Fast Multi-view Consistent 3D Editing with Video Priors

Liyi Chen[1], **Ruihuang Li**[1], **Guowen Zhang**[1], **Pengfei Wang**[1], **Lei Zhang**[1*]

[1] The Hong Kong Polytechnic University
{liyi0308.chen, guowen.zhang, pengfei.wang}@connect.polyu.hk, {csrhli, cslzhang}@comp.polyu.edu.hk

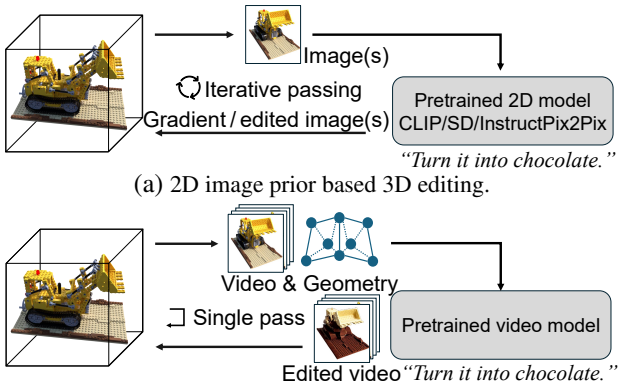
Abstract

Text-driven 3D editing enables user-friendly 3D object or scene editing with text instructions. Due to the lack of multi-view consistency priors, existing methods typically resort to employ 2D generation or editing models to process per-view individually, followed by iterative 2D-3D-2D updating. However, these methods are not only time-consuming but also prone to yielding over-smoothed results, since iterative process averages the different editing signals gathered from different views. In this paper, we propose, an early and pioneering work of generative **Vi**deo **P**rior based **3D E**diting, **ViP3DE** in short, to repurpose the temporal consistency priors from pre-trained video generation models to achieve consistent 3D editing within a single forward pass. Our key insight is to condition the video generation model on a single edited view to generate other consistent edited views for 3D updating directly, thereby bypassing iterative editing paradigm. First, 3D updating requires edited views to be paired with specific camera poses. To this end, we propose *motion-preserved noise blending* for the video model to generate edited views at predefined camera poses. In addition, we introduce *geometrically aware denoising* to further enhance multi-view consistency by integrating 3D geometric priors into video models. Extensive experiments demonstrate that our proposed ViP3DE can achieve high-quality 3D editing results even within a single forward pass, significantly outperforming existing methods in both editing quality and editing time cost.

Project page — https://mt-cly.github.io/ViP3DE

Introduction

3D editing aims to achieve high-quality personalized editing of 3D objects or scenes by modifying their shape and content. Traditional 3D editing requires professional skills and tools to manipulate 3D mesh (Yu et al. 2004; Sorkine et al. 2004) or point clouds (Zwicker et al. 2002), which is time-consuming and labor-intensive. Recent development of 3D representations and multi-modality 2D models has revolutionized the way of 3D editing. By integrating NeRF (Mildenhall et al. 2021) or 3D Gaussians Splatting (GS) (Kerbl et al. 2023) with off-the-shelf 2D multi-modal models such as CLIP (Radford et al. 2021) and Stable Diffusion (SD) (Rombach et al. 2022), a user-friendly interface can be built to edit 3D content using textual instructions.



(a) 2D image prior based 3D editing.

(b) ViP3DE integrates video priors and 3D priors for 3D editing.

Figure 1: **The motivation of ViP3DE.** (a) Most existing studies (Haque et al. 2023; Chen et al. 2024b; Wang et al. 2022; Liu et al. 2024; Chen, Laina, and Vedaldi 2024; Li et al. 2024a; Chen et al. 2024a) employ pre-trained 2D models to iteratively update 3D assets, suffering from slow convergence and over-smooth textures. (b) ViP3DE integrates video priors and source 3D priors to achieve multi-view consistent editing with a single pass.

Most existing methods employ 2D models to perform 3D editing with an iterative updating paradigm, as shown in Fig. 1(a). In each iteration, images from a randomly selected camera pose are rendered and edited individually using models like CLIP and SD (Kamata et al. 2023; Wang et al. 2022, 2023; Chen et al. 2024b; Liu et al. 2024; Li et al. 2024c) via Score Distillation, *e.g.*, SDS, DDS, and SSD (Hertz, Aberman, and Cohen-Or 2023; Zhu et al. 2025) or 2D editor to output edited view (Haque et al. 2023; Chen et al. 2024c). The source 3D representation is updated with edited views through differential volume rendering (Mildenhall et al. 2021) or rasterization (Kerbl et al. 2023). Due to the lack of multi-view consistency priors, these per-image editing-based methods usually demand hundreds or thousands of iterations to average out inconsistent gradient signals or edit pixel values, resulting in slow convergence and

*Corresponding author

over-smoothed texture. Although recent efforts have been made to synchronize different views by introducing extrapolated cross-attention (Chen, Laina, and Vedaldi 2024; Chen et al. 2024a) or point correspondence (Li et al. 2024a; Song et al. 2023), they fail to achieve editing in a single pass, and these issues remain.

Motivated by the capability of pre-trained video models to generate inter-frame continuous videos, we propose to leverage pre-trained video priors to achieve multi-view consistent 3D editing in a single pass, as shown in Fig. 1(b). However, generative video models cannot be applied directly for 3D editing due to two issues. Firstly, 3D updating requires pairs of camera poses and edited images, while existing video models cannot produce edited images corresponding to precise camera poses (He et al. 2024; Ku et al. 2024). Secondly, video models have limited understanding of 3D geometry and physics (Brooks et al. 2024; Kang et al. 2024). Therefore, the edited views often suffer from shape deformation or color shifts, leading to unwanted 3D editing.

