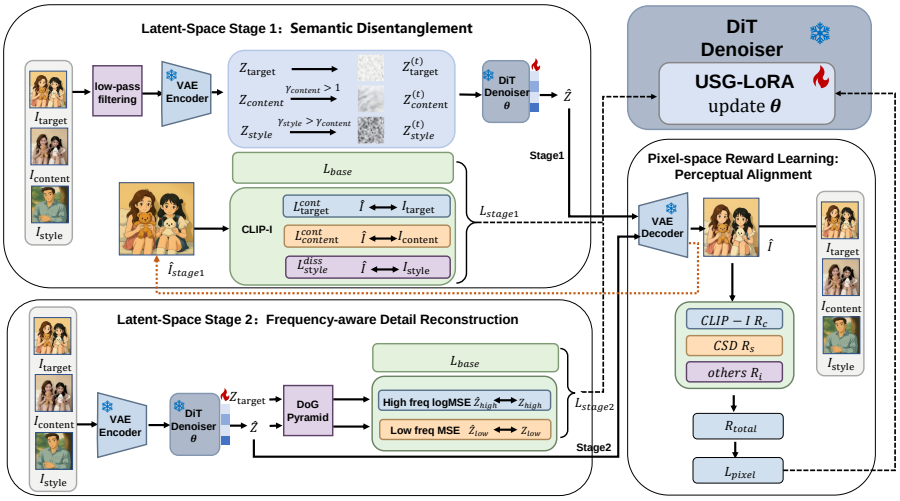


**Fig. 2:** Overview of UniCSG. We use staged semantic–frequency disentanglement in latent space and reward-guided perceptual alignment in pixel space. Stage 1 (semantic disentanglement): low-frequency preprocessing plus conditioning corruption builds a robust content–style scaffold. Stage 2 (frequency-aware detail reconstruction): multi-scale frequency supervision refines high-frequency textures atop the scaffold. Pixel-space reward learning further matches latent objectives to perceptual quality after decoding.

This operation encourages the model to focus on low-frequency structure early in training, supporting stable semantic learning before high-frequency refinement [10].

To achieve precise content–style disentanglement, we introduce conditioning corruption under a unified timestep. Specifically, we keep a single timestep condition per training step and apply different corruption strengths to the content and style conditionings. This yields an information hierarchy: the target is the least corrupted, content is moderately corrupted, and style is the most corrupted.

To remain fully compatible with the pretrained noise schedule, we sample a unified timestep $t$ at each training step and use the scheduler-provided noise coefficient $\sigma(t) \in [0, 1]$ (denoted $\sigma_t$). For clarity, we denote the clean target/content/style latents by $\mathbf{z}_{\text{target}}$, $\mathbf{z}_{\text{content}}$, and $\mathbf{z}_{\text{style}}$, and use the superscript $(t)$ to indicate the noised latent at timestep $t$. Under this unified timestep, the target latent follows the standard convex-combination noising process:

$$\mathbf{z}_{\text{target}}^{(t)} = (1 - \sigma_t)\mathbf{z}_{\text{target}} + \sigma_t \boldsymbol{\epsilon}_{\text{target}}, \quad \boldsymbol{\epsilon}_{\text{target}} \sim \mathcal{N}(0, \mathbf{I}). \tag{2}$$

Rather than using different timesteps, we express the desired information hierarchy by applying stronger corruption to the conditioning branches under the same timestep. Specifically, we introduce noise amplification factors $\gamma_i$ for $i \in \{\text{content}, \text{style}\}$, where $\gamma_{\text{style}} > \gamma_{\text{content}} > 1$. These factors are implemented

via higher-variance Gaussian noise. For a given amplification factor $\gamma_i$, we construct the amplified noise as:

$$\begin{aligned} \boldsymbol{\epsilon}_{\gamma_i} &= \boldsymbol{\epsilon} + \sqrt{\gamma_i^2 - 1}\,\boldsymbol{\epsilon}', \quad \boldsymbol{\epsilon}, \boldsymbol{\epsilon}' \sim \mathcal{N}(0, \mathbf{I}), \text{ independent} \\ &\Rightarrow \ \boldsymbol{\epsilon}_{\gamma_i} \sim \mathcal{N}(0, \gamma_i^2 \mathbf{I}). \end{aligned} \tag{3}$$

Under the same timestep $t$, the noised content and style conditionings are defined as:

$$\begin{aligned} \mathbf{z}_{\text{content}}^{(t)} &= (1 - \sigma_t)\mathbf{z}_{\text{content}} + \sigma_t \boldsymbol{\epsilon}_{\gamma_{\text{content}}}, \\ \mathbf{z}_{\text{style}}^{(t)} &= (1 - \sigma_t)\mathbf{z}_{\text{style}} + \sigma_t \boldsymbol{\epsilon}_{\gamma_{\text{style}}}. \end{aligned} \tag{4}$$

Consequently, under the same timestep $t$, the style conditioning is more corrupted than the content conditioning, while the target branch remains the least corrupted. This $\gamma_i$-controlled information hierarchy encourages the model to extract global style cues from heavily corrupted style inputs and structural semantics from less corrupted content inputs, suppressing reference-content leakage within a single-timestep training framework [**?**].

To improve generalization, we introduce a probabilistic replacement mechanism: with probability $p$, we replace the content or style conditioning with pure noise, forcing the model to generate plausible structure under partial information loss. To smoothly transition from Stage 1 disentanglement learning to Stage 2 full-detail reconstruction, we apply a progressive decay schedule to the corruption strength during training.

