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

Supplementary Material

Overview

This supplementary material provides additional details, analyses, and experimental results for our proposed method, Dynamic-eDiTor.

- Fig. 5 and Fig. 6 present additional qualitative results, including extended comparisons with baseline methods.
- Sec. A and Sec. B provide further implementation details and descriptions of all evaluation metrics used in our experiments.
- Sec. C summarizes the user study protocol and provides a detailed analysis of participant responses.
- Sec. D details our Grid-based Spatio-Temporal Propagation mechanism, including Asymmetric Traversal Strategy and the accompanying algorithm in Algorithm 1.
- Sec. E offers an extended analysis of the *vital layer range* for Spatio-Temporal Sub-Grid Attention (STGA).
- Sec. F contains additional ablation studies analyzing Asymmetric Traversal Strategy of Dynamic-eDiTor.
- Sec. G presents additional qualitative results in the monocular video setting of 4D Gaussian Splatting (4DGS) [26].
- Sec. H discusses the limitations of our Dynamic-eDiTor.

A. Implementation Details

For each scene, we first train the source 4D Gaussian Splatting [26] representation for 30,000 iterations using the Adam optimizer [12] with the same learning rate schedule as 4DGS. During the editing stage, we optimize the model for 20,000 iterations using the edited frames, following the original 4DGS hyperparameter configuration. All experiments are conducted on an NVIDIA H100 GPU; however, by employing local caching for Temporal Context Token Replacement, our method also runs efficiently on an NVIDIA A6000 GPU.

For the 2D MM-DiT [6, 25] image editor, we utilize Qwen-Image-Edit [25] from the Diffusers library [23]. To enhance computational efficiency, we incorporate the LoRA [11] weight Qwen-Image-Lightning-8steps-V1.1. All input images are resized to 768 × 768 before processing.

For baseline comparisons, we follow the official implementations of Instruct4D-to-4D [17] and Instruct-4DGS [14], since both are text-driven 4D editing methods comparable to ours. CTRL-D [10] requires an additional edited reference image that must be produced by choosing one of several diffusion-based editing modes, such as image-prompt editing, text-prompt editing, or mask-based editing, before fine-tuning its InstructPix2Pix [3] backbone. To ensure a fair and consistent comparison, we fix this pre-editing stage to use the standard InstructPix2Pix image editor, which is the backbone originally used in CTRL-D, when generating the reference edited image.

Algorithm 1 Asymmetric Sub-Grid Traversal

Require: Camera–time grid $\mathrm{Grid} = \{f_{v,t}\}$ of size $V \times T$
Ensure: Ordered list of sub-grids $\Omega = \{\mathcal{S}^{(k)}\}$

$\Omega \leftarrow [\,]$ ▷ Initialize empty sub-grid sequence
for $v = 0$ **to** $V - 2$ **do**
  if v is even **or** $v = V - 2$ **then** ▷ Temporal sweep
    for $t = 0$ **to** $T - 2$ **do**
      $\mathcal{S}_{v,t} \leftarrow \{f_{v,t}, f_{v+1,t}, f_{v,t+1}, f_{v+1,t+1}\}$
      Append $\mathcal{S}_{v,t}$ to Ω
    end for
  else ▷ Cross-view alignment at $t = 0$
    $\mathcal{S}_{v,0} \leftarrow \{f_{v,0}, f_{v+1,0}, f_{v,1}, f_{v+1,1}\}$
    Append $\mathcal{S}_{v,0}$ to Ω
  end if
end for
return Ω

B. Metric

To evaluate Dynamic-eDiTor, we use a combination of 2D consistency, 4D editing fidelity, and 4D reconstruction fidelity metrics.

For **2D consistency**, we evaluate both temporal and multi-view stability. Our 4D editing baselines such as Instruct4D-to-4D [17] and CTRL-D [10] rely on Iterative Dataset Update (IDU)[9], and Instruct-4DGS[14] uses an SDS-based [20] optimization strategy. However, these approaches do not generate temporally aligned or viewpoint-consistent 2D edited frames. Their updates are stochastic and occur directly in 3D or 4D space, which makes extracting coherent multi-view video sequences infeasible. Therefore, 2D consistency metrics cannot be fairly compared with these baselines and are used only within our ablation studies.

- **MEt3R** [2]: Evaluates multi-view consistency by comparing feature similarity between view-warped images.

Figure 1. **Qualitative comparison in the monocular video setting.** We evaluate our *Dynamic-eDiTor* on monocular sequences from the DyCheck dataset [8], where only a single moving-camera video is available without multi-view redundancy. We compare against Instruct4D-to-4D [17] and CTRL-D [10]; note that Instruct4D-to-4D does not provide an official implementation for monocular datasets. Built upon Deformable 3D Gaussian Splatting [27], our method enables effective text-driven appearance and object manipulations while preserving temporally consistent motion and stable geometry.