In this paper, we propose ViP3DE to overcome these challenges. To obtain paired edited views and camera poses, we first render a source video along a known camera trajectory. We then acquire inverted video noise to guide subsequent edited view generation via an inversion-based paradigm. Note that, different from previous inversion-based video editing methods (Ling et al. 2024; Fan, Bhattad, and Krishna 2024; Ku et al. 2024), which either overestimate or uunderestimatethe importance of this inverted noise and results in unsatisfactory camera motion or visual quality, we propose *motion-preserved noise blending* to produce desired edited views under given 3D perspectives by blending inverted noise with random Gaussian noise as the initial noise. Besides, to further improve 3D consistency of edited views, we propose *geometrically aware denoising* to integrate 3D priors and video priors during diffusion process. We first build latent feature correspondence between the conditional view and other views based on the geometric relations between 3D representation and camera poses, which can then provide explicit constraints in video latent space across each denoising step. These two novel designs enable the video model to produce edited views that are continuous and consistent in 3D geometry. Finally, the source 3D asset is updated using the edited views in a single forward pass. Our contributions are summarized as follows.

- We propose ViP3DE, an early and pioneering work that introduces generative video priors for text-driven 3D editing.
- We introduce two novel designs, *i.e.*, motion-preserved noise blending and geometrically aware denoising, to produce 3D-consistent edited views with high-quality visual results in a single pass.
- Extensive experiments demonstrates ViP3DE significantly outperforms the previous methods in both efficiency and editing quality.

Related Work

Text-driven 3D Generation and Editing. One line of instruction-based 3D editing directly updates 3D representations using gradients from 2D models (Radford et al. 2021; Rombach et al. 2022; Zhang, Rao, and Agrawala 2023). This includes methods using CLIP as a discriminator (Wang et al. 2022, 2023), SDS loss (Chen et al. 2024b), or learned NeRF mappings (Liu et al. 2024), with many others following a similar paradigm (Kamata et al. 2023; Zhuang et al. 2023). Another line performs 2D view editing (Brooks, Holynski, and Efros 2023) to iteratively optimize the 3D representation, enhancing consistency through techniques like synchronized noise (Chen et al. 2024a), pixel correspondence (Li et al. 2024a; Song et al. 2023), or revised cross-attention (Chen, Laina, and Vedaldi 2024; Wang et al. 2024). However, these iterative approaches (Haque et al. 2023) often cause over-smoothed textures and converge slowly (Song et al. 2023). Instead, our ViP3DE leverages video priors for high-quality, multi-view consistent 3D editing in a single pass.

Video Models. Early video generation methods (Blattmann et al. 2023; Yang et al. 2024) adapt pre-trained 2D models by incorporating temporal modules (Çiçek et al. 2016; Vaswani 2017). Following methods like CameraCtrl (He et al. 2024) attempt to introduce camera control. Existing video editing methods, from early training-free approaches (Wu et al. 2023; Geyer et al. 2023) to more recent video models (Wu et al. 2025; Ouyang et al. 2024; Ku et al. 2024), are fundamentally limited by their lack of 3D geometric knowledge. Consequently, they are ill-suited for 3D editing and often rely on complex pre-processing steps such as optical flow and point tracking.

The Proposed Approach

Following common 3D editing protocols, ViP3DE begins with a 3D scene from systems like COLMAP (Schönberger and Frahm 2016) and user-provided instructions. The 3D editing is performed according to the user-provided instructions. ViP3DE employs SVD-XT (Blattmann et al. 2023) for its competitive performance with much faster inference speed compared to large models $e.g.$, Wan2.2 (Wan et al. 2025), CogvideoX (Yang et al. 2024). The ViP3DE workflow is illustrated in Fig. 2. Firstly, we render 3D scene to obtain source views following continuous camera trajectories, termed source video. Then, we propose *motion-preserved noise blending* and *geometrically aware denoising* to achieve geometrically consistent edited multi-views, which are used to update source 3D Gaussians.

Editing Multi-view Images with Video Prior

ViP3DE accomplishes multi-view editing by integrating InstructPix2Pix (Brooks, Holynski, and Efros 2023) with SVD-XT in an inversion-based manner (Fan, Bhattad, and Krishna 2024; Ku et al. 2024; Ouyang et al. 2024). In particular, for a source video $\mathcal{I}_{src} = \{I^1_{src}, ..., I^N_{src}\}$ with N-view, its latent x_0 is first obtained by the VAE encoder, then passed through EDM (Karras et al. 2022) inversion to get the inverted latent noise x_t at step $t = 1, ..., T$. The first frame I^1_{src} and textual instruction are fed into InstructPix2Pix to obtain an edited view I^1_{cond}, which serves as the condition to guides the denoising process to produce

