GEOMETRY-CONSISTENT VIDEO EDITING FOR APPEARANCE CONTROLLABLE GAUSSIAN SCENES

Anonymous authors
Paper under double-blind review



Figure 1: Our proposed SCENEWEAVER supports geometry-consistent scene editing across viewpoint changes, with diverse variations in lighting, environment, materials, and visual styles.

ABSTRACT

Image editing models provide intuitive control over visual changes, but extending this control to 3D also requires edits to remain consistent across viewpoints. We present SCENEWEAVER, a two-stage framework that constructs an editable Gaussian scene from a single photograph. We progressively build a point cloud from generated source-appearance views, then fix it as a geometric prior for multi-view video generation under different appearance references. We distill these videos into a shared Gaussian representation controlled by edit latents extracted from source–edited image pairs. These latents drive the desired edits relative to the original appearance while Gaussian geometry remains shared across edits. The resulting representation further supports zero-shot transfer of appearance edits not used during scene fitting, without additional optimization. We also introduce a scene-editing benchmark spanning indoor and outdoor scenes to evaluate appearance agreement and structural consistency. Extensive experiments demonstrate that our method achieves strong performance in both reconstruction and editing across multiple dimensions. For more results and videos, please refer to our anonymous project page: `https://anonymous-project-9.github.io/anonymous_iclr/`.

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

1 INTRODUCTION

Image editing models (Wu et al., 2025; OpenAI, 2026) provide an intuitive way to specify visual changes through a source image and its edited counterpart. We aim to extend this form of control to a Gaussian scene constructed from a single photograph, allowing edits specified in image space to persist across viewpoints within a shared 3D representation. Alongside the edits used during scene fitting, the representation should support new appearance edits without further optimization.

Constructing such a representation requires both multi-view observations and a mapping from image edits to Gaussian attributes. Existing 3D editing and stylization methods (Haque et al., 2023; Liu et al., 2024) typically operate on an already reconstructed scene, whereas a single photograph leaves much of the scene unobserved. Camera-controlled video generators (Yu et al., 2024b; Ren et al., 2025; Zhang et al., 2026) can supply missing views, but generating videos from different appearance references can produce inconsistent structures and camera motion, making the observations difficult to fit into one scene. Fitting the available appearances alone does not establish how an image pair should control the reconstructed scene or how that control extends to an unseen appearance.

We introduce SCENEWEAVER, a two-stage framework for constructing image-pair-controlled editable Gaussian scenes from a single photograph (Figure 2). Our key idea is to generate multiple appearances under a fixed geometric prior and distill them into a shared Gaussian representation, where edit latents extracted from image pairs drive the desired scene edits.

To generate multi-view observations, we initialize a point cloud from monocular depth of the source photograph (Yang et al., 2024). Its depth projections guide video generation (Zhang et al., 2023; Wan Team, 2025), and depths estimated from generated source-appearance views are fused back to expand scene coverage. Once constructed, the point cloud is fixed and reused along the same prescribed camera trajectories with different appearance references. Keeping the geometric condition fixed while changing the reference image encourages structural agreement across videos.

We distill these observations into an editable 3D Gaussian representation (Kerbl et al., 2023), using the original photograph's appearance as the default. A pretrained encoder maps each source–edited image pair to an edit latent. Residual heads combine this latent with local Gaussian features to predict changes to the default spherical-harmonic (SH) colour coefficients and opacity, while Gaussian positions, rotations, and scales remain shared across edits. Generated multi-appearance observations supervise this mapping, and photometric augmentations of Gaussian renderings further train the edit head. After scene fitting, a new image pair can drive the learned heads without further optimization, enabling *zero-shot transfer* of appearance edits not used during fitting.

Our contributions are:

- We propose a geometry-guided video generation pipeline that progressively constructs a scene point cloud and reuses its fixed geometry to generate multi-view videos under diverse appearances.
- We develop a Gaussian representation that models edits as image-pair-conditioned SH and opacity residuals relative to the original appearance, supporting multiple edits within one scene and zero-shot transfer to unseen appearance edits.
- We introduce a scene-editing benchmark spanning indoor and outdoor scenes to evaluate appearance agreement and structural consistency. We assess reconstruction and editing quality through extensive quantitative and qualitative comparisons. The results show higher held-out PSNR and VBench imaging quality than the evaluated baselines.

2 RELATED WORK

Camera-Controlled Video Generation. Existing methods introduce explicit viewpoint control into diffusion models through camera embeddings (He et al., 2024; Bahmani et al., 2025b) or 3D proxies constructed from input images (Yu et al., 2024b; Ren et al., 2025; Müller et al., 2024; Zhang et al., 2026). Recent approaches further maintain geometric or spatial memories across generated views (Zhang et al., 2026), while single-image scene generators iteratively combine image synthesis with depth-based warping (Sargent et al., 2024; Yu et al., 2024a). These approaches provide effective camera control, but their geometric guidance is often coupled with RGB appearance, making the generated structure sensitive to changes in the reference image. In contrast, SCENEWEAVER maintains a shared point-cloud geometry and uses depth-only projections for conditioning, allowing different appearances to be generated under the same scene structure.