We employ the official MEt3R metric with MASt3R [16], DINOv2 [18] (FeatUp) features, 448 image resolution, and cosine similarity. Lower values indicate more coherent appearance across viewpoints.

- **Warping Error** [15]: Measures temporal consistency by computing the discrepancy between frame f_t and the optical-flow–warped version of frame f_{t-1} using RAFT [22]. Lower scores indicate smoother temporal alignment and fewer motion artifacts.

For **4D editing fidelity**, we adopt CLIP-based metrics [21]. We compute both CLIP text-image directional similarity and CLIP text-image similarity using the *rendered images* produced by the edited 4D scene. The directional similarity evaluates whether the change described by the text prompt corresponds to the transformation from the source image to the edited rendering in CLIP embedding space. The CLIP text-image similarity, on the other hand, directly measures how well the rendered frames semantically align with the target text prompt.

For **4D reconstruction fidelity**, we report PSNR, SSIM [24], and LPIPS [28], following prior works [5, 13, 14, 29]. All three metrics are computed between the edited test-view image and the rendered test-view image from the same camera viewpoint, enabling a direct comparison of reconstruction quality.

C. User Study Detail

To compare the editing performance of Dynamic-eDiTor against baseline methods, we conducted a user study with 150 participants on Amazon Mechanical Turk [1]. Each participant evaluated 14 scenarios, and for each scenario, they compared the 4D rendered video results produced by four systems across six subjective dimensions, as illustrated in Fig. 7. We designed six evaluation questions covering prompt alignment (Q1), temporal consistency (Q2), viewpoint consistency (Q3), motion consistency (Q4), identity preservation (Q5), and overall visual quality (Q6). For each dimension, participants selected the system they judged to perform best. To reduce human bias, the presentation order of the four systems was randomized for every question.

For analysis, we first counted how many times Dynamic-eDiTor was selected across the 14 scenarios and compared it with the best-performing baseline (best baseline) on each dimension. The results show that Dynamic-eDiTor consistently outperformed the best baseline across all six evaluation dimensions. For example, on overall quality (Q6), Dynamic-eDiTor achieved an average selection rate of 0.49, compared to 0.28 for the best baseline (Instruct4d [17]). The advantage is even more pronounced for prompt alignment (Q1), with selection rates of 0.57 vs. 0.22 (Instruct4d). For other questions, Dynamic-eDiTor's average selection rate exceeded the best baseline by approximately 0.17–0.21, demonstrating stable and comprehensive improvements.

To examine whether these differences were statistically significant, we performed a two-sided signed [4] test for each question, pairing each participant's selection ratio for Dynamic-eDiTor with that of the best baseline. All six dimensions yielded p-values far below 0.01, specifically: Q1 ($p = 8.88 \times 10^{-16}$), Q2 ($p = 5.43 \times 10^{-3}$), Q3 ($p = 4.49 \times 10^{-5}$), Q4 ($p = 1.41 \times 10^{-4}$), Q5 ($p = 9.33 \times 10^{-5}$), and Q6 ($p = 2.10 \times 10^{-5}$).

These results confirm that human evaluators consistently prefer Dynamic-eDiTor over the best baseline. The static significance also reveals that the gains are robust rather than

AGT	STGA	2D Consistency				Reconstruction Fidelity			Editing Fidelity	
		Local Warp-Err $_{10^{-3}}\downarrow$	Global Warp-Err $_{10^{-3}}\downarrow$	Local MEt3R $_{10^{-1}}\downarrow$	Global MEt3R $_{10^{-1}}\downarrow$	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP$_{\text{dir}}$ ↑	CLIP$_{\text{sim}}$ ↑
-	-	56.98	58.47	1.0721	1.4312	26.14	0.7445	0.1408	**0.1930**	0.6414
-	✓	**34.56**	42.86	**0.9266**	1.2984	27.84	0.7793	0.1160	0.1876	**0.6499**
✓	✓	38.64	**42.33**	0.9277	**1.2953**	**28.08**	**0.7875**	**0.1122**	0.1872	0.6407

Table 1. **Ablation Study: Asymmetric Sub-Grid Traversal (AGT).** This evaluation is conducted without CTP to isolate the impact of Asymmetric Sub-Grid Traversal (AGT). The results show that sub-grids without AGT achieve slightly better local consistency metrics because all frames within each sub-grid are updated independently. However, the lack of linkage between sub-grids introduces discontinuities, weakening overall 4D reconstruction fidelity. In contrast, applying AGT improves global consistency by overlapping frames across sub-grids, even at the cost of some local editing precision, as it enables effective information propagation. This leads to more stable and reliable 4D edits, demonstrating that global consistency is ultimately more critical for 4D reconstruction fidelity.