The objective in this stage combines a base reconstruction loss $\mathcal{L}_{\text{base}}$ with a semantic loss $\mathcal{L}_{\text{sem}}$ that guides content–style disentanglement:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{base}} + \mathcal{L}_{\text{sem}}. \tag{5}$$

The base loss is the standard velocity-prediction objective:

$$\mathcal{L}_{\text{base}} = w(t) \cdot \mathbb{E}\Big[\Big\|\mathbf{v}_\theta(\mathbf{z}_{\text{target}}^{(t)}, t, \mathbf{z}_{\text{content}}^{(t)}, \mathbf{z}_{\text{style}}^{(t)}) - (\boldsymbol{\epsilon}_{\text{target}} - \mathbf{z}_{\text{target}})\Big\|_2^2\Big]. \tag{6}$$

The semantic loss uses CLIP [27] image features $\phi(\cdot)$ to define three complementary signals. The target-faithfulness term $\mathcal{L}_{\text{target}}^{\text{cont}}$ aligns the generated output with the ground-truth target at a semantic level. The content-preservation term $\mathcal{L}_{\text{content}}^{\text{cont}}$ enforces the core objects and scene layout of the content image. The content-style repulsion term $\mathcal{L}_{\text{style}}^{\text{diss}}$ reduces semantic similarity with the style image, mitigating unintended content transfer from the style reference:

$$\begin{aligned}\mathcal{L}_{\text{sem}} = \lambda^{\text{cont}}_{\text{target}} \cdot \underbrace{\mathbb{E}\Big[1 - \text{sim}_{\text{CLIP-I}}\big(\phi(\hat{\mathbf{I}}),\, \phi(\mathbf{I}_{\text{target}})\big)\Big]}_{\mathcal{L}^{\text{cont}}_{\text{target}}} \\ + \lambda^{\text{cont}}_{\text{content}} \cdot \underbrace{\mathbb{E}\Big[1 - \text{sim}_{\text{CLIP-I}}\big(\phi(\hat{\mathbf{I}}),\, \phi(\mathbf{I}_{\text{content}})\big)\Big]}_{\mathcal{L}^{\text{cont}}_{\text{content}}} \\ + \lambda^{\text{diss}}_{\text{style}} \cdot \underbrace{\mathbb{E}\Big[\text{sim}_{\text{CLIP-I}}\big(\phi(\hat{\mathbf{I}}),\, \phi(\mathbf{I}_{\text{style}})\big)\Big]}_{\mathcal{L}^{\text{diss}}_{\text{style}}}. \end{aligned} \tag{7}$$

In summary, conditioning corruption and semantic supervision promote content–style disentanglement and provide a semantic foundation for Stage 2.

### 3.2 Frequency-aware Detail Reconstruction

This stage focuses on high-fidelity reconstruction of fine-grained textures after the model has learned a stable semantic blueprint. We guide detail synthesis through multi-scale frequency decomposition and a carefully designed objective emphasizing high-frequency refinement.

Once the model has learned the basic structure, we remove the low-pass constraint and train with full-spectrum information. The key is to introduce frequency-domain losses as explicit guidance. Since the denoising objective is defined in latent space, we perform frequency decomposition directly on latents for consistency with the training target. Specifically, we apply a DoG (Difference of Gaussian) pyramid to decompose the predicted latent $\hat{\mathbf{z}}$ and the target latent $\mathbf{z}_{\text{target}}$ into multi-scale components [33, 41]. This separates low-frequency parts $\hat{\mathbf{z}}^{(k)}_{\text{low}}$, $\mathbf{z}^{(k)}_{\text{target, low}}$ and high-frequency parts $\hat{\mathbf{z}}^{(k)}_{\text{high}}$, $\mathbf{z}^{(k)}_{\text{target, high}}$, where $k$ denotes the scale level. Any perceptual gap introduced by operating in latent space is subsequently addressed by the pixel-space reward learning (Sec. 3.3).

Low-frequency components correspond to transferable global structure with relatively uniform error distributions. We therefore use MSE to provide strong gradients for accurate global reconstruction. In contrast, high-frequency components are inherently mixed: they include both stylistic fine textures to be transferred and irrelevant noise or incidental details. For high-frequency terms, we do not enforce strict pixel-wise matching. Instead, we aim for a weaker consistency that preserves beneficial stylization while filtering meaningless interference.

To reduce the influence of outliers introduced by noise and irrelevant details, we adopt LogMSE for high-frequency components. LogMSE down-weights large errors while remaining sensitive to small-to-moderate discrepancies, which helps refine transferable textures [21].

We therefore construct a weighted composite objective by augmenting the base reconstruction loss with a frequency-domain supervision term:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{base}} + \lambda_{\text{freq}} \cdot \sum_{k} \Big( w_{\text{low}} \cdot \big\| \hat{\mathbf{z}}^{(k)}_{\text{low}} - \mathbf{z}^{(k)}_{\text{target, low}} \big\|_2^2 + w_{\text{high}} \cdot \log\big(1 + \big\| \hat{\mathbf{z}}^{(k)}_{\text{high}} - \mathbf{z}^{(k)}_{\text{target, high}} \big\|_2^2\big)\Big). \tag{8}$$

Here, $w(t)$ is the timestep weight given by the scheduler during training. $\lambda_{\text{freq}}$ controls the overall strength of frequency supervision. $w_{\text{low}}$ and $w_{\text{high}}$ balance low-frequency structure preservation against high-frequency detail reconstruction. We typically set $w_{\text{high}} > w_{\text{low}}$ to prioritize fine-grained high-frequency refinement.