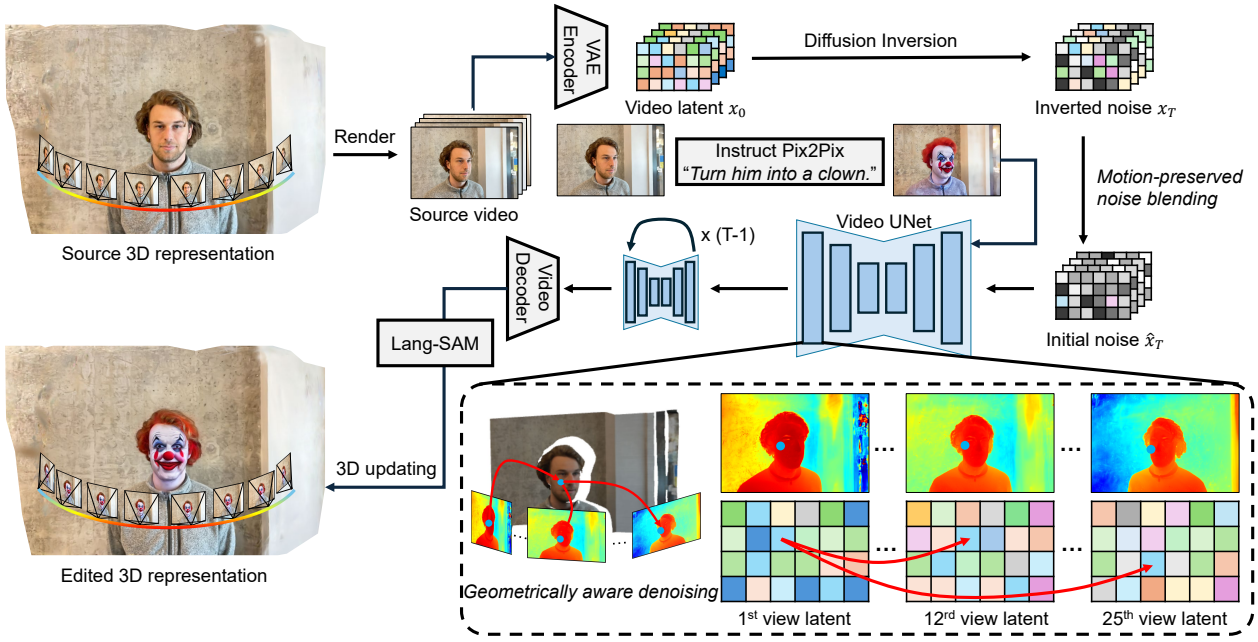


Figure 2: **Workflow of ViP3DE.** First, the contiguous multi-view images are rendered from the source 3D representation as source video. Then, the first frame is edited with InstructPix2Pix as the condition of the video model, and the initial noise of the diffusion process is obtained by *motion-preserved noise blending*. Furthermore, the geometric priors excavated from the source 3D representation are introduced during the video denoising process to improve 3D consistency across views, termed *geometrically aware denoising*. Finally, these edited multi-view images are utilized to update the source 3D representation. Thanks to video priors, ViP3DE achieves 3D consistent editing in a single forward pass.

edited video latent $\widehat{x}_0$. Finally, $\widehat{x}_0$ is converted to edited video $\mathcal{I}_{edit} = \{I^1_{edit}, ..., I^N_{edit}\}$ using the pre-trained video decoder. We fully take advantage of the byproduct of EDM inversion ($i.e.$, inverted noise and attention maps) to inherit the camera motion from the source video to output consistent edited views, thereby they can be directly used to update the 3D representation in known camera poses.

Balance Pose Alignment and Editing Quality

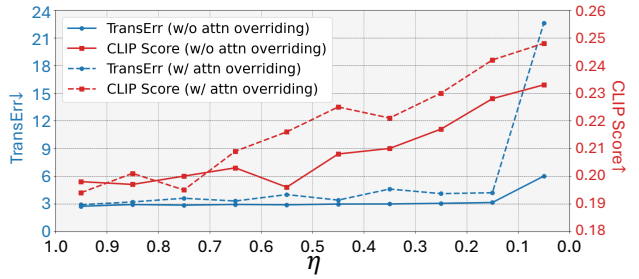
We notice that by starting the generation process from x_T, the edited multi-view images can achieve desired camera pose alignment. However, they exhibit significant artifacts if the instruction indicates substantial modifications, $e.g.$, "*Turn the man into a clown.*", as shown in the first column of Fig. 3(b). We hypothesize that this is because the inverted noise x_T contains not only motion cues but also the appearance cues of the source video. When conditioned on an edited image I^1_{cond} that significantly deviates from the appearance of source video, the network can be confused and produce ambiguous results. Therefore, there is a critical question: *Can we retain the camera pose/motion information while removing unwanted appearance cues in the inverted noise?*. We experimentally find that motion and appearance cues exhibit different behaviors as the initial noise progressively transitions from inverted noise x_T to random Gaussian noise ϵ, enabling the disentanglement of motion cues and appearance cues.

Specifically, we perform a pilot study by randomly selecting 100 off-the-shelf 3D assets (Haque et al. 2023; Mildenhall et al. 2019) and render them into 3D videos. Then we leverage GPT-4o (Achiam et al. 2023) to generate editing instructions and target prompts based on the first frame of the source video, which are utilized to generate the edited video by noise inversion. The initial noise $\widehat{x}_T$ is obtained by blending the inverted noise x_T and random Gaussian noise ϵ with different weights:

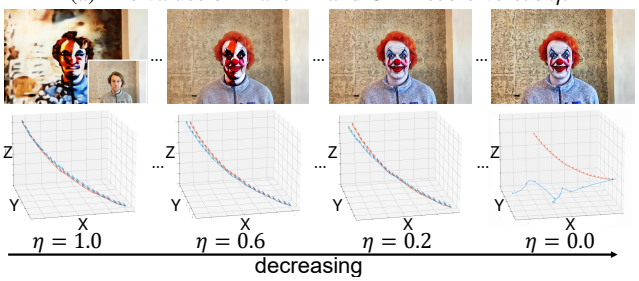
$$\widehat{x}_T = \sqrt{\eta} x_T + \sqrt{1-\eta}\epsilon, \text{where } \epsilon \sim \mathcal{N}(0, \sigma_T^2 \mathbf{I}), \quad (1)$$

where $\eta \in [0, 1]$ controls the intensity of inverted noise. We call this operation *motion-preserved noise blending*.

We employ COLMAP (Schönberger and Frahm 2016) and TransErr (He et al. 2024) to estimate and compare the camera poses of edited and source videos with different $\widehat{x}_T$, and use CLIP text-image score (Radford et al. 2021) to study the faithfulness of edited first frames I^1_{edit}. We analyze both results with and without *attention overriding*, which is a typical technique to improve camera pose alignment (Ku et al. 2024; Ling et al. 2024). The results under different η are plotted in Fig. 3(a). One can observe that as η gradually decreases from 1 to 0, the quality of edited frames improves steadily, with persistent performance gains within the entire range of η. Meanwhile, the camera pose alignment exhibits remarkable robustness, maintaining highly competitive TransErr even when the signal-to-noise ratio is low ($i.e.$, $\eta = 0.1$). Note that $\eta = 0$ corresponds to the completely free generation with random camera trajectories. The editing ex-



(a) The values of TransErr and CLIP score versus η.



