STRUCTURED 3D LATENTS ARE SURPRISINGLY POWERFUL: UNLEASHING GENERALIZABLE STYLE WITH 2D DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

3D asset generation plays a pivotal role in fields such as gaming and virtual reality, enabling the rapid synthesis of high-fidelity 3D objects from a single or multiple images. Building on this capability, enabling style-controllable generation naturally emerges as an important and desirable direction. However, existing approaches typically rely on style images that lie within or are similar to the training distribution of 3D generation models. When presented with out-of-distribution (OOD) styles, their performance degrades significantly or even fails. To address this limitation, we introduce **DiLAST**: 2D Diffusion-based Latent Awakening for 3D Style Transfer. Specifically, we leverage a pretrained 2D diffusion model as a teacher to provide rich and generalizable style priors. By aligning rendered views with the target style under diffusion-based guidance, our method optimizes the structured 3D latent representations for stylization. We observe that this limitation stems not from insufficient model capacity, but from the underutilization of structured 3D latents, which are inherently expressive. Despite being trained on comparatively limited data, 3D generation models can leverage 2D diffusion guidance to steer denoising toward specific directions in latent space, thereby producing diverse, OOD styles. Extensive experiments across diverse data and multiple 3D generation backbones demonstrate the effectiveness and plug-and-play nature of our approach.

1 INTRODUCTION

The rapid advancement of generative models has become an important component of progress in artificial intelligence (AI), giving rise to the paradigm of AI Generated Content (AIGC). In this field, 2D generative models, with diffusion models (Ho et al., 2020; Song et al., 2020a; Rombach et al., 2022) as a representative example, have achieved remarkable success in synthesizing high-quality and diverse images. Recently, 3D generative models have begun to emerge, aiming to produce realistic and controllable 3D assets for applications such as gaming, virtual reality, and digital content creation (Li et al., 2024b). Early 3D generative models (Chen et al., 2024; Poole et al., 2022; Liu et al., 2023b) still leverage the strong generative capabilities of 2D diffusion models, typically through techniques such as score distillation sampling (SDS) or view-conditioned fine-tuning, enabling the generation of 3D assets conditioned on user-provided text or images. More recent efforts have led to the development of true 3D generative models, such as Trellis (Xiang et al., 2025), UniLat (Wu et al., 2025), and Hunyuan3D (Zhao et al., 2025). Unlike earlier approaches that rely on 2D diffusion, these methods perform denoising directly in structured 3D latent spaces, from which the final 3D content is decoded. This leads to more faithful modeling of 3D geometry and appearance, marking a significant step toward native 3D generation.

Building on the aforementioned progress, a natural and practically important task is to enable flexible style manipulation of 3D assets, *i.e.*, transferring arbitrary styles onto 3D content, analogous to the well-studied problem of style transfer in 2D (Chung et al., 2024; Deng et al., 2024; Liu et al., 2025; Wang et al., 2023b; Xu et al., 2025; Zhou et al., 2025). Concretely, given a content image and a style image, the goal is to generate a 3D asset whose geometry is consistent with the content image while its appearance reflects the style image. This problem has attracted increasing attention. StyleSculptor (Qu et al., 2025) enables zero-shot style-guided 3D generation via style-disentangled attention

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

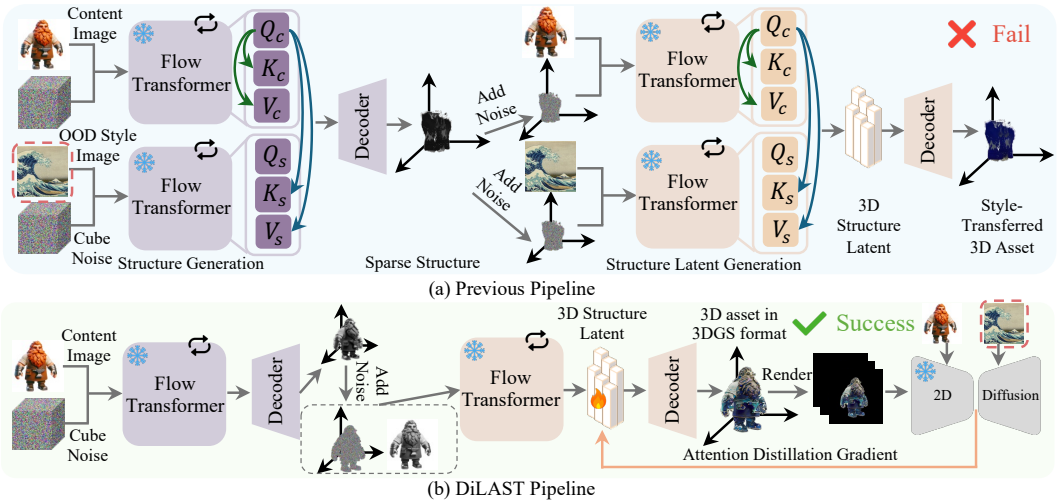


Figure 1: **Comparison of existing 3D style transfer methods and our DiLAST pipeline.** (a) Existing methods rely solely on internal attention of 3D generative models and often fail when handling OOD styles. (b) In contrast, DiLAST leverages a pretrained 2D diffusion teacher, whose attention distillation gradients guide the denoising trajectory of 3D latents, successfully transferring arbitrary OOD styles.

mechanisms, while MorphAny3D (Sun et al., 2026) leverages structured latent representations and attention-based feature blending for high-quality 3D morphing, with extensions to 3D style transfer. However, these methods are largely constrained by the intrinsic capabilities of the underlying 3D generation models. Consequently, the style images are typically required to lie *within or be close to the training distribution*. When given arbitrary out-of-distribution (OOD) styles, their performance degrades significantly or even fails. This limitation significantly restricts their applicability in real-world scenarios, where users often select style images in an unconstrained and arbitrary manner.

To tackle this limitation, we propose **DiLAST**, a framework that leverages 2D diffusion models to awaken the expressive capacity of structured 3D latent representations for style transfer. Specifically, 3D latents can be decoded into various 3D formats. In our implementation, we adopt a 3D Gaussian (3DGS) (Kerbl et al., 2023) decoder to produce 3D Gaussian primitives, which are then efficiently rendered into high-fidelity multi-view images. We then feed these rendered images, together with the content and style images, into a pretrained 2D diffusion model. By computing the gradients of the attention distillation loss with respect to the 3D latents, we guide its denoising trajectory, enabling effective style transfer while preserving the underlying content structure. We observe that, although the latent space of 3D generative models is relatively limited, it possesses substantial expressive potential. With appropriate guidance, the denoising process can be steered along specific directions to reach desired stylized latents, leading to the generation of corresponding 3D. Our contributions are summarized as:

- We **identify a key limitation** of existing 3D style transfer methods, namely their reliance on the intrinsic capabilities of underlying 3D generation models, which restricts them to in-distribution styles and leads to poor generalization to OOD inputs.
- We propose **DiLAST, a test-time optimization framework** that leverages pretrained 2D diffusion models to guide the denoising trajectory of structured 3D latents for style transfer, without modifying or retraining the underlying 3D generation models.
- We provide the **key insight** that structured 3D latent spaces are more expressive than commonly assumed, and can be effectively steered via diffusion guidance along specific denoising directions to achieve desired stylizations.
- We conduct extensive experiments across diverse data and multiple 3D generation backbones to demonstrate the effectiveness, strong generalization to OOD styles, and **plug-and-play** nature of our approach across different latent spaces.

2 Related Work

2.1 2D & 3D Generative Model

2D generation evolved from GANs to diffusion and flow matching. GAN (Goodfellow et al., 2014) enabled realistic synthesis through adversarial learning, while StyleGAN (Karras et al., 2019; 2020) improved controllability and fidelity. Diffusion became dominant with DDPM's (Ho et al., 2020) denoising formulation and score-based models (Song et al., 2020b) unifying diffusion and score matching. Imagen (Saharia et al., 2022), Stable Diffusion (Rombach et al., 2022), and SDXL (Podell et al., 2023) improved text-guided synthesis quality and scalability. Flow Matching (Lipman et al., 2022) offers an efficient diffusion alternative through continuous probability flows, while Rectified Flow (Liu et al., 2022) accelerates sampling with straighter trajectories.