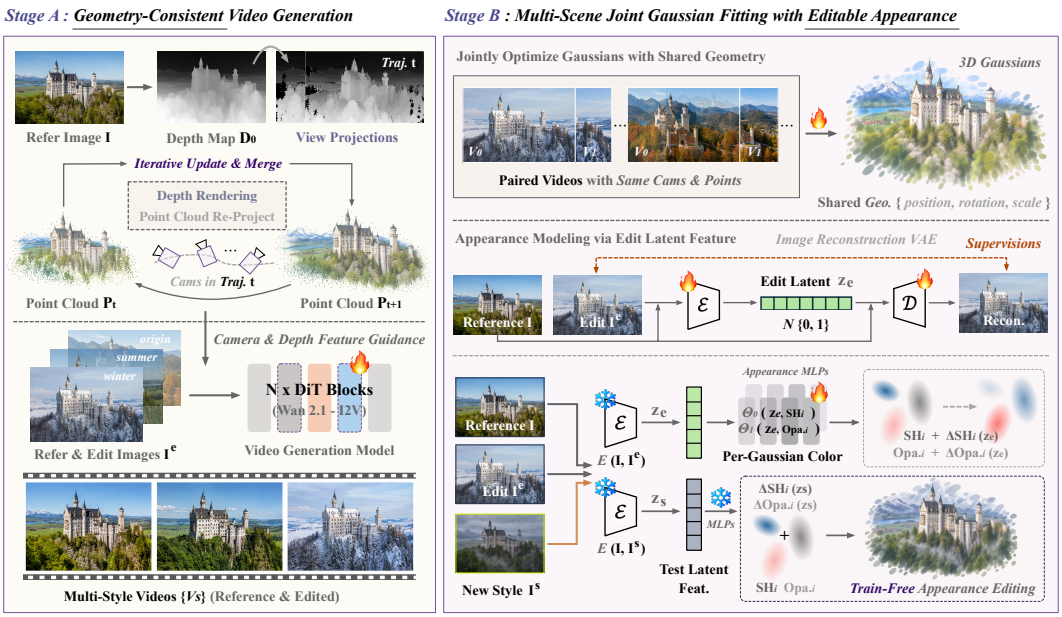

Figure 2: **Overview of SCENEWEAVER. Stage A** constructs a point cloud from generated source-appearance views and fixes it to guide multi-view generation under different appearance references. **Stage B** fits a default Gaussian scene for the source appearance and models the other appearances as edit-conditioned residuals in SH coefficients and opacity, with other Gaussian attributes fixed. A new source–edited image pair drives these residuals without further scene optimization.

Appearance Editing of 3D Scenes. Existing 3D scene editing methods commonly apply 2D editing models, e.g., Brooks et al. (2023), to rendered views and distill the edited observations back into NeRF or Gaussian representations (Haque et al., 2023; Wu et al., 2024b; Chen et al., 2024; Liu et al., 2024). While effective, these methods typically require multi-view editing and scene-specific optimization for each new appearance. Another line of work models appearance variation directly within the 3D representation: NeRF-W (Martin-Brualla et al., 2021) and Ha-NeRF (Chen et al., 2022) associate captured images with optimized appearance codes, while 4DGS(Wu et al., 2024a) models scene variations over time. These representations can separate appearance variation from scene structure, but generally rely on captured observations or optimized latent variables rather than transferring edits from a new style image. In contrast, SCENEWEAVER fits generated multi-appearance observations with a shared default Gaussian and uses image-pair-derived latents to predict appearance residuals, enabling new edits without further scene optimization.

3 METHOD

Given a source photograph A, SCENEWEAVER constructs an editable Gaussian scene in two stages (Figure 2). Stage A expands scene geometry using videos generated with the source appearance, then fixes the resulting point cloud to generate multi-view observations under different appearance references. Stage B distills these observations into a default Gaussian scene and edit-conditioned residuals in its SH coefficients and opacity. All edits share the same default scene and Gaussian geometry. Once fitted, the representation accepts a new source–edited image pair and renders the corresponding edit from novel viewpoints without further scene optimization.

3.1 STAGE A: SHARED-GEOMETRY VIDEO GENERATION

Source-appearance geometry construction. Given the source image A, we estimate its depth D_0 using Depth Anything (Yang et al., 2024) and back-project it into an initial point cloud. To extend the observed geometry, we render depth maps along a prescribed camera trajectory and use them to condition Wan2.1-I2V (Wan Team, 2025). A ControlNet block (Zhang et al., 2023) attached to each of the first 20 transformer blocks receives these depth maps and their validity masks. The source image provides the appearance condition, while camera Plücker embeddings (He et al., 2024) provide the camera condition.

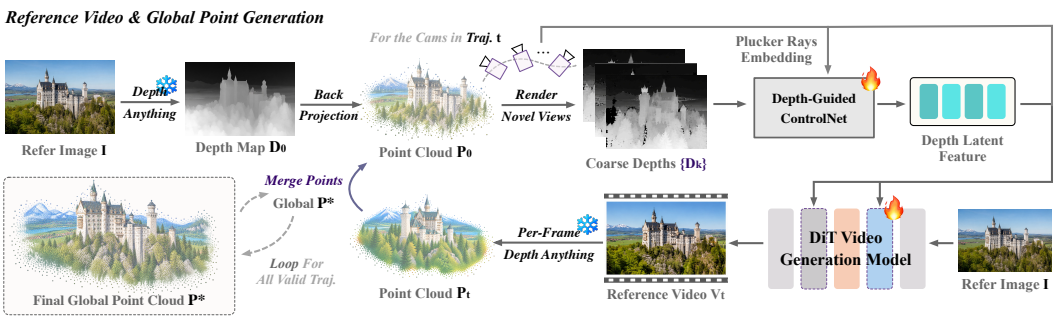

Figure 3: **Source-appearance geometry construction.** Monocular depth initializes a point cloud whose projections guide source-appearance video generation. Depths estimated from the generated views are aligned and fused to expand the point cloud across camera trajectories. The resulting geometry $P^\star$ is then fixed for generating additional appearances.

After generating a source-appearance video, we estimate its frame depths, align their depth scale to the existing point cloud, and back-project them to update the geometry. We repeat this procedure across camera trajectories to expand scene coverage, as illustrated in Figure 3. Depth alignment and point-cloud reduction are described in Appendix B.

Multi-appearance observations from fixed geometry. Let $P^\star$ denote the point cloud after source-appearance geometry construction. We fix $P^\star$ and reuse its depth projections along prescribed trajectories, while replacing the source reference with an appearance reference B_s to generate multi-view images I_v^s, where s indexes appearance and v indexes a trajectory frame. These images and the corresponding depth renderings of $P^\star$ serve as observations for Stage B.

3.2 Stage B: Fitting an Editable Gaussian Scene

Observations and cameras. Stage A provides multi-appearance observations $\{I_v^s\}$ under a shared point-cloud prior, while their actual camera poses may deviate from the prescribed trajectories. We jointly process observations across appearances with VGGT-Ω (Wang et al., 2026) to recover camera poses and geometry, which is fused with the Stage A point cloud $P^\star$ to initialize the Gaussians. During fitting, we refine the per-frame poses and supervise depth using initial renderings of $P^\star$, reprojected and aligned to the refined camera views.

Default Gaussian scene. The appearance of the source photograph A defines the default Gaussian scene. Each Gaussian has a position, rotation, scale, and local feature $f_i \in \mathbb{R}^{16}$ shared across appearances, together with base degree-2 SH coefficients C_i^0 and an opacity parameter o_i^0. Source-appearance observations supervise this default representation; the other appearances are represented by residuals relative to the same base attributes.

From image pairs to edit conditioning. A pretrained image-pair encoder E, kept frozen during Gaussian fitting and editing, represents the change from the source photograph A to an edited reference B_s. Its 128-dimensional posterior mean $h_s = E(A, B_s)$ is mapped by a LayerNorm–linear adapter to 32-dimensional scene conditioning z_s. Together with the local Gaussian features f_i, this conditioning drives the SH and opacity residual heads.