This staged design, together with multi-scale frequency supervision in Stage 2, improves high-frequency detail reconstruction while preserving global structure.

### 3.3 Pixel-space Reward Learning

Even after learning a content–style mapping in latent space through semantic–frequency disentanglement, a key challenge remains: latent-space optimization is not perfectly aligned with human perception in pixel space [10]. Latent reconstruction objectives do not fully capture pixel-level perceptual quality. A solution optimal in latent space may therefore not yield the best decoded perception in terms of content faithfulness, style consistency, or detail sharpness.

To bridge this gap, we incorporate a reward-guided learning mechanism throughout training. A set of multi-dimensional reward models operates on decoded images in pixel space. These models evaluate generated outputs and provide perceptual feedback signals, enabling direct and fine-grained optimization of visual quality and content–style alignment.

Reward-guided learning runs alongside the main training process. Let $S_{\text{total}}$ denote the total training steps and $S_{\text{warmup}}$ the number of Stage 1 steps (default $S_{\text{warmup}} = 0.6 \cdot S_{\text{total}}$). For each mini-batch, we determine the current stage based on step index $s$ and compute the corresponding latent-space loss $\mathcal{L}_{\text{latent}}(s)$.

$$\mathcal{L}_{\text{latent}}(s) = \begin{cases} \mathcal{L}_{\text{stage1}}, & \text{if } s \le S_{\text{warmup}}, \\ \mathcal{L}_{\text{stage2}}, & \text{otherwise.} \end{cases} \tag{9}$$

In parallel, the reward module is activated. We define a multi-dimensional reward function in pixel space, $R(\hat{\mathbf{I}}, \mathbf{I}_{\text{target}}, \mathbf{I}_{\text{content}}, \mathbf{I}_{\text{style}})$, to evaluate the decoded image $\hat{\mathbf{I}}$. The reward aggregates several specialized signals. The content-faithfulness reward $R_c$ measures semantic similarity between the generated image and the content reference via CLIP-I [27] image encoder, encouraging preservation of key content. The style-alignment reward $R_s$ measures stylistic similarity between the generated image and the style reference via CSD [29], encouraging accurate capture and transfer of artistic style. The total reward is a weighted sum of these core rewards together with auxiliary signals including an LPIPS-based perceptual reward and a discriminator-based adversarial reward:

$$R_{\text{total}} = \omega_c R_c + \omega_s R_s + \sum_i \omega_i R_i. \tag{10}$$

Reward learning remains active in both stages: outputs from either Stage 1 or Stage 2 are decoded into pixel space, scored by the reward models, and optimized with a policy-gradient-style objective. To avoid confusion with timestep $t$, we denote the step-wise advantage by $\mathcal{A}(s)$. The pixel-space reward loss is:

$$\mathcal{L}_{\text{pixel}} = -\mathbb{E}_{(\mathbf{z}^{(t)}_{\text{target}}, t, \mathbf{z}^{(t)}_{\text{content}}, \mathbf{z}^{(t)}_{\text{style}}) \sim \mathcal{D}} \left[ \mathcal{A}(s) \cdot \log \pi_\theta \left( \hat{\mathbf{I}} \,\middle|\, \mathbf{z}^{(t)}_{\text{target}}, t, \mathbf{z}^{(t)}_{\text{content}}, \mathbf{z}^{(t)}_{\text{style}} \right) \right]. \tag{11}$$

Here, $\pi_\theta(\cdot)$ denotes the conditional generative distribution induced by the denoiser (without committing to a specific noise parameterization). The advantage is defined as:

$$\mathcal{A}(s) = R_{\text{total}} \left( \hat{\mathbf{I}}, \mathbf{I}_{\text{target}}, \mathbf{I}_{\text{content}}, \mathbf{I}_{\text{style}} \right) - b \left( \mathbf{z}^{(t)}_{\text{target}}, t, \mathbf{z}^{(t)}_{\text{content}}, \mathbf{z}^{(t)}_{\text{style}} \right). \tag{12}$$

$b(\cdot)$ is the baseline function. Finally, we combine the latent-space loss and the pixel-space reward loss with a weighting coefficient to form the overall training objective:

$$\mathcal{L}_{\text{total}}(s) = \mathcal{L}_{\text{latent}}(s) + \lambda_{\text{pixel}} \cdot \mathcal{L}_{\text{pixel}}(s). \tag{13}$$

## 4 Experiments

### 4.1 Experimental Settings

**Setup.** We adopt Qwen-Image-Edit-2509 [37] as the pretrained model. Through the data generation pipeline detailed in the appendix, we construct a training dataset containing 40,000 high-quality four-tuples for model training. Training is conducted on NVIDIA A100 GPUs in two stages: Stage 1 fine-tunes for 12,000 steps on 2 GPUs with a learning rate of $5 \times 10^{-6}$ and batch size 1. Stage 2 fine-tunes for 8,000 steps on 2 GPUs with per-GPU batch size 1 (total batch size 2), using the same learning rate.

**Benchmark.** For fair and comprehensive evaluation, we construct the CSG-Bench benchmark by sampling and augmenting public datasets such as Omni-Consistency [31] and OmniStyle [34]. The test set contains 1,922 content images spanning portraits, architecture, landscapes, cartoons, objects, food, and animals. For styles, we randomly select 6 out of 20 predefined style types, including 3D Chibi, American Cartoon, clay, Ghibli, Paper cutting, and Vector and curate 100 high-quality, clear reference images for each style. Evaluation results on OmniConsistency-bench are provided in the appendix.