3D generation likewise shifted from GAN-based methods (Chan et al., 2022; Deng et al., 2022; Gao et al., 2022; Zheng et al., 2022), whose representational limits constrained quality, to pretrained 2D diffusion priors. These support either SDS-based 3D optimization, pioneered by DreamFusion (Poole et al., 2022) and refined through improved distillation (Liang et al., 2024; Tang et al., 2023a; Lin et al., 2023; Tang et al., 2023b; Wang et al., 2023a), or multi-view synthesis followed by 3D reconstruction (Li et al., 2023; Liu et al., 2023b; 2024b; Long et al., 2024; Shi et al., 2023). Multi-view inconsistency degrades geometry and appearance; LRM (Hong et al., 2023) addresses this through feed-forward reconstruction learned from large-scale data, enabling consistent multi-view generation. Mirroring Stable Diffusion's shift from DDPM's pixel space to latent space, 3D latent generators (Ntavelis et al., 2023; Chen et al., 2025; Lan et al., 2025) improve training efficiency and generalization. Trellis (Xiang et al., 2025) further provides versatile structured 3D latents decodable into multiple asset formats, benefiting downstream tasks including style transfer.

2.2 2D & 3D Style Transfer

2D style transfer, a key generative-model application, includes four approaches. Image-encoder–based methods (Ye et al., 2023; Qi et al., 2024; Li et al., 2024a; Lei et al., 2025) combine explicit reference styles learned by dedicated encoders with text/image conditions. Image-driven methods (Wang et al., 2023b; Deng et al., 2024; Jones et al., 2024; Wang et al., 2025a;b) directly transfer reference appearance through feature matching or attention. Textual inversion (Gal et al., 2022; Zhang et al., 2023) encodes styles in learned tokens, invoked through the text encoder without modifying the backbone. LoRA-based methods (Shah et al., 2024; Liu et al., 2025; Frenkel et al., 2024; Yang et al., 2025) store styles in a small set of adapted low-rank weights rather than prompts or reference features.

In 3D, the widely used NeRF representation (Mildenhall et al., 2021) supports reference-image color and appearance transfer to scenes (Liu et al., 2023a; Fan et al., 2022). As 3DGS increasingly supersedes NeRF, StyleGaussian (Liu et al., 2024a) embeds and transforms 2D style features in Gaussians for real-time stylization through dedicated feature decoding, while StyleSplat (Jain et al., 2024) selectively fine-tunes segmented Gaussians for reference-guided local object stylization. StyleTex (Xie et al., 2024) and TEXTure (Richardson et al., 2023) transfer texture styles to given shapes. Building on recent 3D generators, StyleSculptor (Qu et al., 2025) achieves zero-shot, style-controllable generation through disentangled attention for texture and geometry guidance, while MorphAny3D (Sun et al., 2026) performs training-free style transfer by blending structured latent features through tailored attention within that space. Both remain limited on OOD styles, which our work addresses.

3 Method

3.1 Preliminaries

Latent Diffusion Model (LDM), with Stable Diffusion (SD) (Rombach et al., 2022) being the most representative example, has achieved state-of-the-art performance in 2D image and video generation. LDM incorporates cross-attention mechanisms to inject conditioning information, extending diffusion models into conditional generators. Specifically, an input image x is first encoded into a latent space through a pretrained VAE $\mathcal{E}(\cdot)$. A UNet denoising backbone $\epsilon_\theta(\cdot)$ is then trained to predict the injected noise during the diffusion process by minimizing the mean

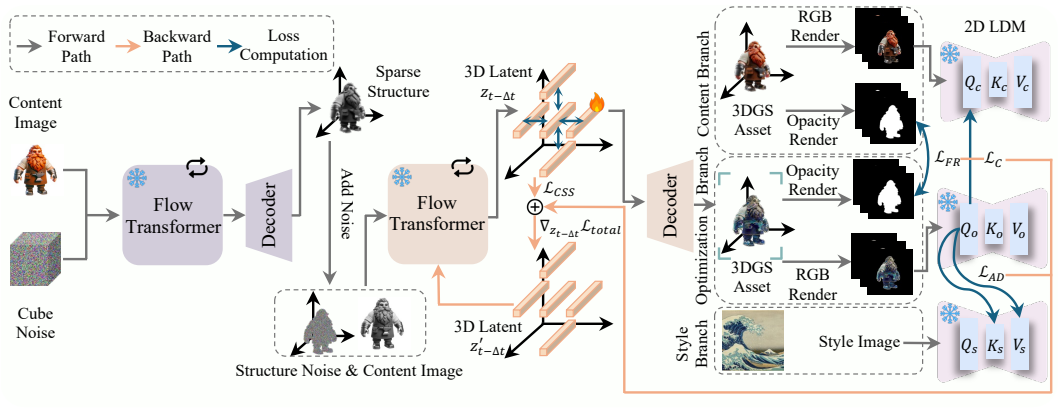


Figure 2: Overview of our DiLAST. We leverage a 2D LDM as a teacher and optimize the 3D latent through three parallel branches, guiding the generated 3D asset toward the target style. The 3D asset highlighted by the green bracket, generated at the final optimization step, is the final output.

squared error between the predicted noise and the added noise ϵ:

$$\mathcal{L}_{\text{LDM}} := \mathbb{E}_{\mathcal{E}(x),\, y,\, \epsilon\sim\mathcal{N}(0,1),\, t}\left[\|\epsilon - \epsilon_\theta(w_t, t, \tau_\theta(y))\|_2^2\right], \tag{1}$$

where y denotes the condition and t represents the timestep, w_t is the noisy latent at t, $\tau_\theta(\cdot)$ is an encoder that projects y to an intermediate representation. This conditioning representation is then embedded into intermediate UNet layers through cross-attention mechanisms:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{2}$$

where $Q = W_Q^{(i)}\varphi_i(z_t),\ K = W_K^{(i)}\tau_\theta(y),\ V = W_V^{(i)}\tau_\theta(y)$. i indexes the i-th cross-attention layer in the UNet, and $\varphi_i(z_t)$ denotes the intermediate latent feature at that layer derived from the noisy latent z_t. Note that for self-attention in LDM, the query, key, and value are all projected from the same latent feature, *i.e.*, $Q = W_Q^{(i)}\varphi_i(z_t)$, $K = W_K^{(i)}\varphi_i(z_t)$, and $V = W_V^{(i)}\varphi_i(z_t)$.

Trellis (Xiang et al., 2025) is a foundation model for 3D generation that can synthesize high-quality 3D assets in multiple formats (*e.g.* 3DGS, NeRF, Mesh) conditioned on user-provided text or images. It adopts a two-stage generation framework based on flow models. In the first stage, DINOv2 is first used to extract image features f, which serve as conditioning for both stages of generation. An initialized 3D latent noise grid s_t is then fed into a flow transformer $v_s(\cdot)$ to predict the velocity field:

$$s_{t-\Delta t} = s_t - v_s(f, s_t, t)\Delta t, \tag{3}$$

where Δt is the step length. At the final denoising step, the resulting s_0 is decoded into a sparse structure represented by active voxels. Each active voxel is injected with an initialized latent noise vector z_t, which is then passed into the flow transformer $v_l(\cdot)$ to predict the velocity field in the second stage:

$$z_{t-\Delta t} = z_t - v_l(f, z_t, t)\Delta t, \tag{4}$$

Finally, the resulting z_0 is decoded into the desired 3D asset in the target output format.

3.2 2D Diffusion-Guided Test-Time Optimization

Given a content image and a style image, our goal is to generate a 3D asset conditioned on the content image, while ensuring that its style is consistent with the style image. Previous works typically rely on a pretrained 3D generative model with separate content and style branches, where stylization is achieved through attention injection or feature recombination within the model. However, current 3D generative models are primarily designed for object-level generation. When arbitrary OOD style images are used to condition the style branch, the resulting distribution mismatch often causes misaligned query-key-value associations in attention layers, leading to unreliable style guidance and poor stylized 3D assets.

In contrast, 2D latent diffusion models are far less susceptible to this issue, as they are trained on much larger-scale and more diverse data, endowing them with stronger robustness to distribution shifts and more reliable query-key-value correspondences in attention layers. However, since the latent spaces of 2D diffusion and 3D generative models are not aligned, directly performing attention injection across them is infeasible. Motivated by this, we adopt the *attention distillation* mechanism, shifting style guidance from the 3D generator to a pretrained 2D latent diffusion teacher, thereby confining all attention interactions to the 2D LDM.

