Style4D-Bench: A Benchmark Suite for 4D Stylization

Beiqi Chen[*1,2], **Shuai Shao**[*2], **Haitang Feng**[3,2], **Jianhuang Lai**[4], **Jianlou Si**[†5], **Guangcong Wang**[† 2]

[1]Harbin Institute of Technology, China
[2]Vision, Graphics, and X Group, Great Bay University, China
[3]Nanjing University, China
[4]Sun Yat-Sen University, China
[5]Alibaba Group, China

Abstract

We introduce **Style4D-Bench**, the first benchmark suite specifically designed for 4D stylization, with the goal of standardizing evaluation and facilitating progress in this emerging area. Style4D-Bench comprises: **1)** a comprehensive evaluation protocol measuring spatial fidelity, temporal coherence, and multi-view consistency through both perceptual and quantitative metrics, **2)** a strong baseline that makes an initial attempt for 4D stylization, and **3)** a curated collection of high-resolution dynamic 4D scenes with diverse motions and complex backgrounds. To establish a strong baseline, we present **Style4D**, a novel framework built upon 4D Gaussian Splatting. It consists of three key components: a basic 4DGS scene representation to capture reliable geometry, a Style Gaussian Representation that leverages lightweight per-Gaussian MLPs for temporally and spatially aware appearance control, and a Holistic Geometry-Preserved Style Transfer module designed to enhance spatio-temporal consistency via contrastive coherence learning and structural content preservation. Extensive experiments on Style4D-Bench demonstrate that Style4D achieves state-of-the-art performance in 4D stylization, producing fine-grained stylistic details with stable temporal dynamics and consistent multi-view rendering. We expect Style4D-Bench to become a valuable resource for benchmarking and advancing research in stylized rendering of dynamic 3D scenes.

Code — https://becky-catherine.github.io/Style4D

Introduction

Recent advances in 4D scene representations (Wu et al. 2024; Luiten et al. 2024; Duan et al. 2024) have enabled high-fidelity modeling of dynamic environments, unlocking new possibilities for immersive content creation in virtual reality and film production. As applications become more demanding, there is a growing need not only for accurate reconstructions but also for controllable appearance and stylization of dynamic 4D content. Users may wish to stylize a 4D scene to reflect specific artistic intents, emotional tones, or narrative contexts—while maintaining both temporal coherence and multi-view consistency.

Despite recent progress in related areas like 2D and 3D stylization, the field of *4D stylization* remains largely underexplored, with no standardized datasets, evaluation metrics, or task definitions. Existing methods adapted from 2D (Gu et al. 2018; Kolkin, Salavon, and Shakhnarovich 2019; Liao et al. 2017), 3D (Chiang et al. 2022; Liu et al. 2024; Saroha et al. 2024; Zhang et al. 2024), or video (Lai et al. 2018; Deng et al. 2021) stylization often fail to meet the unique challenges of 4D settings, such as jointly preserving spatial fidelity, temporal stability, and multi-view consistency under complex motion and occlusion. Moreover, recent dynamic scene representations like 4D Gaussian Splatting (4DGS) (Wu et al. 2024) offer strong rendering capabilities, but have not yet been explored in the context of stylization.

To fill this gap, we introduce **Style4D-Bench**—the first benchmark suite dedicated to 4D stylization. Our benchmark provides: 1) A curated set of high-resolution, complex-background dynamic 4D scenes exhibiting diverse motions, deformations, and view-dependent effects; 2) A comprehensive evaluation protocol, including both perceptual and quantitative metrics, to assess spatial fidelity, temporal coherence, and cross-view consistency. Style4D-Bench is designed to facilitate fair, reproducible, and scalable evaluation of future 4D stylization approaches.

To establish a strong baseline within our benchmark, we also propose **Style4D**, a novel 4D stylization framework based on 4D Gaussian Splatting. Style4D consists of three key components: a basic 4DGS scene representation for geometry modeling, a *Style Gaussian Representation* that incorporates lightweight per-Gaussian MLPs for time- and depth-aware stylization, and a *Holistic Geometry-Preserved Style Transfer* module to enhance spatio-temporal consistency through contrastive learning and content-aware regularization.

Our contributions are summarized as follows: 1) We present **Style4D-Bench**, the first benchmark suite for 4D scene stylization, offering standardized datasets, tasks, and evaluation metrics to drive progress in this emerging area. 2) We define core evaluation challenges in 4D stylization, including spatial fidelity, temporal stability, and multi-view consistency, and design comprehensive protocols to quantify them. 3) We propose **Style4D** as a strong baseline method, which leverages 4D Gaussian Splatting with a style-aware representation and a holistic spatio-temporal transfer mod-

[*]These authors contributed equally.
[†]Corresponding authors

ule. 4) Extensive experiments on Style4D-Bench demonstrate the effectiveness of our method and reveal insights into current limitations and future opportunities in 4D stylization.

Related Works

3D Representations. Neural Radiance Fields (NeRFs) (Mildenhall et al. 2021) represent 3D scenes as continuous volumetric functions using MLPs, which enables high-quality novel-view synthesis. Later work introduced compact representations such as decomposed tensors (Fridovich-Keil et al. 2023), hash tables (Müller et al. 2022), and voxel grids (Fridovich-Keil et al. 2022) to improve efficiency and quality. 3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) offers an explicit 3D Gaussian representation that supports real-time rasterization rendering. This explicit structure also makes scene editing easier (Chen et al. 2024a; Haque et al. 2023; Liu et al. 2023; Wang et al. 2022; Zhuang et al. 2023). However, 3DGS is currently limited to static scenes.

