Dynamic-eDiTor: Training-Free Text-Driven 4D Scene Editing with Multimodal Diffusion Transformer

Dong In Lee[1,2*‡] Hyungjun Doh[1*] Seunggeun Chi[1] Runlin Duan[1]
Sangpil Kim[2†] Karthik Ramani[1†]

[1]Purdue University [2]Korea University

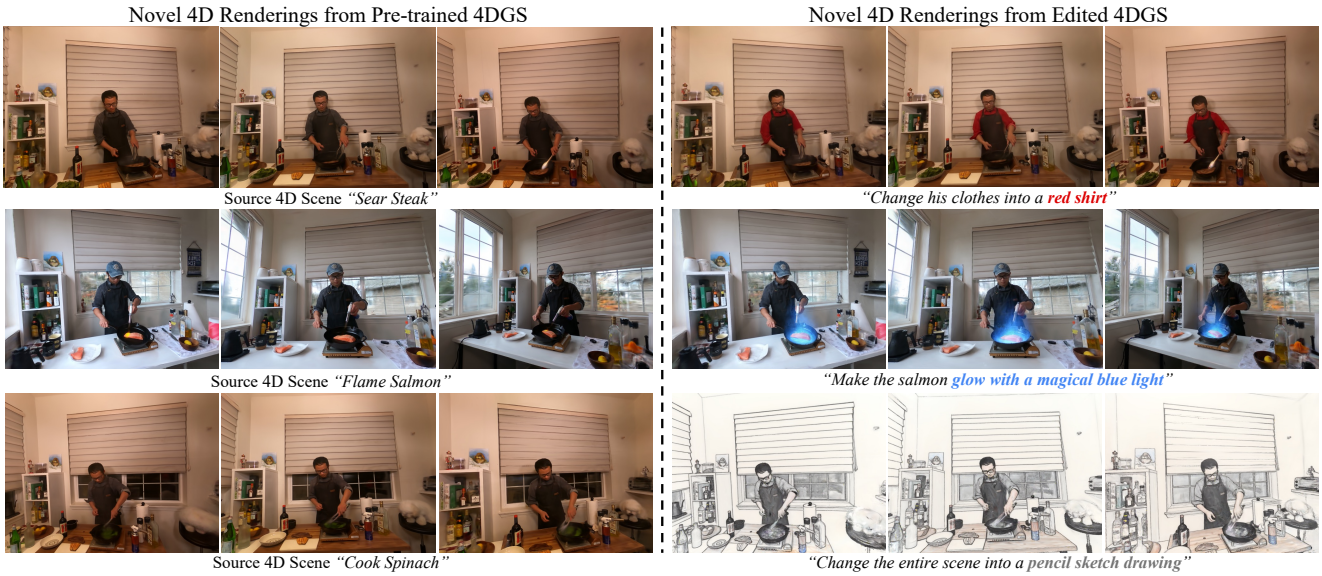

Figure 1. We propose **Dynamic-eDiTor** enables flexible and high-quality editing of pre-trained 4D Gaussian Splatting [46] models leveraging Multimodal Diffusion Transformer [9, 45] guided solely by text instructions. Through its design focused on both multi-view and temporal consistency, our approach demonstrates robust performance, producing realistic and fine-grained 4D scene manipulation.

Abstract

Recent progress in 4D representations, such as Dynamic NeRF and 4D Gaussian Splatting (4DGS), has enabled dynamic 4D scene reconstruction. However, text-driven 4D scene editing remains under-explored due to the challenge of ensuring both multi-view and temporal consistency across space and time during editing. Existing studies rely on 2D diffusion models that edit frames independently, often causing motion distortion, geometric drift, and incomplete editing. We introduce Dynamic-eDiTor, a training-free text-driven 4D editing framework leveraging Multimodal Diffusion Transformer (MM-DiT) and 4DGS. This mechanism consists of Spatio-Temporal Sub-Grid Attention (STGA) for locally consistent cross-view and temporal fusion, and Context Token Propagation (CTP) for global propagation via token inheritance and optical-flow-guided token replacement. Together, these components allow Dynamic-eDiTor to perform seamless, globally consistent multi-view video without additional training and directly optimize pre-trained source 4DGS. Extensive experiments on multi-view video dataset DyNeRF demonstrate that our method achieves superior editing fidelity and both multi-view and temporal consistency prior approaches. Project page for results and code: https://di-lee.github.io/dynamic-eDiTor/

1. Introduction

Recent advances in 3D representations, such as Neural Radiance Field (NeRF) [29] and 3D Gaussian Splatting (3DGS) [19], have achieved significant progress in photo-realistic 3D reconstruction of real-world scenes. More recently, 4D representations such as Dynamic NeRF [33] and 4D Gaussian Splatting (4DGS) [46] extend 3D representations into the time domain, enabling spatio-temporally co-

*Co-first authors.
†Co-corresponding authors.
‡Work done at Purdue University as a visiting scholar.

herent reconstruction. However, text-driven 4D scene editing remains under-explored, primarily due to the difficulty of maintaining both multi-view and temporal consistency across space and time during editing.

In this work, we focus on the multi-view video setting of 4DGS, which provides richer viewpoint coverage but further amplifies the difficulty of achieving both multi-view and temporal consistency during editing. While 3D editing primarily focuses on multi-view consistency, 4D editing introduces the further challenge of ensuring both multi-view and temporal consistency across viewpoints and time.

Recent studies [15, 21, 30] have attempted 4D scene editing by combining 2D diffusion models with 4D representations [38, 46]. However, these methods typically perform frame-wise editing or require per-scene finetuning of the 2D diffusion model [4], lacking a unified mechanism to jointly process information across views and time. Consequently, they struggle with non-rigid content manipulation and are often limited to style-oriented edits, leading to motion distortions, geometric drift, and incomplete editing results.

To address these limitations, we propose **Dynamic-eDiTor**, a novel training-free, text-driven 4D editing framework that leverages Multimodal Diffusion Transformer (MM-DiT) [9, 45] and 4DGS. Our goal is to maintain globally coherent motion and geometry while enabling flexible, semantically grounded edits. To this end, we propose Grid-based Spatio-Temporal Propagation, which represents the entire multi-view video as a unified camera-time grid and jointly aggregates spatial and temporal information and propagates the fused features throughout the grid.