**Table 1:** Quantitative results for text-guided style transfer on CSG-Bench.

| Method | Style consistency | | | Content consistency | | |
|---|---|---|---|---|---|---|
| | FID ↓ | CSD ↑ | CLIP-T ↑ | CLIP-I ↑ | DINO ↑ | DreamSim ↑ |
| OmniConsistency [31] | **117.079** | 0.503 | 0.266 | 0.705 | 0.487 | 0.712 |
| OmniStyle [34] | 134.415 | 0.275 | 0.247 | 0.737 | 0.620 | 0.715 |
| flux1.Kontext-dev [15] | 120.649 | 0.430 | 0.251 | 0.772 | 0.588 | 0.741 |
| Nano-banana [4] | 128.162 | 0.534 | 0.249 | **0.792** | **0.644** | **0.791** |
| DreamOmni2 [40] | 120.697 | 0.388 | 0.252 | 0.757 | 0.531 | 0.704 |
| OmniGen2 [38] | 121.355 | 0.448 | 0.260 | 0.733 | 0.547 | 0.733 |
| BAGEL [5] | 132.504 | 0.525 | 0.265 | 0.696 | 0.425 | 0.638 |
| Ovis-U1 [32] | 136.033 | **0.573** | **0.275** | 0.702 | 0.400 | 0.653 |
| Qwen-Image-Edit [37] (Base) | 120.329 | 0.489 | 0.251 | 0.772 | 0.617 | 0.761 |
| UniCSG (Ours) | **113.940** | **0.541** | **0.267** | **0.797** | **0.701** | **0.816** |

**Table 2:** Quantitative results for reference-guided style transfer on CSG-Bench.

| Method | Style consistency | | | Content consistency | | |
|---|---|---|---|---|---|---|
| | FID ↓ | CSD ↑ | CLIP-T ↑ | CLIP-I ↑ | DINO ↑ | DreamSim ↑ |
| OmniConsistency [31] | 91.140 | 0.635 | 0.269 | 0.666 | 0.408 | 0.679 |
| OmniStyle [34] | 129.775 | 0.393 | 0.256 | 0.732 | **0.615** | 0.670 |
| flux1.Kontext-dev [15] | 91.785 | 0.560 | 0.255 | 0.690 | 0.498 | 0.685 |
| Nano-banana [4] | 127.868 | 0.522 | 0.250 | 0.745 | 0.586 | **0.748** |
| DreamOmni2 [40] | 112.960 | 0.420 | 0.242 | **0.784** | **0.606** | **0.748** |
| OmniGen2 [38] | 90.270 | 0.622 | **0.271** | 0.624 | 0.409 | 0.624 |
| BAGEL [5] | **72.070** | 0.511 | 0.266 | 0.503 | 0.212 | 0.500 |
| Ovis-U1 [32] | 101.215 | **0.703** | **0.277** | 0.552 | 0.274 | 0.555 |
| Qwen-Image-Edit [37] (Base) | 87.922 | 0.577 | 0.253 | 0.570 | 0.328 | 0.565 |
| UniCSG (Ours) | **87.320** | **0.731** | **0.271** | **0.760** | 0.597 | **0.762** |

**Evaluation Metrics.** To comprehensively evaluate generation performance, we adopt a multi-dimensional metric suite covering two core aspects. For content consistency [2], we report CLIP-I [27], DINO [23], and DreamSim [9] as complementary measures. For style consistency [30], we use FID [12], CLIP-T [27], and CSD [29] to assess style alignment from multiple perspectives.

**Baselines.** We compare UniCSG against eight representative and state-of-the-art baselines: OmniConsistency [31], OmniStyle [34], flux1.Kontext-dev [15], Nano-banana [4], DreamOmni2 [40], OmniGen2 [38], BAGEL [5], and Ovis-U1 [32].

### 4.2 Experimental Results

**Quantitative Evaluation.** In all tables, we highlight the best value in **red** and the second-best in **blue** in each column. Quantitative comparisons are summarized in Tables 1 and 2.

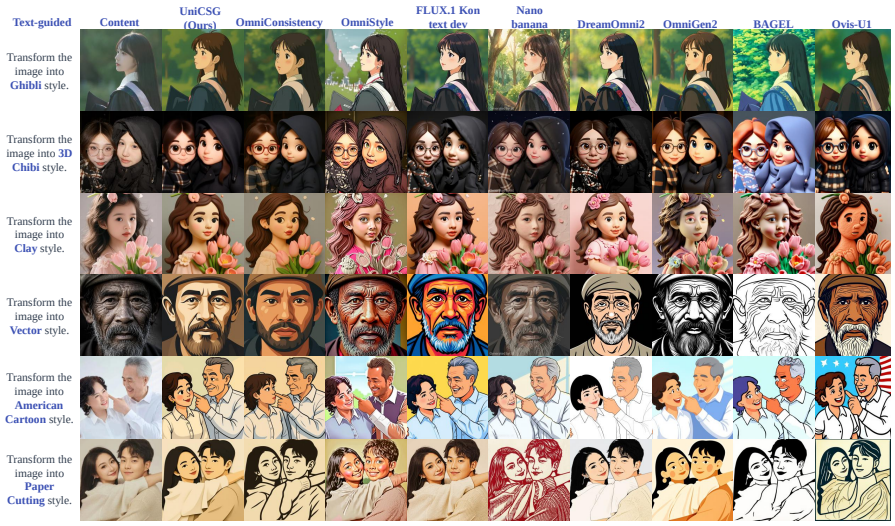


**Fig. 3:** Qualitative results for text-guided style transfer on CSG-Bench.

In the text-guided setting, UniCSG achieves a favorable trade-off between style consistency and content consistency. Although style-consistency metrics are slightly lower than Ovis-U1 [32], it consistently preserves content structure while maintaining competitive stylization strength. Moreover, the FID score of 113.9397 outperforms multiple baselines (e.g., OmniConsistency [31]), indicating that the distribution of generated images is closer to the target style distribution.