4D Representations. Early works extending NeRF to dynamic settings (Pumarola et al. 2021; Park et al. 2021) directly modeled temporal variations, whereas subsequent methods (Cao and Johnson 2023; Shao et al. 2023) improved efficiency via voxel-based decompositions. Despite these advances, achieving real-time rendering while preserving fine-grained geometry under motion remains challenging. 4D Gaussian Splatting (4DGS) (Wu et al. 2024; Duan et al. 2024) addresses dynamic scenes by extending Gaussians into the temporal domain. In contrast to Dynamic3DGS (Luiten et al. 2024), which incurs linear memory growth by storing per-frame parameters, 4DGS employs a deformation network for compact and efficient temporal modeling. However, these 4D representations currently lack mechanisms for stylization and temporal consistency, motivating our benchmark and method for 4D scene stylization.

2D Style Transfer. Neural style transfer, pioneered by (Gatys, Ecker, and Bethge 2016), demonstrated that CNNs can effectively separate and recombine content and style from images. The key insight that second-order statistics of VGG features capture style information led to numerous improvements. Feed-forward methods like AdaIN (Huang and Belongie 2017) significantly accelerated stylization by aligning feature statistics, while recent works have focused on improving semantic consistency and texture preservation (Gu et al. 2018; Kolkin, Salavon, and Shakhnarovich 2019). To maintain temporal coherence across frames, MCCNet (Deng et al. 2021) proposes a Multi-Channel Correlation network. These foundational techniques form the basis for our 4D stylization component, which we enhance with temporal consistency mechanisms to handle dynamic scenes.

3D and Video Stylization. 3D style transfer has attracted significant attention, with NeRF-based methods achieving multi-view consistency via optimization (Wang et al. 2023) or feed-forward networks (Liu et al. 2021; Chiang et al. 2022). With 3D Gaussian Splatting, StyleGaussian (Liu et al. 2024) enables instant stylization by aligning VGG features in Gaussians with style images, while GSS (Saroha et al. 2024) and StylizedGS (Zhang et al. 2024) require per-style optimization. Instruction-driven 3D editing methods like Instruct-GS2GS(Vachha and Haque 2024) are also related. However, these static-scene methods fail to handle 4D temporal dynamics (e.g., motion, deformation, occlusion), causing view inconsistencies and temporal flickering. Existing video stylization (Chen et al. 2024b; Huang et al. 2022; Wang et al. 2023) focus on 2D temporal coherence, lacking 3D understanding, while 3D stylization methods ignore temporal dynamics.

4D Stylization and Benchmark. Despite progress in 2D/3D stylization, *4D scene stylization* remains largely unexplored. The core challenge is ensuring simultaneous multi-view consistency and temporal stability. StyleDyRF (Xu et al. 2024a) stylizes dynamic NeRFs via a canonical style transformation, but inherits NeRF's high rendering cost and low geometric precision. Concurrent work like 4DStyleGaussian (Liang et al. 2024) attempts 4D stylization but is limited in content preservation, geometric fidelity, and temporally adaptive control. Diffusion-based methods like Instruct 4D-to-4D (Mou, Chen, and Wang 2024) apply 2D diffusion per-view to pseudo-3D sequences, but suffer from geometric distortion and high computational cost. CLIP-Gaussian (Howil et al. 2025) guides 2D-4D GS stylization with CLIP/VGG, but modifies both color and geometry, lacking explicit temporal modeling. Our baseline addresses these challenges by introducing per-Gaussian MLPs for fine-grained, temporally-aware appearance modulation, and integrating an enhanced 2D stylization module to preserve spatio-temporal consistency.

Designing a benchmark for 4D stylization poses fundamental challenges. The absence of ground truth and the subjective nature of style make quantitative evaluation inherently difficult. Moreover, measuring temporal coherence and multi-view consistency is non-trivial due to complex spatial-temporal dynamics and the lack of reference sequences. Existing evaluation protocols from 2D, 3D, or video stylization are insufficient for capturing these aspects, highlighting the need for dedicated datasets and tailored metrics.

Style4D-Bench

Existing 4D stylization tasks lack a unified quantitative evaluation standard. Some studies (Liang et al. 2024) measure consistency by calculating frame-wise PSNR, MSE, and other metrics, using style loss to assess stylization degree. In contrast, others (Liu et al. 2021) overly rely on user studies. While these metrics simplify the evaluation process, they introduce several issues. First, simple style losses like MSE overly focus on pixel-level differences between stylized and reference images, failing to accurately reflect overall similarity in semantic and textural aspects. Second, 4D stylization is a complex and multidimensional concept where individual preferences may prioritize different aspects. For instance, some emphasize stylization degree, considering blurriness of characters or objects as part of style, while others prioritize preserving the structure and details of the original scene. To address these challenges, we propose a decomposition method that breaks down the evaluation of 4D stylization into multiple dimensions for a more comprehensive and

nuanced assessment.

We divide the evaluation of 4D stylization into two aspects: 4D stylization quality and 4D stylization consistency. For 4D stylization quality, we focus on assessing the frame-by-frame quality of stylized videos rendered at arbitrary times and angles, without considering similarity to a reference style. Regarding 4D stylization consistency, we distinguish between Temporal Quality and Stylization Quality. Our evaluation method encompasses a total of six specific dimensions, comprising 12 detailed metrics in total.