As its foundation, we introduce Spatio-Temporal Sub-Grid Attention (STGA), which extends MM-DiT's dual-stream self-attention to operate on localized spatio-temporal sub-grids. By jointly attending to adjacent viewpoints and neighboring time steps, STGA enables coherent local feature fusion without additional training. We additionally identify a vital layer range in MM-DiT where incorporating STGA yields the strongest improvements in multi-view and temporal consistency.

To ensure that the fused information is globally propagated throughout the multi-view video, we further introduce Context Token Propagation (CTP), an explicit propagation mechanism that transfers fused tokens along a structured traversal path over the entire multi-view video. Tokens in overlapping regions are fully inherited, while non-overlapping temporal regions are updated through flow-guided token warping using optical flow [39]. This unified propagation strategy ensures coherent feature flow across views and time, reinforcing multi-view and temporal consistency and enabling stable, high-fidelity 4D optimization.

Finally, the edited frames are directly used to optimize the pre-trained 4DGS without the Iterative Dataset Update (IDU) [15, 30], resulting in globally consistent 4D content that faithfully reflects the desired edits.

We validate Dynamic-eDiTor on the multi-view video dataset DyNeRF [26], achieving superior editing fidelity, temporal smoothness, and robustness compared to state-of-the-art methods. Our key contributions are as follows:

- We present Dynamic-eDiTor, a novel training-free, text-driven 4D editing framework that leverages MM-DiT [9, 45] and 4DGS [46] to enable spatially and temporally consistent dynamic 4D scene editing.
- We propose Spatio-Temporal Sub-Grid Attention (STGA), which jointly attends across adjacent viewpoints and neighboring time steps to integrate spatial and temporal features on a vital layer range in MM-DiT.
- We introduce Context Token Propagation (CTP), an explicit propagation mechanism that distributes fused spatio-temporal information across the entire multi-view video by inheriting tokens in overlapping regions and replacing non-overlapping regions via flow-based warping.
- Through extensive qualitative and quantitative experiments, we demonstrate that Dynamic-eDiTor significantly outperforms existing methods in 4D editing fidelity, spatio-temporal stability, and robustness.

2. Related Work

2.1. 2D Editing

2D diffusion models have demonstrated remarkable generalization and controllability for text-guided image editing. U-Net-based [35] models such as Prompt-to-Prompt [16], SDEdit [28], and InstructPix2Pix [4] enable text-guided image manipulation via controlled denoising. More recently, Diffusion Transformers (DiT) [31] replace the U-Net backbone with a Transformer architecture [8, 40], offering improved scalability and visual coherence. This approach has evolved into multimodal variants such as Multimodal Diffusion Transformer (MM-DiT) [9]. MM-DiT employs a dual-stream architecture, processing text and image tokens in parallel streams fused via joint attention. Building on this trend, MM-DiT-based editors such as FLUX [22, 23], HiDream [5], and Qwen-Image-Edit [45] achieve precise instruction-driven image editing. Leveraging the MM-DiT, our Dynamic-eDiTor extends a MM-DiT-based image editor to maintain consistency across time and viewpoints within a unified 4D framework.

2.2. 3D Scene Editing

Neural Radiance Fields (NeRF) [29] and 3D Gaussian Splatting (3DGS) [19] have enabled high-fidelity 3D reconstruction and inspired extensive research on 3D scene editing. Instruct-NeRF2NeRF [14] introduces the Iterative Dataset Update (IDU) that edits the rendered image using a 2D diffusion model [4] while optimizing the underlying 3D model, NeRF. GaussianEditor [41] adopts IDU on

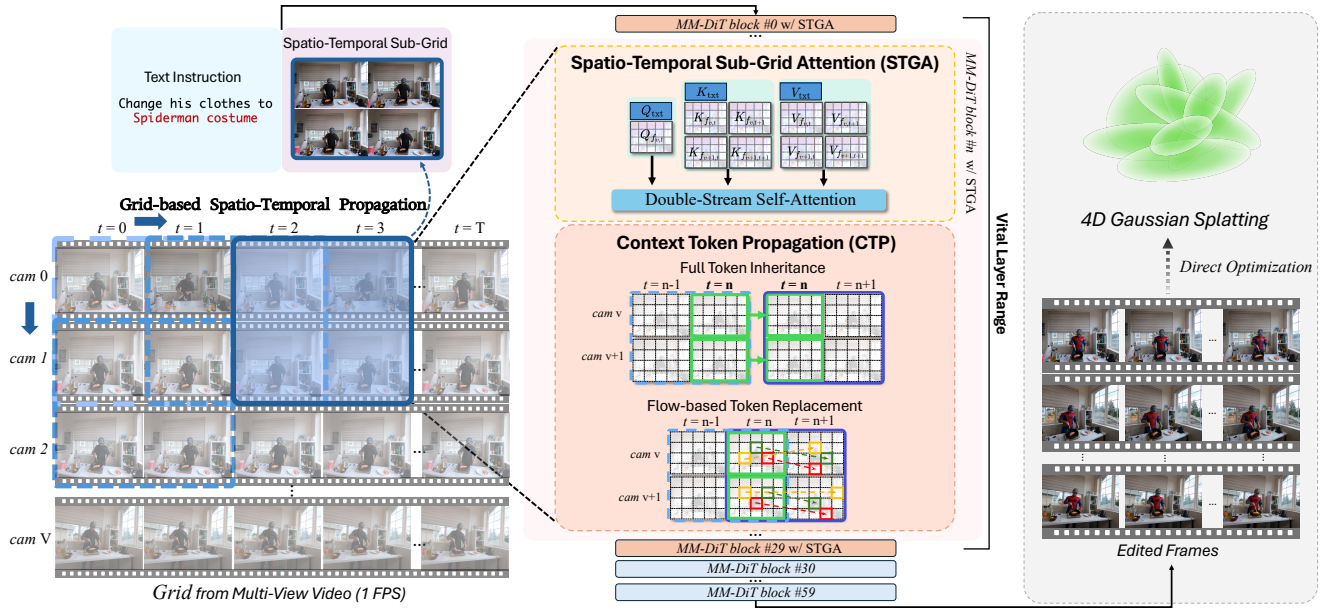

Figure 2. **Dynamic-eDiTor Overview.** We represent the multi-view video as a unified camera–time grid. Dynamic-eDiTor combines Spatio-Temporal Sub-Grid Attention (STGA), which performs locally coherent cross-view and temporal fusion within each sub-grid, with Context Token Propagation (CTP), which globally propagates the aggregated features across the grid via Full Token Inheritance and Flow-guided Token Replacement for robust spatio-temporal consistency enforcement. Together, these modules enable seamless, globally consistent multi-view video editing without additional training, while directly optimizing the pre-trained 4DGS.

