# StyleMe3D: Stylization with Disentangled Priors by Multiple Encoders on 3D Gaussians

Cailin Zhuang<sup>1,2,3</sup> Yaoqi Hu<sup>3</sup> Xuanyang Zhang<sup>2†</sup> Wei Cheng<sup>2</sup> Jiacheng Bao<sup>1</sup> Shengqi Liu<sup>2</sup> Yiying Yang<sup>2</sup> Xianfang Zeng<sup>2</sup> Gang Yu<sup>2‡</sup> Ming Li<sup>4‡</sup> <sup>1</sup> ShanghaiTech University <sup>2</sup> StepFun <sup>3</sup> AIGC Research <sup>4</sup> Guangming Laboratory https://styleme3d.github.io/



Stylization with the Same Style

Stylization on the Same Object

#### Figure 1: Our StyleMe3D approach enables versatile, high-quality 3D stylization across diverse styles.

#### Abstract

3D Gaussian Splatting (3D GS) excels in photorealistic scene reconstruction but struggles with stylized scenarios (e.g., cartoons, games) due to fragmented textures, semantic misalignment, and limited adaptability to abstract aesthetics. We propose StyleMe3D, a holistic framework for 3D GS style transfer that integrates multi-modal style conditioning, multi-level semantic alignment, and perceptual quality enhancement. Our key insights include: (1) optimizing only RGB attributes preserves geometric integrity during stylization; (2) disentangling low-, medium-, and high-level semantics is critical for coherent style transfer; (3) scalability across isolated objects and complex scenes is essential for practical deployment. StyleMe3D introduces four novel components: Dynamic Style Score Distillation (DSSD), leveraging Stable Diffusion's latent space for semantic alignment; Contrastive Style Descriptor (CSD) for localized, content-aware texture transfer; Simultaneously Optimized Scale (SOS) to decouple style details and structural coherence; and 3D Gaussian Quality Assessment (3DG-QA), a differentiable aesthetic prior trained on human-rated data to suppress artifacts and enhance visual harmony. Evaluated on NeRF synthetic dataset (objects) and tandt db (scenes) datasets, StyleMe3D outperforms state-of-the-art

methods in preserving geometric details (e.g., carvings on sculptures) and ensuring stylistic consistency across scenes (e.g., coherent lighting in landscapes), while maintaining real-time rendering. This work bridges photorealistic 3D GS and artistic stylization, unlocking applications in gaming, virtual worlds, and digital art.

### **CCS** Concepts

• Computing methodologies → Computer vision tasks.

#### Keywords

3D gaussian splatting, style transfer, visual priors

# 1 Introduction

The advent of 3D Gaussian Splatting (3D GS) [29] has transformed 3D scene representation, offering high reconstruction fidelity and real-time rendering through explicit, anisotropic Gaussian modeling. However, its application remains largely confined to photorealistic domains, as existing methods rely heavily on real-world 3D data or multi-view 2D captures, leaving stylized scenarios—such as cartoons, anime, games, and virtual worlds—underserved. These

<sup>†</sup> Xuanyang Zhang is the project lead; ‡ Corresponding Authors.

domains demand not only geometric precision but also artistic expressiveness, where naive extensions of 3D GS often fail due to inadequate style-texture alignment, semantic incoherence, and limited adaptability to abstract aesthetics. While recent works explore 3D GS stylization via 2D priors (e.g., VGG [60] for texture transfer [12, 40, 84], CLIP [54] for semantic guidance [34]), their reliance on simplistic feature extraction and trial-and-error optimization leads to fragmented stylization, over-smoothed details, and inconsistent results across objects and scenes.

To address these challenges, we present StyleMe3D, a holistic framework for 3D GS style transfer that systematically integrates multi-modal style conditioning, multi-level semantic alignment, and perceptual quality enhancement. Our work is grounded in three critical insights:

- Geometric preservation: optimizing only the RGB attributes of 3D Gaussians preserves structural integrity while enabling stylization, avoiding the instability of geometry-altering methods.
- Semantic-aware stylization: effective style transfer requires disentangling and aligning features at low-, medium-, and high-semantic levels, which existing single-prior approaches (e.g., VGG or CLIP alone) cannot achieve.
- Scalability: a robust solution must generalize across isolated 3D objects (e.g., virtual assets) and complex scenes (e.g., openworld environments), a capability absent in prior art.

Furtehr, StyleMe3D introduces four key components to address these stylization challenges following the above insights, namely Dynamic Style Score Distillation (DSSD), Simultaneously Optimized Scale (SOS), Contrastive Style Descriptor (CSD) and 3D Gaussian Quality Assessment (3DG-QA). Leveraging Stable Diffusion (SD) [55] as a semantic prior, DSSD dynamically aligns style patterns from text prompts or reference images with 3D content through gradientbased score matching. To our knowledge, this is the first work to exploit SD's latent space for 3D GS style transfer, overcoming the limitations of VGG/CLIP in capturing nuanced artistic semantics. Existing methods often homogenize style application due to overdependence on low-level features (VGG) or global semantics (CLIP). CSD introduces a contrastively trained encoder that extracts medium-level style descriptors from a curated style dataset, enabling localized, content-aware stylization (e.g., applying distinct textures to buildings vs. vegetation in a scene). We propose multiscale optimization to decouple style-texture details (via VGG's shallow layers) and structural coherence (via deeper layers), preserving high-frequency artistic patterns without distorting geometry. Inspired by conventional image quality assessment (IQA) metrics (e.g., CLIP-IQA [69]), 3DG-QA serves as an aesthetic prior explicitly designed for 3D GS optimization. Trained on human-rated stylized 3D scenes, 3DG-QA encodes holistic aesthetic criteria-composition harmony, texture sharpness, depth-aware color consistency-into a differentiable loss. During optimization, 3DG-QA guides the model to suppress artifacts (e.g., over-saturation in occluded regions) while enhancing artistic appeal, acting as a "virtual art director" for 3D style transfer.

We validate StyleMe3D on 3D object dataset NeRF synthetic dataset [47] and 3D scene dataset tandt db [29], demonstrating its universality across geometric complexities and artistic styles.

StyleMe3D achieves superior stylization fieldlity compared with several state-of-the-art methods StyleGaussian [40], ARF [85] and SGSST [15]. For objects, our framework preserves fine details (e.g., intricate carvings on sculptures) while transferring styles with high fieldlity precision. For scenes, it ensures holistic stylistic consistency—maintaining coherent lighting and color palettes across various settings—without sacrificing real-time rendering capabilities.

We summary our contributions as follows:

- A systematic framework for 3D GS style transfer, resolving geometric preservation, multi-modal style conditioning, and multi-level feature alignment.
- First integration of Stable Diffusion into 3D GS optimization, enabling semantically coherent stylization beyond VGG/CLIP priors.
- Novel technical components (DSSD, CSD, SOS and 3DG-QA) that collectively address style localization, detail preservation, and aesthetic quality.
- By bridging photorealistic 3D reconstruction and artistic stylization, StyleMe3D unlocks new possibilities for immersive, stylized environments while preserving the core advantages of 3D GS: precision, scalability, and real-time performance.

# 2 Related Works

# 2.1 2D Generation and Stylization

2D generation has rapidly advanced across generative modeling, customization, conditional control, editing, and stylization. Initial breakthroughs in 2D synthesis with VAEs and GANs [2, 20, 28] were furthered by diffusion models [35, 55, 78, 83], enhancing image quality and diversity for complex manipulation. For efficiency, frequency-based fine-tuning and wavelet VAEs have enabled lightweight models [18, 57]. Personalized generation has also progressed, focusing on customized images [27, 83], video [3, 26], and motion [33]. Text-driven editing now offers extensive control frameworks [1, 11, 23, 32, 45, 56, 62], with character consistency essential for coherent multi-image outputs [21, 38, 71, 82, 89].

Stylization advances emphasize style-content separation, with cross-attention-based transfer [10, 67, 82] and shared attention mechanisms for coherence [77]. Frequency-domain techniques aid diffusion control [17], while Aligning style with textual cues [37], cross-domain fusion [52] and FFT-based transfer [22] expand style applications. In this paper, we aim to style 3D GS and these 2D methods give us a lot of insights and priors that can be reused in the 3D field.

