
Pro3D-Editor: A Progressive-Views Perspective for Consistent and Precise 3D Editing (Supplementary Material)

Anonymous Author(s)

Affiliation

Address

email

1 In this appendix, we provide detailed implementation details and additional visualizations of compar-
2 ative experiments with existing methods in Appendix A. For an intuitive comparison, please refer
3 to the local anonymous HTML file in our supplementary material. Then, to better demonstrate the
4 effectiveness of our proposed paradigm and the essential components in improving editing consistency
5 and quality, we provide more visualizations of ablation experiments and user studies in Appendix B.
6 Finally, we analyze the limitations and broader impacts of our work in Appendix C.

7 A Implementation Details and Comparative Experiments

8 A.1 Implementation Details

9 We use the MV-Adapter SDXL checkpoint as our multi-view diffusion model. In our pipeline, we
10 fine-tune the multi-view attention layers within the MV-Adapter network. For different views, we
11 set distinct B matrices and identical A matrices, with the lora_rank set to 32 and lora_alpha set
12 to 16. During training, the parameters of the A matrix are updated only by the gradients from the
13 primary view. We fine-tune the model for 800 steps, which takes 45 minutes on an A100 GPU.
14 During inference, we set the classifier-free guidance to 2. For 3D editing and refining, we first use a
15 leave-one-out strategy to train the original 3DGS object for 10k steps, resulting in a degraded 3DGS.
16 We then render the degraded views corresponding to the target perspectives and use them as the
17 condition for ControlNet-Tile. Using the generated multi-views as the target, we add LoRA with a
18 rank of 64 to all attention layers of the controlnet and fine-tune for 1800 steps. Finally, we use the
19 fine-tuned ControlNet-Tile to repair the rendered images of new perspectives and train the degraded
20 3DGS for an additional 10k steps. The entire 3D editing and refining process takes about 45 minutes.

21 A.2 Explanation of Quantitative Evaluation Metrics

22 In terms of evaluating editing quality, FID assesses the overall visual similarity between the edited
23 result and the original object. LPIPS measures perceptual similarity, while PSNR reflects changes in
24 detail. FVD evaluates the temporal continuity and stability across multi-views. The 3D plausibility
25 score and texture details score proposed by GPTEval3D specifically measure the structural rationality
26 and texture detail of 3D editing results. In terms of edit controllability, the text-asset alignment score
27 from GPTEval3D and CLIP-T measure the similarity between the editing results and the editing text.
28 DINO-I measures the similarity between the editing results and the original object. Since our task
29 focuses on localized 3D editing, DINO-I can reflect the accuracy of the edits to some extent. Overall,
30 these metrics provide a comprehensive quantitative evaluation of both the editing quality and editing
31 accuracy from different perspectives, collectively reflecting the overall performance of the 3D editing
32 method. However, when it comes to view consistency in the editing results, these metrics fall short
33 of accurately reflecting it. Therefore, we provide additional visualizations to fully demonstrate the
34 improvements of our method compared with existing baselines.