3DGS and explicitly controls 3D Gaussians. Recently, EditSplat [25] achieves high-fidelity 3D edits by integrating multi-view information into the diffusion process and explicitly pruning 3DGS. Despite these advances, existing 3D editing methods [6, 7, 18, 42, 47, 51, 52] primarily focus on enforcing multi-view consistency within static scenes, and cannot handle temporal dynamics across frames.

2.3. 4D Scene Editing

Extending 3D scene editing to 4D representations introduces the additional challenge of maintaining temporal consistency while preserving multi-view coherence. Instruct 4D-to-4D [30] first applied diffusion-based 2D editing to sequentially rendered frames while optimizing the underlying NeRF-based 4D model, while 4D-Editor [17] incorporates spatial segmentation and motion-aware propagation for object-level editing. Control4D [37] enables 4D portrait editing by distilling the knowledge from a 2D diffusion into a 4D GAN [13] generator. CTRL-D [15] finetunes InstructPix2Pix [4] with prior-preservation loss [36] per scene for consistent 2D edits and iteratively optimizes 4DGS. Instruct4DGS [21] edits canonical Gaussians first and employs score-distillation-based [32] refinement for temporal smoothness. While these methods demonstrate notable progress, they still edit frames independently without simultaneously processing information across views and time, often causing motion or geometric distortion. In contrast, our Dynamic-eDiTor achieves consistent 4D editing by jointly editing cross-view and temporal frames, and by propagating these context tokens to the entire multi-view video.

3. Preliminary

3.1. 4D Scene Representation

3D Gaussian Splatting (3DGS) [19] represents a scene as a set of anisotropic Gaussian primitives $\mathcal{G} = \{(\mu_i, \Sigma_i, c_i, \alpha_i)\}_{i=1}^{N}$, each defined by its position, covariance, color, and opacity. Rendering is performed via differentiable alpha compositing. To model dynamic scenes, 4D Gaussian Splatting (4DGS) [46] extends this representation with a deformation field that maps canonical Gaussians to their deformed states over time. The field predicts offsets for position, rotation, and scale using MLPs ϕ_x, ϕ_r, and ϕ_s: $\Delta x = \phi_x(z)$, $\Delta r = \phi_r(z)$, and $\Delta s = \phi_s(z)$, where z is a temporal feature encoding the dynamic state of the scene. The final deformed Gaussian parameters are obtained as:

$$(x', r', s') = (x + \Delta x,\ r + \Delta r,\ s + \Delta s), \tag{1}$$

yielding the time-varying Gaussian set $\mathcal{G}'$.

4. Method

We propose a novel training-free 4D editing framework, **Dynamic-eDiTor**, carefully designed to achieve spatially and temporally consistent 4D scene editing leveraging Multimodal Diffusion Transformer (MM-DiT) [9, 45]. Initially, our approach edits source multi-view video frames at 1 FPS corresponding to a pre-trained 4D Gaussian Splatting (4DGS) [46], ensuring both multi-view and temporal consistency. The edited frames are then used to directly optimize the underlying pre-trained 4DGS representation.

4.1. Grid-based Spatio-Temporal Propagation

To ensure both multi-view and temporal consistency in multi-view video editing, we introduce Grid-based Spatio-Temporal Propagation, which enables coherent feature flow across the entire scene through two components: (1) Spatio-Temporal Sub-Grid Attention (STGA) for local fusion across adjacent views and neighboring time steps and (2) Context Token Propagation (CTP) for globally propagating the fused information through the entire multi-view video.

We begin by representing all multi-view video frames as a unified camera–time grid:

$$Grid = \{f_{v,t} \mid v \in [0, \dots, V],\ t \in [0, \dots, T]\}, \quad (2)$$

where $f_{v,t}$ is the frame captured by viewpoint v at time index t. The Grid's rows correspond to different viewpoints and columns represent temporally aligned frames.

To enable localized spatio-temporal fusion, we partition the $Grid$ into overlapping 2×2 sub-grid $\mathcal{S}_{v,t}$ at position (v, t) on $Grid$ defined as:

$$\mathcal{S}_{v,t} = \{f_{v,t},\ f_{v+1,t},\ f_{v,t+1},\ f_{v+1,t+1}\}, \quad (3)$$

each covering adjacent views and neighboring time steps. These sub-grids serve as the atomic processing units for STGA, allowing each local region to aggregate and share information across both the view and temporal axes.

To propagate information across the entire $V \times T$ $Grid$, we process the sub-grids sequentially using an asymmetric sliding pattern. We first sweep vertically along the spatial axis at t=0 to establish multi-view alignment, and then slide horizontally along the temporal axis to propagate consistency over time. The induced overlaps between neighboring sub-grids provide the structural linkage for STGA and CTP to effectively distribute information, enforcing globally coherent spatio-temporal editing.

4.1.1. Spatio-Temporal Sub-Grid Attention (STGA)

Grid-based Spatio-Temporal Propagation's foundation is the fusion of information within local neighborhoods. We propose Spatio-Temporal Sub-Grid Attention (STGA), which extends the dual-stream self-attention mechanism in MM-DiT architecture [9] to jointly attend adjacent views and temporally neighboring frames.