# 2.2 3D Generation

Native 3D generation has progressed significantly with core representations such as meshes [5, 75, 79] and point clouds [50, 59, 81]. Meshes enable continuous surface modeling, while point clouds allow flexible spatial detail. This field now includes single-view 3D generation [42, 43] for full reconstructions from minimal input and multi-view methods [4, 41, 58, 63, 70] that ensure cross-view consistency. Texture synthesis, particularly with advanced UV mapping [7, 30, 44], enhances realism and surface detail in 3D models.

StyleMe3D: Stylization with Disentangled Priors by Multiple Encoders on 3D Gaussians



Figure 2: Overview of our 3D stylization framework (StyleMe3D): (a) Style Purification: Extracts and refines style representations via Style Cleaning in CLIP space, removing content interference from reference images. (b) Multi-Expert Stylization: The Dynamic Style Score Distillation (DSSD) module employs dynamic noise scheduling and adaptive style guidance, integrating latent losses to achieve consistent stylization step by step. Integrates three specialized components within the Dynamic Style Score Distillation (DSSD) framework: Simultaneously Optimized Scale (SOS): Adaptive noise scheduling for texture preservation. Contrastive Style Descriptor (CSD): Separates style and content via contrastive learning for style similarity score. CLIP-IQA: Quality-guided refinement using antonymic semantic prompts. (c) Progressive Consistency Optimization (Style Outpainting): Progressive outpainting achieves multi-view style propagation. Ensures coherent through iterative latent alignment, eliminating multi-view dependencies.

Text-guided 3D generation has also advanced with Score Distillation Sampling (SDS) [53] and its variants [66, 73], enabling controllable, diverse 3D synthesis. These techniques support artistic scene generation [36] and multimodal inputs (text and image) [63, 64, 74, 80]. Recent improvements in latent diffusion models further enhance the expressiveness and creative potential of text-to-3D generation [80, 88], and more and more multi-view [6, 8, 51] and 3D dataset [13, 76] still stimulate the development of this field. While 3D generation is not our core task in this work and we define the 3D GS stylization as a post-training task which further boradcasts 3D GS to more various applications.

#### 2.3 3D Style Transfer

For 3D stylization, methods like [25, 48] embed styles directly into 3D structures, while radiance field-based methods [49, 65, 85] achieve style transfer through optimization for enhanced scene realism. Though HyperNet [9] enables arbitrary style embedding in MLPs, it suffers from slow rendering and detail loss, while StyleRF [39] offers zero-shot stylization by transforming radiance field features but lacks adaptability and control.

Recent advances in 3D stylization have explored various techniques to embed artistic styles into 3D content, with reference-based methods like [46, 87] for controlled stylization and arbitrary reference techniques [40, 84] for flexible style transfer. Scalable 3D style transfer brings the 3d stylized resolution up to 4K by SOS Loss [15]. Stylized Score Distillation [31] and 3D-aware diffusion models [74] further expand these capabilities. Different from previous works, we systematically analyzed the 3D GS stylization task, proposed a more comprehensive approach to allivate the core challenges within this task and achieved superior performance.

#### 3 Method

In this section, we elaborate on our comprehensive algorithmic framework for 3D style transfer using 2D priors. We first formally define our core task: performing style transfer on reconstructed 3D Gaussian Splatting (3D GS) representations while preserving structural fidelity in 3.1. To address the inherent challenges in crossdimensional style adaptation, we propose StyleMe3D - a systematic framework comprising mixture of four encoders that collectively resolve critical challenges in 3D style migration from Sec.3.2 to Sec.3.5 and is unified in Sec.3.6.

### 3.1 The definination of Stylizing 3D GS

We define initial 3D GS as a pre-trained task, while redefining 3D gaussian stylization as a post-training task. Unlike conventional 3D generation tasks that begin from scratch, our approach applies stylization to pre-reconstructed 3D gausion for both 3D objects and scenes, allowing for enhanced control over style application while preserving the underlying geometry.

Firstly, we define the 3D gaussian reconstruction process as:

$$\min_{\Theta} \frac{1}{N} \sum_{v=1}^{N} MSE(\mathcal{R}(C_v; \Theta), I_v^{gt})$$
(1)

where  $\Theta = \{(u_i, \Sigma_i, \alpha_i, c_{i,0}, (c_{i,j,k})_{j,k}))\}_{i=1}^{N^{Gaussians}}$  represents the 3D gaussian,  $c_{i,0}$  is the main color and  $c_{i,j,k}$  is the coefficient.  $\mathcal{R}(C_v; \Theta)$  means render 3D gaussian and  $I_i^{gt}$  means the ground truth image from the viewpoint  $C_v$  respectively.

After obtaining the optimized 3D gaussian, we further formulate the 3D gaussian style transfer process with 2D prior as follows:

$$\min_{\Theta} \frac{1}{N} \sum_{v=1}^{N} \mathcal{L}(\mathcal{R}(C_v; \Theta); \phi, R)$$
(2)

where  $\phi$  means the 2D prior and *R* means the reference prompt, like text prompts or image prompts.  $\mathcal{L}$  means the loss function to further optimize the 3D gaussian which is initialized with  $\Theta$  from Eq.1.

In the style transfer task, we aim to only change the 3D gaussian stylization rather than the geometry content. We achieve geometry-style decoupling in 3D gaussian by leveraging the inherent separation of geometric and color parameters in its parametric representation. Specifically, our style transfer framework exclusively optimizes the color parameters  $\Theta_{color}$  while maintaining frozen geometric attributes during the stylization process as:

$$\min_{\Theta_{color}} \frac{1}{N} \sum_{v=1}^{N} \mathcal{L}(\mathcal{R}(C_v; \Theta); \phi, R)$$
(3)

We further discuss how to instantiate the  $\mathcal{L}$ ,  $\phi$  and R with different formulations and jointly improve the stylization effectiveness in the following sections.

# 3.2 Dynamic Style Score Distillation

In this section, we distill the prior from the 2D stable diffusion model [55] and use both text and image prompt for style transfer. **Style Cleaning.** Inspired by InstantStyle [67], we use a pre-trained CLIP model for Style Cleaning to isolate pure style information. In CLIP space, we filter out style-irrelevant details by subtracting content descriptors from style embeddings. Specifically, descriptions of the style reference image are generated using a captioning model (e.g., GPT-4V) to distinguish content-related descriptors. The CLIP Text Encoder extracts a *Content Text Embedding* (or both content and style) from these descriptors, while the CLIP Image Encoder produces a *Style Image Embedding*. Subtracting *Content Text Embedding* yields a *Final Style Embedding* containing only style-related information. The style clean process is shown in Fig. 2.

**Progressive Style Outpainting (PSO).** PSO is a novel style guidance method for consistent and detailed style propagation in multiview 3D stylization (see Fig. 2). Using 2D style priors provided by an image stylization diffusion model [16], we redefine multi-view guidance as a progressive outpainting task. By integrating sparseview RGB loss with dense-view SDS loss, PSO ensures consistent 3D stylization across views. Instead of random view selection, our method incrementally propagates style information to adjacent views, enhancing style coherence with each step. Specifically, PSO consists of two primary guidance modes, namely gobal guidance and local guidance.

**Global Guidance**. In the global gudance stage, a uniform noise level is applied to all views before stepwise reduction, defined as:

$$\alpha_{\text{step}} = \frac{\left( \left\lfloor \frac{i_{\text{step}}}{n_{\text{view}}} \right\rfloor \mod n_{\text{opt}} \right)}{n_{\text{opt}}} \tag{4}$$

where  $n_{\text{view}}$  represents the total number of rendering views and  $n_{\text{opt}}$  denotes the required optimizations per view, managed iteratively by  $i_{\text{step}}$ .

**Local Guidance.** Local guidance focuses on single-view optimization, maximizing stylization quality for individual views, albeit at the potential expense of global consistency. The local guidance schedule is defined as:

$$\alpha_{\text{step}} = \frac{i_{\text{step}} \mod n_{\text{opt}}}{n_{\text{opt}}} \tag{5}$$

The effectiveness of these modes in balancing stylization strength and consistency is discussed in Sec. 4.3. To maximize the stylization outcome, we combine both guidance modes for complementary strengths.