4D Stylization Quality

Frame-Wise Quality - Imaging Quality. Imaging quality refers to the distortions (e.g., blurring, high noise, overexposure) present in each frame after 4D stylization. We assess this quality using three metrics: UIQM, Clipiqa+(Wang, Chan, and Loy 2023), and Musiq(Ke et al. 2021).

- UIQM: Integrates three dimensions of image quality—color, sharpness, and contrast—and computes the overall image quality through a weighted averaging approach, aligning with human visual perception.
- Clipiqa+: A CLIP-based image quality assessor fine-tuned using CoOp, designed to evaluate overall image quality.
- Musiq: A multi-scale image quality assessor trained on the KonIQ dataset , capable of capturing and evaluating image quality at various granularities.

Frame-Wise Quality - Aesthetic Quality. Aesthetic quality refers to the artistic and aesthetic value of each video frame, reflecting aspects such as layout, color richness and harmony, photorealism, naturalness, and artistic quality. We assess aesthetic quality using the Qalign(Wu et al. 2023) and Musiq-paq2piq metrics.

- Qalign: A pre-trained large multimodal model (LLM) capable of simultaneously assessing both image quality and aesthetic performance.
- Musiq-paq2piq: A multi-scale image aesthetic assessor trained on the PaQ-2-PiQ dataset, capable of capturing image quality and performing aesthetic evaluation at various granularities.

4D Stylization Consistency

Temporal Quality - Spatiotemporal Consistency. Temporal consistency refers to maintaining continuity between frames and consistency across multiple viewpoints as time and perspectives change. We evaluate multiple helical trajectory videos rendered for each scene using Dists and Warp loss metrics to assess cross-frame consistency.

- Dists: Comprehensively measures both the structural and textural similarity between two images to evaluate their overall quality and perceptual consistency.
- Warp loss: Using RAFT to compute optical flow, mapping the next frame to the current frame, and calculating the L1 error between the current frame and the mapped frame effectively evaluates spatial and motion consistency of images in a temporal sequence

Temporal Quality - Subject Consistency. For stylized videos rendered from a fixed viewpoint, we assess whether the appearance of a subject (e.g., a person, curtains, etc.) remains consistent throughout the entire video. To achieve this, we compute the inter-frame DINO(Caron et al. 2021) feature similarity.

Stylization Quality - Style Consistency. Style consistency refers to the degree of stylization in 4D stylization, evaluated using CKDN(Zheng et al. 2021) and LPIPS metrics to assess the similarity between video frames and style images, representing the level of stylization.

- CKDN: Utilizes learned representations from degraded images to assess style similarity, enabling comprehensive evaluation of both image quality and style similarity.
- LPIPS: A metric based on features extracted from VGG, measuring perceptual similarity between images, effectively assessing similarity to style images.

Stylization Quality - Content Consistency. Content Consistency refers to the semantic and content similarity between stylized 4D scenes and their original counterparts. We employ SSIM and LPIPS metrics to compute the similarity of each frame to the original scene.

User Study Design

To thoroughly evaluate our method, we conducted a two-part user study with 34 participants (N=34). The first part assessed 4D stylization, presenting participants with five video pairs (approx. 10s, 300 frames each) rendered from test and arbitrary viewpoints, along with four selected frames for detailed qualitative assessment. The second part validated our proposed HGST method for video stylization, featuring two 10-second video pairs (300 frames each) and six selected frames for fine-grained evaluation.

Metric. For the extracted single frames, we evaluated two metrics: stylization quality and image quality. Stylization quality measures the extent to which edges and textures are transformed to reflect the target style without compromising the original image structure. Image quality assesses the clarity of object boundaries and facial details.

For the continuous long videos, we adopted three metrics: stylization quality, spatiotemporal consistency, and video quality. Stylization quality evaluates how well the video's style matches the reference while preserving the original structure. Spatiotemporal consistency measures the coherence of the video across temporal progression and viewpoint changes. Video quality assesses the clarity of fine details and textures, as well as overall visual preference.

Style4D: A Strong Baseline

Overview of Style4D. In 4D stylization, the direct use of style transformation leads to low multi-view consistency and significant blurry artifacts. To address these issues, we propose Style4D, a new dynamic scene stylization framework, based on the decoupling of geometry learning and style learning. The framework of Style4D is illustrated in Figure 1. Style4D consists of three key components, a basic 4DGS representation, a Style Gaussian Representation, and a Holistic Geometry-preserved Style Transfer.