SH and opacity updates. For Gaussian i and SH coefficient index k, the conditioned residuals update the default attributes as:

$$\begin{aligned} C_{ik}(z_s) &= C_{ik}^0 + \Delta C_{ik}(z_s), \\ o_i(z_s) &= o_i^0 + \Delta o_i(z_s). \end{aligned} \tag{1}$$

Here ΔC_{ik} and Δo_i are the predicted SH and opacity-parameter residuals. Rendering these edited attributes with the shared Gaussian geometry produces a view at the chosen camera.

Pretraining the edit encoder. To learn this image-pair representation, we pretrain E across scenes using source–edited pairs (A, B). The encoder combines frozen DINOv2 features (Oquab et al.,

2023) with explicit colour statistics, while a FiLM-conditioned decoder (Perez et al., 2018) reconstructs B from A and the edit representation. Training pairs comprise synthetic photometric edits, cross-appearance generated frames, and internet-collected seasonal references, with generated pairs aligned by nominal frame index.

For a batch of source–target pairs $\{(A_j, B_j)\}_{j=1}^n$, the encoder predicts a diagonal Gaussian $q(h \mid A_j, B_j) = \mathcal{N}(\mu_j, \mathrm{diag}(\sigma_j^2))$ over $d = 128$ latent dimensions. We draw reparameterised samples h_j during training and reconstruct $\widehat{B}_j = D(A_j, h_j)$. When extracting an edit code, we use the posterior mean, denoted by $h = E(A, B)$.

The reconstruction loss combines a pixelwise error with a perceptual term:

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{n} \sum_{j=1}^{n} \left[\ell_1(\widehat{B}_j, B_j) + 0.5 \ \mathrm{LPIPS}_{\mathrm{VGG}}(\widehat{B}_j, B_j) \right], \tag{2}$$

where ℓ_1 averages absolute differences over pixels and channels. Input images lie in $[0, 1]$ and are rescaled to $[-1, 1]$ for LPIPS. The perceptual term uses VGG features. We regularise the latent distribution towards a standard Gaussian distribution with the KL divergence term:

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{2n} \sum_{j=1}^{n} \sum_{k=1}^{d} \left(\mu_{jk}^2 + \sigma_{jk}^2 - \log \sigma_{jk}^2 - 1 \right). \tag{3}$$

For batches sharing the same intended edit, we further enforce consistency across image pairs. Such batches contain either different images under the same sampled photometric transform or trajectory frames sharing the same source-to-target appearance change. Let $\bar{\mu} = n^{-1} \sum_{j=1}^{n} \mu_j$. We impose:

$$\mathcal{L}_{\mathrm{cons}} = \frac{1}{n} \sum_{j=1}^{n} \|\mu_j - \bar{\mu}\|_2^2, \tag{4}$$

$$\mathcal{L}_{\mathrm{shared}} = \frac{1}{n} \sum_{j=1}^{n} \ell_1(D(A_j, \bar{\mu}), B_j). \tag{5}$$

The consistency loss encourages similar representations of the same edit. The shared-code loss requires one code to reproduce the edit across different source images, encouraging the decoder to retain image-specific content from A_j. These losses are omitted for internet-collected seasonal pairs without a shared edit. The complete encoder objective is:

$$\mathcal{L}_{\mathrm{enc}} = \mathcal{L}_{\mathrm{rec}} + 10^{-3} \mathcal{L}_{\mathrm{KL}} + 0.1 \mathcal{L}_{\mathrm{cons}} + 0.5 \mathcal{L}_{\mathrm{shared}}. \tag{6}$$

We optimize the encoder head and decoder while keeping DINOv2 frozen.

Scene fitting and optimization. With the pretrained encoder fixed, Gaussian fitting combines supervision from generated observations with augmented edit-head steps. A scene-fitting step samples appearance s and camera v, and supervises the Gaussian rendering against I_v^s with $0.8 \cdot \ell_1 + 0.2 \cdot (1 - \mathrm{SSIM})$. A log-depth loss of weight 0.2 uses the aligned Stage A depth targets. The scene-fitting configuration also includes a scene-code regularizer of weight 10^{-4}. Camera poses are refined jointly with the scene. Augmented steps use photometrically transformed Gaussian renderings as described below. Densification statistics are collected only during geometry optimization, with a doubled growth threshold to suppress excessive Gaussian growth.

Training on augmented Gaussian renderings. To train the edit head on additional photometric changes, we apply random photometric transformations to current Gaussian renderings and use the transformed images as augmented targets. These steps update the edit head while keeping Gaussian geometry, the base colour network, and the pretrained encoder fixed.

Editing the fitted scene. Given a new source–edited image pair (A, B_e), the frozen encoder extracts $h_e = E(A, B_e)$, from which we obtain the scene conditioning z_e. The edit heads predict SH and opacity residuals, which are added to the default Gaussian attributes using Eq. (1). We render the resulting scene with its fitted positions, rotations, and scales, without updating the scene or network parameters. For an appearance edit not used during scene fitting, this provides zero-shot transfer relative to the same source appearance as the fitted edits.

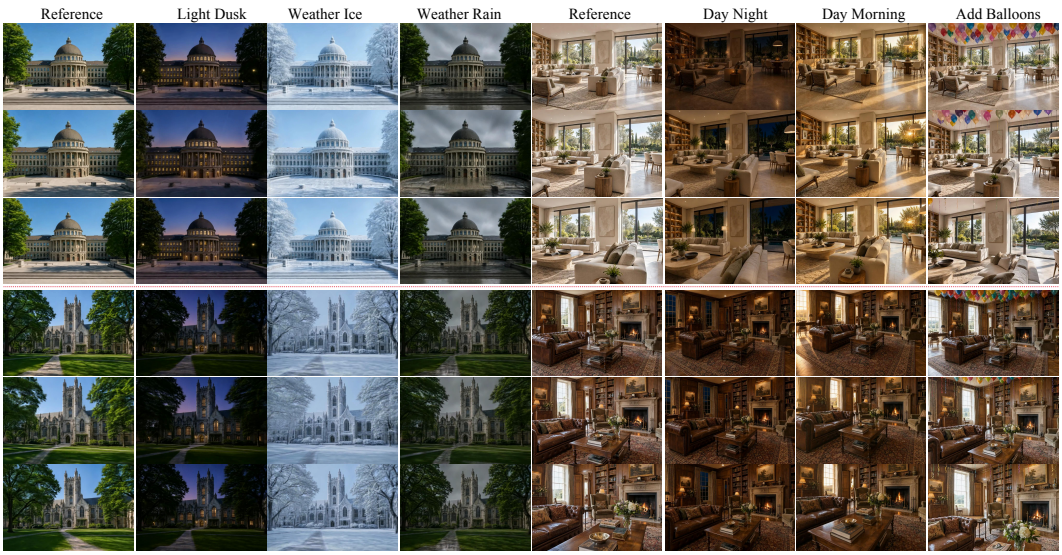

Figure 4: **Various scene edits.** SCENEWEAVER supports scene appearance edits under diverse styles, while preserving consistent scene structure across varying viewpoints.