In the reference-guided setting, UniCSG likewise achieves a strong style-content trade-off. A high CSD score indicates effective style transfer, and a high DreamSim score indicates improved content preservation with reduced reference-content leakage. DINO and CLIP-T scores are comparable to leading models, suggesting competitive visual quality and semantic content retention even when some baselines achieve lower FID at the cost of weaker content control.

**Qualitative Evaluation.** Representative visual comparisons are shown in Fig. 3 and Fig. 4.

In the text-guided setting, UniCSG preserves content well and produces strong details. Fine-grained inspection shows better preservation of hairstyle, semantic identity, object count, and background layout consistency, while still maintaining clear stylization. We observe slightly weaker stylization strength than OmniConsistency [31] and Ovis-U1 [32] in some cases, which may contribute to the lower style-consistency scores.

In the reference-guided setting, our results differ substantially from the text-guided outputs, indicating that the model can learn style characteristics from the reference image. Meanwhile, BAGEL [5] suffers from severe reference-content leakage, making its style distribution closer to the original content distribution.

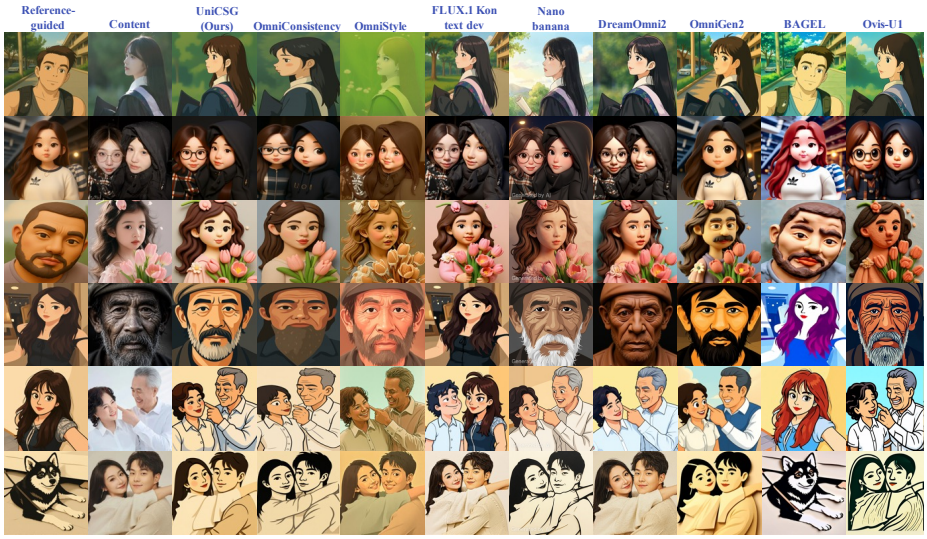


**Fig. 4:** Qualitative results for reference-guided style transfer on CSG-Bench.

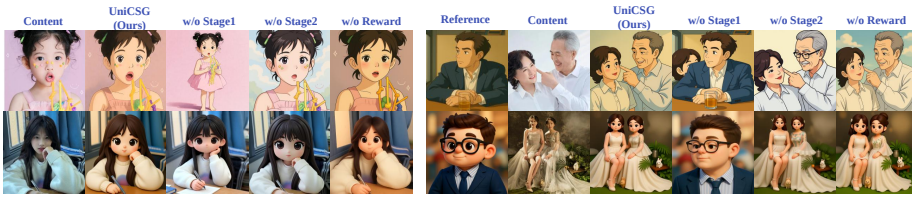


**Fig. 5:** Ablation study results. The first row shows the Ghibli style and the second row shows the 3D Chibi style. Left: text-guided setting. Right: reference-guided setting.

OmniStyle [34] and DreamOmni2 [40] exhibit transfer failures in some cases, producing results closer to the original content images. Compared with these baselines, UniCSG better preserves semantic identity and scene structure while reducing the unintended transfer of reference-specific content.

**User Study.** We conduct a user study to further demonstrate UniCSG's advantages. Thirty evaluators assess 48 image cases covering both text-guided and reference-guided tasks. For each case, evaluators select the best result among nine models in terms of content consistency and style consistency. This human evaluation complements automatic metrics, which may not fully reflect subtle structural drift or reference-content leakage under fine-grained inspection. The preference statistics are reported in Table 3, where UniCSG receives higher user preference in both settings.