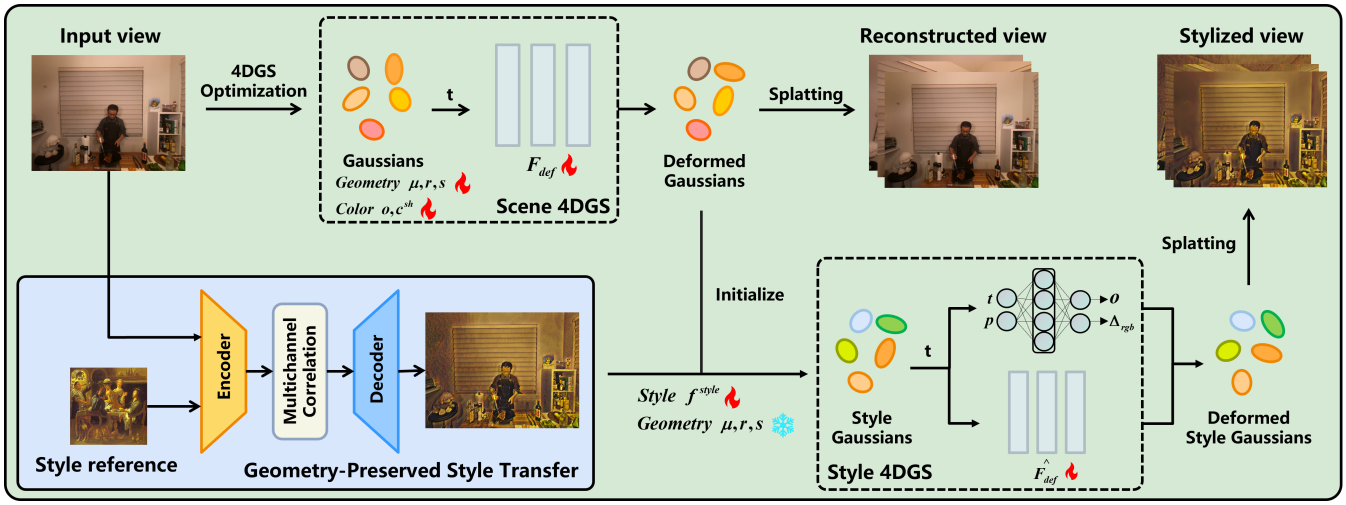

Figure 1: Overview of Style4D. Style4D consists of three key components, a basic 4DGS representation, a Style Gaussian Representation, and a Holistic Geometry-preserved Style Transfer. We first train a basic 4DGS representation with the content image to obtain 4D scene geometry. Then we propose a new Style Gaussian Representation for 4D stylization. We also introduce a Holistic Geometry-preserved Style Transfer module to improve consistency and quality of stylization.

We first train a basic 4DGS representation given multiple views I_{content} to obtain the static Gaussian sequence $\mathcal{G}_i = \{\boldsymbol{\mu}_i, \boldsymbol{r}_i, \boldsymbol{s}_i, o_i, c_i^{\text{sh}}\}$ and the corresponding Gaussian Deformation Field Network F_{def}, capturing the geometry of a dynamic scene. To stylize a 4D scene represented by 4DGS, we propose a Style Gaussian Representation method. It is a novel type of Gaussian ellipsoid, with attributes defined as $\mathcal{G}_i = \left\{\boldsymbol{\mu}_i, \boldsymbol{r}_i, \boldsymbol{s}_i, \mathbf{f}_i^{\text{style}}\right\}$. We design a tiny MLP as part of the style attribute f_i^{style}. The design provides pixel-level stylization mapping, and thus achieves finer color expression while balancing local and global consistency, which significantly improves multi-view consistency. Finally, we introduce a geometry-preserving style transfer approach that integrates an attention-guided 2D stylization module with contrastive coherence learning, enabling the generation of high-quality and temporally consistent training frames.

Style Gaussian Representation. Inspired by SuperGaussians (Xu et al. 2024b), which enhances 2DGS using bilinear interpolation and spatially varying features, we introduce a *Style Gaussian Representation* for 4D scenes. Extending kernel functions to four-dimensional interpolation does not effectively capture the four-dimensional variations in scenes. Therefore, we introduce Gaussian MLP features. It is worth noting that directly mapping intersection coordinates to color and opacity can lead to overfitting. Thus, we map color and opacity variations based on intersection depth. Specifically, each Gaussian is assigned a tiny MLP and a style code f^{style} to modulate color and opacity over space and time. Given the camera pose $M = [R, T]$, we compute the ray-ellipsoid intersection point p_t from pixel p and time t, following (Yu, Sattler, and Geiger 2024). The color is then rendered as:

$$c(p) = \sum_{i \in \mathcal{N}(p)} (c_i + F_c(p_t, t)) \cdot F_\alpha^i \prod_{j=1}^{i-1} (1 - \alpha_j(p)), \quad (1)$$

where c_i is the view-dependent base color, $F_c(p_t, t)$ is the style-driven color increment from the MLP, and $\alpha_j(p)$ denotes opacity. All F_c and F_α^i terms are predicted per-Gaussian via MLPs, with p_t as the ray-Gaussian intersection depth. This formulation preserves 3D geometry, allows precise temporal control, and enhances multi-view consistency.

Holistic Geometry-preserved Style Transfer. Long video stylization remains challenging, especially for high-resolution sequences. Diffusion-based methods (Feng et al. 2024; Kara et al. 2024) often suffer from structural distortions and temporal inconsistency. Optical flow-based constraints (Liu and Zhu 2021) improve coherence but are computationally expensive and scale poorly. Self-supervised approaches (Kong et al. 2024; Wu et al. 2022) reduce flickering but may introduce artifacts such as hollow textures and sharp pixel boundaries due to lack of semantic guidance.