Instead of processing each frame's feature independently as in standard MM-DiT, STGA operates on a local sub-grid $\mathcal{S}_{v,t}$, enabling each frame to aggregate features from its cross-view and temporal neighbors. Each sub-grid contains four frames, and every frame $f_i \in \mathcal{S}_{v,t}$ is sequentially processed as a query in turn. Following MM-DiT's dual-stream attention design, the attention calculation involves two parts–the text stream and image-feature stream. For given frame f_i, we use its image query Q_{f_i}. We then extend the image-feature stream by concatenating all frame-level features within the sub-grid $\mathcal{S}_{v,t}$ to construct joint key and value sets $K_{\mathcal{S}_{v,t}}$ and $V_{\mathcal{S}_{v,t}}$:

$$\begin{aligned} K_{\mathcal{S}_{v,t}} &= [K_{f_{v,t}}, K_{f_{v+1,t}}, K_{f_{v,t+1}}, K_{f_{v+1,t+1}}], \\ V_{\mathcal{S}_{v,t}} &= [V_{f_{v,t}}, V_{f_{v+1,t}}, V_{f_{v,t+1}}, V_{f_{v+1,t+1}}]. \end{aligned} \quad (4)$$

The STGA for each frame f_i integrates the text stream $Q_{\text{txt}}, K_{\text{txt}}, V_{\text{txt}}$ with our modified, spatio-temporal image stream. We then apply Rotary Position Embeddings (RoPE) to the image queries $Q_{f_i}, K_{\mathcal{S}_{v,t}}, V_{\mathcal{S}_{v,t}}$ to provide positional information before the softmax operation:

$$\begin{aligned} \text{STGA}(\mathcal{S}_{v,t}) = \text{softmax}\Big([Q_{\text{txt}}, \text{RoPE}(Q_{f_{v,t}})] \cdot \\ [K_{\text{txt}}, \text{RoPE}(K_{\mathcal{S}_{v,t}})]^\top / \sqrt{d_k}\Big) \cdot [V_{\text{txt}}, V_{\mathcal{S}_{v,t}}], \end{aligned} \quad (5)$$

where d_k denotes the dimensionality of the key vectors.

STGA encourages cross-view and temporal coherence, forming the foundation for globally consistent multi-view video editing. Unlike previous temporal-only extensions of self-attention [12, 48], our STGA enables each frame query to attend both spatially adjacent views and temporally neighboring frames within its spatio-temporal patch. While STGA operates locally within each sub-grid, the overlapping sliding pattern naturally leads to implicit propagation of fused information across adjacent sub-grids.

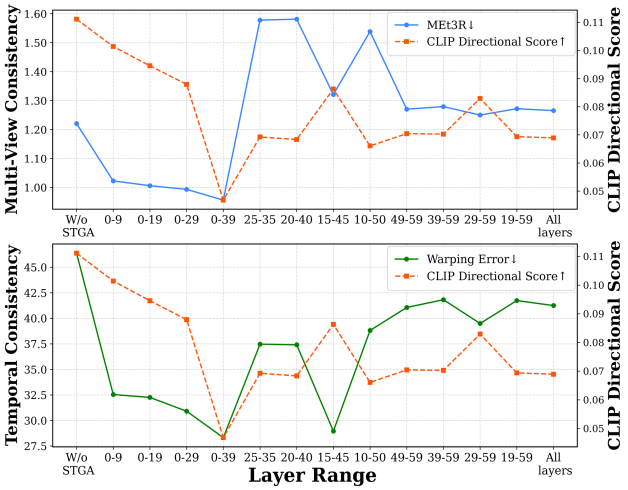

Figure 3. **Vital Layer Range Analysis.** We analyze the impact of applying Spatio-Temporal Sub-Grid Attention (STGA) across different layer ranges in MM-DiT [9, 45] during the multi-view video editing process. Performance is evaluated by temporal consistency (*Warping Error* [24]), multi-view consistency (*MEt3R* [2]), and editing fidelity (*CLIP Text-Image Directional Similarity [34]*). Applying STGA to the early ∼30 layers provides the best trade-off between consistency and editing fidelity.

Vital Layer Range Selection. As illustrated in Fig. 7, applying STGA to all self-attention layers of MM-DiT [9, 45] leads to visual artifacts, as the STGA tends to over-attend within the local spatio-temporal patch, resulting in texture repetition and view-dependent inconsistencies [10]. While prior studies [3, 11, 20, 44] analyze layer importance in

Figure 4. **Qualitative Comparison.** Dynamic-eDiTor enables more robust non-rigid content manipulation and achieves more complete edits of the 4D scene. The top-row displays the original rendered frames, while the following rows show the edited 4DGS renderings produced by each baseline. Our method (bottom-row) outperforms all baselines in both text alignment and overall editing fidelity, while maintaining strong temporal and spatial consistency.

DiT-based models [31] by ranking individual layers. Instead, we investigate applying STGA across continuous layer ranges to capture this cumulative effect. We empirically observe that there is a vital layer range to apply STGA for effectively enforcing multi-view and temporal consistency while alleviating editing-quality degradation. As shown in Fig. 3, applying STGA to the first 30 layers achieves the best trade-off between consistency and fidelity, providing significant improvements in both multi-view and temporal coherence while maintaining editing quality.

4.1.2. Context Token Propagation (CTP)

While STGA achieves local cross-view and temporal coherence by jointly attending adjacent views and neighboring frames within the sub-grid, Context-Aware Propagation ensures this coherence is globally distributed across entire $V \times T$ *Grid*. As the sub-grid slides along the defined traversal path, we introduce Context Token Propagation (CTP), which explicitly propagates context information. This ensures that the coherent feature representations that contain spatial and temporal information computed by STGA in the previous sub-grid S_{prev} are injected into the current sub-grid S_{curr}, ensuring coherent feature flow, enforcing global consistency and preventing information loss.

In this process, the token representation is defined as $\phi(\mathcal{S}_{v,t}) = \mathrm{STGA}(\mathcal{S}_{v,t})$. We employ two Context Token Propagation strategies: Full Token Inheritance and Flow-guided Token Replacement. Full Token Inheritance is applied when the current sub-grid S_{curr} shares frames with the previous sub-grid S_{prev} along the temporal axis ($t = 1 \rightarrow T-1$) or the spatial axis ($v = 1 \rightarrow V-1$). We directly replace the entire current token $\phi(\mathcal{S}_{curr})$ in these overlapped frames with previous token $\phi(\mathcal{S}_{prev})$.