**Fine Timestep Sampling.** Fine timestep sampling enhances temporal resolution by focusing on low-noise intervals for more granular optimization, with noise progressively decreasing from high to low levels. This sampling strategy is formulated as:

$$t = Round((1 - \alpha_{step}^{0.5}) \cdot T).clip(T_{min}, T_{max})$$
(6)

where T denotes the total timesteps, with  $T_{\min}$  and  $T_{\max}$  setting the bounds. Higher noise initialization effectively eliminates outlier Gaussian, refining the stylization outcome.

**Dynamic Style Score Distillation (DSSD).** As shown in Fig. 2(b). DSSD further extends score distillation by applying a dynamic CFG (Classifier-Free Guidance) [24] scale coefficient to optimize the intensity of style guidance. Fixed CFG values can lead to oversmoothing (low CFG) or oversaturation (high CFG). To counter this, we introduce a dynamic guidance coefficient that adaptively balances fixed CFG values throughout optimization. The adaptive coefficient is defined as:

$$\Delta \lambda = \max\left(7.5, \, \lambda_{\max} \cdot \left(\alpha_{\text{step}}^2\right)\right) \tag{7}$$

With this method, we extend the SSD proposed by [31], and define the style loss in latent space as:

$$\text{DSSD}_{2\text{D}}^{z} = (1 - \Delta\lambda_s)\epsilon_{\phi 2\text{D}}(z_t_s | y, t_s) + \Delta\lambda_s \hat{\epsilon}_{\phi 2\text{D}}(z_t_s | y, s, t_s) - \epsilon_s, \quad (8)$$

where  $\epsilon_{\phi 2D}()$  is the predicted noise by the style-based 2D diffusion prior  $\phi$ .

The latent space loss aligns abstract style features, whereas pixelspace loss emphasizes visible characteristics. For stylizing a given 3D model, latent loss ensures style feature transfer, while pixel loss provides further reliability in visual output. Defining  $x_{t_s}$  = Decoder( $z_{t_s}$ ), the pixel-space loss is given by:

 $DSSD_{2D}^{x} = (1 - \Delta\lambda_s)\epsilon_{\phi 2D}(x_{t_s}|y, t_s) + \Delta\lambda_s\hat{\epsilon}_{\phi 2D}(x_{t_s}|y, s, t_s) - \epsilon_s$ (9)

Our Dynamic Style Score Distillation (DSSD) objective function integrates latent DSSD and pixel DSSD:

٦

$$\mathbb{Z}_{\Theta_{color}} \mathcal{L}_{\text{DSD}}(x = \mathcal{R}(C_v; \Theta_{color}); \phi, R) = \mathbb{E}_{t_s^z, t_s^x, \epsilon_s^z, \epsilon_s^x} \left[ \omega(t) \left( \lambda_{\text{2D}}^z \text{DSSD}_{\text{2D}}^z + \lambda_{\text{2D}}^x \text{DSSD}_{\text{2D}}^x \right) \frac{\partial x}{\partial \theta} \right]$$
(10)

where  $\omega(t)$  is a weighting function regulating timestep contributions.

Further, we optimize the stylized multi-view image  $I_{rgb}$  and the associated mask  $I_{mask}$  for alignment with the input data. If required, additional loss terms such as SSIM loss [72] or LPIPS loss [86] may be integrated to enhance alignment. Thus, our final objective function is:

$$\mathcal{L}_{style} = \lambda_{DSSD} \mathcal{L}_{DSSD} + \lambda_{RGB} \mathcal{L}_{RGB} + \lambda_{mask} \mathcal{L}_{mask}$$
(11)

This setup ensures multi-view consistency in 3D stylization, achieving refined style expression and geometric fidelity through the dynamic coefficient adjustment and adaptive optimization strategy.

# 3.3 Simultaneously Optimized Scale (SOS)

To further enhance the texture details of 3D gaussian, multiscale stylization strategy is introduced into the optimization process. Following the silimar approach from [15, 19], we employ VGG-19 [60] to extract high-resolution texture features through its shallow convolutional layers. We use N rendered images (each image is represented as  $I_v$ ) from the source 3D gaussian and style reference image  $I_{ref}$  to compute multi-scale Gram matrix correlations and formulate the style objectiveness as follows:

$$\mathcal{L}_{\text{SOS}} = \frac{1}{N} \sum_{v=1}^{N} \sum_{l \in L_s} \|G(\phi_{\text{VGG}}^l(I_v)) - G(\phi_{\text{VGG}}^l(I_{\text{ref}}))\|_2^2$$
(12)

where  $G(\cdot)$  denotes Gram matrix computation,  $\phi_{VGG}^{l}$  represents features from the *l*-th VGG layer and  $L_{s} = \{ReLU_{k}_{1}, k \in 1, 3, 5\}$ .

#### 3.4 Contrastive Style Descriptor (CSD)

CSD[61] aims to build a high-performance model (variants of ViT [14], like ViT-B and ViT-L) for the representation of the image style. The ViT is trained with both self-supervised learning and supervised objectives. As a result, the ViT can extract image descriptors with concise and effective style information. To further align to the style between the 3D gaussian and the given reference image, we leverage the ViT to extract style feature from rendered images and reference image respectilvely and then calculate the pairwise cosine silimarity score. Finally, the CSD loss term reduces to:

$$\mathcal{L}_{\text{CSD}} = \frac{1}{N} \sum_{v=1}^{N} (1 - \cos(\phi_{\text{ViT}}(I_v), \phi_{\text{ViT}}(I_{\text{ref}})))$$
(13)

### 3.5 3D gaussian Quality Assessment (3DG-QA)

In addition to preserving the original content and migration style of 3D gaussian, we also need to ensure the overall quality between the migrated style and content. CLIP-IQA [68] has been developed to evaluate the look or quality of an image. CLIP-IQA leverages CLIP for perception assessment and calculate the cosine similarity between the feature embeddings of the given text promt and image as follows:

$$s = \frac{x \odot t}{||x|| * ||t||}$$
(14)

where  $x \in \mathbb{R}^C$  and  $t \in \mathbb{R}^C$  represents the image embedding and text embedding, C is the embedding channel dimension. CLIP-IQA further introduces antonym prompts (e.g., "Good photo." and "Bad photo.") to address the linguistic ambiguity.  $t_1$  and  $t_2$  are obtained

from text prompts with good quality and bad quality respectively and the  $s_i$  can be obtained with the corresponding  $t_i$ , then the final

$$\overline{s} = \frac{e^{s_1}}{e^{s_1} + e^{s_2}} \tag{15}$$

We adopt the CLIP-IQA property and extend CLIP-IQA to the 3D style transfer field to ensure the perception quality of 3D gaussian. More specifically, we define the 3D gaussian Quality Assessment (3DG-QA) as a objective term as:

CLIP-IQA score can be formulate as:

$$\mathcal{L}_{3\text{DG-QA}} = \frac{1}{N} \sum_{v=1}^{N} (1 - \bar{s}_v)$$
(16)

where v means the viewpoint index rendered from the 3D gaussian representation.

# 3.6 Stylizing 3D GS with mixture of encoders

The StyleMe3D approach systematically addresses five fundamental aspects in 3D gaussian stylization: (1) Style-content decoupling, (2) Adaptive style conditioning, (3) Multi-scale feature alignment, (4) Texture detail enhancement, and (5) Global aesthetic optimization with four principal components. The DSSD stablishes effective style conditioning through high-level semantic alignment, leveraging score-based stable diffusion to extract and transfer domaininvariant style features. SOS addresses low-level feature alignment via multi-scale optimization, preserving stylistic textures through scale-aware importance sampling and geometric consistency constraints. CSD facilitates mid-level style-content harmonization using contrastive learning to disentangle and recompose style attributes while maintaining content integrity. At last, 3DG-QA enhances global aesthetic quality through metric-guided refinement, employing perceptual quality evaluation to optimize both local textural coherence and global visual appeal.

We integrate the whole optimization goal as:

$$\mathcal{L}_{final} = \lambda_1 \mathcal{L}_{style} + \lambda_2 \mathcal{L}_{SOS} + \lambda_3 \mathcal{L}_{CSD} + \lambda_4 \mathcal{L}_{3DG-QA} \quad (17)$$

where  $\lambda_1, \lambda_2, \lambda_3$  and  $\lambda_4$  stand for the multi-task coefficients. For shortly, this multi-faceted approach ensures semantic-aware style, fine-grained style, style fidelity and global aesthetics quality.