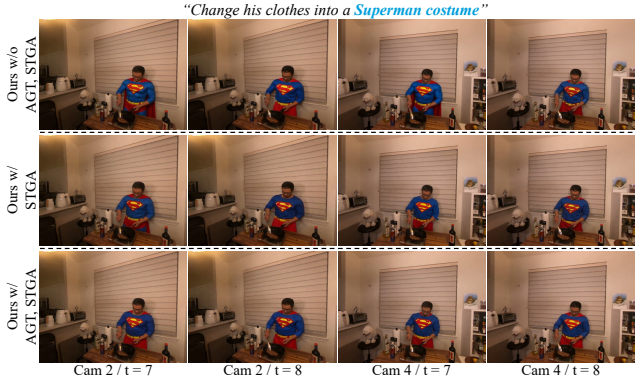

Figure 2. **Ablation Study: Asymmetric Sub-Grid Traversal (AGT).** This qualitative result clearly demonstrates that AGT preserves global multi-view and temporal consistency. Without AGT, noticeable discontinuities appear between sub-grids.

due to random variation.

D. Asymmetric Sub-Grid Traversal with Overlapping Structure

In the main paper, Grid-based Spatio-Temporal Propagation is introduced as a mechanism that performs local fusion via Spatio-Temporal Sub-Grid Attention (STGA) and global propagation via Context Token Propagation (CTP). In this section, we provide additional details on how the camera–time grid $Grid$ is traversed and how overlapping sub-grids are constructed to enable stable spatio-temporal propagation. Algorithm 1 formalizes the sub-grid generation process used in our implementation.

D.1. Overlapping Sub-Grids

Since neighboring sub-grids share frames on their boundaries, they form an overlapping tiling of the camera–time grid. This overlap is crucial for CTP: the shared regions act as "anchors" through which coherent token representations can be propagated from one sub-grid to the next.

D.2. Asymmetric Traversal Strategy

Rather than processing all $\mathcal{S}_{v,t}$ in a simple raster-scan order, we adopt an *asymmetric traversal* strategy that balances temporal propagation and cross-view alignment.

Concretely, we iterate over camera indices $v = 0, \ldots, V-2$, and for each v we choose a different temporal traversal pattern:

- For **even** camera indices (and the last camera pair $v = V-2$), we perform a *full temporal sweep*, generating sub-grids $\mathcal{S}_{v,t}$ for all $t \in [0, T-2]$. This encourages strong temporal propagation along the time axis.
- For **odd** camera indices, we generate only $\mathcal{S}_{v,0}$ (i.e., using the first two time steps $t = 0, 1$). This enforces cross-view alignment between neighboring cameras while avoiding redundant temporal passes.

The resulting traversal order can be summarized as follows: even-indexed camera rows perform dense temporal coverage, while odd-indexed rows act as cross-view bridges at the initial time step. This pattern yields an overlapping chain of sub-grids that spans the entire $V \times T$ grid, ensuring that information fused by STGA in one region can be propagated to distant regions through CTP.

D.3. Effect on STGA and CTP

This asymmetric, overlapping traversal has two key effects:

- **Local fusion via STGA.** Within each $\mathcal{S}_{v,t}$, STGA jointly attends over adjacent views and neighboring time steps, producing locally coherent spatio-temporal features. Due to the overlapping structure, boundary frames participate in multiple sub-grids, implicitly coupling neighboring regions.
- **Global propagation via CTP.** CTP operates along the traversal order Ω, propagating tokens from $\mathcal{S}_{\text{prev}}$ to $\mathcal{S}_{\text{curr}}$ through inherited and flow-guided token replacement. Since the sub-grids overlap in both view and time, this propagation forms a connected path over the entire grid, enabling the fused information to spread globally while respecting camera–time structure.