For a sub-grid along the temporal axis, a defined traversal path yields non-overlapped regions in the rightmost column of the sub-grid. Thus, we apply Flow-guided Token Replacement to these regions, in which the tokens are replaced with tokens warped from the corresponding rightmost column regions of the previous sub-grid. To ensure temporal alignment during warping, we estimate forward and backward optical flow between frames f_t and f_{t-1} using RAFT [39], and downsample the flow fields to match the token resolution. For each spatial location (x, y), we use the downsampled forward flow $\mathbf{F}_{t\rightarrow t-1}(x, y)$ to backward-warp the tokens from the previous patch:

$$\hat{\phi}_{\mathrm{r}}(\mathcal{S}_{v,t}) = \mathrm{Warp}\big(\mathbf{F}_{t\rightarrow t-1}(x, y),\, \phi_{\mathrm{r}}(\mathcal{S}_{v,t-1})\big). \quad (6)$$

where $\hat{\phi}_{\mathrm{r}}(\mathcal{S}_{v,t})$ denotes the warped tokens in the rightmost column of the patch. During the replacement, we compute a validity mask $\mathrm{M(x, y)}$ via a forward–backward consistency check, inspired by [27, 49], to ensure precise replacement.

Method	Editing Fidelity		User Study						Reconstruction Fidelity		
	$\text{CLIP}_{\text{dir}}\uparrow$	$\text{CLIP}_{\text{sim}}\uparrow$	Overall Quality (%)	Motion Consist. (%)	Temporal Consist. (%)	Multi-view Consist. (%)	Prompt Align. (%)	Identity Preserv. (%)	PSNR ↑	SSIM ↑	LPIPS ↓
Instruct4D-to-4D [30]	0.1077	0.6308	27.57	27.72	28.00	27.19	22.14	27.48	21.86	0.6978	0.2145
Instruct-4DGS [21]	0.1501	0.6342	10.48	10.52	11.05	10.48	9.29	11.05	20.62	0.6252	0.2869
CTRL-D [15]	0.1498	0.6141	13.00	14.62	15.57	14.14	11.71	13.95	**31.06**	**0.8498**	**0.0970**
Ours	**0.1849**	**0.6397**	**48.95**	**47.14**	**45.38**	**48.19**	**56.86**	**47.52**	29.25	0.8064	0.1006

Table 1. **Quantitative Comparison.** The evaluation spans three aspects: editing fidelity, user preference, and reconstruction fidelity. CLIP-based metrics [34] show that Dynamic-eDiTor achieves strong alignment with the editing prompts across 4D scenes, and user studies indicate a clear preference for our results over the baselines in terms of semantic alignment, perceptual realism, and coherent motion. Although reconstruction metrics (PSNR, SSIM [43], LPIPS [50]) are slightly lower, they remain competitive and do not detract from the method's overall superiority in semantic accuracy and perceptual edit quality.

With the mask M, tokens in valid regions are replaced by the warped tokens, while those in invalid regions retain the current frame's tokens:

$$\phi_{\text{r}}(\mathcal{S}_{v,t}) = \text{M} \odot \hat{\phi}_{\text{r}}(\mathcal{S}_{v,t}) + (1 - \text{M}) \odot \phi_{\text{r}}(\mathcal{S}_{v,t}), \quad (7)$$

where $\odot$ denotes element-wise multiplication.

This unified propagation mechanism enables efficient and robust propagation across both spatial and temporal dimensions. STGA provides locally coherent spatial-temporal feature aggregation, while CTP propagates this coherence globally across the entire $Grid$. As a whole, they ensure consistent motion and geometry in multi-view videos, significantly improving stability in 4D editing.

4.2. Direct 4D Scene Optimization

Our approach produces multi-view and temporally consistent edited video frames, which can be directly used to optimize the pre-trained 4D representation. In contrast to prior works [15, 30] that rely on the Iterative Dataset Update (IDU), we directly optimize the 4D Gaussian model $\mathcal{G}'_{\text{edit}}$ using all edited frames $f^{\text{edit}}_{v,t}$ across the entire $Grid$. The optimization objective is defined as:

$$\mathcal{G}'_{\text{edit}} = \arg\min_{\mathcal{G}} \sum_{v,t \in V,T} \left\| \hat{f}_{v,t} - f^{\text{edit}}_{v,t} \right\| + \mathcal{L}_{\text{tv}}, \quad (8)$$

where both loss terms follow the 4DGS [46].

5. Experiment

5.1. Experimental Setup

Dataset. We evaluate our method on the real-world multi-view video dataset DyNeRF [26], which contains six dynamic scenes with 16-21 camera views capturing 10-second videos at 30 FPS. To evaluate editing consistency under sparse temporal sampling, we uniformly sample frames at 1FPS (160-210 frames per scene, compared to 4,800-6,300 frames at 30FPS). All experiments are conducted using 14 prompts covering all six scenes in the dataset.

Baselines. We compare our method against state-of-the-art 4D scene editing approaches, including Instruct 4D-to-4D [30], CTRL-D [15], and Instruct 4DGS [21]. Since our task focuses on text-based scene editing, we reproduce all baseline results using text prompt input only.

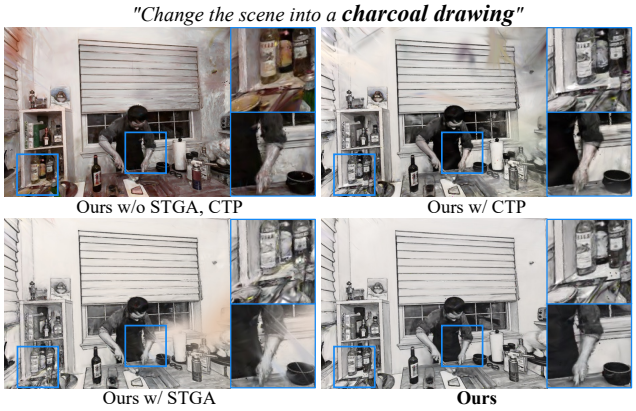

Figure 5. **Qualitative Ablation Results.** The model lacking both components (top-left) suffers from severe artifacts and geometric drift. Adding only STGA or only CTP progressively improves the result, but still leaves residual motion blur and geometric drift. Our full method (bottom-right) successfully ensures the spatio-temporal consistency to produce a stable and complete edit.

