GeM💎NR: Geometry-Aware Multi-View Editing for Nonrigid Scene Changes

Josef Bengtson* Yaroslava Lochman* Fredrik Kahl

Chalmers University of Technology

https://gem-nr.github.io

Abstract. Recent developments in multi-view image editing with generative models have brought us a step closer toward general 3D content generation and customization. Most existing works focus on *rigid* or *appearance-only* edits by utilizing the geometry of the unedited scene. This naturally limits these methods to edits that preserve the underlying scene structure. Other approaches are trained for specific image editing tasks, such as object removal and addition. Despite this progress, general nonrigid edits, *i.e.*, edits that substantially change the scene geometry, remain challenging for existing methods. We propose GeM-NR, a fast and flexible training-free approach for *general* multi-view consistent image editing, including edits that drastically change the geometry and appearance of the scene. Given an anchor image edited with a chosen backbone editor (such as FLUX, Qwen, BrushNet) and a query unedited image, GeM-NR edits the query image consistently with the anchor edit. The method incorporates multiple stages: (i) depth map estimation, where we propose a strategy to maximize the alignment between the 3D point clouds of the edited and unedited scenes, (ii) projection onto a query viewpoint, and (iii) refinement of the obtained image conditioned on the unedited query. The conditioning-based formulation scales well from two to many views of an object. We demonstrate the ability of our method to handle edits with significant changes in geometry and appearance, something that existing methods struggle with. We perform an extensive evaluation showing that our method improves consistency for a wide variety of edit tasks, including generating 3D representations of the edited scene. Both quantitative and qualitative results indicate the state-of-the-art performance of our method in terms of edit quality as well as geometric and photometric consistency across multiple views.

1 Introduction

Consistent multi-view image editing is a core capability for 3D editing, with a large number of possible applications. A major challenge is to handle nonrigid edits, which drastically change the geometry of the scene. One reason for difficulties with these types of edits is that the geometry of the original unedited

* Equal contribution

Fig. 1: The edited images produced by GeM-NR. We handle various types of edits including significant changes in scene geometry. The edits are both photometrically and geometrically consistent across the views.

scene no longer holds for the edited scene, and another is that these more drastic edits often lead to significant inconsistencies across different views, making the task of enforcing consistency more difficult. One approach for solving this is to train a method using paired multi-view images with these types of edits [37], but limitations in available paired data lead to these methods being limited in scope to specific types of edits, such as object addition and removal.

One solution to the issue with handling significant inconsistencies in the edits is to warp an edited image into a target view and use that as conditioning when editing an image at that view, providing the edit of that with clear guidance on how that view should look to be consistent with the previous edit. Existing works [16, 31] use depth maps from a 3D representation acquired from the original scene, limiting the ability to handle nonrigid edits. In contrast we propose estimating the scene geometry from an edited view of the scene with a depth estimator [33], since these depth will then be valid also for nonrigid edits. We formulate this problem as the task of given an initial edited image perform consistent editing of a number of other images of the same scene. This problem could be solved by performing single image novel view synthesis from the initial edited image by rendering new views at the poses of the given views. An approach to solving this is to estimate geometry from the initial image and use this to warp the initial image into the target views. The problem with doing this in the setting of multi-view editing is that this does not take into account the information in the additional views of the scene. We instead propose a way to include these warping in the editing process, so that both the unedited views and the warping of the initial edit are taken into account.

Recent developments in editing methods allow for flexible and multi-reference editing [6,7,60], which could be used for multi-view editing. It is possible to condition these methods on several images and ask to perform multi-view consistent editing, but this is unstable and often leads to failed edits. Our contribution is that by conditioning these multi-reference methods with both the unedited image and the warping of an existing edit, it is possible to reliably perform multi-view consistent editing.

While we rely on the foundation 3D models for reconstruction and multi-reference image generation models for editing, we discover novel ways of leveraging their power. In particular, our first finding is that the unedited and edited scenes together can be cast as one dynamic scene, hence can be handled by a dynamic scene reconstruction model such as Depth Anything 3. Our second finding is that the unedited image together with its partially filled edited version are a better input for the multi-reference image generation model such as FLUX.2 than a fully edited image but from another viewpoint. It has been shown that the image generation models have limited ability in performing geometric tasks. However, if provided with the right geometry, they succeed at preserving it.

In summary, GeM-NR is able to perform multi-view consistent editing with respect to the provided anchor image edit done by a preferred backbone editor. GeM-NR is flexible and can handle widely varied and drastic edits, including significant changes to geometry as seen in Fig. 1. It does not require any per-scene optimization and the runtime is only limited by the time it takes to perform depth estimation and multi-reference editing, which for the methods we choose can be done in $\sim$ 3s per image. Our contributions can be summarized as follows:

- We propose a flexible pipeline for multi-view image editing based on reconstructing the geometry of the edited scene.
- We discover and combine novel ways of leveraging the power of foundation models, in particular dynamic reconstruction models and multi-reference image editing models, to maximize the editing quality *together* with multi-view consistency.
- We conduct extensive evaluation showing that our method can handle edits significantly changing geometry and appearance. We propose a more detailed evaluation pipeline integrating epipolar geometry evaluation, showing improved consistency across a variety of different editing tasks and image editing methods.
- Our pipeline is very fast allowing us to generate edited 3D Gaussians in less than a minute for sparse scenes.