As demonstrated in Sec. 4.3, the combined losses enable simultaneous preservation of geometric integrity and artistic expression while suppressing common artifacts like over-stylization and texture flickering.

# 4 Experiment

#### 4.1 Visual Result

As shown in Fig. 3, we applied six styles to showcase our experimental results on both object and scene datasets (NeRF synthetic dataset [47] and tandt db [29]). The style references fall into two main categories: non-photorealistic art styles (*e.g.*, vangogh, cartoon, sketch, hand-drawing, watercolor, painting) and state-based styles (*e.g.*, fire, water, clouds, hair). These categories highlight our method's ability to handle traditional art styles and capture realistic physical characteristics in 3D. To highlight our method's advantage in preserving detail textures and shadows, we zoom in on details like the legs and detail texture of the chair, texture of the fire on the



Figure 3: Visual Result. Demonstration of our method's performance across five styles (vangogh wheat field, star night, fire nezha, colorful oil, and lighting tiger) applied to five objects(chair, ship, hotdog, lego and mic) and two scenes (man face and train). The results illustrate our model's capability to handle two main categories of styles: (1) Non-photorealistic Art Styles (*e.g.,* cartoon, drawing), showcasing traditional artistic expressions, and (2) State-based Styles (*e.g.,* fire, oil), which capture physical properties. This figure demonstrates our method's versatility and semantic-aware ability in stylizing 3D models while preserving style fidelity and geometric consistency across diverse artistic and physical characteristics. For Example, semantic separation of the legs of the chair from the seat cushion, detail texture of chair, texture of the fire on the hot dog, and metallic sheen on the mic are all effectively preserved.

hot dog, and metallic sheen on the mic. Experimental results indicate that Gaussian Splatting effectively enhances non-photorealistic and state-based style representations, showing strong adaptability in diverse stylized scenarios. Additional results are provided in the Supplementary.

#### 4.2 Comparison Studies

**Qualitative Result.** we show objects and scene stylization comparisons in Fig.4 and Fig.5 respectively. For objects, we applied vangogh, fire nezha, and sketch styles to chair, hotdog and mic. For scene stylization, we select truck and train from tandt db dataset using landscope and lighting tiger styles. We evaluate our method against others, including SGSST [15], StyleGaussian [40] and ARF [85]. The horizontal axis lists competing methods and the vertical axis denotes datasets.

Different from traditional methods based only on VGG networks like SGSST [15], StyleGaussian [40] and ARF [85], which focus on simple style transfer, our approach prioritizes vivid, expressive and semantic-aware stylization. They relies on VGG networks [60] with empirical-based style decoupling, which limiting style extraction with customized references, our diffusion-based and multi-Expert method, pre-trained on large-scale style image-text data, captures style features with greater fidelity. Moreover, training on image-text data enhances semantic understanding, allowing content filtering in CLIP space for precise style extraction. Unlike ARF [85], which depends on carefully pre-stylized views for effective color matching and risks texture drift if the initial view is misaligned, our method only requires a single arbitrary style reference image. While we incorporate pre-stylized multi-views, they serve solely for pixel-level style guidance in our outpainting process rather than relying on single-view matching, establishing a distinct way from that of ARF.

Table 1: Quantitative comparison with competing methods

Method	<b>PSNR</b> ↑	SSIM↑	LPIPS↓
ARF	17.537	0.802	0.188
SGSST	11.963	0.678	0.306
StyleGaussian	7.279	0.129	0.558
Ours	18.015	0.830	0.174

**Quantitative Evaluation** We evaluate our method with three standard image quality metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [72], and Learned Perceptual Image Patch Similarity (LPIPS) [86]. PSNR quantifies pixel-level accuracy, indicating how closely the stylized image matches the original. SSIM measures structural similarity, capturing perceptual features like textures and edges. LPIPS assesses perceptual

StyleMe3D: Stylization with Disentangled Priors by Multiple Encoders on 3D Gaussians



Figure 4: Qualitative Comparisons on Object Level Stylization. We compare our method against other SOTA (SGSST [15], StyleGaussian [40] and ARF [85]) on nerf synthetic dataset (selected chair, hotdog, and mic) using vangogh wheat field, fire nezha, and sketch styles. The horizontal axis represents the compared methods, and the vertical axis displays different data. Our method effectively retains semantic and details of original model and style feature of reference image, such as semantic separation of the legs of the chair from the seat cushion, texture of the fire on the hot dog, and metallic sheen on the mic. Compared to others, our method exhibits stronger semantic understanding, clearly distinguishing elements like the cushions, backrest and legs on the chair.



Figure 5: Qualitative Comparisons on Scene Level Stylization. We compare our method against other SOTA (SGSST [15], StyleGaussian [40] and ARF [85]) on tandt db dataset (selected truck and train) using landscope and lighting tiger styles. The horizontal axis represents the compared methods, and the vertical axis displays different data. Our method effectively retains semantic and details of original model and style feature of reference image, such as the truck wheel and train fence (as shown in Zoom-in). Compared to others, our method exhibits stronger semantic understanding, clearly distinguishing elements like the fence, tire and rail.



Figure 6: Ablation study on style outpainting guidance mode. (a) Baseline without style outpainting exhibits limited stylization scope and view-dependent artifacts (red boxes). (b) Local Guidance enables single-view enhancement but causes multiview inconsistencies. Global-Local Fusion achieves crossview style propagation through adaptive attention weighting, improving style consistency while preserving view-specific details.



Figure 7: Ablation study on dynamic noise scheduling. Low Scale (7.5) produces incomplete stylization with missing texture details. High Scale (50) introduces oversaturation artifacts and structural distortions. Dynamic Scale (7.5-30) adaptively balances detail preservation and style intensity.

quality based on deep network features, emphasizing visual similarity as perceived by humans.

As shown in Tab. 1, our method achieves significantly higher SSIM and PSNR scores, demonstrating enhanced structural and perceptual fidelity compared to SGSST, StyleGaussian and ARF. Our higher PSNR and SSIM score indicates better fidelity in color and texture reproduction while preserving structural details. Furthermore, the LPIPS score, measuring perceptual similarity, supports our method's superior style consistency and stylization quality across multiple viewpoints.

Table 2: Quantitative comparison with DSSD version andMulti-Expert version

Method	PSNR↑	SSIM↑	LPIPS↓
Ours (DSSD)	17.270	0.776	0.181
Ours (Multi-Expert)	18.015	0.830	0.174

DSSD DSSD + SOS DSSD + SOS + CSD + CLIP-IO/

Figure 8: Ablation study loss design. (a) DSSD-only initialization yields semantically coherent but texture-deficient results with color shifts (see missing curvilinear patterns in Van Gogh stylization). (b) DSSD+SOS achieves texturegeometry equilibrium through gradient mutual regularization, recovering fine details while suppressing oversmoothing. (c) Full Model (DSSD+SOS+CSD+CLIP-IQA) enhances perceptual quality via knowledge-driven style assessment, achieving remarkable improvement over baseline (Table 2).

#### 4.3 Ablation Study

Style

We conducted ablation studies to assess the impact of various components and parameters in our method, focusing on style outpainting mode, DDSD and multi-expert module.

**Ablation on Style Outpainting.** As shown in Fig. 6. We present an ablation study on the impact of Style Outpainting. Without it, the degree of stylization is visibly limited, whereas applying Style Outpainting allows effective style propagation across views. We compares different guidance schemes: local mode & global-local mode. Local mode shows inconsistencies, resulting in artifacts and missing details in certain views. In contrast, global-local mode enhances stylization intensity and detail refinement, achieving more coherent stylization across views.

**Ablation on DSSD.** As shown in Fig. 7. We conducted an ablation study on the effectiveness of dynamic guidance scale in DSSD. Comparing results at a low scale of 7.5, a high scale of 50, and a dynamic scale ranging from 7.5 to 30, we observed that the dynamic scale approach consistently outperforms static setting.