Since 2D latent diffusion models take images as input, they naturally serve as a reliable bridge between 2D and 3D models. Leveraging the versatile structured 3D latents provided by Trellis, we adopt the 3DGS decoder to obtain Gaussian primitives representing the generated 3D asset, due to its favorable properties, including efficient rendering, high-fidelity synthesis, and differentiable rasterization:

$$\mathcal{D}_{GS} : \{(z_i, p_i)\}_{i=1}^{L} \mapsto \left\{ \{(o_i^k, c_i^k, s_i^k, \alpha_i^k, r_i^k)\}_{k=1}^{K} \right\}_{i=1}^{L}, \tag{5}$$

where $\mathcal{D}_{GS}$ denotes the 3D Gaussian decoder, p_i represents the i-th active voxel, z_i is the latent vector associated with that voxel, and L is the number of active voxels. Each active voxel is decoded into K Gaussians, parameterized by $(o_i^k, c_i^k, s_i^k, \alpha_i^k, r_i^k)$, where o_i^k denotes position, c_i^k color, s_i^k scale, α_i^k opacity, and r_i^k rotation. Then, given a specific camera pose π, the color $I_\pi(x, y)$ at each pixel (x, y) in the rendered 2D image is computed through alpha-blending over the set of Gaussians $\mathcal{N}$ that are visible under the view and contribute to the pixel due to spatial overlap:

$$I_\pi(x, y) = \sum_{u \in \mathcal{N}} c_u \alpha_u \prod_{v=1}^{u-1} (1 - \alpha_v). \tag{6}$$

After obtaining the rendered images, we initialize three parallel branches in the same 2D LDM. The first is a *content branch*, whose input consists of rendered views from a 3D asset generated by Trellis in 3DGS format, conditioned on the content image. The second is a *style branch*, which directly takes the style image as input. The third is an *optimization branch*, whose input consists of rendered views decoded from the 3D asset at a given timestep during optimization. For this branch, we preserve the first-stage sparse structure generated by Trellis under the content image condition. In the second stage, we instead use the grayscale content image as conditioning to reduce structural drift while preventing the original content style from biasing the learning of the target style. This design is well motivated, as style transfer should not induce large structural changes to the underlying 3D asset. Moreover, stylization in the 3DGS representation modifies Gaussian parameters during optimization, which is sufficiently expressive to capture both geometric and appearance-level style variations. With the three branches defined, the attention distillation loss can be formulated as follows:

$$\mathcal{L}_{AD} = \left\| \mathrm{SelfAttn}(Q_o, K_o, V_o) - \mathrm{SelfAttn}(Q_o, K_s, V_s) \right\|_1, \tag{7}$$

where Q_o, K_o, and V_o denote the query, key, and value from the optimization branch, while K_s and V_s denote the key and value from the style branch. At a high level, this objective encourages the optimization branch to produce attention responses that match those obtained when querying the style branch. Intuitively, once querying the style branch yields the same response as querying itself, the optimization branch has acquired the desired style characteristics. For content preservation, we adopt the content loss proposed in (Zhou et al., 2025), defined as follows:

$$\mathcal{L}_C = \left\| Q_o - Q_c \right\|_1, \tag{8}$$

where Q_c is the query from the content branch. With the attention distillation loss and content preservation loss defined, we compute the gradient of the total loss with respect to the 3D latents and use it to steer the optimization trajectory during denoising:

$$z_{t-\Delta t} := z_{t-\Delta t} - \eta \nabla_{z_{t-\Delta t}} \mathcal{L}_{total}, \tag{9}$$

where η controls the guidance strength. $\mathcal{L}_{total} = \mathcal{L}_{AD} + \lambda \mathcal{L}_C$, where λ is the weight of content loss. More specifically, we do not modify the parameters of the flow transformer, since this would complicate optimization and potentially degrade velocity estimation. Instead, the transformer first predicts a standard flow update from z_t to $z_{t-\Delta t}$, after which the gradient of the objective is used to steer the direction of this update. The guided latent is then fed back into the flow transformer for the next prediction, iteratively refining the denoising trajectory until the final latent z_0 is reached.

3.3 FLOATER REMOVAL

The Gaussian parameters are decoded from the optimized 3D latent representations. During style-driven optimization, without additional regularization, some parameters, including positions, scales, and rotations, may undergo excessive deviations. In 3D space, these deviations manifest as floating Gaussians (floaters), producing artifacts that degrade the quality of the generated 3D asset. Inspired by (Hakie et al., 2025; Rogge & Stricker, 2025), we introduce a floater removal loss to regularize such undesirable deviations:

$$\mathcal{L}_{FR} = \frac{1}{NHW} \sum_{n=1}^{N} \sum_{h=1}^{H} \sum_{w=1}^{W} \left\| A^{(n)}(h,w) - M^{(n)}(h,w) \right\|_1, \tag{10}$$

where N denotes the number of rendered views, and H and W denote the height and width of each image, respectively. A denotes the alpha compositing map rendered from the currently optimized 3D asset, where each pixel accumulates the opacity contributions of visible Gaussians. M denotes the corresponding reference alpha map rendered from the content-conditioned 3D asset, serving as a silhouette mask.

3.4 COLOR SPECKLE SUPPRESSION

During style-driven optimization, without proper regularization, Gaussian color parameters may drift beyond a stable range, after which they become clipped by the rendering process. In RGB space, this often manifests as saturated extreme values (*e.g.*, near 0 or 255 in individual channels), producing highly saturated spurious color patches that degrade visual quality. Inspired by (Fridovich-Keil et al., 2022), we introduce a color speckle suppression loss to regularize such undesirable color deviations:

$$\mathcal{L}_{\text{CSS}} = \frac{1}{|\Gamma|} \sum_{d \in \Gamma} \frac{1}{|\Omega_d|} \sum_{(i,j) \in \Omega_d} \sqrt{\|z_i - z_j\|_2^2}, \tag{11}$$

where $\Gamma \subseteq \{l, m, n\}$ denotes the set of valid spatial directions containing at least one neighboring voxel pair, and $|\Gamma|$ is the number of such directions. For each direction d, Ω_d denotes the set of adjacent active voxel pairs along direction d, with cardinality $|\Omega_d|$. Specifically, for voxel coordinates $p_i = (l_i, m_i, n_i)$, a pair $(i, j) \in \Omega_d$ satisfies $p_j = p_i + \delta_d$, where $\delta_l = (1, 0, 0), \delta_m = (0, 1, 0), \delta_n = (0, 0, 1)$. This loss promotes local smoothness in the structured 3D latent field by penalizing large discrepancies between neighboring voxels, thereby discouraging abrupt variations in latent features and suppressing spurious artifacts.

3.5 OVERALL GUIDANCE

Based on the preceding subsections, the total loss for optimizing the 3D latents is redefined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{AD} + \lambda_1 \mathcal{L}_C + \lambda_2 \mathcal{L}_{FR} + \lambda_3 \mathcal{L}_{CSS}, \tag{12}$$

where λ_1, λ_2, and λ_3 are the weighting coefficients associated with the corresponding loss terms, respectively. Note that Trellis is used here only as an illustrative example to present our framework, rather than a restriction of its applicability. Our method is inherently plug-and-play and can be integrated with other 3D generation backbones, as will be validated in the experimental section.

4 EXPERIMENTS

In this section, we conduct extensive experiments to answer the following research questions: **RQ1:** Can DiLAST achieve high-quality 3D style transfer across diverse styles? **RQ2:** How does DiLAST compare with existing methods? **RQ3:** What is the contribution of each loss component (ablation study)? **RQ4:** Does our method generalize effectively to other 3D generation backbones?

4.1 EXPERIMENT SETTINGS

Datasets & Baselines. For evaluation, we sample content images from the TRELLIS-500K (Xiang et al., 2025) dataset and style images from InstantStyle-Plus (Wang et al., 2024). Since style transfer in structured 3D latent based 3D generative models remains relatively underexplored, we select two