To address these issues, we propose a *Holistic Geometry-preserved Style Transfer* (HGST) module based on an encoder-transformer-decoder architecture. We fuse style and content features using Multichannel Correlation to ensure global consistency, and decode stylized outputs with better structural integrity. However, the encoder-decoder pipeline still leads to temporal instability, particularly on large unseen frames. Inspired by (Wu et al. 2022), we introduce a dual constraint: an attention-guided local contrastive loss $\mathcal{L}_{lcl}$ and a global feature consistency loss $\mathcal{L}_{content}$, which jointly enhance local coherence and global structure, effectively mitigating flickering and spatial artifacts. To enhance local and global coherence, we extract multi-scale features from I_{cs}, $I_{content}$, and I_{style} via an encoder, denoted as f_i^{cs},

Method	Dataset	Imaging Quality			Aesthetic Quality		Spatio-Temp Consistency		Subject Consistency	Style Consistency		Content Consistency	
		UIQM↑	Clipiqa+↑	Musiq↑	Qalign↑	Musiq(paq2piq)↑	Dists↓	Warp Loss↓	DINO Score↑	CKDN↑	LPIPS↓	SSIM↑	LPIPS↓
AdaIN	cook_spinach	1.2995	0.4209	49.4095	2.8665	60.2165	0.0114	0.0091	0.9309	0.2084	0.6913	0.4763	0.4898
AdaAttN		1.7240	0.3770	36.1325	2.1298	50.7028	0.0215	0.0117	0.9078	0.2173	0.6904	0.6444	0.2841
4DSGaussian		1.0834	0.4267	44.2200	2.7019	55.2280	0.0121	**0.0053**	**0.9403**	0.1978	0.7106	0.7646	0.2159
Style4D(Ours)		**1.9290**	**0.4437**	**53.4681**	**3.2072**	**65.8520**	**0.0112**	0.0058	0.9395	**0.2290**	**0.6866**	**0.7771**	**0.1834**
AdaIN	flame_salmon	1.6918	0.3336	41.3119	2.8497	51.6363	0.0171	0.0141	0.9267	0.2193	0.7605	0.5201	0.4030
AdaAttN		1.2928	0.3120	39.9117	2.2357	54.5593	0.0271	0.0169	0.9072	0.1966	0.7861	0.6054	0.2944
4DSGaussian		1.5488	0.3218	51.3875	3.2182	61.4087	0.0140	0.0074	0.9402	0.1897	0.7701	0.5081	0.6051
Style4D(Ours)		**1.7529**	**0.3962**	**55.2302**	**3.6030**	**63.5178**	**0.0138**	**0.0067**	**0.9415**	**0.2354**	**0.7602**	**0.6963**	**0.2704**
AdaIN	sear_steak	1.3544	0.3430	51.6330	2.6115	62.2037	0.0129	0.0078	0.9463	0.2352	0.7114	0.6000	0.2987
AdaAttN		1.1910	0.3996	43.6474	2.0062	63.0785	0.0234	0.0126	0.9021	0.3204	0.7050	0.5153	0.4768
4DSGaussian		1.3843	0.3613	42.1204	2.5841	56.5093	0.0131	**0.0050**	0.9530	0.2722	0.7161	0.6557	0.4819
Style4D(Ours)		**1.6818**	**0.4176**	**53.3443**	**2.8820**	**68.0488**	**0.0108**	0.0066	**0.9564**	**0.3239**	**0.7014**	**0.7503**	**0.2146**

Table 1: Quantitative comparisons of our proposed Style4D against state-of-the-art methods on Style4D-Bench.

f_i^c, and f_i^s for layers $i = 1$ to 5. For $i = 3$ to 5, we apply CBAM (Woo et al. 2018) to enhance salient features and randomly sample N locations G_x and their 8-nearest neighbors $G_{x,y}$ with small perturbations. Local differences are defined as $d^g_{x,y} = G_x^{f_i^{cs}} - G_{x,y}^{f_i^{cs}}$, $d^c_{x,y} = G_x^{f_i^{c}} - G_{x,y}^{f_i^{c}}$. We employ a contrastive loss to maximize similarity between aligned local differences:

$$\mathcal{L}_{lcl} = \sum_{m=1}^{8N} -\log \frac{\exp(d^g_m \cdot d^c_m/\tau)}{\sum_{n=1}^{8N} \exp(d^g_m \cdot d^c_n/\tau)}, \quad \tau = 0.07. \tag{2}$$

To mitigate artifacts and preserve global structure, we introduce a global content loss:

$$\mathcal{L}_{content} = \frac{1}{N} \sum_{i=1}^{N} \|f_{csi} - f_{ci}\|_2^2. \tag{3}$$

The final consistency loss combines both terms:

$$\mathcal{L}_{consistency} = \mathcal{L}_{lcl} + \mathcal{L}_{content}. \tag{4}$$

Training Objective. We first train a 4D Gaussian Splatting model using multi-view content images $I_{content}$ to reconstruct the scene geometry, yielding a static Gaussian sequence $\mathcal{G}_i$ and a deformation field network F_{def}. We then train a holistic geometry-preserved style transfer module with the following overall objective:

$$\begin{aligned}\mathcal{L}_{total} = \lambda_{consistency}\mathcal{L}_{consistency} + \lambda_{style}\mathcal{L}_{style} \\ + \lambda_{id}\mathcal{L}_{id} + \lambda_{illum}\mathcal{L}_{illum} + \lambda_{ins}\mathcal{L}_{ins},\end{aligned} \tag{5}$$

where $\mathcal{L}_{consistency}$ ensures spatio-temporal coherence (see Eq. 4), and the remaining terms follow conventional stylization objectives (Kong et al. 2024): $\mathcal{L}_{style}$ for perceptual style alignment, $\mathcal{L}_{id}$ for content preservation, $\mathcal{L}_{illum}$ for illumination stability, and $\mathcal{L}_{ins}$ for intra-channel coherence. See supplementary material for definitions. After obtaining high-quality stylized frames, we train a Style Gaussian Representation supervised by these frames. The optimization objective is:

$$\mathcal{L} = \left\| \hat{I} - S_t(I) \right\|_1 + \mathcal{L}_{tv}, \tag{6}$$

where $\hat{I}$ is the rendered image, $S_t(I)$ is the corresponding stylized frame, and $\mathcal{L}_{tv}$ denotes total variation regularization for spatial smoothness.