5.2. Implementation Details

Our method leverages the MM-DiT-based [9] image editing model, Qwen-Image-Edit [45] and 4D Gaussian Splatting [46]. For evaluation, we use all camera views in the dataset and hold out the final frame of each view as the test set. We evaluate our rendered results on this test set. The full 4D editing process takes approximately 51 minutes for the "coffee martini" scene on a single NVIDIA H100 GPU.

5.3. Qualitative Results

Fig. 1 illustrates Dynamic-eDiTor's ability to perform multi-view temporal scene editing. Our model effectively edits diverse scenes and local objects while maintaining strong temporal and spatial consistency. Dynamic-eDiTor is able to perform non-rigid appearance editing, semantic local editing, and consistent stylization while still preserving motion consistency across viewpoints and over time.

In Fig. 4, we compare Dynamic-eDiTor with recent 4D scene editing baselines [15, 21, 30]. We observe that most

STGA	CTP	2D Consistency		Reconstruction Fidelity			Editing Fidelity	
		Warp-Err $_{10^{-3}}\downarrow$	MEt3R $_{10^{-1}}\downarrow$	PSNR ↑	SSIM ↑	LPIPS ↓	$\text{CLIP}_{\text{dir}}\uparrow$	$\text{CLIP}_{\text{sim}}\uparrow$
-	-	56.98	1.0721	26.14	0.7445	0.1408	0.1930	0.5414
✓	-	38.64	0.9277	28.08	0.7875	0.1122	0.1872	0.6407
-	✓	29.44	1.0695	28.74	0.8013	0.1165	**0.1944**	**0.6418**
✓	✓	**28.94**	**0.9074**	**29.25**	**0.8064**	**0.1006**	0.1849	0.6397

Table 2. **Ablation Study.** Each component, Spatio-Temporal Sub-Grid Attention (STGA) and Context Token Propagation (CTP), helps preserve temporal and multi-view consistency, improving the 4D reconstruction quality. Our method prioritizes a globally stable 4D structure, yielding consistent temporal and spatial behavior and thus more robust reconstruction fidelity. Although CLIP-based metrics [34] show a slight drop due to the trade-off between semantic alignment and spatio-temporal coherence, our method still produces more stable and reliable 4D edits, avoiding the geometric and temporal artifacts seen in the ablated variants.

CTP-Full	CTP-Flow	2D Consistency		Reconstruction Fidelity			Editing Fidelity	
		Warp-Err $_{10^{-3}}\downarrow$	MEt3R $_{10^{-1}}\downarrow$	PSNR ↑	SSIM ↑	LPIPS ↓	$\text{CLIP}_{\text{dir}}\uparrow$	$\text{CLIP}_{\text{sim}}\uparrow$
-	-	38.64	0.9277	28.08	0.7875	0.1122	**0.1872**	**0.6407**
-	✓	29.79	0.9205	28.97	0.7990	0.1034	0.1852	0.6402
✓	-	33.22	0.9094	28.19	0.7906	0.1089	0.1865	0.6400
✓	✓	**28.94**	**0.9074**	**29.25**	**0.8064**	**0.1006**	0.1849	0.6397

Table 3. **Ablation Study: Context Token Propagation (CTP).** This ablation study is conducted with STGA included to isolate the impact of CTP. Full Token Inheritance (CTP-Full) and Flow-Guided Token Replacement (CTP-Flow) play a critical role in reinforcing temporal and multi-view consistency, enabling more accurate reconstruction of the edited dynamic scene. Despite a slight trade-off in CLIP-based metrics [34], CTP substantially improves spatio-temporal coherence and overall 4D editing fidelity.

baseline methods fail to handle non-rigid content manipulation and are often limited to style-oriented edits. This core limitation leads to significant artifacts, such as motion distortions, geometric drift, and incomplete editing results. These failures are evident across the examples. When editing the scene into a "fire emergency", all baselines fail to generate plausible emergency-related elements, revealing weak text–scene alignment and incomplete editing. In the second column, Instruct-4DGS struggles with non-rigid editing, causing clear motion distortions around the hand. Meanwhile, Instruct 4D-to-4D and CTRL-D introduce noticeable artifacts such as facial color shifts and blurring. In the third column, CTRL-D further demonstrates viewpoint inconsistencies and produces blurred edited regions, while other baselines result in incomplete edits. Instruct 4D-to-4D fails to modify the target scene, incorrectly altering surrounding objects. This indicates weak text alignment despite sharing the same diffusion backbone such as InstructPix2Pix [4] as other baselines. Overall, Dynamic-eDiTor outperforms all previous 4D scene editing methods by achieving superior editing completeness and effectively preserving temporal coherence and multi-view consistency, resulting in high-quality dynamic scene edits.

5.4. Quantitative Results

Tab. 1 presents a quantitative comparison with prior 4D scene editing methods, focusing on the 4D rendered image quality. Our evaluation is structured into two categories: text-prompt alignment and reconstruction fidelity.

To evaluate text-prompt alignment, we use CLIP-based [34] metrics. Specifically, the CLIP text-image directional similarity captures how changes in text captions correspond to changes between the source and rendered images in CLIP embedding space, while the CLIP text-image similarity directly measures alignment between the target text and rendered frames. Our method surpasses all baselines in these CLIP metrics, demonstrating that our rendered results are significantly better aligned with the user's intended edit.

To assess reconstruction fidelity, we report PSNR, SSIM [43], and LPIPS [50]. These metrics are computed between the final rendered test frames and the corresponding 2D edited target frames, measuring how faithfully the 4D model reconstructs the target edits. Although Dynamic-eDiTor obtains slightly lower values than CTRL-D on these reconstruction metrics, this highlights that our method achieves a better balance by prioritizing faithful text alignment and reliable 4D scene editing. This trade-off is further supported by our vital layer analysis in Fig. 3.