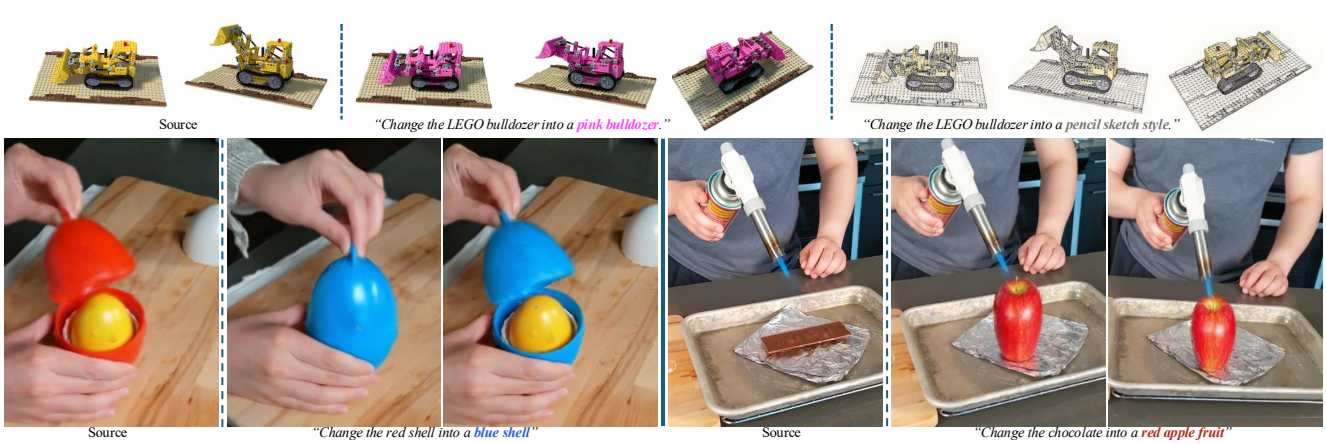

Figure 3. **Qualitative Results in Monocular video setting.** We evaluate the applicability of Dynamic-eDiTor on the challenging monocular video dataset HyperNeRF dataset [19], where only a single moving-camera sequence is available and no multi-view redundancy exists. Using Deformable 3D Gaussian Splatting [27] as the underlying 4D representation, our method successfully performs text-driven appearance and object manipulations while maintaining stable geometry and consistent motion over time.

Together, the asymmetric traversal and overlapping subgrids provide a principled backbone for Grid-based Spatio-Temporal Propagation, ensuring that local STGA fusion and global CTP propagation jointly enforce consistent editing across all views and time steps.

E. Vital Layer Range Analysis for STGA

In the main paper, we apply Spatio-Temporal Sub-Grid Attention (STGA) only to a vital layer range of MM-DiT in order to enhance spatio-temporal consistency without overly harming editing fidelity. Here, we provide a more detailed quantitative analysis of this design choice.

Experimental setup. We conduct a systematic study on the DyNeRF dataset using 3 scenes sampled at 1 FPS and 5 editing prompts per scene (15 sequences in total). For each configuration, we enable STGA on a different continuous range of MM-DiT layers and keep all other components fixed. We report (i) *Warping Error*↓ [15] for temporal consistency, (ii) *MEt3R*↓ [2] for multi-view consistency, and (iii) *CLIP Text-Image Directional Similarity*↑ (CLIP_{dir}) [21] for editing fidelity.

Effect of early-layer STGA. Without STGA, both Warping Error and MEt3R are the worst (46.37 and 1.22), indicating pronounced temporal flicker and cross-view inconsistency. As we progressively introduce STGA from shallow layers (0–9, 0–19, 0–29), both consistency metrics steadily improve. In particular, enabling STGA on the first 30 layers (0–29) yields a strong reduction in temporal and multi-view error (Warping Error 30.90, MEt3R 0.99), while preserving a relatively high CLIP_{dir} score (0.088). This configuration achieves the best overall trade-off: it significantly enhances spatio-temporal coherence compared to the baseline, yet maintains competitive editing fidelity.

Layer Range	Warp-Err $_{10^{-3}}$ ↓	MEt3R $_{10^{-1}}$ ↓	CLIP_{dir} ↑
W/o STGA	46.37	1.221	0.1111
0–9	32.54	1.022	0.1014
0–19	32.25	1.006	0.0946
0–29	30.90	0.993	0.0879
0–39	28.32	0.956	0.0468
25–35	37.47	1.578	0.0693
20–40	37.41	1.581	0.0683
15–45	28.96	1.321	0.0863
10–50	38.81	1.538	0.0661
49–59	41.06	1.270	0.0704
39–59	41.82	1.279	0.0703
29–59	39.49	1.250	0.0829
19–59	41.74	1.272	0.0694
All layers	41.25	1.265	0.0689

Table 2. **Detailed vital layer range analysis for STGA.** This table reports the exact numerical values corresponding to the trend shown in Figure 3 of the main paper.

Applying STGA too deep. Extending STGA too aggressively into deeper layers (e.g., 0–39 or mid-to-deep ranges such as 25–35, 20–40, 10–50, or 19–59) further reduces or even oscillates consistency metrics, but at the cost of a substantial drop in CLIP_{dir}. For example, 0–39 attains the lowest Warping Error and MEt3R among all settings, but its CLIP_{dir} score collapses to 0.047, indicating that over-attending within local spatio-temporal neighborhoods can oversmooth edits, weaken text alignment, and lead to texture repetition or view-dependent artifacts, as shown in Fig. 7 of the main paper. Similarly, configurations that only activate STGA in deeper blocks (e.g., 39–59, All layers) nei-

ther recover the consistency of early-layer STGA nor preserve high editing fidelity, suggesting that late-stage modifications are less effective for enforcing stable geometry and motion.

Chosen configuration. Based on these observations, we adopt the 0–29 configuration as our default choice in the main paper. This vital layer range provides a balanced compromise: it substantially improves temporal and multi-view consistency over the baseline and deep-only variants, while incurring only a modest decrease in CLIP_{dir} relative to the no-STGA setting. In practice, we find that this trade-off yields visually smoother 4D reconstructions and more reliable 4D scene editing, whereas configurations with either no STGA or overly deep STGA tend to produce flickering, geometric drift, or oversmoothed, weakly edited results.