Ablation on Multi-Expert. As shown in Fig. 8, we analyze the impact of SOS, CSD and 3DG-QA on stylization quality. our analysis reveals that initial stylization using DSSD alone produces semantically coherent results but suffers from two critical limitations: 1) Insufficient low-level texture details (e.g., missing curvilinear patterns in Van Gogh-inspired wheat field renderings), and 2) Systematic color deviation artifacts. The introduction of SOS loss establishes a dual-optimization framework where DSSD and SOS operate concurrently within single-view projections. This configuration enables mutual regularization of their gradient optimization directions - DSSD's tendency toward over-smoothing is counterbalanced by SOS's capacity for detail enhancement, while SOS's potential over-emphasis on low-level features is constrained by DSSD's semantic guidance.

Subsequent integration of CSD and 3DG-QA implements knowledgedriven perceptual assessment through CLIP-space cosine similarity

Cailin Zhuang, Yaoqi Hu, Xuanyang Zhang, Wei Cheng et al.

metrics. The CSD module specializes in style authenticity evaluation through learned artistic aesthetics criteria, while 3DG-QA provides quality-focused guidance via antonymic text prompts. Quantitative analysis shows this combined approach achieves remarkable improvement in human perceptual quality scores compared to baseline configurations (see Table 2).

#### 5 Conclusion

We redefine the 3D Gaussian Splatting (3D GS) stylization task through comprehensive analysis and propose the StyleMe3D framework, establishing a novel paradigm for artistic 3D scene stylization. StyleMe3D enables artistic 3D Gaussian Splatting stylization via Stable Diffusion-guided score distillation (DSSD), contrastive style descriptors (CSD), and multi-scale optimization (SOS). The 3DG-QA module ensures aesthetic coherence while preserving geometry. Experiments show superior detail retention, style consistency across various objects and scenes.

#### References

- Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. 2022. Text2live: Text-driven layered image and video editing. In *European conference* on computer vision. Springer, 707–723.
- [2] Andrew Brock. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. arXiv preprint arXiv:1809.11096 (2018).
- [3] Hila Chefer, Shiran Zada, Roni Paiss, Ariel Ephrat, Omer Tov, Michael Rubinstein, Lior Wolf, Tali Dekel, Tomer Michaeli, and Inbar Mosseri. 2024. Still-moving: Customized video generation without customized video data. arXiv preprint arXiv:2407.08674 (2024).
- [4] Hao Chen, Jiafu Wu, Ying Jin, Jinlong Peng, Xiaofeng Mao, Mingmin Chi, Mufeng Yao, Bo Peng, Jian Li, and Yun Cao. 2024. VI3DRM: Towards meticulous 3D Reconstruction from Sparse Views via Photo-Realistic Novel View Synthesis. arXiv preprint arXiv:2409.08207 (2024).
- [5] Sijin Chen, Xin Chen, Anqi Pang, Xianfang Zeng, Wei Cheng, Yijun Fu, Fukun Yin, Billzb Wang, Jingyi Yu, Gang Yu, et al. 2024. Meshxl: Neural coordinate field for generative 3d foundation models. Advances in Neural Information Processing Systems 37 (2024), 97141–97166.
- [6] Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, et al. 2023. Dna-rendering: A diverse neural actor repository for high-fidelity human-centric rendering. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 19982– 19993.
- [7] Wei Cheng, Juncheng Mu, Xianfang Zeng, Xin Chen, Anqi Pang, Chi Zhang, Zhibin Wang, Bin Fu, Gang Yu, Ziwei Liu, et al. 2024. MVPaint: Synchronized Multi-View Diffusion for Painting Anything 3D. arXiv preprint arXiv:2411.02336 (2024).
- [8] Wei Cheng, Su Xu, Jingtan Piao, Chen Qian, Wayne Wu, Kwan-Yee Lin, and Hongsheng Li. 2022. Generalizable neural performer: Learning robust radiance fields for human novel view synthesis. arXiv preprint arXiv:2204.11798 (2022).
- [9] Pei-Ze Chiang, Meng-Shiun Tsai, Hung-Yu Tseng, Wei-Sheng Lai, and Wei-Chen Chiu. 2022. Stylizing 3d scene via implicit representation and hypernetwork. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 1475–1484.
- [10] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. 2024. 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. 8795–8805.
- [11] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance. In European Conference on Computer Vision. Springer, 88–105.
- [12] Valentin De Bortoli, Agnès Desolneux, Alain Durmus, Bruno Galerne, and Arthur Leclaire. 2021. Maximum entropy methods for texture synthesis: theory and practice. SIAM Journal on Mathematics of Data Science 3, 1 (2021), 52–82.
- [13] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. 2023. Objaverse-XL: A Universe of 10M+ 3D Objects. arXiv preprint arXiv:2307.05663 (2023).
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg

Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

- [15] Bruno Galerne et al. 2024. SGSST: Scaling Gaussian Splatting StyleTransfer. arXiv preprint arXiv:2412.03371 (2024).
- [16] Junyao Gao, Yanchen Liu, Yanan Sun, Yinhao Tang, Yanhong Zeng, Kai Chen, and Cairong Zhao. 2024. Styleshot: A snapshot on any style. arXiv preprint arXiv:2407.01414 (2024).
- [17] Xiang Gao, Zhengbo Xu, Junhan Zhao, and Jiaying Liu. 2024. Frequency-Controlled Diffusion Model for Versatile Text-Guided Image-to-Image Translation. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38. 1824–1832.
- [18] Ziqi Gao, Qichao Wang, Aochuan Chen, Zijing Liu, Bingzhe Wu, Liang Chen, and Jia Li. 2024. Parameter-Efficient Fine-Tuning with Discrete Fourier Transform. arXiv preprint arXiv:2405.03003 (2024).
- [19] Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. 2017. Controlling perceptual factors in neural style transfer. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3985–3993.
- [20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. Advances in neural information processing systems 27 (2014).
- [21] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, and Qian He. 2024. PuLID: Pure and Lightning ID Customization via Contrastive Alignment. arXiv preprint arXiv:2404.16022 (2024).
- [22] Feihong He, Gang Li, Mengyuan Zhang, Leilei Yan, Lingyu Si, Fanzhang Li, and Li Shen. 2024. Freestyle: Free lunch for text-guided style transfer using diffusion models. arXiv preprint arXiv:2401.15636 (2024).
- [23] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022).
- [24] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022).
  [25] Hsin-Ping Huang, Hung-Yu Tseng, Saurabh Saini, Maneesh Singh, and Ming-
- [25] Hsin-Ping Huang, Hung-Yu Tseng, Saurabh Saini, Maneesh Singh, and Ming-Hsuan Yang. 2021. Learning to stylize novel views. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 13869–13878.
- [26] Zhitong Huang, Mohan Zhang, and Jing Liao. 2024. LVCD: Reference-based Lineart Video Colorization with Diffusion Models. arXiv preprint arXiv:2409.12960 (2024).
- [27] Kyungmin Jo and Jaegul Choo. 2024. Skip-and-Play: Depth-Driven Pose-Preserved Image Generation for Any Objects. arXiv preprint arXiv:2409.02653 (2024).
- [28] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF* conference on computer vision and pattern recognition. 4401–4410.
- [29] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. ACM Trans. Graph. 42, 4 (2023), 139–1.
- [30] Jangyeong Kim, Donggoo Kang, Junyoung Choi, Jeonga Wi, Junho Gwon, Jiun Bae, Dumim Yoon, and Junghyun Han. 2024. RoCoTex: A Robust Method for Consistent Texture Synthesis with Diffusion Models. arXiv preprint arXiv:2409.19989 (2024).
- [31] Hubert Kompanowski and Binh-Son Hua. 2024. Dream-in-Style: Text-to-3D Generation using Stylized Score Distillation. arXiv preprint arXiv:2406.18581 (2024).
- [32] Gwanhyeong Koo, Sunjae Yoon, Ji Woo Hong, and Chang D Yoo. 2024. Flexiedit: Frequency-aware latent refinement for enhanced non-rigid editing. arXiv preprint arXiv:2407.17850 (2024).
- [33] Divya Kothandaraman, Kuldeep Kulkarni, Sumit Shekhar, Balaji Vasan Srinivasan, and Dinesh Manocha. 2024. ImPoster: Text and Frequency Guidance for Subject Driven Action Personalization using Diffusion Models. arXiv preprint arXiv:2409.15650 (2024).
- [34] Áron Samuel Kovács, Pedro Hermosilla, and Renata G Raidou. 2024. Style: Stylized Gaussian Splatting. In *Computer Graphics Forum*, Vol. 43. Wiley Online Library, e15259.
- [35] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. 2024. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. arXiv preprint arXiv:2402.17245 (2024).
- [36] Pengzhi Li, Chengshuai Tang, Qinxuan Huang, and Zhiheng Li. 2024. Art3d: 3d gaussian splatting for text-guided artistic scenes generation. arXiv preprint arXiv:2405.10508 (2024).
- [37] Wen Li, Muyuan Fang, Cheng Zou, Biao Gong, Ruobing Zheng, Meng Wang, Jingdong Chen, and Ming Yang. 2024. StyleTokenizer: Defining Image Style by a Single Instance for Controlling Diffusion Models. arXiv preprint arXiv:2409.02543 (2024).
- [38] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. 2024. Photomaker: Customizing realistic human photos via stacked id embedding. In Proceedings of the IEEE/CVF Conference on Computer Vision and

Cailin Zhuang, Yaoqi Hu, Xuanyang Zhang, Wei Cheng et al.