3.3 IMPLEMENTATION DETAILS

Stage-A modeling and training. We build the video editing framework based on Wan2.1-I2V with 20 ControlNet blocks and a global point-cloud memory. Depth is represented in log space, replicated to three channels, and encoded by the VAE to match the original RGB conditioning interface. During training, we sample 4-6 frames from trajectories different from the target trajectory, back-project their depths into a point cloud, and render it along the target views for geometric conditioning. We fine-tune the model on 81-frame clips using a flow-matching velocity loss. Training data are sampled from SpatialVID-HQ with MegaSaM-estimated depth (Wang et al., 2025; Li et al., 2025), InteriorGS-360(SpatialVerse Research Team, 2025), DL3DV-10K (Ling et al., 2024), and TartanAir (Wang et al., 2020) with probabilities $0.55/0.30/0.10/0.05$. We use AdamW with a learning rate of 10^{-5} and 100 warm-up steps, together with EMA and bf16 mixed precision. Training uses FSDP2 with gradient checkpointing on 64 A100 GPUs and an effective batch size of 16.

Stage-B Gaussian fitting. We use the gsplat rasterizer (Ye et al., 2025a) and initialize Gaussians from the fused Stage A and VGGT-Ω geometry. The base-color MLP contains two hidden layers of width 128, with SH cached for a fixed edit. We adopt absolute-gradient densification with a threshold of 4×10^{-4}, starting at step 300 and accumulating statistics only during geometry optimization. The edit head is trained with 32 random photometric augmentations of current Gaussian renderings per step, applied with probability 0.5. The pretrained image-pair encoder remains frozen. Per-camera and per-(camera, season) SE(3) corrections are jointly optimized with separate learning rates. We train for 15k steps at 720p resolution, requiring approximately seven minutes on a single A100 GPU.

4 EXPERIMENTS

4.1 BENCHMARK AND EVALUATION

Scene-editing benchmark. We construct a benchmark of 20 landmarks (*e.g.*, Great Pyramid, Eiffel Tower, Taj Mahal, Colosseum, and Great Wall), 20 universities (*e.g.*, Harvard, MIT, and Stanford), and 20 indoor scenes. Each outdoor scene is collected from a fixed viewpoint across four seasons and three lighting conditions at 1536×1024, with summer serving as the geometry source and default appearance for seasonal editing. For each scene, we generate ten 81-frame trajectories at 1088×720: two reference paths transferred from a table-top scene and scaled by the estimated depth d (motion scale 0.4), and eight analytic paths covering dolly ($0.35d$), truck ($0.25d$), pan (20°), orbit (30°), and spiral motions.

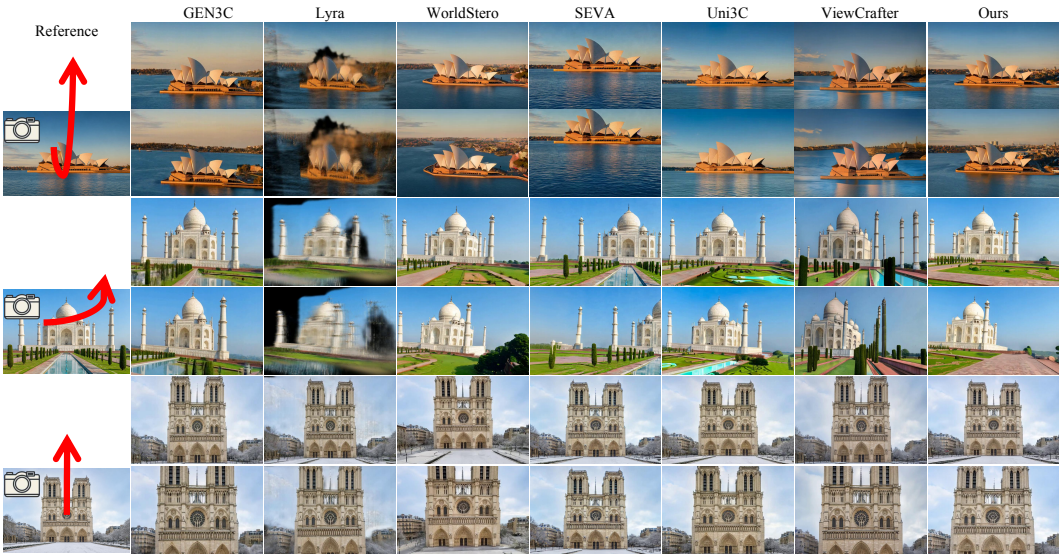

Figure 5: **Camera trajectory control.** Our method supports flexible camera trajectories while preserving consistent scene geometry across viewpoint changes.

Table 1: Single-image novel-view synthesis on Mip-NeRF 360 and Tanks and Temples, with one reference photograph per scene. The two groups evaluate cameras along the generated paths and held-out cameras, respectively; all SCENEWEAVER outputs are Gaussian renderings. The vis. columns report PSNR on pixels covered by the global-memory point cloud.

	Views in Generated Frames				Held-out Views			
	PSNR ↑	SSIM ↑	LPIPS ↓	vis.	PSNR ↑	SSIM ↑	LPIPS ↓	vis.
ViewCrafter	13.78	0.297	0.567	14.43	10.88	0.279	0.585	10.27
WorldStero	15.15	0.436	0.386	14.04	12.75	0.311	0.536	11.97
Stable Virtual Camera	13.07	0.262	0.595	14.17	10.69	0.275	0.605	10.82
Uni3C	14.45	0.301	0.474	13.72	10.74	0.296	0.557	9.94
GEN3C	15.45	0.369	0.538	**15.38**	12.54	0.371	0.560	11.48
Lyra (feed-forward 3DGS)	13.11	0.267	0.634	11.76	9.97	0.261	0.628	7.73
SCENEWEAVER (Ours)	**16.03**	**0.452**	**0.317**	14.63	**14.49**	**0.384**	**0.479**	**14.38**

Metrics. We measure appearance agreement with CIELAB moment distance d_{Lab}, its anchor-to-source ratio, and CLIP anchor retrieval; structural consistency with cross-appearance edge IoU and scale-invariant log-depth error (SI-log); and reconstruction with PSNR, SSIM, and LPIPS (Wang et al., 2004; Zhang et al., 2018). VBench (Huang et al., 2024) assesses video consistency, motion, and visual quality. Appendix C defines these metrics.