F. Additional Ablation Study

We conduct additional ablation study to evaluate the impact of Asymmetric Sub-Grid Traversal (AGT) in Dynamic-eDiTor.

Asymmetric Sub-Grid Traversal (AGT) For the ablation of AGT, we perform experiments without CTP, as it relies on the sliding mechanism introduced by AGT. AGT is designed to create overlapping regions between adjacent sub-grids, promoting smoother transitions and improving global consistency. We hypothesize that removing this overlap will yield results that remain locally consistent within each sub-grid but exhibit severe temporal and multi-view discontinuities across sub-grid boundaries.

To fairly assess global consistency, we introduce two new metrics: **Global Warping Error** and **Global MEt3R**. Unlike their standard versions, these metrics are computed only between the left boundary frames of adjacent sub-grids, directly quantifying the discontinuities that AGT aims to mitigate. Additionally, we define per-frame consistency metrics as **Local Warping Error** and **Local MEt3R**, which are the original 2D consistency metric used in the main paper.

As shown in Tab. 1, removing AGT yields slightly higher local consistency because each sub-grid updates all its frames during every iteration. However, without any connection between sub-grids, clear discontinuities emerge between them, degrading overall 4D reconstruction quality. In contrast, AGT introduces overlap across sub-grids, enabling information to propagate between them and producing more coherent results, despite a slight reduction in local consistency, as only the non-overlapping frames are updated. Overall, these findings confirm that AGT plays a crucial role in preserving global coherence, enabling higher-quality and more faithful 4D reconstruction than using STGA alone.

We also observe that the non-sliding variant yields slightly higher CLIP scores, reflecting the inherent trade-off between consistency and text alignment. Without sliding, the model can more aggressively edit each isolated sub-grid to match the prompt, but this comes at the cost of failing to produce a globally coherent 4D scene, which is the primary objective of our method.

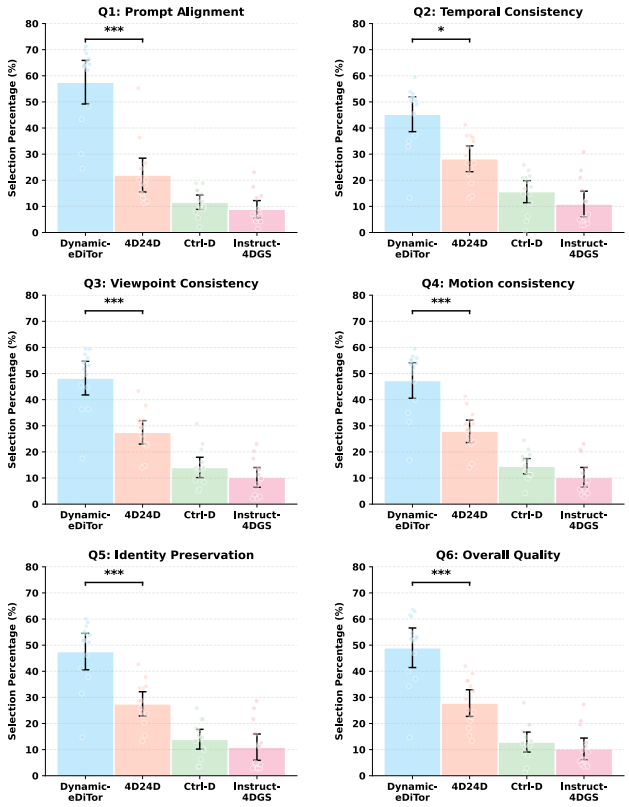

Figure 4. **User Study.** The user study results indicate a human preference for Dynamic-eDiTor, with superior ratings in both consistency and edited-quality categories compared to all baselines

G. Monocular Video Setting

4D dynamic scene editing typically refers to a multi-view video setting where sufficient spatio-temporal information is captured. However, to explore the applicability of Dynamic-eDiTor in a monocular setting, we evaluate our method on the DyCheck [7] dataset using Deformable 3D Gaussian Splatting model [27]. Since a monocular dataset contains only a single camera, we modify the camera–time grid to a purely temporal grid:

$$Grid_{temp} = \{f_t \mid t \in [0, \ldots, T]\}, \tag{1}$$

Accordingly, each sub-grid $\mathcal{S}_t$ consists of consecutive frames along the temporal axis:

$$\mathcal{S}_t = \{f_t,\ f_{t+1},\ f_{t+2},\ f_{t+3}\}. \tag{2}$$

Based on this modified sub-grid, we apply same Spatio-

STGA	CTP	2D Consistency				Reconstruction Fidelity			Editing Fidelity	
		Local Warp-Err $_{10^{-3}}\downarrow$	Global Warp-Err $_{10^{-3}}\downarrow$	Local MEt3R $_{10^{-1}}\downarrow$	Global MEt3R $_{10^{-1}}\downarrow$	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP$_{\text{dir}}$ ↑	CLIP$_{\text{sim}}$ ↑
-	-	56.98	58.37	1.0721	1.4312	26.14	0.7445	0.1408	0.1930	0.5414
✓	-	38.64	42.33	0.9277	1.2953	28.08	0.7875	0.1122	0.1872	0.6407
-	✓	29.44	31.63	1.0695	1.4364	28.74	0.8013	0.1165	**0.1944**	**0.6418**
✓	✓	**28.94**	**30.69**	**0.9074**	**1.2657**	**29.25**	**0.8064**	**0.1006**	0.1849	0.6397