Experiments

Datasets. We evaluate our model on the real-world Neu3D dataset (Li et al. 2022) to benchmark its performance in realistic scenarios. Neu3D comprises six dynamic scenes, each observed by 15–20 static cameras distributed in space. The dataset features long video sequences (300 frames), with complex scene dynamics and nontrivial viewpoint variations. The videos are recorded at a resolution of 1352 × 1014 pixels, with 300 frames per sequence.

Experiment settings. We build our implementation of Style4D based on the publicly available 4DGS codebase (Wu et al. 2024). During training, we adopt the Adam optimizer with hyperparameters following those used in 4DGS. The batch size is set to 2, and each lightweight per-Gaussian MLP consists of two layers. For each scene, we train the model for up to 14,000 iterations. Both training and inference are performed on a single NVIDIA A40 GPU with 48GB of memory. In practice, training a single scene takes approximately 2 hours, with peak GPU memory usage around 10GB.

Evaluation on Style4D-Bench

Qualitative results. We compare our method with existing 4D stylization methods including 4DStyleGaussian, as well as baseline 4DGS models trained with AdaIN and AdaAttN stylized images, to assess stylization quality. As shown in Figure 2, our method demonstrates stronger temporal consistency compared to 4DGS with AdaIN and AdaAttN, exhibiting fewer artifacts and blurriness on background objects. Moreover, compared to 4DStyleGaussian, our approach significantly enhances stylization effects while maintaining consistency and effectively preserving structural details of the original scenes without excessive smoothing. Due to space limitations, we only present results from fixed test viewpoints here. Stylization results from spiral viewpoints are provided in the appendix, where our method demonstrates stronger consistency and enhanced stylization quality.

Meanwhile, we also compare our proposed Holistic Geometry-preserved Style Transfer model (HGST), with several state-of-the-art 2D stylization methods. As shown in Figure 3, our method outperforms AdaIN, AdaAttN, and

Figure 2: 4D Stylization Comparison: (a) Original scene image, (b) 4DGS with AdaIN, (c) 4DGS with AdaAttN, (d) 4DStyleGaussian, (e) Style4D (Ours).

Method	Frame	
	Stylization Quality	Image Quality
4DGS(AdaIN)	11.76%	2.94%
4DGS(AdattN)	14.70%	8.82%
4DStyleGaussian	11.76%	17.64%
Style4D(Ours)	**61.76%**	**70.58%**

Table 2: Single-frame image performance: results of user study voting

Method	Video		
	Stylization Quality	Video Quality	Spatio-Temp Cons
4DStyleGaussian	23.52%	30.12%	44.11%
Style4D(Ours)	**76.47%**	**69.87%**	**55.88%**
4DGS(AdaIN)	20.58%	11.76%	14.70%
Style4D(Ours)	**79.41%**	**88.23%**	**85.29%**
4DGS(AdaAttN)	14.70%	17.64%	14.70%
Style4D(Ours)	**85.29%**	**82.35%**	**85.29%**

Table 3: User study results for video performance evaluation

MCCNet by improving temporal consistency while maintaining stylization quality, without exhibiting the blocky pixel artifacts observed in CCPL.

Quantitative Evaluation and Analysis. As shown in **Table 1**, we conduct comprehensive quantitative comparisons across three dynamic 4D scenes (*cook_spinach*, *flame_salmon_1*, and *sear_steak*), evaluating multiple aspects of stylization performance, including imaging quality, aesthetic perception, spatial-temporal consistency, and fidelity to both content and style.

Across all datasets, **Style4D** consistently achieves top performance in most metrics. For imaging quality, we observe clear improvements in UIQM, Clipiqa+, and Musiq scores, suggesting that our method preserves structural clarity and visual quality more effectively than all baselines. In particular, Style4D achieves a UIQM of 1.9290 on *cook_spinach* and 1.7529 on *flame_salmon_1*, surpassing prior methods such as 4DGS(AdaIN) and 4DStyleGaussian by a large margin.

In terms of aesthetic quality, our method attains the highest Qalign and Musiq-PAQ2PIQ scores across all scenes, indicating better perceptual stylization aligned with artistic intent. For example, on *flame_salmon_1*, Style4D yields a Qalign of 3.6030 and a Musiq-PAQ2PIQ of 63.5178, outperforming the second-best baseline by significant margins.

Regarding spatial-temporal consistency, our model shows robust performance with the lowest DISTS and Warp Loss values, confirming its ability to produce temporally stable and coherent results. On *sear_steak*, Style4D reduces Dists to 0.0108 and Warp Loss to 0.0066, improving both perceptual smoothness and geometric coherence.