Pattern Recognition. 8640-8650.

- [39] Kunhao Liu, Fangneng Zhan, Yiwen Chen, Jiahui Zhang, Yingchen Yu, Abdulmotaleb El Saddik, Shijian Lu, and Eric P Xing. 2023. Stylerf: Zero-shot 3d style transfer of neural radiance fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8338–8348.
- [40] Kunhao Liu, Fangneng Zhan, Muyu Xu, Christian Theobalt, Ling Shao, and Shijian Lu. 2024. StyleGaussian: Instant 3D Style Transfer with Gaussian Splatting. arXiv preprint arXiv:2403.07807 (2024).
- [41] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. 2024. 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. 10072–10083.
- [42] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In Proceedings of the IEEE/CVF international conference on computer vision. 9298– 9309.
- [43] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. 2024. Wonder3d: Single image to 3d using cross-domain diffusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 9970–9980.
- [44] Jiawei Lu, Yingpeng Zhang, Zengjun Zhao, He Wang, Kun Zhou, and Tianjia Shao. 2024. GenesisTex2: Stable, Consistent and High-Quality Text-to-Texture Generation. arXiv preprint arXiv:2409.18401 (2024).
- [45] Yang Luo, Yiheng Zhang, Zhaofan Qiu, Ting Yao, Zhineng Chen, Yu-Gang Jiang, and Tao Mei. 2024. Freeenhance: Tuning-free image enhancement via contentconsistent noising-and-denoising process. arXiv preprint arXiv:2409.07451 (2024).
- [46] Yiqun Mei, Jiacong Xu, and Vishal M Patel. 2024. Reference-based Controllable Scene Stylization with Gaussian Splatting. arXiv preprint arXiv:2407.07220 (2024).
- [47] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.
- [48] Fangzhou Mu, Jian Wang, Yicheng Wu, and Yin Li. 2022. 3d photo stylization: Learning to generate stylized novel views from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 16273– 16282.
- [49] Thu Nguyen-Phuoc, Feng Liu, and Lei Xiao. 2022. Snerf: stylized neural implicit representations for 3d scenes. arXiv preprint arXiv:2207.02363 (2022).
  [50] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen.
- [50] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. 2022. Point-e: A system for generating 3d point clouds from complex prompts. arXiv preprint arXiv:2212.08751 (2022).
- [51] Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, et al. 2023. RenderMe-360: a large digital asset library and benchmarks towards high-fidelity head avatars. Advances in Neural Information Processing Systems 36 (2023), 7993–8005.
- [52] Kien T Pham, Jingye Chen, and Qifeng Chen. 2024. TALE: Training-free Crossdomain Image Composition via Adaptive Latent Manipulation and Energy-guided Optimization. In Proceedings of the 32nd ACM International Conference on Multimedia. 3160–3169.
- [53] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022).
- [54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning. PMLR, 8748–8763.
- [55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 10684–10695.
- [56] Nataniel Ruiz, Yuanzhen Li, Neal Wadhwa, Yael Pritch, Michael Rubinstein, David E Jacobs, and Shlomi Fruchter. 2024. Magic Insert: Style-Aware Drag-and-Drop. arXiv preprint arXiv:2407.02489 (2024).
- [57] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M Weber. 2024. LiteVAE: Lightweight and Efficient Variational Autoencoders for Latent Diffusion Models. arXiv preprint arXiv:2405.14477 (2024).
- [58] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. 2023. Mvdream: Multi-view diffusion for 3d generation. arXiv preprint arXiv:2308.16512 (2023).
- [59] Dong Wook Shu, Sung Woo Park, and Junseok Kwon. 2019. 3d point cloud generative adversarial network based on tree structured graph convolutions. In Proceedings of the IEEE/CVF international conference on computer vision. 3859– 3868.
- [60] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
- [61] Gowthami Somepalli, Ayan Bansal, Micah Goldblum, Jonas Geiping, Tom Goldstein, et al. 2024. Measuring Style Similarity in Diffusion Models. arXiv preprint arXiv:2404.01292 (2024).

- [62] Zihan Su, Junhao Zhuang, and Chun Yuan. 2024. TextureDiffusion: Target Prompt Disentangled Editing for Various Texture Transfer. arXiv preprint arXiv:2409.09610 (2024).
- [63] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. 2025. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In European Conference on Computer Vision. Springer, 1–18.
- [64] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. 2023. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023).
- [65] Can Wang, Ruixiang Jiang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2023. Nerf-art: Text-driven neural radiance fields stylization. *IEEE Transactions on Visualization and Computer Graphics* (2023).
- [66] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. 2023. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 12619–12629.
- [67] Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. 2024. Instantstyle: Free lunch towards style-preserving in text-to-image generation. arXiv preprint arXiv:2404.02733 (2024).
- [68] Jianyi Wang, Kelvin C.K. Chan, and Chen Change Loy. 2022. Exploring CLIP for Assessing the Look and Feel of Images. arXiv preprint arXiv:2207.12396 (2022).
- [69] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. 2023. Exploring clip for assessing the look and feel of images. In Proceedings of the AAAI conference on artificial intelligence, Vol. 37. 2555–2563.
- [70] Peng Wang and Yichun Shi. 2023. Imagedream: Image-prompt multi-view diffusion for 3d generation. arXiv preprint arXiv:2312.02201 (2023).
- [71] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. 2024. Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024).
- [72] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions* on image processing 13, 4 (2004), 600–612.
- [73] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. 2024. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in Neural Information Processing Systems 36 (2024).
- [74] Zhenwei Wang, Tengfei Wang, Zexin He, Gerhard Hancke, Ziwei Liu, and Rynson WH Lau. 2024. Phidias: A Generative Model for Creating 3D Content from Text, Image, and 3D Conditions with Reference-Augmented Diffusion. arXiv preprint arXiv:2409.11406 (2024).
- [75] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. 2024. Unique3D: High-Quality and Efficient 3D Mesh Generation from a Single Image. arXiv preprint arXiv:2405.20343 (2024).
- [76] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. 2023. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 803–814.
- [77] Zongze Wu, Yotam Nitzan, Eli Shechtman, and Dani Lischinski. 2021. Stylealign: Analysis and applications of aligned stylegan models. arXiv preprint arXiv:2110.11323 (2021).
- [78] Ximing Xing, Haitao Zhou, Chuang Wang, Jing Zhang, Dong Xu, and Qian Yu. 2024. SVGDreamer: Text guided SVG generation with diffusion model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4546–4555.
- [79] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. 2024. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. arXiv preprint arXiv:2404.07191 (2024).
- [80] Zizheng Yan, Jiapeng Zhou, Fanpeng Meng, Yushuang Wu, Lingteng Qiu, Zisheng Ye, Shuguang Cui, Guanying Chen, and Xiaoguang Han. 2024. DreamDissector: Learning Disentangled Text-to-3D Generation from 2D Diffusion Priors. arXiv preprint arXiv:2407.16260 (2024).
- [81] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. 2019. Pointflow: 3d point cloud generation with continuous normalizing flows. In Proceedings of the IEEE/CVF international conference on computer vision. 4541–4550.
- [82] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023).
- [83] Yu Zeng, Vishal M Patel, Haochen Wang, Xun Huang, Ting-Chun Wang, Ming-Yu Liu, and Yogesh Balaji. 2024. JeDi: Joint-Image Diffusion Models for Finetuning-Free Personalized Text-to-Image Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 6786–6795.
- [84] Dingxi Zhang, Yu-Jie Yuan, Zhuoxun Chen, Fang-Lue Zhang, Zhenliang He, Shiguang Shan, and Lin Gao. 2024. Stylizedgs: Controllable stylization for 3d gaussian splatting. arXiv preprint arXiv:2404.05220 (2024).
- [85] Kai Zhang, Nick Kolkin, Sai Bi, Fujun Luan, Zexiang Xu, Eli Shechtman, and Noah Snavely. 2022. Arf: Artistic radiance fields. In European Conference on Computer

StyleMe3D: Stylization with Disentangled Priors by Multiple Encoders on 3D Gaussians

Vision. Springer, 717-733.