Scene reconstruction and geometry consistency. We evaluate novel-view synthesis on Tanks and Temples (Knapitsch et al., 2017) and Mip-NeRF 360 (Barron et al., 2022), comparing with ViewCrafter, WorldStereo (Zhang et al., 2026), Stable Virtual Camera (Zhou et al., 2025), Uni3C (Cao et al., 2025), GEN3C (Ren et al., 2025), and Lyra (Bahmani et al., 2025a). Following a shared reference-image and camera protocol (Appendix A and D), we evaluate against captured images along generated trajectories and at 20 held-out views per scene. As shown in Table 1, SCENEWEAVER achieves the best PSNR, SSIM, and LPIPS in both settings. Figure 5 further shows consistent scene geometry across camera trajectories.

Appearance agreement across views. We next evaluate appearance agreement and structural consistency on the scene-editing benchmark (Table 4 in the appendix). SCENEWEAVER achieves the lowest CIELAB moment distance and anchor-to-source ratio and the highest anchor retrieval. The structural metrics favor different methods: Uni3C obtains the highest edge IoU, WorldStereo the lowest proxy depth error, and GEN3C the lowest temporal warp error.



Figure 6: **Comparison with video-editing methods.** We compare summer-to-winter editing of the Colosseum and noon-to-evening editing of a campus scene. SCENEWEAVER better preserves target appearance and scene geometry across viewpoints.

Table 2: **Video editing quality on our benchmark.** VBench scores are scaled by 100. Dynamic degree reports motion activity; higher values are preferred for the other metrics. All SCENEWEAVER outputs are final Gaussian renderings.

	Season Changes					Time-of-Day Changes				
Method	bg. cons.	motion sm.	dynamic	aesthetic	imaging	bg. cons.	motion sm.	dynamic	aesthetic	imaging
AnyV2V	93.35	96.91	55.17	65.42	76.46	90.61	93.08	62.06	61.95	76.62
VideoPainter	**96.71**	91.24	20.68	54.81	71.43	**97.15**	92.16	17.24	57.83	69.71
StyleMaster	93.44	97.89	66.67	65.33	72.37	92.34	96.11	48.27	62.11	65.32
UniVideo	95.54	98.03	55.56	66.69	77.83	95.89	**98.26**	41.67	64.17	75.94
SCENEWEAVER(ours)	95.78	**98.14**	**79.31**	**66.93**	**78.96**	96.02	96.51	**65.51**	**65.03**	**78.75**

Comparisons of scene editing. As shown in Fig. 4, our method supports diverse reference-guided edits while preserving scene structure across viewpoints. To assess video quality, we compare SCENEWEAVER with AnyV2V (Ku et al., 2024), VideoPainter (Bian et al., 2025), StyleMaster (Ye et al., 2025b), and UniVideo (Wei et al., 2026) using five VBench (Huang et al., 2024) metrics: background consistency, motion smoothness, dynamic degree, aesthetic quality, and imaging quality. We use released models at their supported resolutions and clip lengths, providing the edited anchor as the first-frame or reference condition when applicable. In Table 2, SCENEWEAVER achieves higher imaging and aesthetic quality in both editing settings, along with higher dynamic degree. VideoPainter obtains higher background consistency in both settings, while UniVideo has higher motion smoothness for time-of-day editing. Figure 6 compares the edited appearances across viewpoints.

4.2 ABLATION STUDY

Depth conditioning. We compare depth concatenation and depth cross-attention in Wan2.1-I2V with the full SCENEWEAVER training configuration. The full configuration improves appearance agreement and temporal warp error, while cross-attention obtains higher edge IoU. Figure 7 illustrates artifacts in the alternative conditioning configurations.

Transfer to unseen edits. For edits not used in scene fitting, we compare target-image DINO conditioning with our edit encoder and decoupled edit head. The seasonal test uses a leave-one-out setting on the Colosseum with a generated video as the reference; the lighting test uses nine



Figure 7: **Qualitative ablations.** Visual comparisons corresponding to the ablations in Tab. 3.

Table 3: Ablations of depth conditioning, zero-shot editing, and modulated Gaussian attributes. P/S/L denotes PSNR↑/SSIM↑/LPIPS↓.

(a) Depth injection	d_{Lab} ↓	ratio ↓	retrieval ↑	edge ↑	depth ↓	warp ↓
Wan2.1-I2V + depth concatenation	17.681	0.694	0.593	0.457	0.374	15.287
Wan2.1-I2V + depth cross-attention	14.902	0.704	0.543	0.493	0.370	15.443
SCENEWEAVER (full schedule, reference)	3.886	0.388	0.857	0.445	0.284	4.734
(b) Zero-shot editing	Seasons: P / S / L		d_{Lab} ↓	CLIP pref. ↑	Lighting: P / S / L	
DINO features → Gaussian colours	13.3731 / 0.467 / 0.564		17.281	-0.011	13.1730 / 0.324 / 0.751	
Edit encoder + decoupled edit head (ours)	13.1630 / 0.473 / 0.607		19.333	-0.003	14.6935 / 0.391 / 0.654	
(c) Latent modulates	PSNR ↑	sharpness ↑				
Latent → SH colours only	20.514	0.654				
Latent → opacity only	19.185	0.605				
Latent → SH colours + opacity/scale residual	20.967	0.692				

unseen anchors. Our model improves PSNR, SSIM, and LPIPS for lighting. On the seasonal split, it slightly improves SSIM and CLIP preference, while DINO conditioning yields better PSNR, LPIPS, and colour-statistics distance.

Editable Gaussian attributes. Table 3(c) compares SH-only and opacity-only modulation with an extended variant that also predicts scale residuals alongside SH and opacity. SH-only modulation outperforms opacity-only modulation in reconstruction PSNR and sharpness; the extended variant improves both metrics further. Figure 7 illustrates how opacity modulation changes appearance-dependent visibility. The final representation described in Section 3.2 edits SH and opacity with shared positions, rotations, and scales.

5 Conclusion and Limitation

We present SCENEWEAVER, a framework for constructing an image-pair-controlled editable Gaussian scene from a single photograph. The central idea is to decouple scene geometry from appearance variation: geometry-guided video generation establishes a shared 3D structure, while Gaussian distillation learns appearance changes from multi-view observations and image-pair edit representations. The resulting representation supports new appearance edits without further optimization while maintaining geometric consistency across viewpoints. We further introduce a scene-editing benchmark covering 60 indoor and outdoor scenes. Experiments demonstrate improved reconstruction and perceptual quality over existing baselines, validating the effectiveness of shared geometry and image-pair conditioning for editable 3D scene representations.