(b) Edited first frame and camera poses with different η.

Figure 3: **Demonstration of motion-preserved noise blending.** Appearance and pose alignment exhibit different levels of robustness to noise.

ample "*Turn the man into a clown*" is shown in Fig. 3(b).

The above findings can be interpreted from two perspectives. Firstly, setting initial noise close to random Gaussian noise allows better appearance manipulation, which is identical to the conclusion in many image editing works (Mokady et al. 2023; Li et al. 2024b). Secondly, we posit that camera motion constitutes global low-frequency information, making it relatively insensitive to additive white Gaussian noise. Based on the different behaviors between motion and appearance cues, we set $\eta = 0.15$ to balance camera pose alignment and editing quality.

Integrating 3D Priors and Video Priors

Equipped with motion-preserved noise blending, our model can generate high-quality edited views under specific camera poses. However, we observe that parts of edited views suffer from structural deformation and color shifts. This is because while the generated videos exhibit inter-frame continuity, they are not consistent in 3D space. This discrepancy stems from the fact that current generative video models lack a comprehensive understanding of the real-world 3D geometry (Kang et al. 2024; Brooks et al. 2024). To address this issue, we propose *geometrically aware denoising* to exploit the geometric priors from the source 3D Gaussians to guide the denoising process of video generation.

We begin by rendering the depth of the source Gaussian to obtain depth maps $\mathcal{D} = \{D_1, ..., D_N\}$ from N views. Based on known camera poses, we identify the corresponding pixel at $(\hat{u}, \hat{v})$ in the first conditional frame for the pixel at (u, v) in the i-th frame through 3D projection:

$$[\hat{u}, \hat{v}, \widehat{D}_1(\hat{u}, \hat{v})]^{\mathrm{T}} = \mathrm{K}_1 \mathrm{Rt}_1 \mathrm{Rt}_i^{-1} \mathrm{K}_i^{-1} [u, v, D_i(u, v)]^{\mathrm{T}}, \quad (2)$$

Algorithm 1: 3D Consistent Editing with ViP3DE

Input: Source 3D views $\mathcal{I}_{src} = \{I^1_{src}, ..., I^N_{src}\}$, hyper-parameter η, τ, classifier guidance w, geometric constraint index M, and instruction $\mathcal{P}$.
Output: Edited 3D views $\mathcal{I}_{edit} = \{I^1_{edit}, ..., I^N_{edit}\}$.
$\{x_1, x_2, ...x_T\} \leftarrow$ EDM-Inv$(\mathcal{I}, I^1_{src})$
$I^1_{cond} \leftarrow Edit(I^1_{src}; \mathcal{P})$
$\widehat{x}_T = \sqrt{\alpha} x_T + \sqrt{1-\alpha}\epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_T^2 \mathbf{I})$
for $t = T, T-1, \ldots, 1$ **do**
  $\widehat{x}^{uc}_t \leftarrow \widehat{x}_t$
  $\widehat{x}^{c}_t \leftarrow \widehat{x}_t[M]$ # geometrically aware denoising
  $\widehat{x}_{t-1} \leftarrow$ $(1+w)$GVM$(\widehat{x}^c_t, I^1_{cond}; t) - wGVM(\widehat{x}^{uc}_t, \varnothing; t)$ # classifier-free guidance
end
$\mathcal{I}_{edit} \leftarrow \widehat{x}_0$
Return $\mathcal{I}_{edit}$

where K_i and Rt_i are intrinsic and extrinsic matrices of the i-th camera pose, and $\widehat{D}_1(\hat{u}, \hat{v})$ is the projected Gaussian depth at $(\hat{u}, \hat{v})$ of the first frame. Considering that some projected Gaussians may not be visible in the first frame, we filter out occluded pixel correspondences by comparing the projected depth $\widehat{D}_1$ with the rendered depth map D_1. The mapping M for (u, v) of the i-th frame is defined as follows:

$$M(i, u, v) = \begin{cases} (\hat{u}, \hat{v}) & \text{if } |\widehat{D}_1(\hat{u}, \hat{v}) - D_1(\hat{u}, \hat{v})| < \tau \\ \varnothing & \text{otherwise,} \end{cases} \quad (3)$$

where τ is the threshold to filter occluded correspondences.

The computed geometric correspondence is then leveraged to provide additional guidance during the diffusion denoising process. Specifically, we downsample the pixel correspondences to the latent resolution. At each denoising step, the features of all frames in the last layer of the UNet are replaced by those from the corresponding positions in the first frame if valid correspondence exists. Furthermore, we fully exploit the advantage of classifier-free guidance by performing feature overriding only in conditional latent features and omitting the unconditional features so that rich texture can be generated from the unconditional prediction term, achieving multi-view consistent 3D editing results.

Pseudo Code. Formally, let $Edit(I, \mathcal{P})$ be the image editing function with input image I and editing instruction prompt $\mathcal{P}$. We employ InstructPix2Pix (Brooks, Holynski, and Efros 2023) to align with previous work. Denote by $GVM(x_t, I; t)$ the computation of a single diffusion step t in the generative video model with noisy video latent x_t conditioned on Image I. Let M record the indices of features from each view's latent to corresponding first view latent features, used for providing 3D geometric constraint. The algorithm of ViP3DE is summarized in Alg. 1.