Table 3. **Ablation Study: Local and Global Consistency.** Our method improves both local and global 2D consistency, ensuring that each sub-grid remains coherent both internally (local) and with its neighbors (global). Each component, STGA and CTP, helps maintain temporal and multi-view consistency, improving overall 4D reconstruction. By enforcing a globally stable 4D structure, our method achieves more consistent spatio-temporal behavior and higher reconstruction fidelity. Although CLIP-based metrics [21] show a slight drop due to the trade-off between semantic alignment and spatio-temporal coherence, our approach still delivers 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	
		Local Warp-Err $_{10^{-3}}\downarrow$	Global Warp-Err $_{10^{-3}}\downarrow$	Local MEt3R $_{10^{-1}}\downarrow$	Global MEt3R $_{10^{-1}}\downarrow$	PSNR ↑	SSIM ↑	LPIPS ↓	CLIP$_{\text{dir}}$ ↑	CLIP$_{\text{sim}}$ ↑
-	-	56.98	58.37	1.0721	1.4312	26.14	0.7445	0.1408	**0.1930**	**0.6407**
-	✓	29.79	32.66	0.9205	1.2813	28.97	0.7990	0.1034	0.1852	0.6402
✓	-	33.22	37.36	0.9094	1.2736	28.19	0.7906	0.1089	0.1865	0.6400
✓	✓	**28.94**	**30.69**	**0.9074**	**1.2657**	**29.25**	**0.8064**	**0.1006**	0.1849	0.6397

Table 4. **Ablation Study: Context Token Propagation (CTP).** This ablation study is conducted with STGA included to isolate the effect of CTP. The results show that our method maintains both local (within each sub-grid) and global consistency. Full Token Inheritance (CTP-Full) and Flow-Guided Token Replacement (CTP-Flow) play a crucial role in reinforcing temporal and multi-view coherence, enabling more accurate reconstruction of edited dynamic scenes. Although CLIP-based metrics [21] show a slight trade-off, CTP significantly enhances spatio-temporal consistency and overall 4D editing fidelity.

Temporal Sub-Grid Attention (STGA) mechanism. However, because only temporally adjacent frames are available, the key and value sets $K_{\mathcal{S}_t}$ and $V_{\mathcal{S}_t}$ become:

$$\begin{aligned} K_{\mathcal{S}_t} &= [K_{f_t}, K_{f_{t+1}}, K_{f_{t+2}}, K_{f_{t+3}}], \\ V_{\mathcal{S}_t} &= [V_{f_t}, V_{f_{t+1}}, V_{f_{t+2}}, V_{f_{t+3}}]. \end{aligned} \tag{3}$$

Thus, STGA in the monocular setting becomes:

$$\begin{aligned} \mathrm{STGA}(\mathcal{S}_t) = \mathrm{softmax}\Big([Q_{\text{txt}},\, \mathrm{RoPE}(Q_{f_t})] \cdot \\ [K_{\text{txt}},\, \mathrm{RoPE}(K_{\mathcal{S}_t})]^\top / \sqrt{d_k}\Big) \cdot [V_{\text{txt}},\, V_{\mathcal{S}_t}], \end{aligned} \tag{4}$$

where d_k denotes the dimensionality of the key vectors.

For Context Token Propagation (CTP), where the token representation is defined as $\phi(\mathcal{S}_t) = \mathrm{STGA}(\mathcal{S}_t)$, we employ two Context Token Propagation strategies: Full Token Inheritance and Flow-guided Token Replacement. Since the sub-grids overlap by two temporal frames, we directly replace the entire current token $\phi(\mathcal{S}_{curr})$ in these overlapped frames with the previous token $\phi(\mathcal{S}_{prev})$. For the non-overlapped region, which corresponds to the two rightmost frames of the sub-grid, we apply flow-guided token replacement. To propagate the most recent temporal information, we warp tokens from the rightmost frame of the overlapped region and replace the tokens of the two non-overlapping frames:

$$\hat{\phi}_{\text{r}}(\mathcal{S}_t) \;=\; \mathrm{Warp}\big(\mathbf{F}_{t\to t-1}(x,y),\, \phi_{\text{r}}(\mathcal{S}_{t-1})\big), \tag{5}$$

where $\hat{\phi}_{\text{r}}(\mathcal{S}_t)$ denotes the warped tokens in the rightmost column of the patch and $\mathbf{F}_{t\to t-1}(x,y)$ represents the downsampled forward flow. To ensure precise replacement, we compute a validity mask M(x,y) and replace only tokens in valid regions with warped tokens.