Limitations. Our method may degrade in large outdoor scenes, where monocular depth errors affect point-cloud construction and subsequent Gaussian fitting. Stronger depth priors or multi-view constraints could improve reconstruction in these settings. Opacity updates alter the visibility of existing Gaussians, while shared positions, rotations, and scales limit substantial topology changes. Extending the representation to introduce new geometry is a direction for future work.

AI USE STATEMENT

Generative AI tools assisted with initial manuscript drafting, translation, and language editing. During revision, these tools also supported literature verification and consistency checks across the manuscript, implementation, and saved experimental results. References and numerical claims were verified against primary publications and project artifacts. AI tools were additionally used to monitor experimental runs, conduct scheduled checks of the server cluster, and assist with debugging the experiments in conjunction with command-line tools. The authors are responsible for the final text, claims, and research artifacts.

ETHICS STATEMENT

The method can synthesize plausible depictions of real places under conditions that did not occur. Generated images and videos should be identified as synthetic, particularly when they depict altered buildings, environmental conditions, or people. Internet availability of the reference images does not establish redistribution rights. Any release of these images must respect source permissions, and use of the pretrained models and training data remains subject to their respective terms.

REPRODUCIBILITY AND DATA AVAILABILITY

Section 3 specifies the architecture, training objectives, and fitting procedure. Section 4 and Appendices A–D describe the evaluation data, camera protocols, metrics, and baseline configurations. Additional visual results are available at the anonymous project page, `https://anonymous-project-9.github.io/anonymous_iclr/`.

REFERENCES

Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B. Lindell, Zan Gojcic, Sanja Fidler, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra: Generative 3D scene reconstruction via video diffusion model self-distillation. *arXiv:2509.19296*, 2025a.

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. VD3D: Taming Large Video Diffusion Transformers for 3D Camera Control. In *ICLR*, 2025b.

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022.

Yuxuan Bian, Zhaoyang Zhang, Xuan Ju, Mingdeng Cao, Liangbin Xie, Ying Shan, and Qiang Xu. VideoPainter: Any-length video inpainting and editing with plug-and-play context control. In *SIGGRAPH*, 2025.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 2023.

Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3C: Unifying precisely 3D-enhanced camera and human motion controls for video generation. In *SIGGRAPH Asia*, 2025.

Minghao Chen, Iro Laina, and Andrea Vedaldi. DGE: Direct Gaussian 3D Editing by Consistent Multi-view Editing. In *ECCV*, 2024.

Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, and Jue Wang. Hallucinated Neural Radiance Fields in the Wild. In *CVPR*, 2022.

Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions. In *ICCV*, 2023.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. AnyV2V: A tuning-free framework for any video-to-video editing tasks. *Transactions on Machine Learning Research*, 2024.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, Fast and Robust Structure and Motion from Casual Dynamic Videos. In *CVPR*, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. DL3DV-10K: A Large-Scale Scene Dataset for Deep Learning-based 3D Vision. In *CVPR*, 2024.

Kunhao Liu, Fangneng Zhan, Muyu Xu, Christian Theobalt, Ling Shao, and Shijian Lu. StyleGaussian: Instant 3D Style Transfer with Gaussian Splatting. In *SIGGRAPH Asia Technical Communications*, 2024.

Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 2021.

Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. MultiDiff: Consistent Novel View Synthesis from a Single Image. In *CVPR*, 2024.

OpenAI. GPT-Image-2, 2026. URL `https://platform.openai.com/docs/models/gpt-image-2`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. In *AAAI*, 2018.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control. In *CVPR*, 2025.

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Fei-Fei Li, Deqing Sun, and Jiajun Wu. ZeroNVS: Zero-Shot 360-Degree View Synthesis from a Single Image. In *CVPR*, 2024.

Manycore Tech Inc. SpatialVerse Research Team. Interiorgs: A 3d gaussian splatting dataset of semantically labeled indoor scenes. `https://huggingface.co/datasets/spatialverse/InteriorGS`, 2025.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020. URL `https://arxiv.org/abs/2003.12039`.

Wan Team. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, Xiaoxiao Long, Hao Zhu, Zhaoxiang Zhang, Xun Cao, and Yao Yao. SpatialVID: A Large-Scale Video Dataset with Spatial Annotations. *arXiv preprint arXiv:2509.09676*, 2025.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-ω, 2026. URL https://arxiv.org/abs/2605.15195.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A Dataset to Push the Limits of Visual SLAM. In *IROS*, 2020.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://ieeexplore.ieee.org/document/1284395/.

Cong Wei, Quande Liu, Zixuan Ye, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, and Wenhu Chen. UniVideo: Unified understanding, generation, and editing for videos. In *ICLR*, 2026.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. In *CVPR*, 2024a.

Jing Wu, Jia-Wang Bian, Xinghui Li, Guangrun Wang, Ian Reid, Philip Torr, and Victor Adrian Prisacariu. GaussCtrl: Multi-View Consistent Text-Driven 3D Gaussian Splatting Editing. In *ECCV*, 2024b.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024.

Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An Open-Source Library for Gaussian Splatting. *Journal of Machine Learning Research*, 2025a.

Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang, and Wenhan Luo. StyleMaster: Stylize your video with artistic generation and translation. In *CVPR*, 2025b.

Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T. Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, and Charles Herrmann. WonderJourney: Going from Anywhere to Everywhere. In *CVPR*, 2024a.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. ViewCrafter: Taming Video Diffusion Models for High-fidelity Novel View Synthesis. *arXiv preprint arXiv:2409.02048*, 2024b.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, pp. 586–595, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Yisu Zhang, Chenjie Cao, Tengfei Wang, Xuhui Zuo, Junta Wu, Jianke Zhu, and Chunchao Guo. WorldStereo: Bridging Camera-Guided Video Generation and Scene Reconstruction via 3D Geometric Memories. *arXiv preprint arXiv:2603.02049*, 2026.

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv:2503.14489*, 2025.

A SINGLE-IMAGE NOVEL-VIEW PROTOCOL ON CAPTURED SCENES

Scenes and cameras. We use the COLMAP reconstruction shipped with the data (the `images_4` downsampled photographs for Mip-NeRF 360) and centre-crop every photograph to 3:2 before resizing to 720×480, adjusting the intrinsics accordingly. The reference is the registered image whose position is closest to the median camera position. Candidates are the 60 cameras nearest to the reference whose viewing direction deviates by less than 60°; they are ordered by greedily visiting the nearest unvisited camera, which keeps consecutive frames close. Cameras 1–20 and 21–40 form the two clips (each clip starts with the reference as frame 0, so 41 distinct views are generated); cameras 41–60 are held out and never shown to any method.