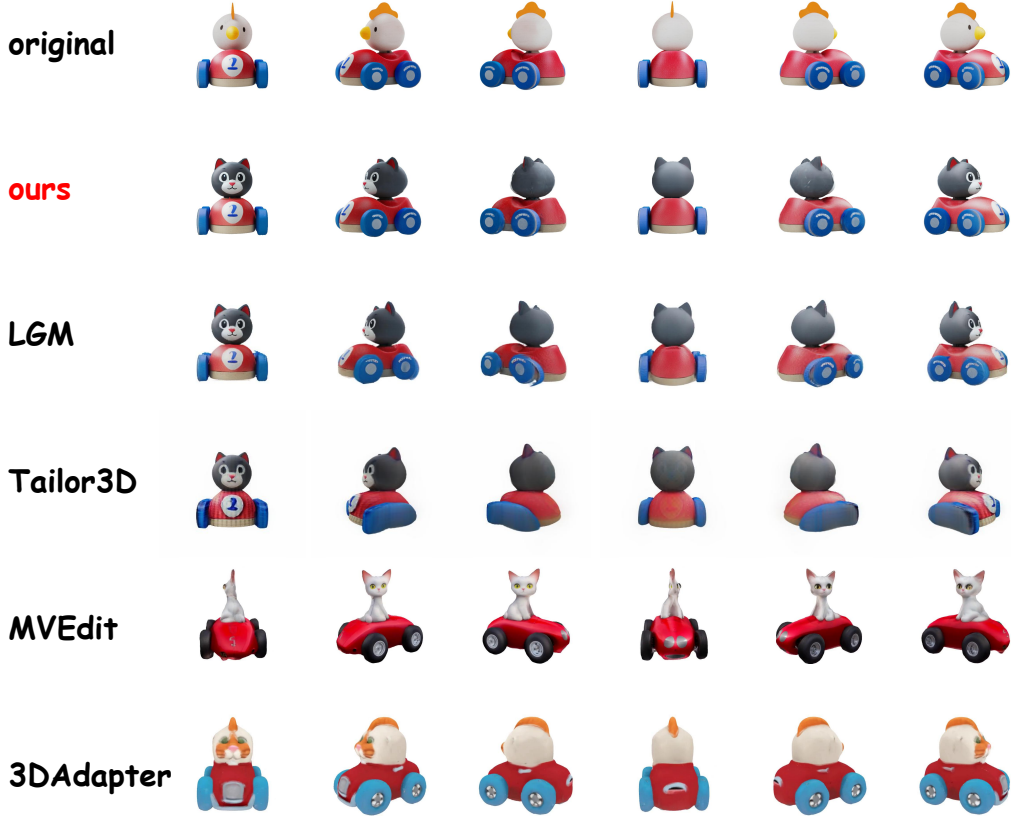


Figure 1: **Qualitative comparison** with existing methods. It can be observed that our method achieves precise and high-quality local 3D editing while addressing the issue of spatial inconsistency.

35 A.3 Comparison with Existing Methods

36 In Fig. 1, we show a detailed editing example. Existing methods often edit the entire object and fail
 37 to preserve local regions that are semantically irrelevant to the editing text. Even though LGM and
 38 Tailor3D use multi-views generated from our method, they still modify semantically irrelevant regions.
 39 Moreover, existing methods such as MVEdit often generate spatially inconsistent 3D objects. In
 40 contrast, our method achieves consistent, precise, and high-quality text-guided 3D editing. For more
 41 comparison results, please refer to the HTML file provided in our supplementary materials, which
 42 contains multiple orbiting videos that demonstrate the improvements of our method in text-guided
 43 3D editing.

44 B More Ablation Experiments and User Studies

45 B.1 Ablation of Each Component

46 **Effectiveness of Primary-view Sampler.** In Fig. 2 and Fig. 3, we highlight the importance of
 47 the Primary-view Sampler. When the primary view is randomly selected and editing semantics are
 48 propagated from an editing-sparse view to editing-salient views, it results in inter-view inconsistency
 49 (*i.e.*, lack of spatial coherence across views) and intra-view indiscrimination (*i.e.*, poor control over
 50 editing-salient views). These issues are clearly illustrated by the inconsistent beard appearance across
 51 views in Fig. 2 and the unreasonable editing of the cat’s head in certain views in Fig. 3. It underscores
 52 the necessity of our progressive-views paradigm, which directs semantic flow from editing-salient
 53 to editing-sparse views. The precise and consistent 3D editing achieved by our method stems not
 54 merely from fine-tuning, but from this carefully designed paradigm.

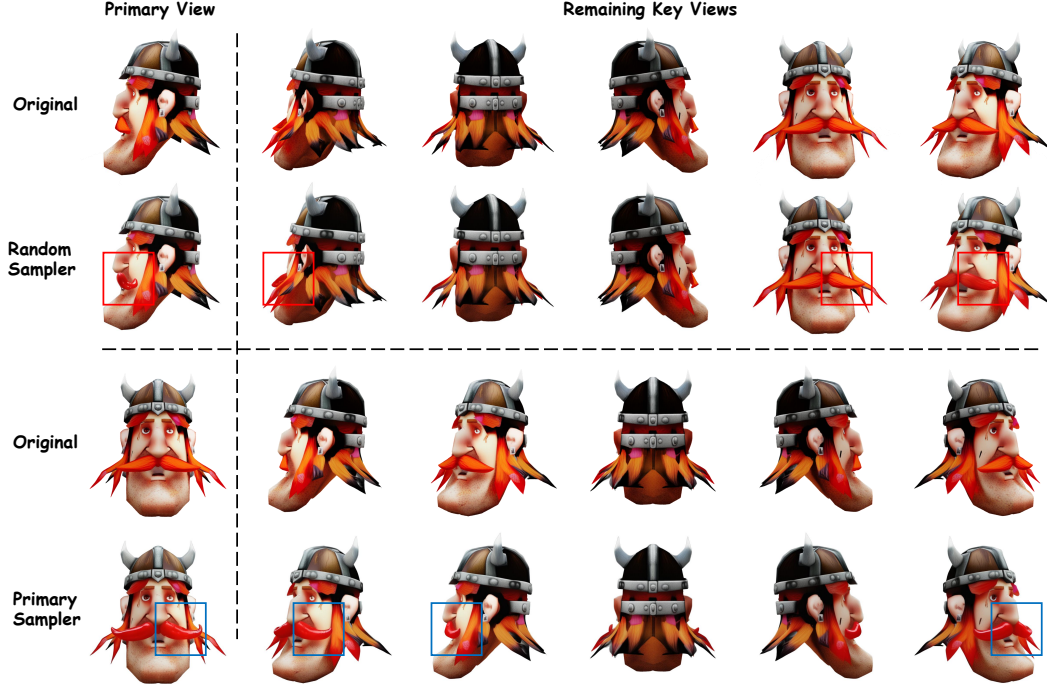


Figure 2: **Ablation studies** of Primary-view Sampler. When the randomly selected view is not the most editing-salient view, the editing information from this editing-sparse view may fail to propagate effectively to the editing-salient views, leading to spatially inconsistent across multi-views.

