VDFE: Difference-Aware 3D Scene Editing with Non-Intrusive Video Diffusion Priors for Multi-View Consistency and Efficiency

Supplementary Material

1. Appendix A

1.1. Hyperparameters

To ensure the fairness and reproducibility of our results, we implement a unified hyperparameter configuration across all experiments, thereby eliminating confounding variables and enabling direct performance comparison. The total number of iterations is set between 2000 and 3000, a range that aligns with established baseline methods. Adhering to the best practices outlined in IN2N, we train the 3DGS model using loss function that combines LPIPS and L1 loss. Furthermore, to introduce controlled stochasticity and enhance model robustness without destabilizing the underlying geometry, we apply a subtle yet critical perturbation ratio of 0.33% during the Parameter Perturbation stage. The flow-difference maps, a key component for aligning edits across frames, is derived from the weighted summation of the first 12 steps of the flow model's forward process.

1.2. Evaluation

While the evaluation of 3D and video editing is inherently subjective, we use a comprehensive set of quantitative metrics to assess the quality and controllability of the editing outcomes, thus providing a rigorous validation of our method's effectiveness.

For 3D editing tasks, we employ an evaluation strategy that combines quantitative semantic metrics with qualitative human perception. We utilize CLIP text-image directional similarity (CLIP-dir) to verify that the edit's semantic direction aligns with the instruction, and assess the alignment of the final result using CLIP text-image similarity (CLIP-sim). Furthermore, to guarantee the geometric integrity of the rendered views, we introduce A-LPIPS and Met3R for multi-view consistency evaluation. Specifically, A-LPIPS is utilized to measure multi-view perceptual consistency by assessing texture alignment across different viewpoints, while Met3R is employed to evaluate 3D geometric consistency by quantifying feature coherence in 3D space, ensuring the edited structure maintains robust spatial coherence. To capture aspects of realism and aesthetic quality that automated metrics overlook, we incorporate User Study as the gold standard, ensuring the result is not only technically correct but also subjectively compelling. The User Study is conducted with 35 participants to vote.

For video editing, as the inherent complexity of the temporal dimension requires a more specialized protocol, we adopt the comprehensive FIVE-Benchmark. This protocol systematically evaluates multiple aspects: Structural Distance and Background Preservation Capability (quantified by PSNR, LPIPS, SSE, and SSIM outside the edited mask) ensure that edits are localized and do not damage the video's overall quality. Consistency of Edited Content is verified using CLIP-sim to guarantee that the edited region remains semantically aligned with the target description across all frames. Image Quality is measured by the no-reference metric NIQE to ensure each frame is visually clear and natural, while Temporal Consistency, assessed by the Motion Fidelity Score, is important for ensuring smooth, stable, and flicker-free motion. Finally, to assess the model's high-level understanding and control, we utilize FIVE-acc, which measures the model's precise adherence to fine-grained, complex instructions, distinguishing between general compliance and accurate, localized execution.

1.3. Others

Score Distillation: We omit Score Distillation Sampling to avoid prohibitive computational costs. Current VDMs employ VAEs for end-to-end video modeling. For 3D scenes exceeding 80 frames, gradient backpropagation requires simultaneous VAE processing across all frames, leading to a surge in GPU memory usage. Although splitting long sequences into shorter clips can alleviate memory pressure, this approach would introduce inter-clip coherence issues. Furthermore, excessive resource consumption makes it difficult to distinguish whether performance improvements stem from methods or additional resources.

Trajectories: Camera trajectory adjustment employs an interpolation algorithm based on camera pose distances. In case of occlusion, three adjacent poses are discarded to bypass the region, followed by re-interpolation.

Dependency and Robustness: We aim to harness the potential of VDMs for editing tasks. Experiments indicate that our proposed module significantly reduces background drift and yields smoother results, thereby minimizing inconsistencies and artifacts compared to image diffusion-based methods and VDM-only methods. Moreover, DAGE guides 3DGS optimization via localization, reducing sensitivity to a single frame and avoiding over-updating irrelevant regions. Frame-level errors (e.g., inconsistencies or artifacts) minimally affect the results.

Additional Results: To better evaluate the effectiveness of our method, we present additional visualization results of editing effects, as shown in Figures 1, 2 and 3.