B GENERATION AND MEMORY UPDATES

Inference uses 40 denoising steps and classifier-free guidance 5. During source-appearance geometry construction, DepthAnything estimates depth for generated frames. The median log-depth ratio on overlapping valid regions aligns new estimates to the existing memory; unreliable estimates use a batch-median fallback. Voxel downsampling is applied when the point cloud exceeds 1.5M points. Once source-appearance geometry construction is complete, the resulting point cloud $P^\star$ is fixed and reused to generate additional appearances, without appearance-specific point-cloud updates. For the qualitative long-orbit examples, four overlapping segments extend the trajectory; the nominal 360° endpoint need not close the visual loop.

C METRIC DEFINITIONS

Evaluated outputs. All SCENEWEAVER results are final Gaussian renderings. We evaluate generalization to held-out camera views in the reconstruction study and to appearance edits excluded from scene fitting in the zero-shot study.

Camera paths. The ten-path protocol includes two transferred reference paths and eight analytic paths. Reference translations are scaled by the ratio of source and reference subject depths and then by 0.4. Subject depth d is the median DepthAnything depth in the central 40% image box. Analytic paths include dolly motion of $0.35d$, truck motion of $0.25d$, 20° pans, and 30° orbits. Their scales inherit uncertainty from monocular geometry. The four-path comparison uses a common subset for each method, with model-specific temporal resampling described below.

Color statistics. Let $m(I)$ concatenate the three CIELAB channel means and standard deviations of image I. We convert OpenCV uint8 Lab to $L \in [0, 100]$ by multiplying by $100/255$ and subtract 128 from a and b. Frames are sampled at stride two, including frame zero:

$$d_{\mathrm{Lab}}(V, A) = \left\| \frac{1}{|T|} \sum_{t \in T} m(I_t) - m(A) \right\|_2, \qquad \rho(V, A, G) = \frac{d_{\mathrm{Lab}}(V, A)}{\max(d_{\mathrm{Lab}}(V, G), 10^{-6})}. \tag{7}$$

We average frame-level CIELAB moments before taking the norm. Benchmark summaries average video-level distances or ratios, including identity cases.

Anchor retrieval. We use the pretrained image tower associated with Wan 2.1, normalize its pooled features, and compare five uniformly spaced frames from indices 1 through 20 with the four seasonal references for the same landmark. Retrieval measures anchor preference: the fraction of sampled frames for which the intended anchor has the highest feature similarity. Frame zero is excluded.

Edges and depth. For edge IoU, images are resized to 768×512, converted to grayscale, and processed with Canny thresholds 60/150 and 5×5 dilation. Corresponding appearance and summer frames at indices $0, 4, 8, 12, 16, 20$ are compared. This measure is sensitive to appearance changes as well as geometric ones. For valid positive finite depths, let $\delta_u = \log \widehat{d}_u - \log d_u$. We compute

$$\text{SI-log}(\widehat{d}, d) = \sqrt{\frac{1}{|U|}\sum_{u \in U} \delta_u^2 - \left(\frac{1}{|U|}\sum_{u \in U} \delta_u\right)^2}. \tag{8}$$

At least 500 valid pixels are required. Six evaluated frames (indices $1, 4, 8, 12, 16, 20$) supply the depth comparisons. Cross-season SI-log measures consistency between estimated appearance and summer depths; proxy SI-log measures consistency between estimated output depth and source-point-cloud depth rendered at the nominal camera.

Temporal warp and fitting sharpness. RAFT-large (Teed & Deng, 2020) estimates forward/backward optical flow between consecutive frames. We warp the next frame into the current frame and measure mean absolute RGB error in 8-bit levels where the forward–backward residual is below one pixel. Sharpness divides the sum of per-image mean Sobel-gradient magnitudes across evaluated renders by the corresponding sum across generated targets. It measures relative gradient strength and is sensitive to oversharpening.

VBench. We use custom-input evaluation for background consistency, motion smoothness, dynamic degree, aesthetic quality, and imaging quality. These dimensions respectively use CLIP similarity, interpolation-based smoothness, flow-based motion detection, an aesthetic predictor, and MUSIQ. High consistency can coincide with limited movement; dynamic degree measures motion activity rather than camera-path accuracy.

Table 4: Appearance and structural metrics on the scene-editing benchmark: CIELAB moment distance d_{Lab}, edit ratio, CLIP anchor retrieval, cross-season edge IoU, SI-log to the source proxy, and temporal warp error. All SCENEWEAVER outputs are final Gaussian renderings.

Method	$d_{\text{Lab}} \downarrow$	ratio $\downarrow$	retrieval $\uparrow$	edge $\uparrow$	depth $\downarrow$	warp $\downarrow$
ViewCrafter	9.373	0.606	0.8125	0.3436	0.297	6.151
Stable Virtual Camera	5.687	0.429	0.8129	0.3705	0.345	4.751
Uni3C	6.469	0.460	0.8286	**0.4827**	0.388	6.572
GEN3C	6.929	0.477	0.7982	0.3744	0.308	**4.159**
Lyra (feed-forward splat)	11.954	0.670	0.7036	0.2801	0.311	5.317
WorldStero	7.180	0.527	0.7357	0.4503	**0.240**	6.057
Depth baseline	5.743	0.451	0.8107	0.4739	0.274	5.183
SCENEWEAVER	**3.845**	**0.395**	**0.8571**	0.4411	0.305	4.734

D COMPARISON CONFIGURATIONS

Camera Control Guided Scene Construction. ViewCrafter uses the released 25-frame 576×1024 model and a DUSt3R point cloud from a 16:9 center crop of the edited anchor. Camera translation scale is aligned using the median ratio of DepthAnything and DUSt3R depths. Wan 2.1 uses its 14B 480P checkpoint, 81 frames at 832×480, 40 UniPC steps, guidance 5, and flow shift 3. InstructPix2Pix uses framewise instructions, 20 Euler-ancestral steps, text guidance 7.5, and image guidance 1.5. Uni3C and GEN3C receive the source geometry and scale-aligned camera paths. Uni3C uses its Wan 2.1-I2V-14B-720P controller with 81 interpolated cameras, 40 steps, and guidance 5; frame $4k$ corresponds to prescribed camera k. GEN3C uses its Cosmos-7B multi-view interface with 121 interpolated cameras; frame $6k$ corresponds to camera k. Stable Virtual Camera (1.3B, v1.1) uses its single-input sparse-view interface and a two-pass trajectory prior, with guidance 3/2 and the default camera-scale setting. That setting preserves path shape rather than the common metric translation scale. Lyra predicts a Gaussian scene using its preset trajectories; we render it at the evaluation cameras after aligning the depth scale. It does not take the evaluation trajectory as a generation input.