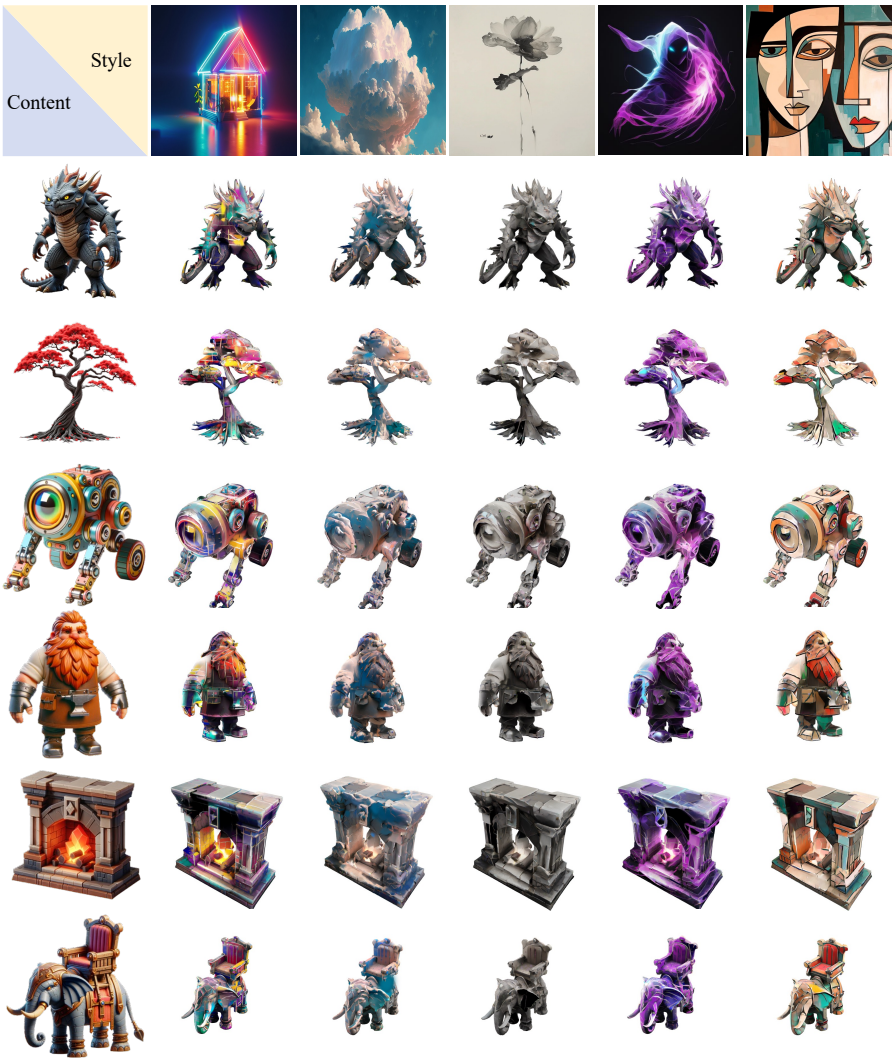


Figure 3: Qualitative results of style transfer using our method.

representative state-of-the-art methods, MorphAny3D (Sun et al., 2026) and StyleSculptor (Qu et al., 2025), as baselines.

Metric. We first adopt HPSv3 (Human Preference Score v3) (Ma et al., 2025), a learned metric that predicts human preference scores by aligning model outputs with human aesthetic and semantic judgments. As HPSv3 requires a text prompt, we design two types of prompts for evaluation. The first is a generic prompt ("a stylized object"), which minimizes potential bias introduced by prompt semantics. The second is a specific prompt: for each style image, we use a vision-language model (VLM) (OpenAI, 2026) to generate a textual description of its style in the form of "¡style name¿ style." Based on this description, we then construct the final prompt as "a ¡object name¿ in ¡style name¿ style." This design allows us to better assess style consistency between the generated 3D assets and the target styles. We further introduce an aesthetic score by leveraging VLMs to assess the visual quality of rendered multi-view images from stylized 3D assets across multiple perceptual dimensions. To improve reliability, we incorporate both closed-source (OpenAI, 2026) and open-source (Bai et al., 2025) VLMs. We also conducted a user preference study with 10 volunteers using the same evaluation criteria. Detailed prompts can be found in Appendix A.

Implementation Details. All experiments are conducted on NVIDIA A100 GPUs. We use TRELLIS-image-large (Xiang et al., 2025) as the 3D generation model and Stable Diffusion v1-5 as the 2D latent diffusion model. All model parameters are frozen during optimization. For the 2D LDM, we use an empty text condition, since our framework does not rely on text prompts.

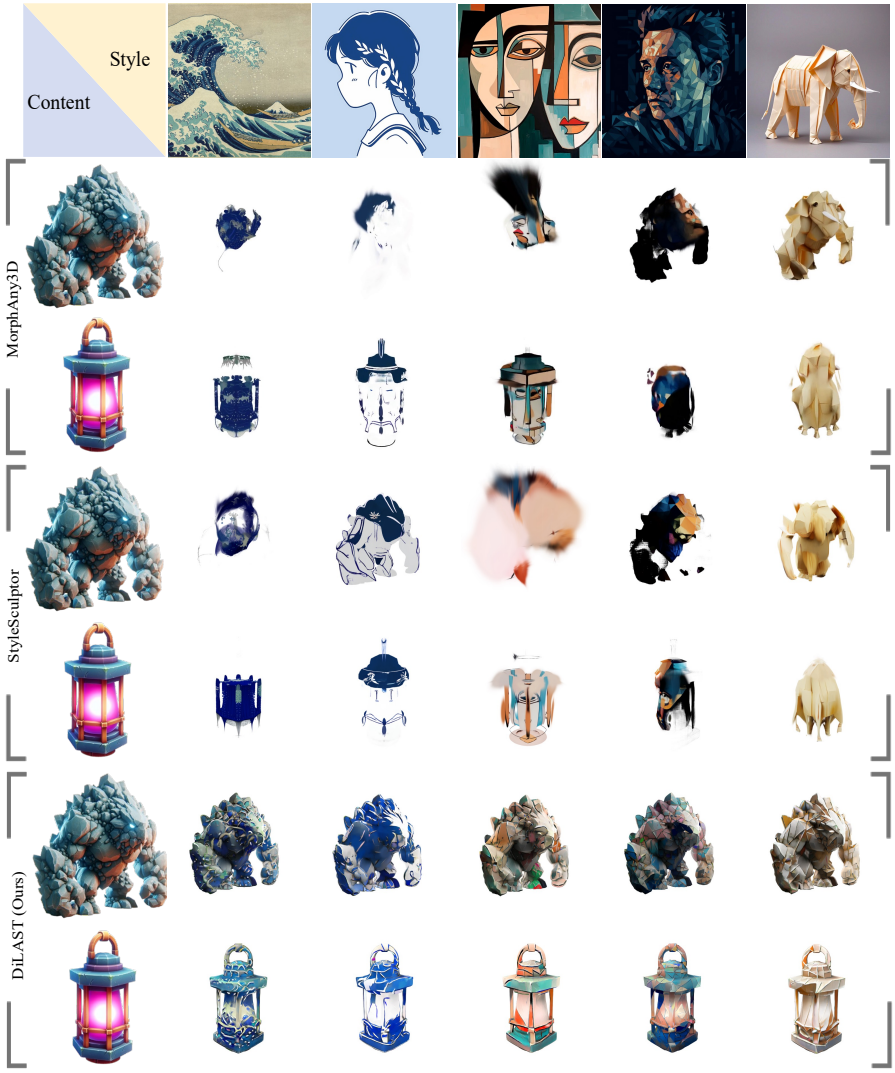


Figure 4: Qualitative comparison of style transfer results across different methods.

Unless otherwise specified, we set the loss weights to $\lambda_1 = 0.2$, $\lambda_2 = 20$, and $\lambda_3 = 10$. The number of sampling steps for the flow transformer is set to 500 to allow fine-grained control over the denoising trajectory, and the guidance strength η is set to 5. To reduce memory consumption and improve efficiency, we use rendered images from 5 fixed viewpoints at each optimization step. More implementation details are provided in Appendix B.

4.2 Experiment Results

Effectiveness of DiLAST (RQ1). We present the qualitative results in Figure 3. Due to space limitations, we only show a subset of content-style combinations. More results are provided in Appendix C. As shown in the figure, DiLAST successfully transfers diverse styles to multiple 3D assets while preserving the underlying content structure. Since we use 3DGS as the 3D asset representation, the textures and brushstroke patterns of the style images can still be effectively reflected even without modifying the first-stage sparse

Table 1: Quantitative comparison. The best results are in **bold**. GP: general prompt, SP: specific prompt, AS denotes aesthetic score (averaged on two VLMs), and UP denotes user preference.

Method	HPSv3 (GP)↑	HPSv3 (SP)↑	AS (%)↑	UP (%)↑
MorphAny3D	-4.5784	-4.7300	1.43	0.00
StyleSculptor	0.7625	-0.7858	14.29	10.00
DiLAST (Ours)	**3.0264**	**3.3314**	**84.28**	**90.00**

structure. This is because changes in the optimized 3D latents induce corresponding updates to the parameters of the 3D Gaussian primitives.

Comparison between DiLAST and Baseline Methods (RQ2). We comprehensively compare DiLAST with the baseline methods through both quantitative and qualitative results. Table 1 reports the quantitative results, where our method outperforms the baselines across all metrics, especially on AS and UP, which provide more comprehensive evaluations. For HPSv3, our method achieves the best performance under both the generic and specific prompts. Notably, when switching from the generic prompt to the specific prompt, only our method shows an improvement, indicating that DiLAST receives higher scores under more fine-grained prompts. This further demonstrates that the stylized 3D assets generated by our method achieve the best balance between content preservation, reflected by the object description in the prompt, and style transfer, reflected by the style description in the prompt. Figure 4 shows the qualitative results. As can be observed, the stylized 3D assets generated by the baseline methods contain noticeable artifacts or even fail completely. In contrast, our method effectively handles OOD-style images and achieves a better balance between content preservation and style transfer.