Finally, Style4D also excels in content and style consistency. It achieves the best DINO scores, CKDN accuracy, and LPIPS/SSIM metrics across scenes. These re-

Figure 3: Visualization of HGST stylization method (Ours) compared with other 2D style transfer approaches: (a) Content image; (b) AdaIN; (c) AdaAttN; (d) CCPL; (e) MCCNet; (f) HGST (Ours).

sults demonstrate our model's ability to retain scene semantics while applying stylization, balancing content preservation and stylized appearance. For instance, on *cook_spinach*, it attains a SSIM of 0.7771 and a content LPIPS of 0.1834—clearly outperforming other methods.

Tables 2 and 3 present the results of our user study. It can be observed that our method is consistently preferred in terms of both image quality and overall video quality, while simultaneously maintaining strong stylization effects. These findings align well with the quantitative results reported in Table 1.

These results collectively confirm that Style4D delivers a strong balance of style fidelity, spatial-temporal coherence, and content preservation, establishing new state-of-the-art performance in 4D stylization.

Further Analyses and Ablation Studies. We perform extensive ablations to validate the necessity and effectiveness of each component in Style4D. We further evaluate our method on synthetic data and the D-NeRF benchmark, and include additional comparisons with Instruct-Pix2Pix–based baselines and Instruct-4D-to-4D approaches. Extended results and discussions are provided on our project page.

Conclusion

We present **Style4D-Bench**, the first benchmark for 4D stylization, featuring: **1)** a strong baseline method, **2)** a unified evaluation protocol covering spatial fidelity, temporal coherence, and multi-view consistency, and **3)** a curated set of high-resolution dynamic scenes. To establish the baseline, we propose **Style4D**, a novel 4D stylization framework that combines reliable scene representation, per-Gaussian MLPs for appearance control, and a geometry-preserved stylization module for spatio-temporal consistency. Extensive experiments demonstrate that Style4D outperforms existing methods, delivering high-quality stylization with stable dynamics and coherent multi-view rendering. We expect Style4D-Bench to promote future research in 4D stylized scene synthesis. In future work, we plan to further improve the quality of stylization and broader user-controllable style manipulation for interactive applications.

Limitation. Style4D achieves high-quality stylization with strong spatiotemporal consistency, yet limitations remain. Its multi-stage pipeline and per-Gaussian MLPs incur high training costs, and the framework supports only a single style per scene, without fast style switching.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62506063), the Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515110178), and the Start-up Funding (No. YJKY230070). Computational resources were provided by the Songshan Lake High Performance Computing Center (SSL-HPC) at Great Bay University. This work was also supported by the Guangdong Research Team for Communication and Sensing Integrated with Intelligent Computing (Project No. 2024KCXTD047).

References

Cao, A.; and Johnson, J. 2023. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 130–141.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, Y.; Chen, Z.; Zhang, C.; Wang, F.; Yang, X.; Wang, Y.; Cai, Z.; Yang, L.; Liu, H.; and Lin, G. 2024a. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21476–21485.

Chen, Y.; Yuan, Q.; Li, Z.; Liu, Y.; Wang, W.; Xie, C.; Wen, X.; and Yu, Q. 2024b. Upst-nerf: Universal photorealistic style transfer of neural radiance fields for 3d scene. *IEEE Transactions on Visualization and Computer Graphics*.

Chiang, P.-Z.; Tsai, M.-S.; Tseng, H.-Y.; Lai, W.-S.; and Chiu, W.-C. 2022. Stylizing 3d scene via implicit representation and hypernetwork. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1475–1484.

Deng, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; and Xu, C. 2021. Arbitrary video style transfer via multi-channel correlation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1210–1217.

Duan, Y.; Wei, F.; Dai, Q.; He, Y.; Chen, W.; and Chen, B. 2024. 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes. In *ACM SIGGRAPH 2024 Conference Papers*, 1–11.

Feng, R.; Weng, W.; Wang, Y.; Yuan, Y.; Bao, J.; Luo, C.; Chen, Z.; and Guo, B. 2024. Ccedit: Creative and controllable video editing via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6712–6722.

Fridovich-Keil, S.; Meanti, G.; Warburg, F. R.; Recht, B.; and Kanazawa, A. 2023. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12479–12488.

Fridovich-Keil, S.; Yu, A.; Tancik, M.; Chen, Q.; Recht, B.; and Kanazawa, A. 2022. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5501–5510.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423.

Gu, S.; Chen, C.; Liao, J.; and Yuan, L. 2018. Arbitrary style transfer with deep feature reshuffle. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8222–8231.

Haque, A.; Tancik, M.; Efros, A. A.; Holynski, A.; and Kanazawa, A. 2023. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19740–19750.

Howil, K.; Waczyńska, J.; Borycki, P.; Dziarmaga, T.; Mazur, M.; and Spurek, P. 2025. CLIPGaussian: Universal and Multimodal Style Transfer Based on Gaussian Splatting.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.

Huang, Y.-H.; He, Y.; Yuan, Y.-J.; Lai, Y.-K.; and Gao, L. 2022. Stylizednerf: consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18342–18352.

Kara, O.; Kurtkaya, B.; Yesiltepe, H.; Rehg, J. M.; and Yanardag, P. 2024. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6507–6516.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. MUSIQ: Multi-scale Image Quality Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5148–5157.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Kolkin, N.; Salavon, J.; and Shakhnarovich, G. 2019. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10051–10060.

Kong, X.; Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Chen, Y.; He, Z.; and Xu, C. 2024. Exploring the Temporal Consistency of Arbitrary Style Transfer: A Channelwise Perspective. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6): 8482–8496.

Lai, W.-S.; Huang, J.-B.; Wang, O.; Shechtman, E.; Yumer, E.; and Yang, M.-H. 2018. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, 170–185.