Scene video editing. Each video-editing baseline receives the same 81-frame base-appearance video (summer or noon) and the target appearance reference or its supported instruction equivalent. Outputs are resampled to 81 frames at 720×480. AnyV2V uses I2VGen-XL, resampling to 16 frames, 500-step DDIM inversion, 50 generation steps, guidance 9, and feature/spatial/temporal plug-and-play thresholds 0.2/0.2/0.5. VideoPainter uses CogVideoX-5B-I2V with the inpainting branch, 49 frames at 480×720, 50 steps, and dynamic guidance 6. We fix the edited first frame and apply a full-frame mask to the remaining frames, leaving no unmasked spatial observations for the context branch.

StyleMaster uses Wan 2.1 1.3B with a gray-tile ControlNet, 81 frames at 832×480, 50 steps, guidance 8, and style guidance 3; ControlNet is active for the first 70% of steps. UniVideo uses Qwen2.5-VL-7B and HunyuanVideo, with a target reference and season/time-of-day instruction, 81 frames at 480×720, 50 steps, guidance 7, and image guidance 2. Seeds are 8888, 42, 0, and 42 for AnyV2V, VideoPainter, StyleMaster, and UniVideo, respectively. The comparison uses these model-specific resolutions, crops, input interfaces, and model scales.

For SCENEWEAVER, image-pair edit latents control the fitted Gaussian scene, which is rendered along the evaluation trajectory with all scene and network parameters fixed.



Figure 8: Campus content edits



Figure 9: Campus season edits



Figure 10: Campus time edits

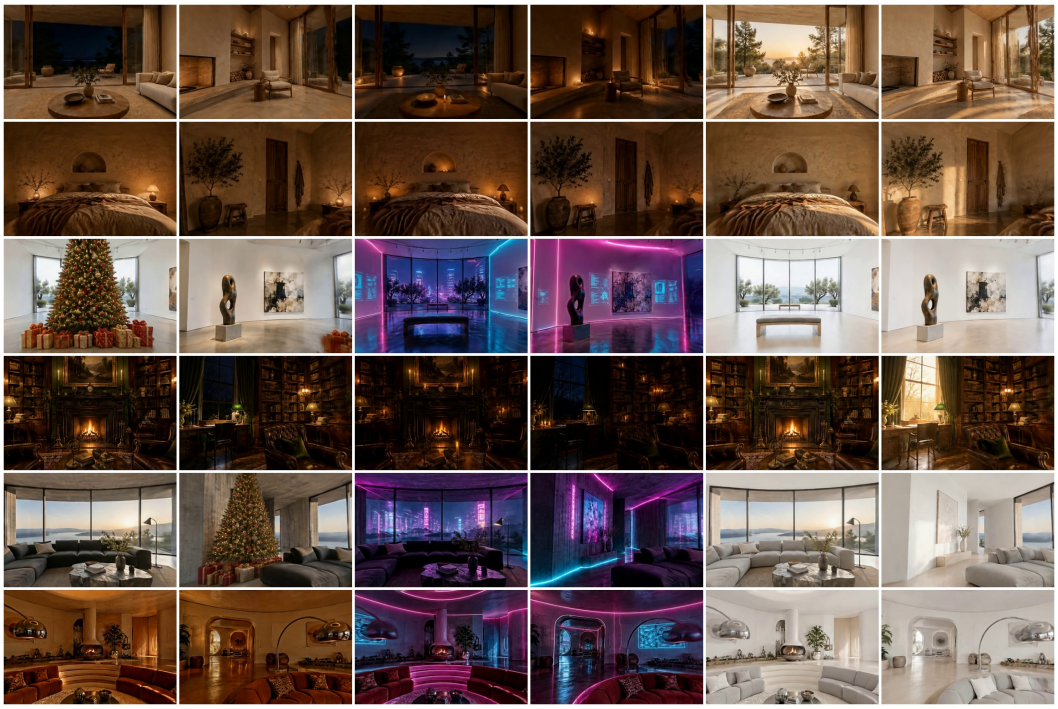

Figure 11: Indoor lighting edits

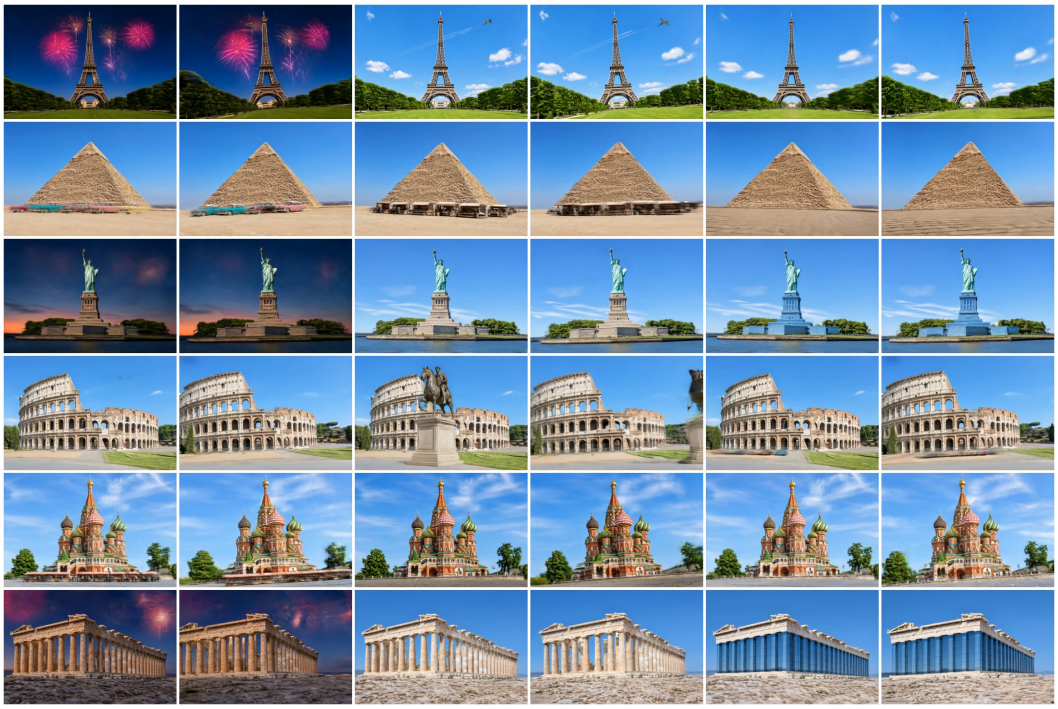

Figure 12: Various landmarks content edits



Figure 13: Various landmarks time changes