**Table 3:** User preference rates(%) for content and style consistency across methods.

| Method | Text-guided | | Reference-guided | |
|---|---|---|---|---|
| | Content ↑ | Style ↑ | Content ↑ | Style ↑ |
| OmniConsistency [31] | 6.9 | 9.9 | 3.2 | 5.6 |
| OmniStyle [34] | 1.8 | 1.2 | 2.0 | 2.0 |
| flux1.Kontext-dev [15] | **26.6** | 22.0 | 11.5 | 7.1 |
| Nano-banana [4] | 25.8 | **22.8** | **30.4** | **24.2** |
| DreamOmni2 [40] | 4.0 | 3.6 | 11.3 | 8.7 |
| OmniGen2 [38] | 3.4 | 5.8 | 0.2 | 3.2 |
| BAGEL [5] | 0.6 | 1.8 | 0.0 | 0.8 |
| Ovis-U1 [32] | 1.4 | 2.8 | 3.2 | 3.8 |
| UniCSG (Ours) | **29.6** | **30.2** | **38.3** | **44.6** |

**Table 4:** Ablation results for text-guided style transfer on CSG-Bench.

| Method | Style consistency | | | Content consistency | | |
|---|---|---|---|---|---|---|
| | FID ↓ | CSD ↑ | CLIP-T ↑ | CLIP-I ↑ | DINO ↑ | DreamSim ↑ |
| UniCSG (Ours) | **113.940** | **0.541** | **0.267** | **0.797** | **0.701** | **0.816** |
| w/o Stage 1 | **121.855** | **0.524** | **0.258** | 0.748 | 0.590 | 0.763 |
| w/o Stage 2 | 124.718 | 0.516 | 0.255 | 0.755 | **0.630** | 0.770 |
| w/o Reward | 123.331 | 0.521 | 0.255 | **0.758** | 0.622 | **0.775** |

**Table 5:** Ablation results for reference-guided style transfer on CSG-Bench.

| Method | Style consistency | | | Content consistency | | |
|---|---|---|---|---|---|---|
| | FID ↓ | CSD ↑ | CLIP-T ↑ | CLIP-I ↑ | DINO ↑ | DreamSim ↑ |
| UniCSG (Ours) | **87.320** | **0.731** | **0.271** | **0.760** | **0.597** | **0.762** |
| w/o Stage 1 | **90.708** | 0.645 | **0.266** | 0.716 | 0.417 | 0.691 |
| w/o Stage 2 | 92.317 | 0.619 | 0.262 | 0.714 | **0.452** | 0.701 |
| w/o Reward | 92.335 | **0.661** | 0.264 | **0.723** | 0.451 | **0.706** |

**Ablation Study.** We visualize ablation trends in Fig. 5 and report detailed numbers in Tables 4 and 5. We conduct ablations on three components: Stage 1, Stage 2, and the reward model. Removing Stage 1 (low-frequency training with conditioning corruption) significantly degrades content-consistency metrics: in the text-guided setting, content becomes uncontrolled with hallucinations, while in the reference-guided setting, severe reference-content leakage occurs, demonstrating the importance of learning low-frequency structure for content constraints. Removing Stage 2 (frequency supervision) reduces style-consistency performance, including weaker geometric deformation and less accurate color


**Fig. 6:** Qualitative results on unseen styles during training.

**Table 6:** Quantitative results under seen and unseen settings on CSG-Bench.

| Task-Style setting | Style consistency | | | Content consistency | | |
|---|---|---|---|---|---|---|
| | FID ↓ | CSD ↑ | CLIP-T ↑ | CLIP-I ↑ | DINO ↑ | DreamSim ↑ |
| Text-Seen | 113.940 | 0.541 | 0.267 | 0.797 | 0.701 | 0.816 |
| Text-Unseen | 118.303 | 0.541 | 0.259 | 0.734 | 0.546 | 0.726 |
| Reference-Seen | 87.320 | 0.731 | 0.271 | 0.760 | 0.597 | 0.762 |
| Reference-Unseen | 113.532 | 0.701 | 0.257 | 0.728 | 0.536 | 0.718 |

tones, suggesting that high-frequency refinement improves stylization by recovering fine-grained textures and geometric cues. Removing the reward model yields a smaller but consistent drop, consistent with its role in perceptual alignment, particularly refining background and clothing details. Gains are larger in the reference-guided setting, indicating that UniCSG is particularly effective for reference-based stylization.

## 5 Discussion and Limitations

We discuss the generality and limitations of UniCSG from three perspectives: generalization, style transfer granularity, and efficiency.

**Generalization to unseen styles.** Our staged training explicitly separates semantic consistency (Stage 1) from texture stylization (Stage 2), and randomly sampling from the style pool during training exposes the model to diverse styles. Together, these design choices improve generalization to unseen style. Fig. 6 and


**Fig. 7:** Qualitative results across diverse style categories.

Table 6 show that performance degrades on unseen styles, which is expected because the unseen setting provides a harder transfer scenario without exposure to that style category during training. Nevertheless, UniCSG retains relatively strong content-related metrics, suggesting that it learns transferable content-preservation capability rather than merely overfitting to seen-style distributions.

**Image-level vs. category-level stylization.** Our style pool covers diverse artistic categories, including atmosphere and palette changes, stroke- and texture-driven rendering, and geometric or shape-deforming stylization. This diversity highlights a key distinction from image-level transfer, which is often closer to atmosphere or appearance translation tied to a single reference image. In contrast, UniCSG targets category-level stylization: beyond color tone and rendering patterns, it also learns transferable stylistic transformations shared across a style category, including geometric deformation and structural restylization, while preserving the content semantics of the source image. As a result, the method is better suited to content-constrained stylization settings where avoiding reference-specific leakage is as important as achieving strong visual style transfer. Fig. 7 illustrates representative results across different style categories.