Remarks. Existing methods rely on epipolar constraints (Chen, Laina, and Vedaldi 2024; Huang et al. 2024) or semantic correspondence (Luo et al. 2024), which omit 3D Gaussian geometric priors and suffer from

Method	Multi-modal Models	Consistency Mechanism	CLIP T-I Sim.	CLIP Direction Sim.	Forward Pass	Time
NeRF-Art (Wang et al. 2023)	CLIP	Iterative Updating	0.243	0.121	400	$>$ 8 hours
ViCA-NeRF (Dong and Wang 2024)	InstructPix2Pix	Depth Constraint	0.274	0.183	2	$\sim$ 25 min
GaussCtrl (Wu et al. 2024)	ControlNet	Depth Constraint	0.266	0.170	1	$\sim$ 10 min
InstructN2N (Haque et al. 2023)	InstructPix2Pix	Iterative Updating	0.262	0.145	5000	$\sim$ 28 min
GaussianEditor (Chen et al. 2024c)	InstructPix2Pix	Iterative Updating	0.272	0.187	1500	$\sim$ 8 min
DGE (Chen, Laina, and Vedaldi 2024)	InstructPix2Pix	Extrapolated Attention	0.269	0.180	3	$\sim$ 4 min
AnyV2V (Ku et al. 2024)	I2VGen-XL	Video Editing	0.230	0.114	1	$\sim$ 12 min
InsViE (Wu et al. 2025)	CogVideoX	Video Editing	0.244	0.132	1	$\sim$ 2 min
I2VEdit (Ouyang et al. 2024)	SVD-XT	Video Editing	0.264	0.178	1	$\sim$ 44 min
ViP3DE (Ours)	SVD-XT	Consistent Video Prior	**0.284**	**0.197**	1	$\sim$ 3 min

Table 1: **Quantitative comparison with previous methods.** ViP3DE achieves more faithful results to the texture instruction with higher CLIP scores. In addition, ViP3DE costs less time to converge.

inaccurate matching. Studies (Li et al. 2024a; Song et al. 2023) enforce pixel-wise constraints, introducing noticeable artifacts and requiring multiple forward passes. In contrast, our method addresses these limitations through latent-space integration of 3D priors and video priors, achieving high-fidelity editing in a single forward pass.

Implementation Details

Sub-video Parallel Inference. We perform individual view editing for all video frames and select one with the highest CLIP similarity score as the condition of SVD-XT. To adapt the first-frame conditional paradigm, we temporally partition the video into two subsequences at the conditional frame while reversing the temporal order of the preceding sub-video. These two sub-videos share the same condition and are edited in parallel to reduce time costs. An autoregressive manner (Ouyang et al. 2024) is adopted if the frames of sub-video exceed the model's context length.

Views Continuity. We place continuous cameras around the 3D scene to obtain source video. Specifically, we first follow DGE (Chen, Laina, and Vedaldi 2024) to sort the training view cameras according to their view changes. Then, we randomly select several cameras from the sorted training views as the key cameras, which are used to obtain interpolated cameras based on Slerp (Shoemake 1985; Dam, Koch, and Lillholm 1998) and linear interpolation.

Updating 3D Representation. We find that InstructPix2Pix tends to override the whole image even if the editing instruction specifies partial localization. Therefore, we follow GaussianEditor (Chen et al. 2024c) to employ SAM (Kirillov et al. 2023) to calculate 2D masks, which prevent the 3D updating from occurring in unwanted regions. Following the typical 3D Gaussians reconstruction (Kerbl et al. 2023), we update source 3D under the supervision of editing results with L_1 and LPIPS losses (Zhang et al. 2018).

Experiments

Dataset and Metrics. To compare ViP3DE with previous methods, we collect 3D scenes and objects assets from diverse datasets. Considering the significant time cost in previous methods like (Wang et al. 2023), we conduct fair comparative evaluations using a subset of Mip-NeRF360 (Barron et al. 2022) and LLFF (Mildenhall et al. 2019). We follow common practice (Haque et al. 2023; Chen, Laina, and Vedaldi 2024) using CLIP text-image similarity and directional similarity to evaluate the alignment of editing and instructions, CLIP temporal consistency (Haque et al. 2023) is adopted to study the cross-view consistency.

Experimental Setting. The CFG of textual and image conditions in InstructPix2Pix (Brooks, Holynski, and Efros 2023) are set to 7.5 and 1.5. The numbers of inversion steps and denoising steps in video models are set to 25. The τ in Eq. 3 is set to 0.5. We use edited multi-view images to perform 750 updating iterations. All experiments are conducted on RTX A6000. We collect assets of 3D scenes and objects from diverse datasets. Considering the significant time cost in previous methods (Wang et al. 2023), we conduct fair comparison using a subset of Mip-NeRF360 (Barron et al. 2022) and LLFF (Mildenhall et al. 2019).

Comparison to 3D Editing Methods

Quantitative Comparison. We We demonstrate the effectiveness and efficiency of ViP3DE by comparing it to previous 3D editing methods and recent video editing methods. The numerical results are reported in Table 1. Early work NeRF-Art (Wang et al. 2023) employs CLIP as a discriminator to edit a 3D object with VolSDF as representation, suffering from low convergence since rendering is time-consuming. ViCA-NeRF (Dong and Wang 2024) forces the consistency of features in InstructPix2Pix, resulting in over-smoothed editing results with relatively low CLIP scores. Although GaussianCtrl (Wu et al. 2024) can achieve 3D editing in a single iteration, it uses ControlNet (Zhang, Rao, and Agrawala 2023) to perform editing, which often produces unfaithful results to the given text prompt. InstructN2N (Haque et al. 2023) and GaussianEditor (Chen et al. 2024c) take advantage of InstructPix2Pix to achieve high-quality editing. However, they edit each view independently, resulting in cross-view inconsistency. In addition, they rely on multiple forward passes, causing slow