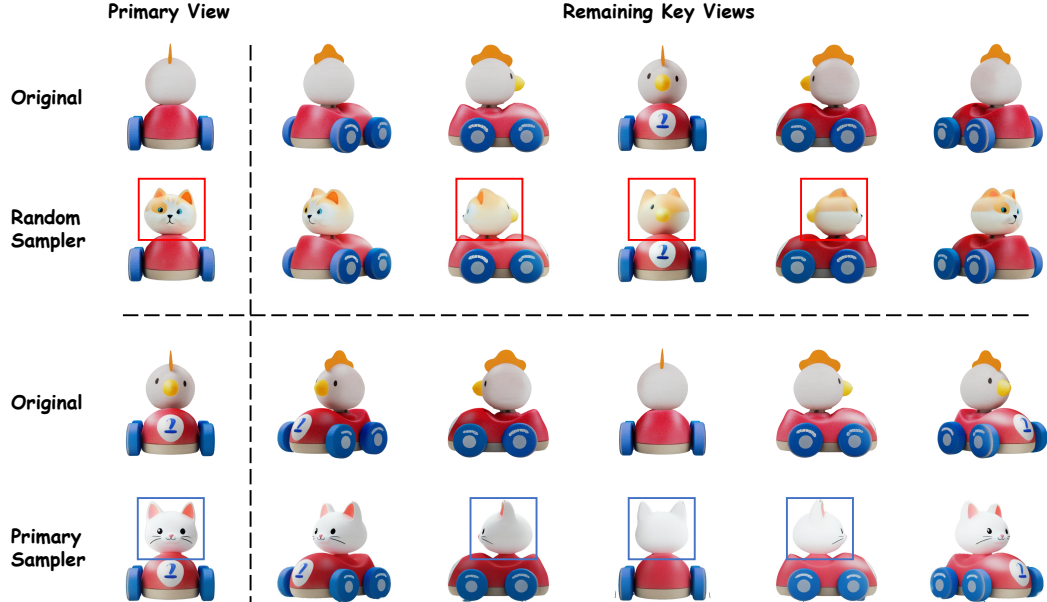


Figure 3: **Ablation studies** of Primary-view Sampler. Since editing-salient views are difficult to be precisely controlled by editing-sparse views, when the randomly selected view is not the most editing-salient view, the other editing-salient views may produce unreasonable editing results.

55 **Effectiveness of MoVE-LoRA.** In Fig. 4, we present qualitative results to demonstrate the effective-
 56 ness of MoVE-LoRA in enhancing multi-view editing consistency, as it is difficult to accurately
 57 evaluate such consistency using existing quantitative metrics. Here, "Shared LoRA" refers to a setting
 58 where the same LoRA matrices A and B are applied to the latent of multi-views. As shown in the
 59 figure, Shared LoRA fails to accurately preserve the original object features (e.g., incorrect object
 60 colors) and leads to spatially inconsistent edits (e.g., misaligned ears across views). In contrast,
 61 our MoVE-LoRA not only better preserves the original object features but also ensures spatially
 62 consistent editing across multi-views.