**Efficiency.** UniCSG is built on the Qwen-Image-Edit-2509 backbone and further accelerated with Qwen-Image-Lightning, reducing per-image inference latency to approximately 4s for text-guided tasks and 10s for reference-guided tasks at 1024 resolution.

## 6 Conclusion

We propose UniCSG, a unified training framework for high-fidelity, content-constrained, style-driven generation. With a staged design semantic disentanglement, frequency-aware reconstruction, and pixel-space reward learning. UniCSG improves style-transfer stability and fine-detail quality while preserving content structure and substantially mitigating reference-content leakage. Future work will explore stronger cross-domain generalization and more efficient reward modeling to further improve controllable generation in complex scenes.

## References


1. Cao, S., Ma, N., Li, J., Li, X., Shao, L., Zhu, K., Zhou, Y., Pu, Y., Wu, J., Wang, J., Qu, B., Wang, W., Qiao, Y., Yao, D., Liu, Y.: Artimuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding (2025), https://arxiv.org/abs/2507.14533
2. Chen, B., Zhao, B., Xie, H., Cai, Y., Li, Q., Mao, X.: Consislora: Enhancing content and style consistency for lora-based style transfer. arXiv preprint arXiv:2503.10614 (2025) 2, 10
3. Chen, Z., Zhu, J., Chen, X., Zhang, J., Hu, X., Zhao, H., Wang, C., Yang, J., Tai, Y.: Dip: Taming diffusion models in pixel space (2025), https://arxiv.org/abs/2511.18822 4
4. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025), https://arxiv.org/abs/2507.06261 10, 13
5. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., Shi, G., Fan, H.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025) 10, 11, 13
6. Deng, Y., Tang, F., Dong, W., Ma, C., Pan, X., Wang, L., Xu, C.: Stytr2: Image style transfer with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11326–11336 (June 2022) 3
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021), https://arxiv.org/abs/2010.11929
8. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first International Conference on Machine Learning (2024), https://openreview.net/forum?id=FPnUhsQJ5B 2
9. Fu, S., Tamir, N.Y., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: learning new dimensions of human visual similarity using synthetic data. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (2023) 10
10. Gao, Z., Song, J., Zhang, Z., Deng, J., Patras, I.: Frequency-guided diffusion for training-free text-driven image translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19195–19205 (October 2025) 5, 8
11. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016) 2, 3
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6629–6640. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017) 10
13. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV (2017) 3

14. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024) 2
15. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742 10, 13
16. Lai, B., Wang, X., Rambhatla, S., Rehg, J.M., Kira, Z., Girdhar, R., Misra, I.: Toward diffusible high-dimensional latent spaces: A frequency perspective (2025), https://arxiv.org/abs/2511.22249 4
17. Lee, D.J., Hur, J., Choi, J., Yu, J., Kim, J.: Frequency-aware token reduction for efficient vision transformer (2025), https://arxiv.org/abs/2511.21477 4
18. Lei, M., Song, X., Zhu, B., Wang, H., Zhang, C.: Stylestudio: Text-driven style transfer with selective control of style elements. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 23443–23452 (June 2025) 2
19. Li, H., Fan, Z., Wen, Z., Zhu, Z., Li, Y.: Aicomposer: Any style and content image composition via feature integration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16840–16850 (October 2025) 2
20. Li, T., He, K.: Back to basics: Let denoising generative models denoise (2026), https://arxiv.org/abs/2511.13720 4
21. Liu, J., Zhang, Y., Huang, T., Xu, W., Yang, R.: Distilling cross-modal knowledge via feature disentanglement (2025), https://arxiv.org/abs/2511.19887 7
22. Ma, Z., Wei, L., Wang, S., Zhang, S., Tian, Q.: Deco: Frequency-decoupled pixel diffusion for end-to-end image generation (2025), https://arxiv.org/abs/2511.19365 2, 4
23. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=a68SUt6zFt, featured Certification 3, 10
24. Pan, Y., Feng, R., Dai, Q., Wang, Y., Lin, W., Guo, M., Luo, C., Zheng, N.: Semantics lead the way: Harmonizing semantic and texture modeling with asynchronous latent diffusion. arXiv preprint arXiv:2512.04926 (2025) 3
25. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4195–4205 (October 2023) 2
26. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., Sun, Y. (eds.) International Conference on Learning Representations. vol. 2024, pp. 1862–1874 (2024), https://proceedings.iclr.cc/paper_files/paper/2024/file/081b08068e4733ae3e7ad019fe8d172f-Paper-Conference.pdf 2
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/radford21a.html 6, 8, 10