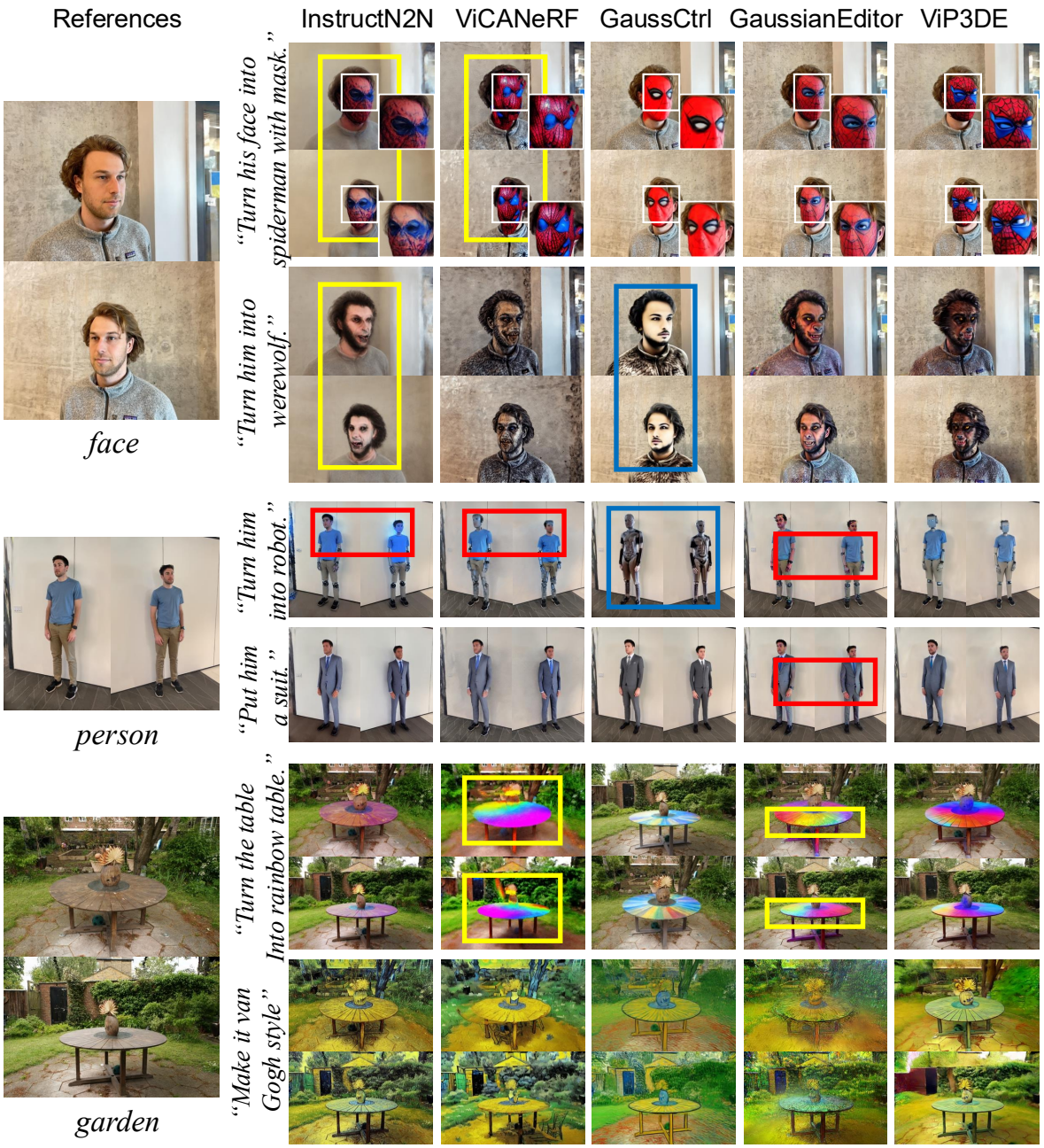


Figure 4: **Qualitative comparison.** We highlight the edited results suffering from inconsistency (red boxes), poor details (yellow boxes), and unfaithfulness (blue boxes). In comparison, ViP3DE obtains consistent results with higher faithfulness to instruction. Besides, rich details are preserved by avoiding multiple iterations that typically cause over-smooth textures.

convergence. DGE (Chen, Laina, and Vedaldi 2024) introduces semantic correspondence from TokenFLow (Geyer et al. 2023) to improve consistency. However, it inherits the 2D model limitation of lacking cross-view consistency priors. AnyV2V (Ku et al. 2024), InsViE (Wu et al. 2025) and I2VEdit (Ouyang et al. 2024) fail to achieve 3D consistent results since they are toward dynamic object editing in pure video space.

Instead, ViP3DE introduces video priors with proposed motion-preserved noise blending and geometrically aware denoising, making edited views consistent and faithful.

Qualitative Comparison. We provide visual comparisons to further illustrate the advantages of ViP3DE. We mainly study three failure cases including inconsistency, poor texture details, and unfaithfulness. As shown in Fig. 4, InstructN2N and GaussianEditor lead to view-inconsistent re-

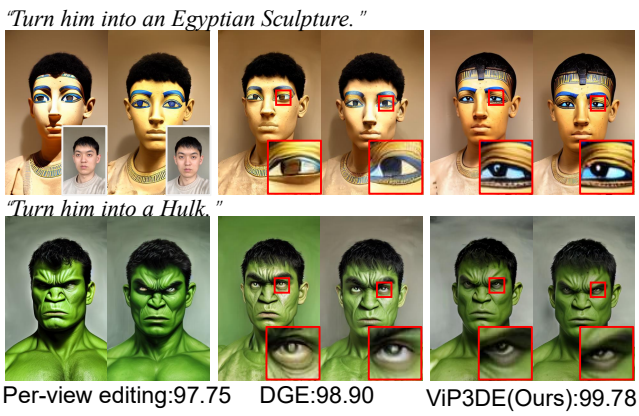


Figure 5: **The editing results with CLIP temporal score in a single forward pass.** ViP3DE achieves better consistency.

sults due to per-image editing. On *robot*, InstructN2N shows inconsistency in person heads. GaussianEditor maintains different hand colors in different views. In addition, it often introduces floating Gaussian artifacts, as shown in the *garden* scene. ViCANeRF fails to produce satisfied appearance with over-smoothed textures (*e.g.*, *rainbow table*). Although GaussCtrl can achieve consistent editing, it suffers from unfaithfulness to the given textural prompt, caused by the use of ControlNet. DGE (Chen, Laina, and Vedaldi 2024) extends InstructPix2Pix to a video model, sharing a similar motivation to ours. We compare the details of edited frames in a single forward pass in Fig. 5. Both DGE and ViP3DE achieve better temporal consistency compared to per-image editing independently. On *fangzhou*, compared to DGE, which fails to maintain detail consistency and demands iterative updating, ViP3DE achieves better consistency and richer textures by fully exploiting video priors from pre-trained video models.