As shown in Fig. 3, Dynamic-eDiTor achieves stable and reliable monocular scene editing. Our model effectively maintains the temporal consistency, and both STGA and CTP contribute significantly to producing temporally coherent non-rigid appearance edits and semantic local editing.

H. Limitation.

While Dynamic-eDiTor effectively enforces multi-view and temporal consistency during text-driven edits, it is less suitable for large-scale geometric alterations such as substantial motion reconfiguration or topology-changing edits. Since our framework operates by propagating spatio-temporal features without modeling geometric deformation, edits that require significant structural changes remain challenging. Extending our approach to handle more drastic motion editing or geometry-changing transformations represents an important direction for future work.

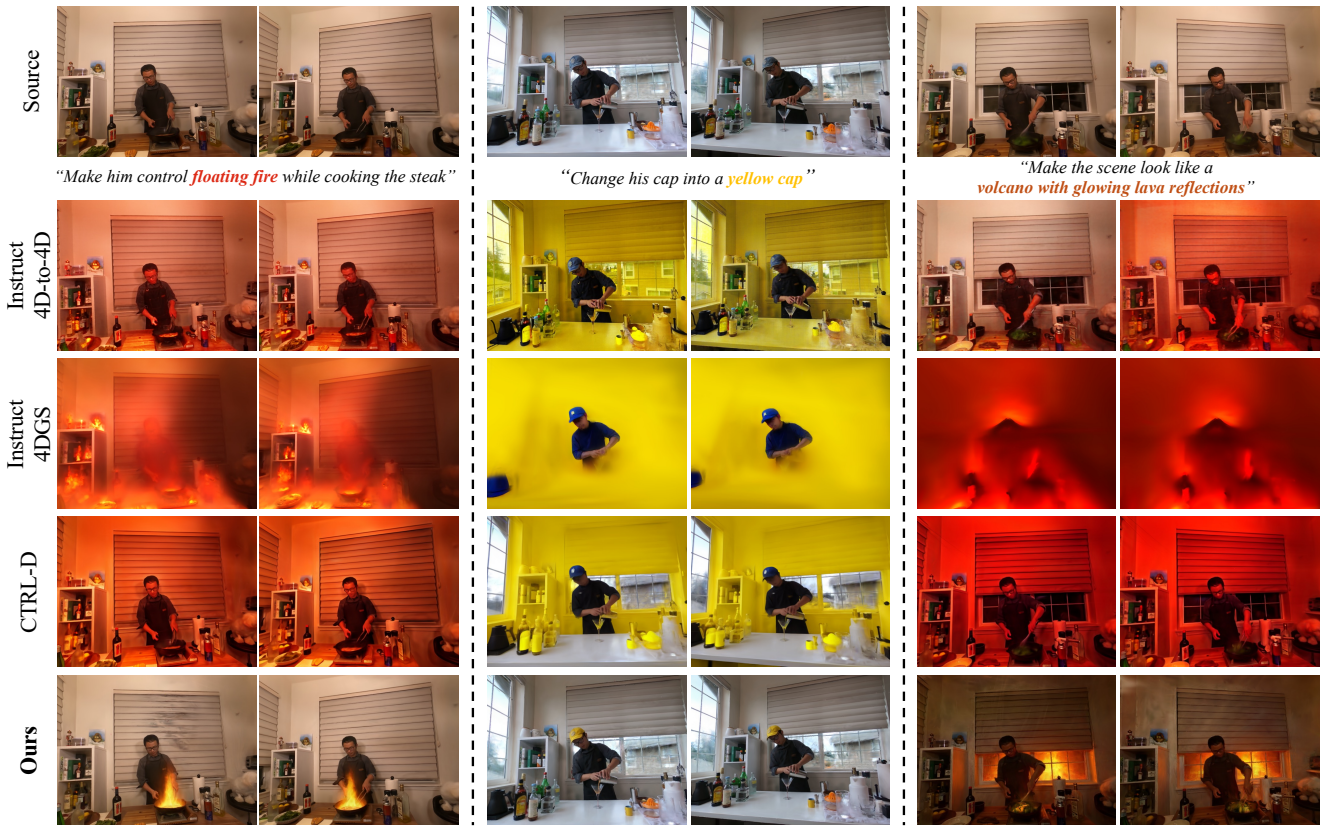

Figure 5. **Qualitative Results.** Dynamic-eDiTor enables higher-quality manipulation of non-rigid content and delivers more complete edits across the 4D scene. The upper row contains the original rendered frames, while the rows beneath show the edited 4DGS results from each baseline approach. Our method (bottom row) demonstrates superior correspondence to the text prompt and achieves strong edit fidelity while maintaining temporal and spatial consistency.