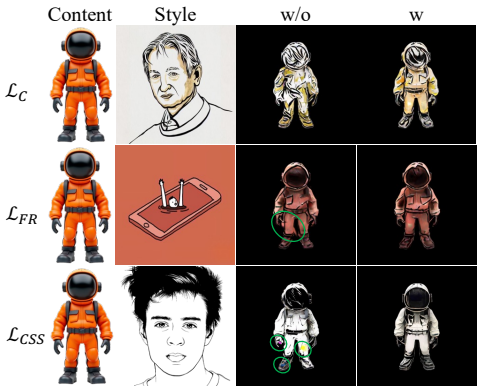


Figure 5: Contribution of Each Loss Component. "w/o" denotes removing the corresponding loss term, while "w" denotes including it.

Ablation Study (RQ3). Figure 5 illustrates the contribution of each loss term. The first row shows the effect of $\mathcal{L}_C$. Without $\mathcal{L}_C$, severe content leakage from the style image is observed, whereas adding this loss effectively alleviates the issue. The second row shows the effect of $\mathcal{L}_{FR}$. Without $\mathcal{L}_{FR}$, floating Gaussians appear in the regions highlighted by the green circles; after adding this constraint, these floaters are suppressed. The third row shows the effect of $\mathcal{L}_{CSS}$. Without $\mathcal{L}_{CSS}$, especially when using light-colored style images, many color artifacts appear on the generated 3D assets. The RGB values of the artifacts highlighted by the green circles are (255,0,255) and (255,255,0), indicating clear overflow and clipping behavior in the 3DGS decoder. Adding this loss term mitigates these artifacts.

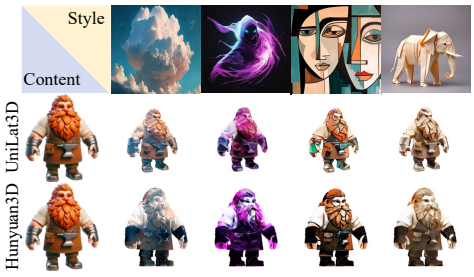


Figure 6: Qualitative results by using other 3D generative models.

Plug-and-Play (RQ4). In this paper, we mainly use Trellis as the 3D generation model. However, this does not mean that our method is limited to Trellis or cannot be adapted to other 3D latent based 3D generation models. In this section, we select two representative 3D generation models, UniLat3D (Wu et al., 2025) and Hunyuan3D (Zhao et al., 2025), and combine them with our method to generate stylized 3D assets. Unlike Trellis, UniLat3D adopts a single-stage generation process, while Hunyuan3D first decodes the 3D latents into multi-view images and then bakes them onto the mesh structure generated in the first stage. As shown in Figure 6, our method exhibits a strong plug-and-play property. This demonstrates not only the flexibility of our framework, but also that 3D latents in other latent spaces likewise possess strong potential for stylization.

5 Conclusion

In this paper, we propose DiLAST, a test-time optimization framework that leverages 2D diffusion guidance to awaken structured 3D latents for generalizable style transfer. By steering the denoising trajectory with attention distillation, DiLAST transfers out-of-distribution styles while preserving content structure. We highlight the untapped potential of 3D latent spaces and provide a flexible solution for OOD 3D stylization. In the future, we hope to extend this framework to broader 3D editing tasks beyond style transfer.

AI USE STATEMENT

Generative AI tools were used to refine the language, readability, and presentation of this manuscript and to support VLM-based evaluation of rendered results according to predefined criteria. These tools were not used to formulate the research problem, design DILAST, implement the system, or produce the method outputs evaluated in our experiments. The authors reviewed and revised all AI-assisted text and manually checked the evaluation procedures and outputs for consistency and correctness. We assume full responsibility for the final manuscript and all associated claims, results, and artifacts, including those prepared with generative AI assistance.

ETHICS STATEMENT

Human participation was limited to a user preference study comparing rendered results from different methods. All participants took part voluntarily and were appropriately compensated for their time. The study required only visual assessments of rendered scenes and presented no foreseeable risks to participants. No personally identifiable information is disclosed, and participant responses are reported exclusively in aggregate form.

REPRODUCIBILITY STATEMENT

The main paper describes the overall framework and key components of DILAST. The appendices and supplementary materials provide further implementation details, hyperparameter settings, evaluation procedures, and qualitative results to facilitate reproduction of our experiments and findings. Upon acceptance, we will make the source code, configuration files, and evaluation scripts publicly available.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16123–16133, 2022.

Yongwei Chen, Tengfei Wang, Tong Wu, Xingang Pan, Kui Jia, and Ziwei Liu. Comboverse: Compositional 3d assets creation using spatially-aware diffusion guidance. In *European Conference on Computer Vision*, pp. 128–146. Springer, 2024.

Zhaoxi Chen, Jiaxiang Tang, Yuhao Dong, Ziang Cao, Fangzhou Hong, Yushi Lan, Tengfei Wang, Haozhe Xie, Tong Wu, Shunsuke Saito, et al. 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26576–26586, 2025.

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*, pp. 8795–8805, 2024.

Yingying Deng, Xiangyu He, Fan Tang, and Weiming Dong. Z*: Zero-shot style transfer via attention reweighting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6934–6944, 2024.

Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10673–10683, 2022.

Zhiwen Fan, Yifan Jiang, Peihao Wang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Unified implicit neural stylization. In *European conference on computer vision*, pp. 636–654. Springer, 2022.

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

Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5501–5510, 2022.

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

Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *Advances in neural information processing systems*, 35:31841–31854, 2022.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Aly El Hakie, Yiren Lu, Yu Yin, Michael Jenkins, and Yehe Liu. Fix false transparency by noise guided splatting. *arXiv preprint arXiv:2510.15736*, 2025.

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

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023.

Sahil Jain, Avik Kuthiala, Prabhdeep Singh Sethi, and Prakanshul Saxena. Stylesplat: 3d object style transfer with gaussian splatting. *arXiv preprint arXiv:2407.09473*, 2024.

Maxwell Jones, Sheng-Yu Wang, Nupur Kumari, David Bau, and Jun-Yan Zhu. Customizing text-to-image models with a single image pair. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–13, 2024.

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*, pp. 4401–4410, 2019.

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*, pp. 8110–8119, 2020.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Yushi Lan, Fangzhou Hong, Shangchen Zhou, Shuai Yang, Xuyi Meng, Yongwei Chen, Zhaoyang Lyu, Bo Dai, Xingang Pan, and Chen Change Loy. Ln3diff++: Scalable latent neural fields diffusion for speedy 3d generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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*, pp. 23443–23452, 2025.

Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023.

Wen Li, Muyuan Fang, Cheng Zou, Biao Gong, Ruobing Zheng, Meng Wang, Jingdong Chen, and Ming Yang. Styletokenizer: Defining image style by a single instance for controlling diffusion models. In *European Conference on Computer Vision*, pp. 110–126. Springer, 2024a.

Xiaoyu Li, Qi Zhang, Di Kang, Weihao Cheng, Yiming Gao, Jingbo Zhang, Zhihao Liang, Jing Liao, Yan-Pei Cao, and Ying Shan. Advances in 3d generation: A survey. *arXiv preprint arXiv:2401.17807*, 2024b.

Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6517–6526, 2024.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 300–309, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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

Kunhao Liu, Fangneng Zhan, Yiwen Chen, Jiahui Zhang, Yingchen Yu, Abdulmotaleb El Saddik, Shijian Lu, and Eric P Xing. Stylerf: Zero-shot 3d style transfer of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8338–8348, 2023a.

Kunhao Liu, Fangneng Zhan, Muyu Xu, Christian Theobalt, Ling Shao, and Shijian Lu. Stylegaussian: Instant 3d style transfer with gaussian splatting. In *SIGGRAPH Asia 2024 Technical Communications*, pp. 1–4. 2024a.

Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10072–10083, 2024b.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9298–9309, 2023b.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9970–9980, 2024.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15086–15095, 2025.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Evangelos Ntavelis, Aliaksandr Siarohin, Kyle Olszewski, Chaoyang Wang, Luc V Gool, and Sergey Tulyakov. Autodecoding latent 3d diffusion models. *Advances in Neural Information Processing Systems*, 36:67021–67047, 2023.