Diagnostic Experiments

Quantitative Ablation Studies. We conduct ablation studies to evaluate the effectiveness of key components in ViP3DE in Tab. 2. We see that the generative video model fails to tackle discrete 3D views and only achieves a CLIP score of 0.198. With continuous views as input, editing 3D multi-views with InstructPix2Pix and generative video models still outputs unfaithful visual results. Our proposed motion-preserved noise blending brings a significant performance improvement. Introducing geometrically aware denoising further enforces cross-view 3D consistency, achieving the best performance.

Qualitative Ablation Studies. We employ face and bear as examples to demonstrate the functionality of each component. The absence of ordered continuous views in input would cause the video model to fail in proper view editing, resulting in a mismatch between edited views and their corresponding camera poses. Consequently, the edited 3D representation exhibits significant quality degradation even with mask constraints. Removing motion-preserved noise blending makes the editing unfaithful to the conditional images, showing confused appearance artifacts. In addition, geometrically aware denoising reinforces the geometric relationships between views while maintaining alignment with the source 3D structure, effectively preventing blurring artifacts in the edited 3D views.

Continuous views	Mition-preserved noise blending	Geometrically aware denoising	CLIP score T-I	CLIP score Direction
			0.198	0.083
✓			0.220	0.109
✓	✓		0.271	0.185
✓		✓	0.258	0.181
✓	✓	✓	**0.284**	**0.197**

Table 2: **Ablation study on ViP3DE.** Removing either motion-preserved noise blending or geometrically aware denoising reduces the performance significantly.

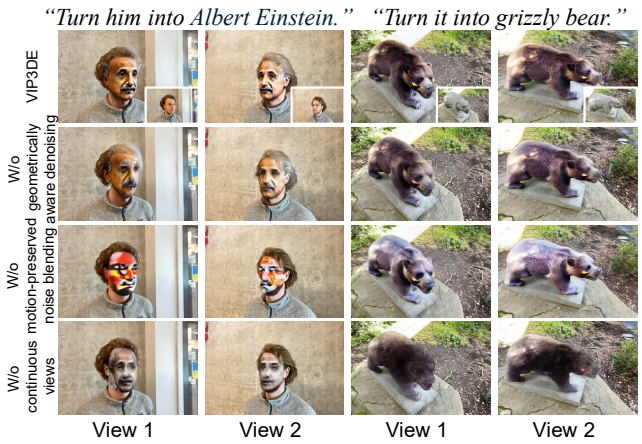


Figure 6: **Qualitative ablation results.** All the three components are important to achieve 3D consistent editing.

Conclusion

In this paper, we proposed ViP3DE, which introduced generative video priors for fast and consistent 3D editing. We first rendered continuous views to bridge the gap between discrete 3D views and video. Then, we proposed motion-preserved noise blending to improve editing quality, and introduced geometrically aware denoising to integrate 3D priors with video priors to enhance cross-view 3D consistency. The edited results could be directly used to update 3D without iterative passes. Extensive experimental results demonstrated the superiority of ViP3DE over previous methods.

Limitation. While ViP3DE can manage 3D editing with certain geometric changes, such as adding glasses to a person, it is limited in editing scenes with significant geometric alterations, such as raising a person's hands. This limitation mainly inherits from InstructPix2Pix, which is used to generate the edited first frame. As the editing and generative capabilities of image and video models continue to evolve, 3D editing with substantial geometric changes will be solved.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5470–5479.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; et al. 2024. Video generation models as world simulators. *OpenAI Blog*, 1(8): 1.

Chen, J.-K.; Bulò, S. R.; Müller, N.; Porzi, L.; Kontschieder, P.; and Wang, Y.-X. 2024a. ConsistDreamer: 3D-Consistent 2D Diffusion for High-Fidelity Scene Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21071–21080.

Chen, M.; Laina, I.; and Vedaldi, A. 2024. Dge: Direct gaussian 3d editing by consistent multi-view editing. In *European Conference on Computer Vision*, 74–92. Springer.

Chen, M.; Xie, J.; Laina, I.; and Vedaldi, A. 2024b. SHAP-EDITOR: Instruction-guided Latent 3D Editing in Seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26456–26466.

Chen, Y.; Chen, Z.; Zhang, C.; Wang, F.; Yang, X.; Wang, Y.; Cai, Z.; Yang, L.; Liu, H.; and Lin, G. 2024c. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21476–21485.

Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, 424–432. Springer.

Dam, E. B.; Koch, M.; and Lillholm, M. 1998. *Quaternions, interpolation and animation*, volume 2. Citeseer.

Dong, J.; and Wang, Y.-X. 2024. Vica-nerf: View-consistency-aware 3d editing of neural radiance fields. *Advances in Neural Information Processing Systems*, 36.

Fan, X.; Bhattad, A.; and Krishna, R. 2024. Videoshop: Localized Semantic Video Editing with Noise-Extrapolated Diffusion Inversion. arXiv:2403.14617.

Geyer, M.; Bar-Tal, O.; Bagon, S.; and Dekel, T. 2023. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*.

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

He, H.; Xu, Y.; Guo, Y.; Wetzstein, G.; Dai, B.; Li, H.; and Yang, C. 2024. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*.

Hertz, A.; Aberman, K.; and Cohen-Or, D. 2023. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2328–2337.

Huang, Z.; Wen, H.; Dong, J.; Wang, Y.; Li, Y.; Chen, X.; Cao, Y.-P.; Liang, D.; Qiao, Y.; Dai, B.; et al. 2024. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9784–9794.