- [86] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition. 586–595.
- [87] Yuechen Zhang, Zexin He, Jinbo Xing, Xufeng Yao, and Jiaya Jia. 2023. Refnpr: Reference-based non-photorealistic radiance fields for controllable scene stylization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4242–4251.
- [88] Junsheng Zhou, Weiqi Zhang, and Yu-Shen Liu. 2024. DiffGS: Functional Gaussian Splatting Diffusion. arXiv preprint arXiv:2410.19657 (2024).
  [89] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou.
- [89] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. 2024. StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation. arXiv preprint arXiv:2405.01434 (2024).

# Appendix

# **A** Preliminary

# A.1 Style-aware Image Customization

In recent advancements in style transfer, StyleShot [16] and IP-Adapter [82] represent two prominent techniques, each employing distinct methods to transfer style from a reference image to a target image. StyleShot emphasizes the extraction of detailed style features using a style-aware encoder, which leverages multiscale patch partitioning to capture both low-level and high-level style cues. Specifically, StyleShot divides the reference image into non-overlapping patches of three sizes, corresponding to different scales. For each patch scale, there is a dedicated ResBlock at different depths.

The following are the key formulas for style injection in StyleShot:

Attention
$$(Q, K_s, V_s) = \operatorname{softmax}\left(\frac{QK_s^T}{\sqrt{d}}\right) \cdot V_s$$
 (S1)

where Q is the query projected from the latent embeddings f, and  $K_s$  and  $V_s$  are the keys and values, respectively, that the style embeddings  $f_s$  are projected onto through independent mapping functions  $W_{K_s}$  and  $W_{V_s}$ . The attention outputs of the text embeddings  $f_t$  and style embeddings  $f_s$  are then combined into new latent embeddings f', which are fed into subsequent blocks of Stable Diffusion:

$$f' = \text{Attention}(Q, K_t, V_t) + \lambda \text{Attention}(Q, K_s, V_s)$$
 (S2)  
where  $\lambda$  represents the weight balancing the two components.

# A.2 Score Distillation Sampling for 3D Generation

Text-guided 3D generation has gained significant attention due to advancements in methods such as Score Distillation Sampling (SDS) [53], which facilitates the optimization of 3D representations using pre-trained diffusion models. SDS optimizes the parameters  $\theta$  of a 3D model  $g(\theta)$  by distilling gradients from a diffusion model  $\phi$ , ensuring that 2D projections generated from  $g(\theta)$  align with a target text prompt. The gradient of the SDS loss is defined as:

$$\nabla_{\theta} L_{\text{SDS}}(\phi, x = g(\theta)) = \mathbb{E}_{t,\epsilon} \left[ \omega(t) \left( \hat{\epsilon}_{\phi}(z_t; y, t) - \epsilon \right) \frac{\partial x}{\partial \theta} \right], \quad (S3)$$

where  $\hat{\epsilon}_{\phi}(z_t; y, t)$  represents the predicted noise residual from the pre-trained diffusion model,  $\epsilon$  is the actual noise used in the forward process,  $z_t$  is the latent variable at timestep t, and  $\omega(t)$  is a timestep weighting function.

These have been extended to artistic scene generation [36] and combined input conditions, including text and images [74, 80]. Recent advances leveraging latent diffusion models have improved the scope and expressiveness of text-to-3D synthesis [80, 88], supporting more nuanced and creative 3D outputs.

# A.3 3D Gaussian Splatting

3D Gaussian Splatting (3D GS) [29] represents a 3D scene using a collection of spatial Gaussians. Each Gaussian  $g_i$  is defined by a mean position  $\mu_i \in \mathbb{R}^3$  and a covariance matrix  $\Sigma_i \in \mathbb{R}^{3\times 3}$ , which determines its shape and orientation. The Gaussian's influence on a point **x** is given by:

$$G(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\mu_i)^{\top} \sum_{i=1}^{n-1} (\mathbf{x}-\mu_i)}$$
(S4)

where  $\Sigma_i = \mathbf{R}\mathbf{S}\mathbf{S}^{\top}\mathbf{R}^{\top}$  is decomposed into a rotation **R** and scaling **S** matrices. Each Gaussian has an opacity  $\alpha_i$  and a view-dependent color  $c_i$ .

During rendering, Gaussians are projected to 2D and blended using alpha compositing. The final pixel color *C* is calculated as:

$$C = \sum_{i=1}^{n} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j)$$
(S5)

Here,  $\alpha_i$  is the effective opacity of the *i*-th Gaussian in sorted depth order. Gaussian Splatting enables real-time, differentiable rendering and can reconstruct scenes with multi-view supervision.

Compared to NeRF [47], 3D Gaussian Splatting is significantly more efficient in both time and memory usage. By representing scenes with Gaussian primitives rather than dense neural networks, it allows for faster rendering and lower computational costs, making it more suitable for real-time applications.

# **B** Implementation Details

**Computational Environment**: All experiments were conducted on a single NVIDIA L40S GPU with 46GB of VRAM.

**Dataset**: NeRF synthetic dataset [47] and tandt db [29], was used for all experiments.

# B.1 Details of Dynamic Style Score Distillation (DSSD)

- (1) Backbone Models: For the style-aware diffusion model, we adopt StyleShot, which builds on IP-Adapter and incorporates a style-aware encoder to enhance style representation, enabling robust style transfer through score distillation guidance.
- (2) **Fine Timestep Sampling**: We employ a fine-grained timestep sampling strategy with a timestep constant T = 1000. Minimum and maximum timesteps were set as  $T_{min} = 0.02 \cdot T$  and  $T_{max} = 0.75 \cdot T$ , respectively. The noise intensity was dynamically reduced to high, medium, and low levels to stabilize the updates during training.
- (3) Dynamic Guidance Coefficients: The dynamic guidance coefficient Δλ was tuned to adapt to varying scales of the dataset and style variations. For the NeRF Synthetic dataset, we selected λ<sub>max</sub> = 20 and confined Δλ within [7.5, 20].
- (4) Guidance Modes and Outpainting Strategy: A total of 2800 steps were employed, segmented into specific guidance modes:
  - Main RGB Loss (Local Mode): Steps 100 to 600.
  - Adaptive Iteration (Global Mode): Steps 1 to 1000, alternating between global RGB and global SDS losses.
  - Fixed or Free Global Modes: Steps 1000 to 1900, alternating between global-fix and global-free modes.
  - Local Mode: Steps 1900 to 2800.
  - This hybrid strategy begins with global optimization before transitioning to local refinement, requiring 1800 iterations for SDS loss.
- (5) Iteration Time and Cost Analysis:

StyleMe3D: Stylization with Disentangled Priors by Multiple Encoders on 3D Gaussians

ArXiv Preprint, 2025,

- Average Time Per Iteration: Single-view RGB loss averaged 0.1 seconds, while SDS loss averaged 2.5 seconds.
- Total Iteration Count and Convergence: Using RGB loss for the initial 1000 steps and SDS loss for the subsequent 2000 steps, convergence was achieved in approximately 2600 seconds. For enhanced local convergence, an additional 500 to 1000 SDS iterations were applied.