Li, T.; Slavcheva, M.; Zollhoefer, M.; Green, S.; Lassner, C.; Kim, C.; Schmidt, T.; Lovegrove, S.; Goesele, M.; Newcombe, R.; et al. 2022. Neural 3d video synthesis from

multi-view video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5521–5531.

Liang, W.; Xu, H.; Chen, W.; Xiao, F.; and Kang, W. 2024. 4DStyleGaussian: Zero-shot 4D Style Transfer with Gaussian Splatting. *arXiv preprint arXiv:2410.10412*.

Liao, J.; Yao, Y.; Yuan, L.; Hua, G.; and Kang, S. B. 2017. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088*.

Liu, K.; Zhan, F.; Chen, Y.; Zhang, J.; Yu, Y.; El Saddik, A.; Lu, S.; and Xing, E. P. 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.

Liu, K.; Zhan, F.; Xu, M.; Theobalt, C.; Shao, L.; and Lu, S. 2024. Stylegaussian: Instant 3d style transfer with gaussian splatting. In *SIGGRAPH Asia 2024 Technical Communications*, 1–4.

Liu, S.; Lin, T.; He, D.; Li, F.; Wang, M.; Li, X.; Sun, Z.; Li, Q.; and Ding, E. 2021. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6649–6658.

Liu, S.; and Zhu, T. 2021. Structure-guided arbitrary style transfer for artistic image and video. *IEEE Transactions on Multimedia*, 24: 1299–1312.

Luiten, J.; Kopanas, G.; Leibe, B.; and Ramanan, D. 2024. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, 800–809. IEEE.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Mou, L.; Chen, J.-K.; and Wang, Y.-X. 2024. Instruct 4D-to-4D: Editing 4D Scenes as Pseudo-3D Scenes Using 2D Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20176–20185.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4): 1–15.

Park, K.; Sinha, U.; Hedman, P.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Martin-Brualla, R.; and Seitz, S. M. 2021. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*.

Pumarola, A.; Corona, E.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10318–10327.

Saroha, A.; Gladkova, M.; Curreli, C.; Muhle, D.; Yenamandra, T.; and Cremers, D. 2024. Gaussian splatting in style. In *DAGM German Conference on Pattern Recognition*, 234–251. Springer.

Shao, R.; Zheng, Z.; Tu, H.; Liu, B.; Zhang, H.; and Liu, Y. 2023. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16632–16642.

Vachha, C.; and Haque, A. 2024. Instruct-GS2GS: Editing 3D Gaussian Splats with Instructions.

Wang, C.; Chai, M.; He, M.; Chen, D.; and Liao, J. 2022. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3835–3844.

Wang, C.; Jiang, R.; Chai, M.; He, M.; Chen, D.; and Liao, J. 2023. Nerf-art: Text-driven neural radiance fields stylization. *IEEE Transactions on Visualization and Computer Graphics*.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring CLIP for Assessing the Look and Feel of Images. In *AAAI*.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. CBAM: Convolutional Block Attention Module. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision – ECCV 2018*, 3–19. Cham: Springer International Publishing.

Wu, G.; Yi, T.; Fang, J.; Xie, L.; Zhang, X.; Wei, W.; Liu, W.; Tian, Q.; and Wang, X. 2024. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20310–20320.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Li, C.; Liao, L.; Wang, A.; Zhang, E.; Sun, W.; Yan, Q.; Min, X.; Zhai, G.; and Lin, W. 2023. Q-Align: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels. *arXiv preprint arXiv:2312.17090*. Equal Contribution by Wu, Haoning and Zhang, Zicheng. Project Lead by Wu, Haoning. Corresponding Authors: Zhai, Guangtai and Lin, Weisi.

Wu, Z.; Zhu, Z.; Du, J.; and Bai, X. 2022. Ccpl: Contrastive coherence preserving loss for versatile style transfer. In *European conference on computer vision*, 189–206. Springer.

Xu, H.; Chen, W.; Xiao, F.; Sun, B.; and Kang, W. 2024a. StyleDyRF: Zero-shot 4D Style Transfer for Dynamic Neural Radiance Fields. *arXiv preprint arXiv:2403.08310*.

Xu, R.; Chen, W.; Wang, J.; Liu, Y.; Wang, P.; Gao, L.; Xin, S.; Komura, T.; Li, X.; and Wang, W. 2024b. SuperGaussians: Enhancing Gaussian Splatting Using Primitives with Spatially Varying Colors. arXiv:2411.18966.

Yu, Z.; Sattler, T.; and Geiger, A. 2024. Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. *ACM Transactions on Graphics (TOG)*, 43(6): 1–13.

Zhang, D.; Yuan, Y.-J.; Chen, Z.; Zhang, F.-L.; He, Z.; Shan, S.; and Gao, L. 2024. Stylizedgs: Controllable stylization for 3d gaussian splatting. *arXiv preprint arXiv:2404.05220*.

Zheng, H.; Fu, J.; Zeng, Y.; Zha, Z.-J.; and Luo, J. 2021. Learning Conditional Knowledge Distillation for Degraded-Reference Image Quality Assessment. *ICCV*.

Zhuang, J.; Wang, C.; Lin, L.; Liu, L.; and Li, G. 2023. Dreameditor: Text-driven 3d scene editing with neural fields. In *SIGGRAPH Asia 2023 Conference Papers*, 1–10.