OpenAI. Chatgpt-5.5. `https://chatgpt.com/`, 2026. Large language model.

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. *arXiv preprint arXiv:2307.01952*, 2023.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Tianhao Qi, Shancheng Fang, Yanze Wu, Hongtao Xie, Jiawei Liu, Lang Chen, Qian He, and Yongdong Zhang. Deadiff: An efficient stylization diffusion model with disentangled representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8693–8702, 2024.

Zefan Qu, Zhenwei Wang, Haoyuan Wang, Ke Xu, Gerhard Petrus Hancke, and Rynson WH Lau. Stylesculptor: Zero-shot style-controllable 3d asset generation with texture-geometry dual guidance. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH 2023 conference proceedings*, pp. 1–11, 2023.

Marcel Rogge and Didier Stricker. Object-centric 2d gaussian splatting: Background removal and occlusion-aware pruning for compact object models. *arXiv preprint arXiv:2501.08174*, 2025.

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*, pp. 10684–10695, 2022.

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.

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

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

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

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*, 2020b.

Xiaokun Sun, Zeyu Cai, Hao Tang, Ying Tai, Jian Yang, and Zhenyu Zhang. Morphany3d: Unleashing the power of structured latent in 3d morphing. *arXiv preprint arXiv:2601.00204*, 2026.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023a.

Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 22819–22829, 2023b.

Haofan Wang, Peng Xing, Renyuan Huang, Hao Ai, Qixun Wang, and Xu Bai. Instantstyle-plus: Style transfer with content-preserving in text-to-image generation. *arXiv preprint arXiv:2407.00788*, 2024.

Ye Wang, Tongyuan Bai, Xuping Xie, Zili Yi, Yilin Wang, and Rui Ma. Sigstyle: Signature style transfer via personalized text-to-image models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8051–8059, 2025a.

Ye Wang, Ruiqi Liu, Jiang Lin, Fei Liu, Zili Yi, Yilin Wang, and Rui Ma. Omnistyle: Filtering high quality style transfer data at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7847–7856, 2025b.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023a.

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*, pp. 7677–7689, 2023b.

Guanjun Wu, Jiemin Fang, Chen Yang, Sikuang Li, Taoran Yi, Jia Lu, Zanwei Zhou, Jiazhong Cen, Lingxi Xie, Xiaopeng Zhang, et al. Unilat3d: Geometry-appearance unified latents for single-stage 3d generation. *arXiv preprint arXiv:2509.25079*, 2025.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21469–21480, 2025.

Zhiyu Xie, Yuqing Zhang, Xiangjun Tang, Yiqian Wu, Dehan Chen, Gongsheng Li, and Xiaogang Jin. Styletex: Style image-guided texture generation for 3d models. *ACM Transactions on Graphics (TOG)*, 43(6):1–14, 2024.

Ruojun Xu, Weijie Xi, XiaoDi Wang, Yongbo Mao, and Zach Cheng. Stylessp: Sampling startpoint enhancement for training-free diffusion-based method for style transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18260–18269, 2025.

Jiahui Yang, Yongjia Ma, Donglin Di, Jianxun Cui, Hao Li, Wei Chen, Yan Xie, Xun Yang, and Wangmeng Zuo. Qr-lora: Efficient and disentangled fine-tuning via qr decomposition for customized generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17587–17597, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

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*, pp. 10146–10156, 2023.

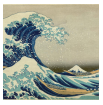
Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

Xinyang Zheng, Yang Liu, Pengshuai Wang, and Xin Tong. Sdf-stylegan: implicit sdf-based stylegan for 3d shape generation. In *Computer Graphics Forum*, volume 41, pp. 52–63. Wiley Online Library, 2022.

Yang Zhou, Xu Gao, Zichong Chen, and Hui Huang. Attention distillation: A unified approach to visual characteristics transfer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18270–18280, 2025.

A Additional Metric Details

Table 2: Style images and corresponding style names.

Style Image	Style Name
	Neon cyberpunk glow style
	Soft atmospheric cloudscape painting style
	Geometric cubist surrealism style
	Monochrome ink wash minimalist style
	Ethereal neon fantasy illustration style
	Japanese ukiyo-e woodblock print style
	Minimal blue line-art anime style
	Modern cubist portrait style
	Low-poly geometric illustration style
	Origami paper sculpture style

To compute the HPSv3 score under the specific prompt, we first use a VLM to assign a style name to each style image. The results are shown in Table 2. When using VLMs to compute the aesthetic score, we use the following prompt:

You are a professional art-style evaluator with strong visual judgment. Your task is to assess the quality of stylized 3D assets generated by different methods.

For each evaluation case, I will provide:

- *1. A content image, which defines the original object, structure, geometry, and semantic identity.*
- *2. A style image, which defines the target artistic appearance, including color palette, texture, brushwork, material impression, and overall visual mood.*
- *3. Rendered images of the generated 3D assets from five different viewpoints.*
- *4. Results from three different methods.*

You must compare the three generated 3D assets and make a single-choice decision: select the best one among the three methods. Please base your judgment on the following criteria:

- *1. Overall visual impression: Evaluate the immediate visual appeal and impact of the asset at first glance. Consider whether the asset looks artistically harmonious, visually coherent, structurally complete, and aesthetically pleasing across different views.*
- *2. Content and geometry preservation: Compare the stylized 3D asset with the content image. Assess whether the generated asset preserves the original object identity, global structure, local geometry, proportions, and important semantic details.*
- *3. Style consistency and transfer strength: Compare the stylized 3D asset with the style image. Assess how well the asset reflects the target artistic style, including colors, textures, patterns, material appearance, brush strokes, and overall artistic atmosphere. Also consider whether the style transfer is sufficiently strong without destroying the content structure.*
- *4. Realism and rendering quality: Evaluate whether the generated 3D asset looks natural, realistic, and high-quality. Consider the quality of materials, surface details, lighting consistency, absence of artifacts, and whether the asset remains visually plausible across the five rendered viewpoints.*

After considering all four criteria, choose the single best result among the three methods. Your decision should be based only on your visual and artistic perception. You do not need to rely on deeper technical knowledge about 3D generation, rendering, or style transfer methods.

Please output only the final choice in the following format: Best Method: [A/B/C].

For the user preference study, we used the same evaluation criteria and instructions as those provided to the VLM. The study was conducted in the form of a questionnaire and posed no risk to participants.

B Additional Implementation Details

Implementation for 2D LDM. In the attention distillation module, self-attention features are extracted from the 10-th and 16-th layers of Stable Diffusion. To better leverage the attention features of Stable Diffusion and align with the diffusion mechanism, we first encode all input images into latent representations using the VAE encoder, and then add random noise corresponding to 200 diffusion steps to these latents by following this equation:

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \tag{13}$$

where z_0 denotes the original clean latent representation, z_t denotes the noisy latent representation at diffusion timestep t, ϵ is Gaussian noise sampled from the standard normal distribution $\mathcal{N}(0, \mathbf{I})$, and $\bar{\alpha}_t$ is the cumulative noise schedule coefficient defined as

$$\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s, \qquad \alpha_s = 1 - \beta_s, \tag{14}$$

where β_s controls the noise variance added at timestep s.

Implementation for Rendered Images. At each optimization step, we render the 3D asset in 3DGS format into images, which are then fed into the 2D LDM. First, following Trellis, we define a camera trajectory using the following equation:

$$\begin{aligned} \theta_i &= \frac{2\pi i}{N-1}, \\ \phi_i &= 0.25 + 0.5 \sin\left(\frac{2\pi i}{N-1}\right), \\ \rho_i &= 0, \qquad i \in \{0, \ldots, N-1\}. \end{aligned} \tag{15}$$

$$\mathbf{c}_i = r \begin{bmatrix} \sin(\theta_i)\cos(\phi_i) \\ \cos(\theta_i)\cos(\phi_i) \\ \sin(\phi_i) \end{bmatrix}, \qquad r = 2. \tag{16}$$

where N denotes the total number of rendered views, and i is the view index. θ_i, ϕ_i, and ρ_i denote the yaw, pitch, and roll angles of the i-th camera view, respectively. The yaw angle θ_i changes linearly from 0 to 2π, forming a full circular trajectory around the object. The pitch angle ϕ_i varies sinusoidally, allowing the camera to observe the object from smoothly changing elevations, while the roll angle ρ_i is fixed to zero for all views. The camera center $\mathbf{c}_i$ is placed on a sphere with radius $r = 2$ around the origin. We render images from 300 viewpoints and then uniformly select 5 views to feed into the 2D LDM, reducing memory consumption and improving optimization efficiency. Moreover, the resolution of all rendered images is 512×512.