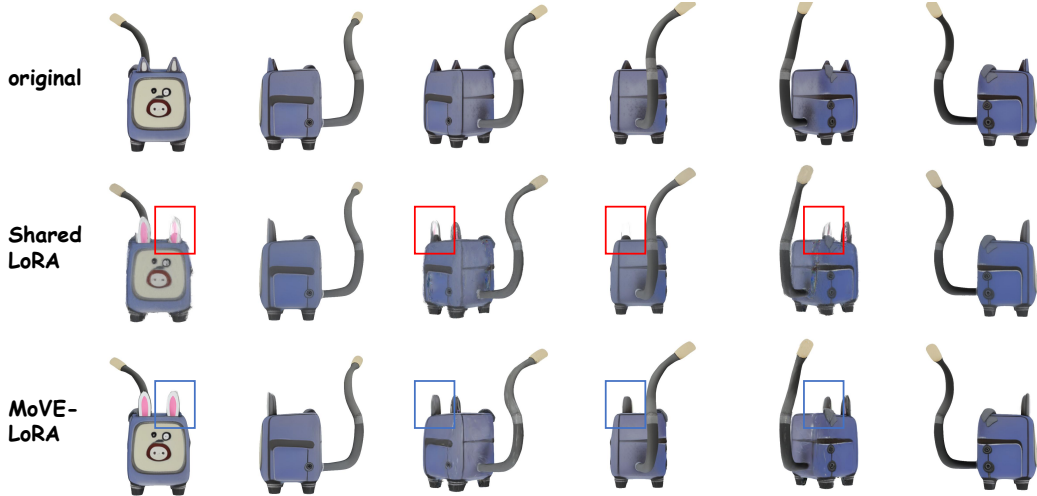


Figure 4: **Ablation studies** of MoVE-LoRA. Compared with Shared LoRA, our MoVE-LoRA not only better preserves the features of the original multi-views, but also ensures spatial consistency of the editing regions, achieving precise and consistent multi-view editing.



Figure 5: **Ablation studies** of Full-view Refiner. Introducing the Full-view Refiner can improve the quality of the final 3D editing results by eliminating some floating discrete Gaussians, addressing fragmentation issues, and ensuring the structural continuity of the edited 3D object.

Effectiveness of Full-view Refiner. As shown in Fig. 5, we compare the 3D editing results with and without Full-view Refiner. Without Full-view Refiner, the edited 3DGS object may become fragmented. For example, in the case of a doll’s mask, the absence of the Full-view Refiner can lead to the generation of numerous floating discrete Gaussians. This is because sparse-view guidance of 3DGS updates prioritizes consistency with the given multi-views at specific angles, potentially neglecting the overall 3D structural continuity. In contrast, introducing Full-view Refiner provides extra 3D structural information, ensuring the surface continuity of the final edited 3DGS object.

B.2 Human Perception Evaluation

We recruit 8 volunteers to evaluate *Pro3D-Editor* under different settings from three aspects: Editing Consistency (EC), Editing Accuracy (EA), and Editing Quality (EQ). The volunteers were asked to rank the editing results under different settings from first to fourth place. Each volunteer is given two different edited objects to assess. As shown in Tab. 1, it can be observed that with the addition of each essential module, the final editing results align more closely with human preferences. Notably, the results in the table represent the average rankings given by all volunteers.

Table 1: **Human perception evaluation** for different settings. The inclusion of each module achieves more effective editing of 3D results that align with human preferences.

ID	Settings	EC (1 ~ 4) ↑	EA (1 ~ 4) ↑	EQ (1 ~ 4) ↑
0	naive method	1.75	1.625	1.625
1	0 + Primary-view Sampler	1.875	1.875	2.125
2	1 + MoVE-LoRA	2.5	2.625	2.4375
3	2 + Full-view Refiner	3.875	3.875	3.8125

77 C Limitations and Broader Impacts

78 C.1 Limitations

79 The *Pro3D-Editor* is computationally demanding and requires substantial GPU memory, primarily
80 due to the fine-tuning process on a high-resolution multi-view generation model. Compared to
81 existing training-free methods, our approach necessitates more computational resources for model
82 training. However, it achieves more precise and consistent 3D editing. The *Pro3D-Editor* framework
83 also differs from existing methods in 2D-guided 3D editing. Existing methods typically generate
84 a new 3D object directly from 2D multi-views without considering the structural features of the
85 original 3D object. In contrast, our method employs the concept of sparse 3DGS reconstruction for
86 3D editing, which is more time-consuming than existing methods in obtaining a refined 3D structure.

87 C.2 Broader Impacts

88 **Positive Societal Impacts.** *Pro3D-Editor* brings several contributions to the field of text-guided
89 3D editing. By enabling semantically accurate and spatially consistent edits across multi-views,
90 it addresses key limitations of existing training-free approaches, which often suffer from view
91 inconsistency and structural degradation. It has the potential to lower the barrier for creating high-
92 quality 3D content, making it easier for designers, artists, and even non-experts to customize 3D assets
93 using intuitive language prompts. This increased accessibility could help foster broader participation
94 in 3D content creation and may contribute to progress in areas such as digital art, gaming, and virtual
95 reality, where interactive and editable 3D representations are becoming increasingly important.

96 **Negative Societal Impacts.** Despite its advantages, the use of AI-driven 3D editing tools may
97 also raise concerns about potential misuse. As the modification of 3D assets becomes easier and
98 more automated, issues related to ownership, copyright infringement, and unauthorized replication
99 of proprietary 3D models may arise. The ability to edit and redistribute high-quality 3D content
100 with minimal expertise could blur the lines between original and derivative works, making it more
101 challenging to protect the intellectual property rights of creators. Currently, the protection of original
102 creators often relies on ethical norms rather than enforceable legal mechanisms, which may be
103 insufficient to deter misuse in practice.