Beyond reconstruction metrics, we evaluate perceptual quality through a user study with 150 participants on Amazon Mechanical Turk [1]. Participants were asked to compare our final 4D rendered videos against baseline methods [15, 21, 30]. As shown in Tab. 1, our method consistently outperforms the baselines in terms of overall visual quality, motion consistency, temporal and multi-view consistency, text-prompt alignment, and identity preservation.

5.5. Ablation Study

We conduct an ablation study on our Grid-based Spatio-Temporal Propagation and Context Token Propagation.

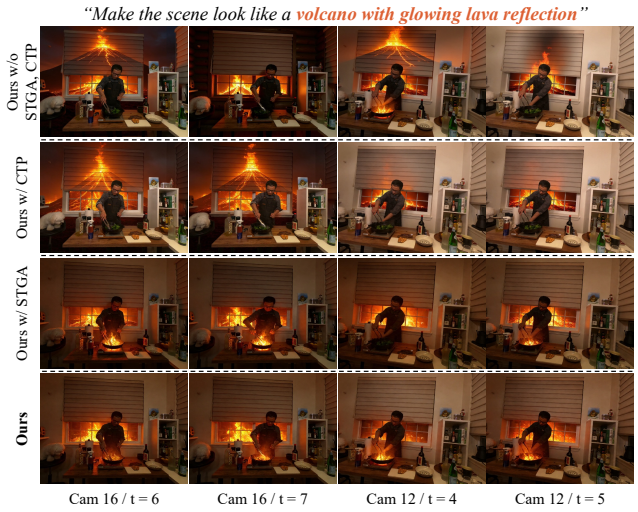

Figure 6. **Ablation Study: 2D Consistency.** Each component in our method strengthens temporal and multi-view consistency in 2D editing. STGA improves semantic alignment and preserves fine details across views, while CTP enhances coherence by propagating information across neighboring frames.

2D Consistency. We first assess each component's impact on 2D temporal and multi-view consistency, a key factor for high-quality 4D reconstruction. As shown in Fig. 6, STGA strengthens semantic alignment and view consistency, whereas CTP improves temporal coherence through information propagation. Collectively, they yield notable improvements in 2D spatial and temporal stability. The quantitative results in Tab. 2 support these findings. Our full method achieves superior spatio-temporal consistency, evidenced by the lowest warping error [24] and MEt3R [2], which further strengthens overall 4D fidelity.

4D Fidelity. Fig. 5 demonstrates how improved 2D consistency translates into higher-quality 4D scene edits. Without our components, the edited scene exhibits severe motion artifacts, especially around the man's hand, along with background degradation and incorrect text alignment, such as failing to produce "charcoal drawing" colors. Adding STGA reduces large-scale artifacts and stabilizes dynamic motion, while incorporating CTP further refines fine-grained details by leveraging temporal and multi-view cues from previous grids. With all components combined, Dynamic-eDiTor achieves consistent motion reconstruction and editing, effectively eliminating geometric drift in the 4D rendered output. Tab. 2 reveals that each component reinforces 4D fidelity. The PSNR, SSIM, and LPIPS scores validate this, demonstrating that Dynamic-eDiTor delivers highly coherent edits. We also note that CLIP-based metrics are slightly higher when STGA is removed. As mentioned in Sec. 4.1.1, this reflects the trade-off between semantic alignment and spatio-temporal coherence. balanced, spatio–temporally coherent 4D reconstruction, rather than optimizing semantic alignment alone. Our method prioritizes a balanced, stable, and coherent 4D reconstruction,

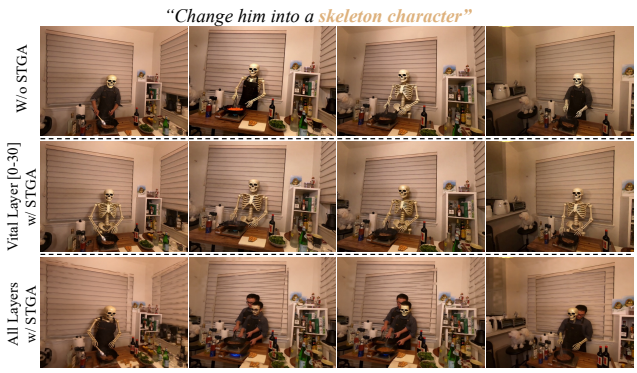

Figure 7. **Qualitative Analysis of Vital Layer Range.** Applying STGA to all layers introduces visual artifacts across views and time, while omitting STGA produces inconsistent multi-view and temporal edits. Restricting STGA to the vital range yields the most coherent and stable multi-view–time editing results.

whereas the ablated variant attains higher CLIP-based metrics by sacrificing this stability, resulting in geometrically and temporally unstable reconstruction.

Context Token Propagation. We further ablate our Context Token Propagation (CTP) components in Tab. 3. First, removing only the Full Token Inheritance leads to reduced 2D multi-view consistency and lower fidelity in the 4D rendered images. Next, removing only the Flow-guided Token Replacement results in a significant drop in temporal consistency. Finally, removing the entire Context Token Propagation mechanism (both components) causes a severe degradation in performance, dramatically worsening both 2D consistency and 4D reconstruction fidelity. Similar to the previous ablation, these results reflect the inherent trade-off between semantic alignment and spatio-temporal coherence, confirming that both the Full Token Inheritance and Flow-guided Token Replacement are essential for achieving high-quality and consistent 4D editing.

6. Conclusion

We presented Dynamic-eDiTor, a training-free framework for text-driven 4D scene editing that achieves spatially and temporally consistent results across multi-view videos, enabling stable optimization of 4D representations with MM-DiT [9, 45] and 4DGS [46]. The core of our approach is Grid-based Spatio-Temporal Propagation, combining Spatio-Temporal Sub-Grid Attention (STGA) for localized view-time fusion and Context Token Propagation (CTP) for explicit global consistency. Together, these components ensure coherent geometry and motion, and high-fidelity dynamic edits. Extensive experiments on DyNeRF [26] demonstrate our method significantly outperforms prior work in editing fidelity, temporal smoothness, and robustness. We believe Dynamic-eDiTor represents a notable progression toward text-driven dynamic scene editing.