Implementation for Overall Optimization. Both the Floater Removal loss and the Color Speckle Suppression loss impose a certain degree of smoothing on the 3D latents, which may hinder style transfer. To avoid interfering with the early guidance of attention distillation on the 3D latents, we introduce these two losses only after 400 sampling steps.

C Additional Results.

C.1 Supplementary Qualitative and Quantitative Results

Since our motivation is to enrich the diversity of 3D asset generation through 3D style transfer, and since 3D assets are primarily used in applications such as gaming and animation, we mainly select content images with artistic styles as generation conditions in the main paper. In this subsection, we select five real-world objects and provide the corresponding style transfer results. Additional experimental results that cannot be included in the main paper due to space limitations are also presented in this subsection. Videos are provided in the supplementary material. Figure

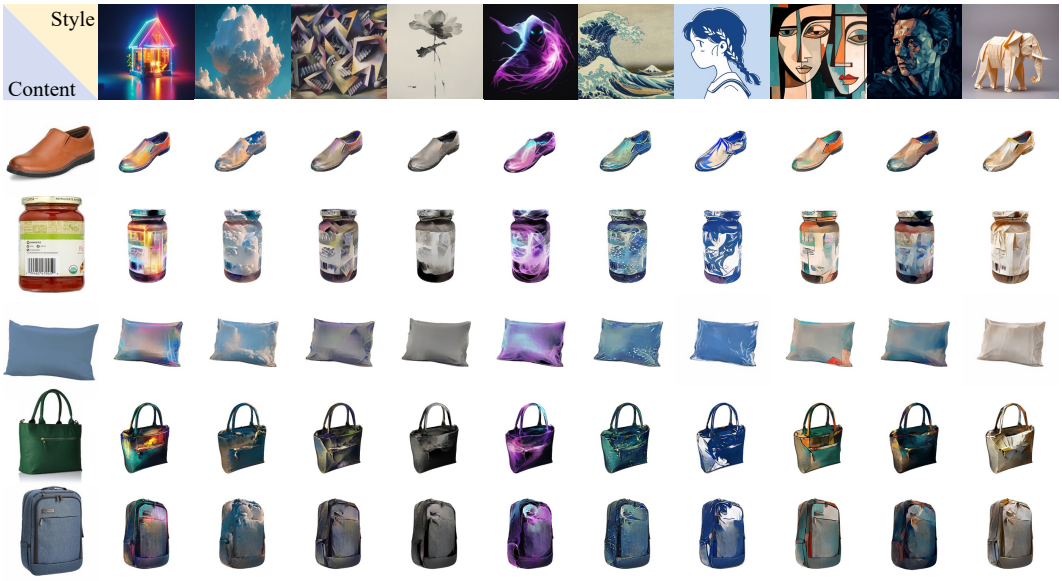


Figure 7: Qualitative results of real-world objects using our method.

Table 3: Quantitative comparison of real-world objects. The best results are in **bold**. GP: general prompt, SP: specific prompt, AS denotes aesthetic score (averaged on two VLMs), and UP denotes user preference.

Method	HPSv3 (GP)↑	HPSv3 (SP)↑	AS (%)↑	UP (%)↑
MorphAny3D	-5.4793	-5.7145	6.43	3.70
StyleSculptor	-2.2589	-2.4302	24.29	11.11
DiLAST (Ours)	**2.7565**	**2.3995**	**69.28**	**85.19**



Figure 8: Supplementary qualitative results for the main paper by using DiLAST.



Figure 9: Selected zoomed-in results.

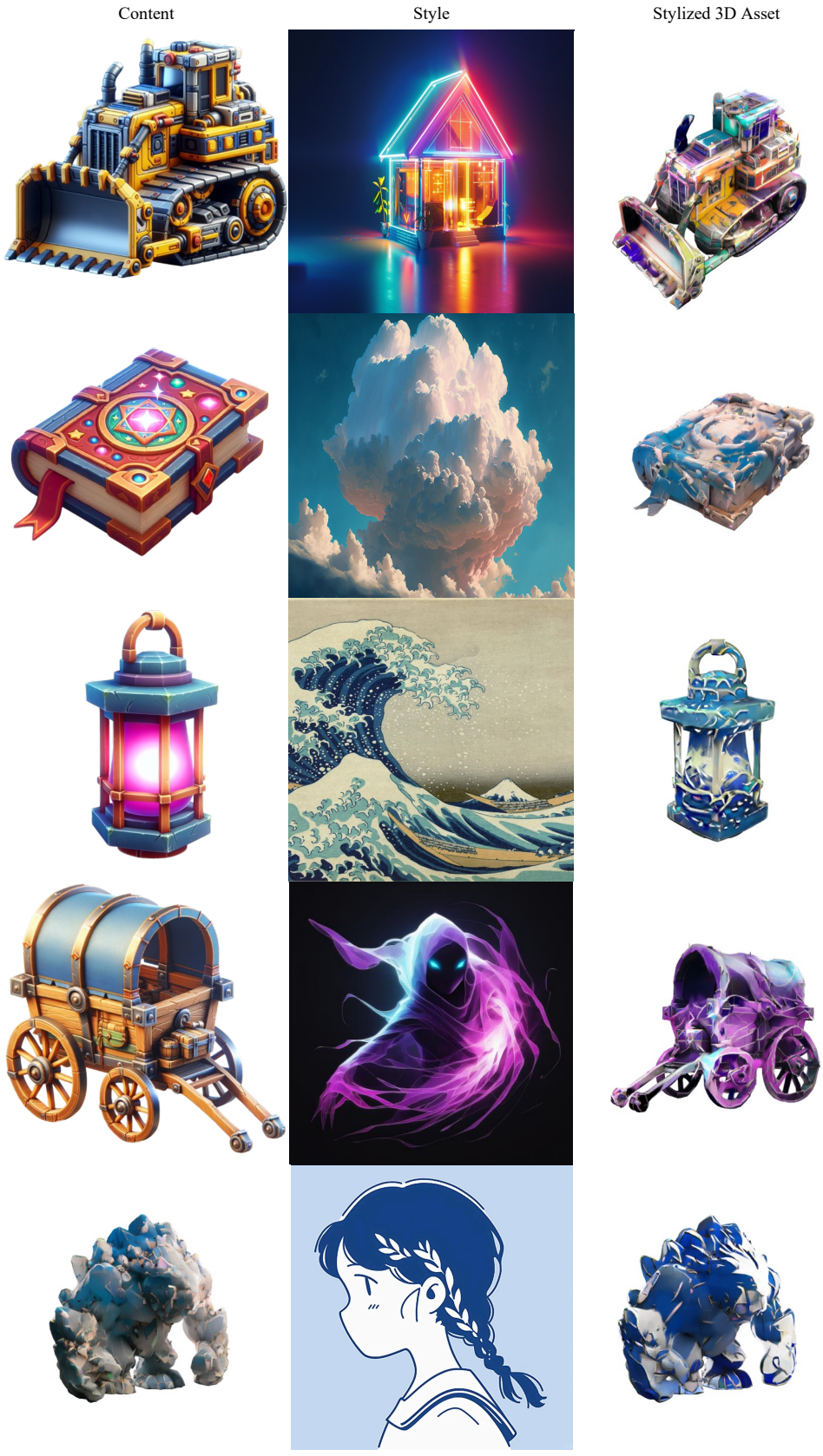


Figure 10: Selected zoomed-in results (part 2).