# B.2 Details of Simultaneously Optimized Scale (SOS)

- (1) VGG Feature Extraction
  - Style layers:
  - ['r11','r21','r31','r41','r51']
  - Content layer: ['r42']
  - Gram matrix weights: [1e3/64<sup>2</sup>, 1e3/128<sup>2</sup>, 1e3/256<sup>2</sup>, 1e3/512<sup>2</sup>, 1e3/512<sup>2</sup>]
- (2) **Two-Phase Optimization** 
  - Pretraining phase:
    - Trigger: optimize\_iteration=10000 and current\_iter
       < 10000</li>
    - Fixed scale: optimize\_size=0.5 (uses minimum resize\_images if unspecified)
    - Downsampling: Bilinear interpolation mode="bilinear"
  - Full multi-scale phase: Activates all resize\_images scales

# B.3 Details of Contrastive Style Descriptor (CSD)

• Deployed CSD ViT-L style encoder pretrained on LAION-Styles dataset.

# B.4 Details of 3D Gaussian Quality Assessment (3DG-QA)

- Integrated CLIP-ViT-B with antonymic prompts: "Good, Sharp, Colorful" vs "Bad, Blurry, Dull", prompts=("quality", "sharpness", "colorfulness")
- loss = 1 (0.4\*scores['quality'] + 0.4\*scores['sharpness'] + 0.2\*scores['colorfullness']).mean(), where  $w_q = whe$ 0.4,  $w_s = 0.4$ ,  $w_c = 0.2$  denote quality, sharpness, and the product of the second second

# C Additional Method Analysis

The challenges of directly transferring 3D generation techniques to 3D stylization stem from the optimization gap between pretraining and post-training stages. This section provides a theoretical and visual analysis of this gap.

# C.1 Misalignment in Optimization Pathways

- **Pre-training Objective**: The goal of 3D reconstruction during pre-training is to capture geometric and photometric properties accurately. This optimization process is typically smooth and guided by explicit ground truth data.
- **Post-training Objective**: In the post-training phase, the focus shifts to aesthetic alignment using style-aware guidance, which lacks explicit supervision and introduces higher uncertainty.



Figure S1: Optimization Pathways for Pre-training vs. Posttraining. The plot illustrates the optimization pathways for *pre-training* (blue solid line) and *post-training* (orange dashed line), highlighting the *optimization gap* (gray shaded area) between 3D reconstruction and stylization. The pretraining pathway shows smooth, steady convergence, while the post-training pathway oscillates due to inherent uncertainty in stylization. The optimization gap represents misalignment between the stages, emphasizing the need for alignment techniques, such as *style-aware priors* and *dynamic guidance*, to achieve stable and consistent 3D stylization.

• **Disjoint Loss Landscapes**: The loss landscapes for pretraining and post-training differ significantly. Pre-training minimizes reconstruction errors, while stylization involves abstract priors from style information, leading to potential misalignment.

The optimization pathways during pre-training and post-training can be represented as two distinct loss functions:

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{recon}}(G_{\text{pre}}(x), x_{\text{gt}}), \tag{S6}$$

where  $\mathcal{L}_{\text{recon}}$  minimizes geometric and photometric errors between the predicted  $G_{\text{pre}}(x)$  and ground truth  $x_{\text{gt}}$ , and:

$$\mathcal{L}_{\text{post}} = \mathcal{L}_{\text{style}}(G_{\text{post}}(x), s_{\text{ref}}), \tag{S7}$$

where  $\mathcal{L}_{style}$  aligns the generated results  $G_{post}(x)$  with a style reference  $s_{ref}$  using abstract priors.

The optimization gap can then be formulated as:

$$\Delta \mathcal{L} = \left| \mathcal{L}_{\text{pre}} - \mathcal{L}_{\text{post}} \right|, \tag{S8}$$

where  $\Delta \mathcal{L}$  quantifies the divergence between the loss landscapes, reflecting the mismatch in optimization objectives.

# C.2 High Uncertainty in Style Information

- Multi-modal Style Representations: Styles are inherently diverse and lack well-defined ground truth, making the optimization process less predictable.
- **Temporal Instability**: Stylization optimization pathways often exhibit oscillations due to conflicts between style priors and geometric constraints.

The uncertainty in style optimization can be modeled as the variance in style priors:

$$\sigma_{\text{style}}^2 = \text{Var}(s_{\text{ref}}),\tag{S9}$$

where *s*<sub>ref</sub> represents multi-modal style representations. Temporal oscillations in optimization can be expressed as:

$$\delta_t = \left| \nabla \mathcal{L}_{\text{post},t+1} - \nabla \mathcal{L}_{\text{post},t} \right|, \qquad (S10)$$

where  $\delta_t$  measures the instability between consecutive timesteps t and t + 1.

# C.3 Visualization Analysis

The graph (Figure S1) visualizes the optimization gap between pre-training and post-training:

- **Pre-training pathway** (blue solid line) shows smooth convergence, reflecting steady optimization for geometric fidelity.
- **Post-training pathway** (orange dashed line) exhibits oscillations, driven by the abstract and subjective nature of style priors.
- **Optimization gap** (gray shaded area) represents the divergence between the two pathways, indicating the challenges of transitioning between the stages.

To bridge the optimization gap, alignment strategies must minimize:

$$\min_{\Omega} \Delta \mathcal{L} + \lambda_{\text{cons}} \mathcal{L}_{\text{consistency}}, \tag{S11}$$

where  $\mathcal{L}_{consistency}$  enforces multi-view consistency, and  $\lambda_{cons}$  is a weighting factor to balance consistency with style fidelity.

# C.4 Key Observations and Insights

(1) **Mismatch in Optimization**: The smooth convergence of pre-training contrasts with the oscillatory adjustments in post-training, reflecting the differences in objectives—geometric accuracy vs. subjective style transfer.

The loss landscapes  $\mathcal{L}_{pre}$  and  $\mathcal{L}_{post}$  differ fundamentally in their curvature:

$$\kappa_{\rm pre} \ll \kappa_{\rm post},$$
 (S12)

(S13)

where  $\kappa$  represents the curvature, indicating smoother optimization for pre-training compared to post-training.

- (2) Impact of the Gap: The optimization gap introduces challenges such as:
  - **Optimization Instability**: Misaligned pathways can lead to instability during post-training.
  - Inconsistent Stylization: Divergent trajectories may result in geometric distortions or incomplete stylization. Misaligned pathways can exacerbate:
  - Instability: ΔL leads to higher gradients:

$$\nabla \mathcal{L}_{\text{post}} \gg \nabla \mathcal{L}_{\text{pre.}}$$

• Inconsistency: Variance in style priors 
$$\sigma_{\text{style}}^2$$
 introduces

inconsistencies in multi-view stylization.

(3) **Bridging the Gap**: Effective strategies such as *style-aware diffusion priors*, *dynamic style score distillation*, and *progressive style outpainting* are critical to aligning pathways and ensuring robust stylization.

Cailin Zhuang, Yaoqi Hu, Xuanyang Zhang, Wei Cheng et al.

Introducing regularization terms:

$$\mathcal{L}_{\text{align}} = \lambda_{\text{prior}} \mathcal{L}_{\text{style}} + \lambda_{\text{geo}} \mathcal{L}_{\text{recon}}, \qquad (S14)$$

where  $\lambda_{\text{prior}}$  and  $\lambda_{\text{geo}}$  balance style fidelity and geometric preservation, helps align the pathways.

# C.5 Conclusion

This analysis highlights the inherent challenges in aligning pretraining and post-training optimization pathways. The visualization emphasizes the need for dedicated techniques to bridge the gap, ensuring high-fidelity and consistent stylization while maintaining geometric coherence.

# **D** More Visual Result

As shown in Figure S2, S3, and S4, we demonstrate our method's performance across nine distinct styles (sky painting, cartoon, watercolor, fire, cloud, Wukong, drawing, color oil, and sketch) on three datasets (chair, hotdog, and mic).

StyleMe3D: Stylization with Disentangled Priors by Multiple Encoders on 3D Gaussians



Figure S2: More visual results. Demonstration of our method's performance across nine distinct styles (sky painting, cartoon, watercolour, fire, cloud, Wukong, drawing, color oil, and sketch) applied to chair.



Figure S3: More visual results. Demonstration of our method's performance across nine distinct styles (sky painting, cartoon, watercolour, fire, cloud, Wukong, drawing, color oil, and sketch) applied to hotdog.

Cailin Zhuang, Yaoqi Hu, Xuanyang Zhang, Wei Cheng et al.



Figure S4: More visual results. Demonstration of our method's performance across nine distinct styles (sky painting, cartoon, watercolour, fire, cloud, Wukong, drawing, color oil, and sketch) applied to mic.