Figure 6. **Qualitative Results.** Dynamic-eDiTor preserves both multi-view and temporal consistency, enabling high-quality text-driven editing of pre-trained 4D Gaussian Splatting. It is capable of performing effective edits across diverse scenes as well as on local objects.

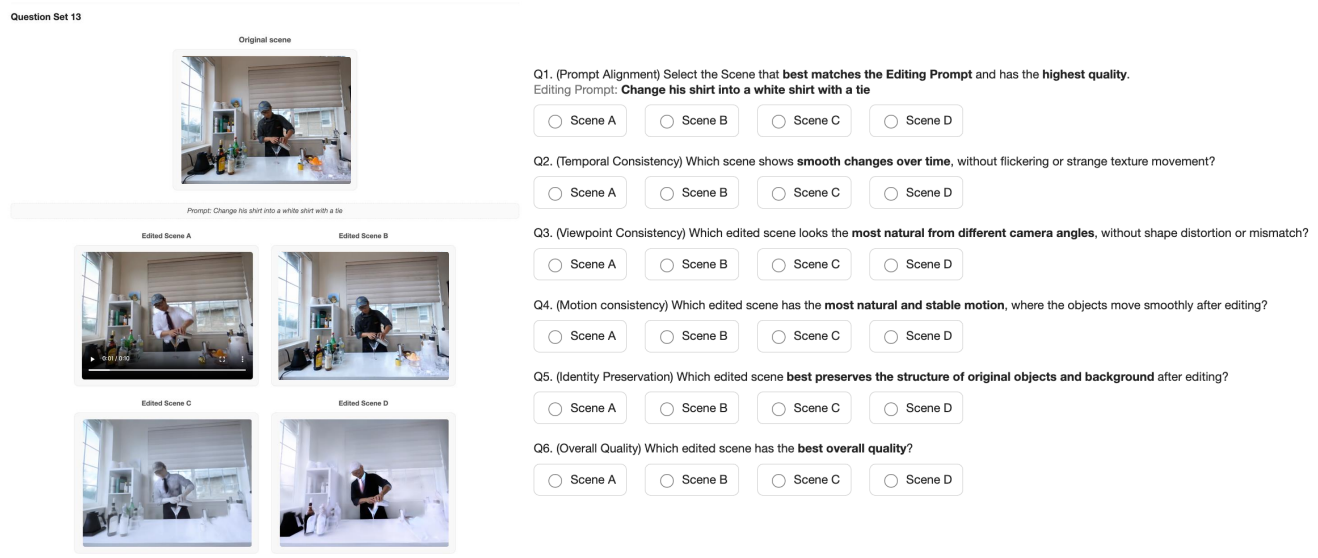

Figure 7. **User study interface and questionnaire.** We illustrate the interface used in our user study. Each participant is first shown the original scene and text prompt, then presented with four edited 4D-rendered video results (A–D) generated by different methods. Participants watch the videos and select the best method for each of the six evaluation criteria: prompt alignment (Q1), temporal consistency (Q2), viewpoint consistency (Q3), motion consistency (Q4), identity preservation (Q5), and overall quality (Q6).

References

[1] Amazon mechanical turk. https://www.mturk.com/, 2005. 2

[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. 1, 4

[3] 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. 1

[4] Wilfrid J Dixon and Alexander M Mood. The statistical sign test. *Journal of the American Statistical Association*, 41(236):557–566, 1946. 2

[5] Hyungjun Doh, Dong In Lee, Seunggeun Chi, Pin-Hao Huang, Kwonjoon Lee, Sangpil Kim, and Karthik Ramani. Occlusion-aware temporally consistent amodal completion for 3d human-object interaction reconstruction. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 52–61, 2025. 2

[6] 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

[7] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Dynamic novel-view synthesis: A reality check. In *NeurIPS*, 2022. 5

[8] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. *Advances in Neural Information Processing Systems*, 35:33768–33780, 2022. 2

[9] 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. 1

[10] 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. 1, 2

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 1

[12] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[13] Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars. *arXiv preprint arXiv:2502.20220*, 2025. 2

[14] 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. 1, 2

[15] 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. 2, 4

[16] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 2

[17] 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. 1, 2

[18] 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. 2

[19] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 4

[20] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1

[21] 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. 2, 4, 6

[22] 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

[23] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 1

[24] 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. 2

[25] 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

[26] 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

[27] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint arXiv:2309.13101*, 2023. 2, 4, 5

[28] 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. 2

[29] Qi Zuo, Xiaodong Gu, Yuan Dong, Zhengyi Zhao, Weihao Yuan, Lingteng Qiu, Liefeng Bo, and Zilong Dong. High-fidelity 3d textured shapes generation by sparse encoding and adversarial decoding. In *European Conference on Computer Vision*, pages 52–69. Springer, 2024. 2