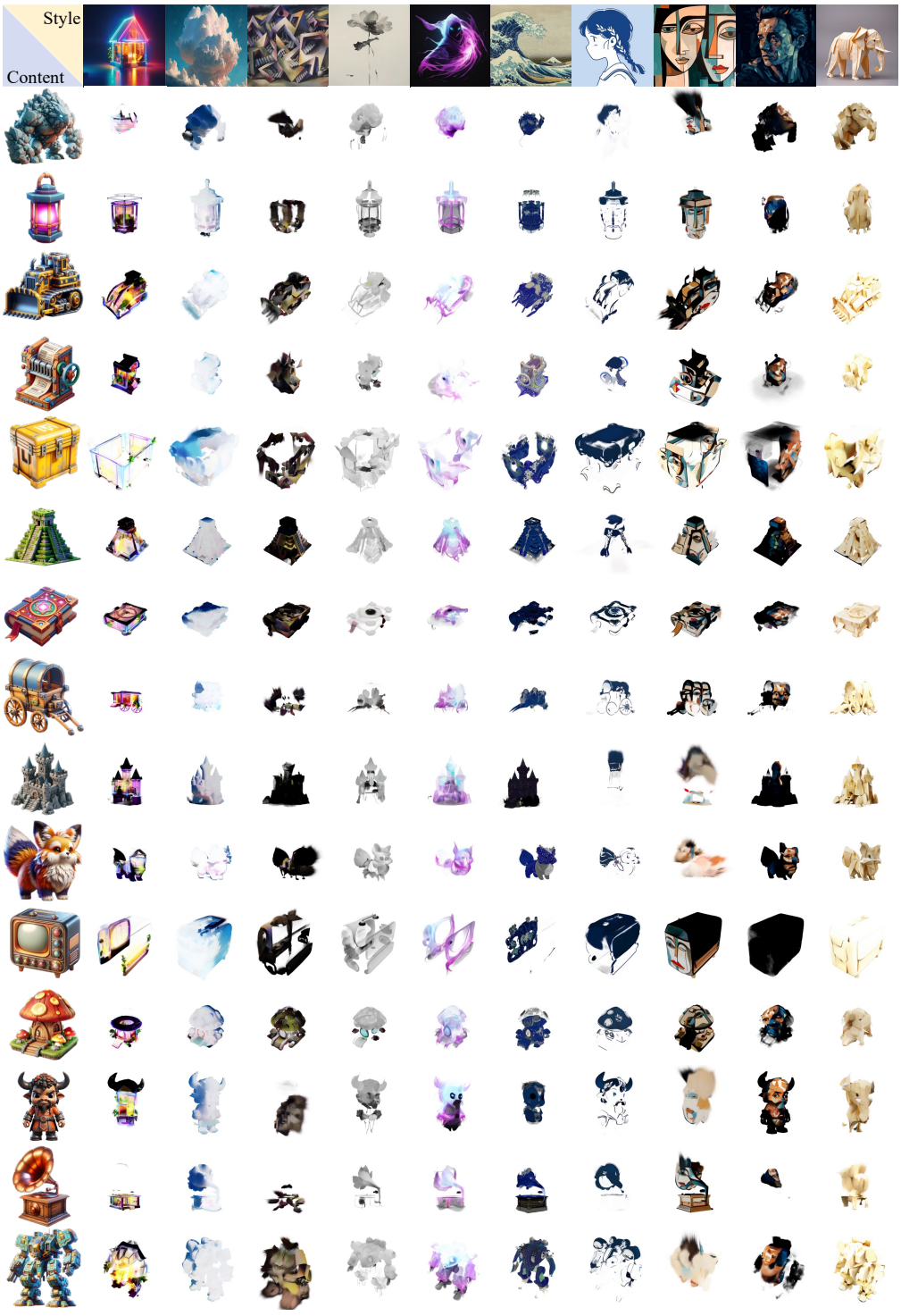


Figure 11: Supplementary qualitative results by using MorphAny3D.

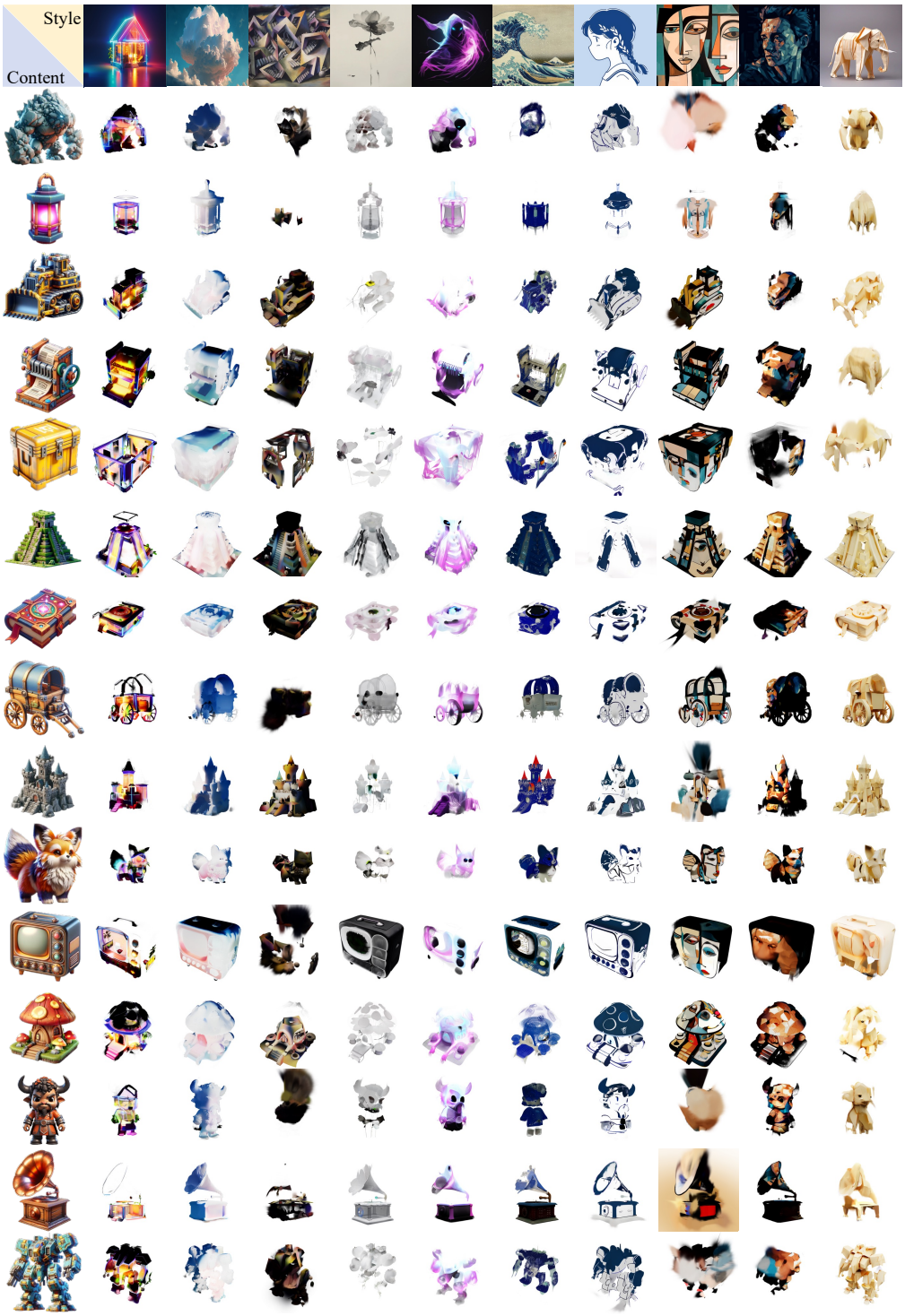


Figure 12: Supplementary qualitative results by using StyleSculptor.

7 and Table 3 present the qualitative and quantitative results of the generated 3D assets for real-world objects, respectively. Figures 8, 11, and 12 show additional stylized 3D assets generated by DiLAST, MorphAny3D, and StyleSculptor, respectively. To better showcase the details of the stylized 3D assets generated by our method, we select several examples and provide zoomed-in visualizations in Figures 9 and 10.

C.2 Analysis of Transfer-then-Generate Approaches

Transfer-then-generate (TTG) approaches first apply a 2D style transfer method to obtain a stylized 2D image, and then use this image as the condition for 3D asset generation. StyleSculptor Qu et al. (2025) has pointed out that such approaches can lead to distortions in 3D geometry and even introduce semantic inaccuracies. In this section, we further illustrate this issue through several examples. As shown in Figure 13, stylized 3D assets generated by transfer-then-generate methods suffer from artifacts, missing details, and inconsistencies with the conditioning images. In contrast, DiLAST avoids these issues and produces more faithful stylized 3D assets. This is mainly because some 2D stylized images serve as out-of-distribution conditions for 3D generative models. Another issue with TTG is that, regardless of whether the content preservation weight is set high or low during 2D stylization, the resulting 3D asset often fails to reflect the target style faithfully. As shown in Figure 14, when content preservation is too strong, the style cannot be effectively transferred. Conversely, when content preservation is too weak, the stylized image deviates further from the training distribution of the 3D generative model, leading to poor generation quality. In practice, it is often difficult for 2D style transfer to find a suitable level of content preservation that remains ”recognizable” to the 3D model. In contrast, our method directly guides the 3D latents, indirectly injecting an appropriate degree of style into the latent space while achieving suitable content preservation, without being affected by the 2D stylized image.



Figure 13: Qualitative comparison between TTG approach and DiLAST.



Figure 14: Qualitative comparison between different TTG approach settings and DiLAST.