28. Shang, C., Wang, Z., Wang, H., Meng, X.: Scsa: A plug-and-play semantic continuous-sparse attention for arbitrary semantic style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13051–13060 (June 2025) 3
29. Somepalli, G., Gupta, A., Gupta, K., Palta, S., Goldblum, M., Geiping, J., Shrivastava, A., Goldstein, T.: Measuring style similarity in diffusion models. arXiv preprint arXiv:2404.01292 (2024) 8, 10
30. Song, X., Wang, L., Wang, W., Li, Z., Sun, J., Zheng, D., Chen, J., Li, Q., Sun, Z.: 3sgen: Unified subject, style, and structure-driven image generation with adaptive task-specific memory (2025), https://arxiv.org/abs/2512.19271 10
31. Song, Y., Liu, C., Shou, M.Z.: Omniconsistency: Learning style-agnostic consistency from paired stylization data (2025), https://api.semanticscholar.org/CorpusID:278905729 9, 10, 11, 13
32. Wang, G.H., Zhao, S., Zhang, X., Cao, L., Zhan, P., Duan, L., Lu, S., Fu, M., Zhao, J., Li, Y., Chen, Q.G.: Ovis-u1 technical report. arXiv preprint arXiv:2506.23044 (2025) 10, 11, 13
33. Wang, X., Xu, W., Zhang, Q., Zheng, W.S.: Domain generalizable portrait style transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15802–15811 (October 2025) 4, 7
34. Wang, Y., Liu, R., Lin, J., Liu, F., Yi, Z., Wang, Y., Ma, R.: Omnistyle: Filtering high quality style transfer data at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7847–7856 (June 2025) 9, 10, 12, 13
35. Wang, Y., Yi, Z., Zhang, Y., Zheng, P., Xie, X., Lin, J., Wang, Y., Ma, R.: Omnistyle2: Scalable and high quality artistic style transfer data generation via destylization. ArXiv **abs/2509.05970** (2025), https://api.semanticscholar.org/CorpusID:286975228
36. Wang, Z., Wang, X., Xie, L., Qi, Z., Shan, Y., Wang, W., Luo, P.: StyleAdapter: A Unified Stylized Image Generation Model. International Journal of Computer Vision **133**(4), 1894–1911 (Apr 2025). https://doi.org/10.1007/s11263-024-02253-x, https://doi.org/10.1007/s11263-024-02253-x 2
37. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), https://arxiv.org/abs/2508.02324 4, 9, 10
38. Wu, C., Zheng, P., Yan, R., Xiao, S., Luo, X., Wang, Y., Li, W., Jiang, X., Liu, Y., Zhou, J., Liu, Z., Xia, Z., Li, C., Deng, H., Wang, J., Luo, K., Zhang, B., Lian, D., Wang, X., Wang, Z., Huang, T., Liu, Z.: Omnigen2: Exploration to advanced multimodal generation (2025), https://arxiv.org/abs/2506.18871 10, 13
39. Wu, S., Huang, M., Cheng, Y., Wu, W., Tian, J., Luo, Y., Ding, F., He, Q.: Uso: Unified style and subject-driven generation via disentangled and reward learning (2025) 2
40. Xia, B., Peng, B., Zhang, Y., Huang, J., Liu, J., Li, J., Tan, H., Wu, S., Wang, C., Wang, Y., Wu, X., Yu, B., Jia, J.: Dreamomni2: Multimodal instruction-based editing and generation (2025), https://arxiv.org/abs/2510.06679 10, 12, 13
41. Xiang, X., Tian, X., Zhang, G., Chen, Y., Zhang, S., Wang, X., Tao, X., Fan, Q.: Denoising vision transformer autoencoder with spectral self-regularization (2025), https://arxiv.org/abs/2511.12633 4, 7

42. Xing, P., Wang, H., Sun, Y., wangqixun, Baixu, Ai, H., Huang, J.Y., Li, Z.: CSGO: Content-style composition in text-to-image generation. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=n4nmiIq3qj 2
43. Yao, J., Song, Y., Zhou, Y., Wang, X.: Towards scalable pre-training of visual tokenizers for generation. arXiv preprint arXiv:2512.13687 (2025) 2, 3
44. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models (2023) 3
45. Yin, B., Hu, X., Zhou, X., Jiang, P.T., Liao, Y., Zhu, J., Zhang, J., Tai, Y., Wang, C., Yan, S.: Fera: Frequency-energy constrained routing for effective diffusion adaptation fine-tuning (2025), https://arxiv.org/abs/2511.17979 4
46. Yu, Y., Xiong, W., Nie, W., Sheng, Y., Liu, S., Luo, J.: Pixeldit: Pixel diffusion transformers for image generation (2025), https://arxiv.org/abs/2511.20645 4
47. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3836–3847 (October 2023) 3
48. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric (2018), https://arxiv.org/abs/1801.03924
49. Zhang, S.: Fast imagic: Solving overfitting in text-guided image editing via disentangled UNet with forgetting mechanism and unified vision-language optimization. In: Fumero, M., Domine, C., Lähner, Z., Crisostomi, D., Moschella, L., Stachenfeld, K. (eds.) Proceedings of UniReps: the Second Edition of the Workshop on Unifying Representations in Neural Models. Proceedings of Machine Learning Research, vol. 285, pp. 232–243. PMLR (14 Dec 2024), https://proceedings.mlr.press/v285/zhang24a.html 2
50. Zhang, S., Chen, Z., Chen, L., Wu, Y.: Cdst: Color disentangled style transfer for universal style reference customization (2025), https://arxiv.org/abs/2506.13770 2
51. Zhang, S., Huang, H., Zhang, C., Li, X.: Qwenstyle: Content-preserving style transfer with qwen-image-edit (2026), https://arxiv.org/abs/2601.06202 2
52. Zhang, S., Xiao, S., Huang, W.: Forgedit: Text guided image editing via learning and forgetting (2024), https://openreview.net/forum?id=nh4vQ1tGCt 2
53. Zhang, Z., Ma, A., Cao, K., Wang, J., Liu, S., Ma, Y., Cheng, B., Leng, D., Yin, Y.: U-stydit: Ultra-high quality artistic style transfer using diffusion transformers (2025), https://arxiv.org/abs/2503.08157 2