Algorithm 1 Implementation of FlowOCE and DFD

Input: Pseudo video P_v, Pre-trained flow model V^θ, VAE encoder/decoder $\mathcal{E}, \mathcal{D}$, Source/Target text c_{src}, c_{tgt}, Custom hyperparameter λ, Flow difference maps fusion steps M

$X_1^{src} \leftarrow \mathcal{E}(P_v), \quad \hat{X}_1^{edit} \leftarrow \mathbf{X}_1^{src}$

for $t : 1 \rightarrow 0$ **do**

  $X_t^{src} \leftarrow (1-t) \cdot X_1^{src} + t \cdot \epsilon, \epsilon \sim \mathcal{N}(0, I)$

  $X_t^{tar} \leftarrow \hat{X}_t^{edit} - X_1^{src} + X_t^{src}$

  $v_q^{src} \leftarrow V^\theta(X_t^{src}, c_{src}), \quad v_p^{src} \leftarrow V^\theta(X_t^{tar}, c_{src}), \quad v_p^{tar} \leftarrow V^\theta(X_t^{tar}, c_{tar})$

  $D_t \leftarrow ||v_p^{tar} - v_p^{src}||_2^2$

  $D_{maps}(i,j) \leftarrow \sum_{t=1}^{M} w_t \cdot \frac{\exp\Big(D_t(i,j)\Big)}{\sum_{k=1}^{H}\sum_{l=1}^{W}\exp\Big(D_t(k,l)\Big)}$

  $M \leftarrow Thresholding(D_{maps})$

  $\mathbb{E}[X_0|X_t] \leftarrow X_t - t \cdot V^\theta(X_t, c), \quad \mathbb{E}[X_1|X_t] \leftarrow X_t - (1-t) \cdot V^\theta(X_t, c)$

  $u_t \leftarrow \lambda\Big[\Big(\mathbb{E}[X_0^{\text{tar}} \mid X_t^{\text{tar}}] - \mathbb{E}[X_0^{\text{src}} \mid X_t^{\text{src}}]\Big) + \Big(\mathbb{E}[X_1^{\text{tar}} \mid X_t^{\text{tar}}] - \mathbb{E}[X_1^{\text{src}} \mid X_t^{\text{src}}]\Big)\Big]$

  $\hat{v}_t^{edit} \leftarrow v_p^{tar} - v_q^{src} + u_t$

  $\hat{X}_t^{edit} \leftarrow \hat{X}_t^{edit} + \hat{v}_t^{edit} \cdot \Delta t$

  $\hat{X}_t^{inject} \leftarrow M \cdot \hat{X}_t^{edit} + (1-M) \cdot X_1^{edit}$

end for

$\mathcal{D}(\hat{X}_t^{inject}), \quad D_{maps}$

Turn this man into Albert Einstein

Transform this man into an anime style character

Turn him into a wooden puppet with joints

Turn him into a women

Figure 1. **Additional extensive results 1.** We present extensive qualitative results to highlight the robustness and versatility of our proposed method. Our VDFE ensures multi-view consistency and provides flexible editing, while excelling particularly in precise local editing.

Corgi Scene

Fangzhou Scene

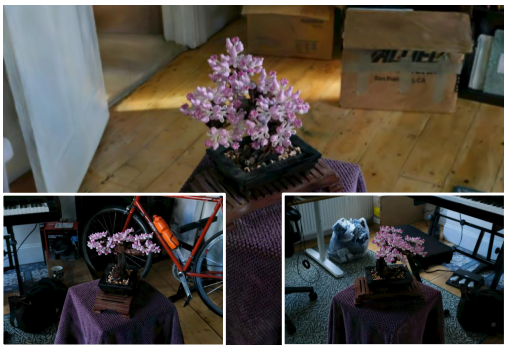
Turn the corgi into a giant panda

Turn this man into an Elven King with long pointed-ears

Turn him into a Joker

Figure 2. **Additional extensive results 2.** We present extensive qualitative results to highlight the robustness and versatility of our proposed method. Our VDFE ensures multi-view consistency and provides flexible editing, while excelling particularly in precise local editing.

Garden Scene



Bonsai Scene

Add a puddle on the table

Replace soft pink blossoms with yellow blossoms

Replace soft pink blossoms with blue-purple Chrysanthemum

Figure 3. **Additional extensive results 3.** We present extensive qualitative results to highlight the robustness and versatility of our proposed method. Our VDFE ensures multi-view consistency and provides flexible editing, while excelling particularly in precise local editing.