7. Acknowledgements

This work is partially supported by the NSF under the Future of Work at the Human-Technology Frontier (FW-HTF) 1839971 and Partnership for Innovation (NSF PFI 2329804). The authors also acknowledge the Feddersen Distinguished Professorship Funds. Additional support for this work is partially provided by the Culture, Sports, and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (International Collaborative Research and Global Talent Development for the Development of Copyright Management and Protection Technologies for Generative AI, RS-2024-00345025), by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00521602), and by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University)) Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the funding agency.

References

[1] Amazon mechanical turk. https://www.mturk.com/, 2005. 7

[2] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6034–6044, 2025. 4, 8

[3] Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7877–7888, 2025. 4

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 2, 3, 7

[5] Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025. 2

[6] Minghao Chen, Iro Laina, and Andrea Vedaldi. Dge: Direct gaussian 3d editing by consistent multi-view editing. In *European Conference on Computer Vision*, pages 74–92. Springer, 2024. 3

[7] Jiahua Dong and Yu-Xiong Wang. Vica-nerf: View-consistency-aware 3d editing of neural radiance fields. *Advances in Neural Information Processing Systems*, 36: 61466–61477, 2023. 3

[8] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 1, 2, 3, 4, 6, 8

[10] Haoran Feng, Zehuan Huang, Lin Li, Hairong Lv, and Lu Sheng. Personalize anything for free with diffusion transformer. *arXiv preprint arXiv:2503.12590*, 2025. 4

[11] Sicheng Gao, Nancy Mehta, Zongwei Wu, and Radu Timofte. Ditvr: Zero-shot diffusion transformer for video restoration. *arXiv preprint arXiv:2508.07811*, 2025. 4

[12] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 4

[13] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3

[14] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19740–19750, 2023. 2

[15] Kai He, Chin-Hsuan Wu, and Igor Gilitschenski. Ctrl-d: Controllable dynamic 3d scene editing with personalized 2d diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26630–26640, 2025. 2, 3, 6, 7

[16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[17] Dadong Jiang, Zhihui Ke, Xiaobo Zhou, Tie Qiu, Xidong Shi, and Hao Yan. 4d-editor: Interactive object-level editing in dynamic neural radiance fields via semantic distillation. In *2025 International Conference on 3D Vision (3DV)*, pages 702–712. IEEE, 2025. 3

[18] Nazmul Karim, Hasan Iqbal, Umar Khalid, Chen Chen, and Jing Hua. Free-editor: zero-shot text-driven 3d scene editing. In *European Conference on Computer Vision*, pages 436–453. Springer, 2024. 3

[19] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 2, 3

[20] Min-Jung Kim, Dongjin Kim, Seokju Yun, and Jaegul Choo. Tv-live: Training-free, text-guided video editing via layer informed vitality exploitation. *arXiv preprint arXiv:2506.07205*, 2025. 4

[21] Joohyun Kwon, Hanbyel Cho, and Junmo Kim. Efficient dynamic scene editing via 4d gaussian-based static-dynamic separation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26855–26865, 2025. 2, 3, 6, 7

[22] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2

[23] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 2

[24] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 170–185, 2018. 4, 8

[25] Dong In Lee, Hyeongcheol Park, Jiyoung Seo, Eunbyung Park, Hyunje Park, Ha Dam Baek, Sangheon Shin, Sangmin Kim, and Sangpil Kim. Editsplat: Multi-view fusion and attention-guided optimization for view-consistent 3d scene editing with 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11135–11145, 2025. 3

[26] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5521–5531, 2022. 2, 6, 8

[27] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 5

[28] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[29] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2

[30] Linzhan Mou, Jun-Kun Chen, and Yu-Xiong Wang. 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*, pages 20176–20185, 2024. 2, 3, 6, 7

[31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2, 5

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

[33] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10318–10327, 2021. 1

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 4, 6, 7

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[36] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3

[37] Ruizhi Shao, Jingxiang Sun, Cheng Peng, Zerong Zheng, Boyao Zhou, Hongwen Zhang, and Yebin Liu. Control4d: Efficient 4d portrait editing with text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4556–4567, 2024. 3

[38] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023. 2

[39] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 2, 5

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[41] Junjie Wang, Jiemin Fang, Xiaopeng Zhang, Lingxi Xie, and Qi Tian. Gaussianeditor: Editing 3d gaussians delicately with text instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20902–20911, 2024. 2

[42] Yuxuan Wang, Xuanyu Yi, Zike Wu, Na Zhao, Long Chen, and Hanwang Zhang. View-consistent 3d editing with gaussian splatting. In *European conference on computer vision*, pages 404–420. Springer, 2024. 3

[43] 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. 6, 7

[44] Tianyi Wei, Yifan Zhou, Dongdong Chen, and Xingang Pan. Freeflux: Understanding and exploiting layer-specific roles in rope-based mmdit for versatile image editing. *arXiv preprint arXiv:2503.16153*, 2025. 4

[45] 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. 1, 2, 3, 4, 6, 8

[46] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20310–20320, 2024. 1, 2, 3, 6, 8

[47] 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 *European Conference on Computer Vision*, pages 55–71. Springer, 2024. 3

[48] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7623–7633, 2023. 4

[49] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8121–8130, 2022. 5

[50] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6, 7

[51] Canyu Zhao, Xiaoman Li, Tianjian Feng, Zhiyue Zhao, Hao Chen, and Chunhua Shen. Tinker: Diffusion's gift to 3d–multi-view consistent editing from sparse inputs without per-scene optimization. *arXiv preprint arXiv:2508.14811*, 2025. 3

[52] Zhe Zhu, Honghua Chen, Peng Li, and Mingqiang Wei. Coreeditor: Consistent 3d editing via correspondence-constrained diffusion. *arXiv preprint arXiv:2508.11603*, 2025. 3