Kamata, H.; Sakuma, Y.; Hayakawa, A.; Ishii, M.; and Narihira, T. 2023. Instruct 3d-to-3d: Text instruction guided 3d-to-3d conversion. *arXiv preprint arXiv:2303.15780*.

Kang, B.; Yue, Y.; Lu, R.; Lin, Z.; Zhao, Y.; Wang, K.; Huang, G.; and Feng, J. 2024. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*.

Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Ku, M.; Wei, C.; Ren, W.; Yang, H.; and Chen, W. 2024. Anyv2v: A plug-and-play framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*.

Li, R.; Chen, L.; Zhang, Z.; Jampani, V.; Patel, V. M.; and Zhang, L. 2024a. SyncNoise: Geometrically Consistent Noise Prediction for Text-based 3D Scene Editing. *arXiv preprint arXiv:2406.17396*.

Li, R.; Li, R.; Guo, S.; and Zhang, L. 2024b. Source prompt disentangled inversion for boosting image editability with diffusion models. In *European Conference on Computer Vision*, 404–421. Springer.

Li, Y.; Dou, Y.; Shi, Y.; Lei, Y.; Chen, X.; Zhang, Y.; Zhou, P.; and Ni, B. 2024c. Focaldreamer: Text-driven 3d editing via focal-fusion assembly. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3279–3287.

Ling, P.; Bu, J.; Zhang, P.; Dong, X.; Zang, Y.; Wu, T.; Chen, H.; Wang, J.; and Jin, Y. 2024. MotionClone: Training-Free Motion Cloning for Controllable Video Generation. *arXiv preprint arXiv:2406.05338*.

Liu, X.; Xue, H.; Luo, K.; Tan, P.; and Yi, L. 2024. GenN2N: Generative NeRF2NeRF Translation. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5105–5114.

Luo, C.; Di, D.; Yang, X.; Ma, Y.; Xue, Z.; Wei, C.; and Liu, Y. 2024. Trame: Trajectory-anchored multi-view editing for text-guided 3d gaussian splatting manipulation. *arXiv preprint arXiv:2407.02034*.

Mildenhall, B.; Srinivasan, P. P.; Ortiz-Cayon, R.; Kalantari, N. K.; Ramamoorthi, R.; Ng, R.; and Kar, A. 2019. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (ToG)*, 38(4): 1–14.

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

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6038–6047.

Ouyang, W.; Dong, Y.; Yang, L.; Si, J.; and Pan, X. 2024. I2VEdit: First-Frame-Guided Video Editing via Image-to-Video Diffusion Models. *arXiv preprint arXiv:2405.16537*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Schönberger, J. L.; and Frahm, J.-M. 2016. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shoemake, K. 1985. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, 245–254.

Song, L.; Cao, L.; Gu, J.; Jiang, Y.; Yuan, J.; and Tang, H. 2023. Efficient-NeRF2NeRF: Streamlining Text-Driven 3D Editing with Multiview Correspondence-Enhanced Diffusion Models. *arXiv preprint arXiv:2312.08563*.

Sorkine, O.; Cohen-Or, D.; Lipman, Y.; Alexa, M.; Rössl, C.; and Seidel, H.-P. 2004. Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, 175–184.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; Zeng, J.; Wang, J.; Zhang, J.; Zhou, J.; Wang, J.; Chen, J.; Zhu, K.; Zhao, K.; Yan, K.; Huang, L.; Feng, M.; Zhang, N.; Li, P.; Wu, P.; Chu, R.; Feng, R.; Zhang, S.; Sun, S.; Fang, T.; Wang, T.; Gui, T.; Weng, T.; Shen, T.; Lin, W.; Wang, W.; Wang, W.; Zhou, W.; Wang, W.; Shen, W.; Yu, W.; Shi, X.; Huang, X.; Xu, X.; Kou, Y.; Lv, Y.; Li, Y.; Liu, Y.; Wang, Y.; Zhang, Y.; Huang, Y.; Li, Y.; Wu, Y.; Liu, Y.; Pan, Y.; Zheng, Y.; Hong, Y.; Shi, Y.; Feng, Y.; Jiang, Z.; Han, Z.; Wu, Z.-F.; and Liu, Z. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*.

Wang, C.; Chai, M.; He, M.; Chen, D.; and Liao, J. 2022. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3835–3844.

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

Wang, Y.; Yi, X.; Wu, Z.; Zhao, N.; Chen, L.; and Zhang, H. 2024. View-consistent 3d editing with gaussian splatting. In *European conference on computer vision*, 404–420. Springer.

Wu, J.; Bian, J.-W.; Li, X.; Wang, G.; Reid, I.; Torr, P.; and Prisacariu, V. A. 2024. GaussCtrl: multi-view consistent text-driven 3D Gaussian splatting editing. *arXiv preprint arXiv:2403.08733*.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. 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*, 7623–7633.

Wu, Y.; Chen, L.; Li, R.; Wang, S.; Xie, C.; and Zhang, L. 2025. InsViE-1M: Effective Instruction-based Video Editing with Elaborate Dataset Construction. *arXiv preprint arXiv:2503.20287*.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072*.

Yu, Y.; Zhou, K.; Xu, D.; Shi, X.; Bao, H.; Guo, B.; and Shum, H.-Y. 2004. Mesh editing with poisson-based gradient field manipulation. In *ACM SIGGRAPH 2004 Papers*, 644–651.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhu, H.; Xu, Y.; Xu, C.; Shen, T.; Liu, W.; Du, Y.; Yu, J.; and He, S. 2025. Stable Score Distillation. *arXiv preprint arXiv:2507.09168*.

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

Zwicker, M.; Pauly, M.; Knoll, O.; and Gross, M. 2002. Pointshop 3D: An interactive system for point-based surface editing. *ACM Transactions on Graphics (TOG)*, 21(3): 322–329.