2 Related Work

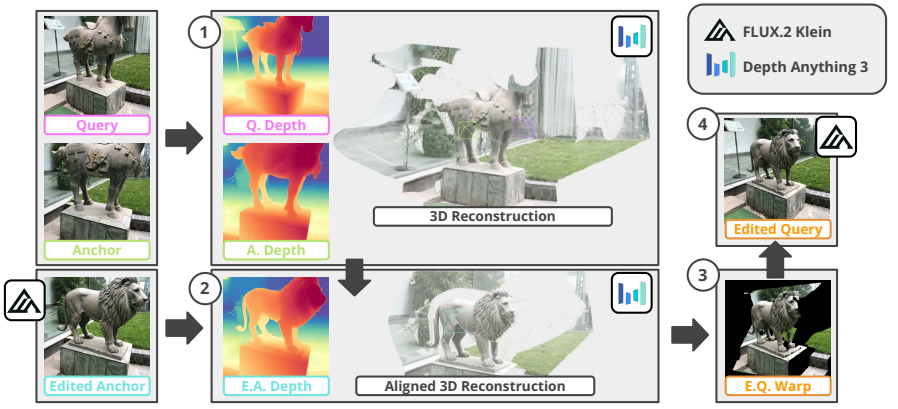
Image editing using generative models. The development of diffusion models trained on vast amounts of data have led to significant advancements in both image generation [25, 45, 49, 51] and editing [8, 23, 39, 50]. This has enabled different types of editing tasks, including instruction-based editing [8, 20, 70], image inpainting [27, 38, 75], style transfer [10, 14, 24, 72] and multi-reference

editing [6, 7, 60, 63]. Recent developments in flow matching [34, 36] and multimodal transformer architectures [19] have led to high quality unified editing models [6, 7, 29, 60] that can apply significant changes and perform very precise edits. Our work focuses on how to use these powerful 2D image editing methods to perform consistent multi-view editing.

Multi-view consistent image editing. A main line of research in multi-view consistent editing focuses on utilizing 2D image editing models to perform editing of an existing 3D representation, such as NeRFs or 3D Gaussians. Instruct-NeRF2NeRF [22] presents the idea of iterative dataset update (IDU) that utilizes the consistency from a 3D representation to achieve consistent edits by iteratively editing views and updating the 3D representation, which has been adopted in [13, 16, 40, 55, 59]. Other approaches use different properties of an existing 3D Gaussian Splatting (3DGS) model to achieve consistency, such as utilizing the geometry of the 3D representation of the unedited scene to guide the editing [28, 31, 61, 74], rendering smooth camera trajectories as in video editing methods [11,12], or using edited multi-views to update an existing 3DGS model, inheriting consistency from the underlying 3D representation [11,12,30,31,59,61]. These approaches require dense views to obtain a high-quality 3D representation of the unedited scene. One approach that does not require this is using correspondences from the unedited images to direct the editing such that corresponding points are edited in a consistent way [2, 4, 53]. This approach directly encourages multi-view consistency, but has limitations in the case of edits that significantly change the geometry of the scene, leading to the correspondences from the unedited images not holding after the edit. Another approach is to perform direct 3D editing, where a method directly edits a 3D representation or a set of multi-view images. A challenge with this approach is the need for paired 3D data, limiting most methods to training on single objects and synthetic scenes [5, 32, 62, 67]. Recent methods extend this approach to real images, either by proposing a paired multi-view data generation pipeline [37] for specific types of edits based on recent developments in 2D image editing and visual-language models (VLMs) [15, 64], or performing reinforcement learning-based fine-tuning and leveraging the 3D foundation model VGGT [56] as a 3D consistency reward [57]. In contrast our method is training-free, making it possible to take any initial edit and consistently edit another image to be consistent with the initial edit. We achieve this by estimating the geometry of the edited scene from the initial edit, warping the obtain 3D representation into the desired viewpoint and conditioning the query image editing process with the obtained warp, extending the capability of multi-view consistent image editing to nonrigid edits.

Warping based conditioning. Improving multi-view consistency by warping an input view into a target view has been studied in the past. Such an approach is used for single image novel view synthesis, where a given image is warped to a target pose and provided as conditioning when generating novel views for the target poses [18, 35, 42, 43, 52, 54, 68, 69]. Several works also use warping to improve consistency in 3D editing, leveraging depth from the unedited 3D

Fig. 2: Method overview. GeM-NR estimates the geometry of the edited scene globally aligned with the unedited scene, projects the edited scene point cloud onto the new view and conditions editing of the next image with the resulting warp.



representation to blend edits across views [16, 31], directly condition the editing process on the estimated depths [48, 61], or warp attention feature maps used during editing [21]. These approaches are limited to rigid edits that roughly preserve the scene geometry, since the depths are based on the geometry of the unedited scene. We propose to instead utilize the geometry of the edited scene which remains valid for nonrigid edits. One way of achieving this is to use recent models [33, 46, 47, 65] that can estimate geometry from a single view, in our case the initial edited image, that can then be used to warp the initial edited image into a target view. We take a step further and use a dynamic scene reconstruction approach [33] to address the alignment of the edited scene geometry with respect to the unedited scene geometry, a crucial part affecting the alignment of the warped edit with respect to the unedited query image. Our approach adapts this methodology for general, including nonrigid, multi-view editing, where we provide both the unedited query image and the warped edit to the multi-reference image editing model to preserve multi-view details and consistency even when the edit changes scene geometry.

3 Method

3.1 Problem formulation

Given N source views $\{I^1_{src}, I^2_{src}, ..., I^N_{src}\}$ and a text description T of the desired edit, the goal is to edit the images, obtaining $\{I^1_{edited}, I^2_{edited}, ..., I^N_{edited}\}$, such that the appearance and geometry are consistent across all the views, and $\{I^n_{edited}\}$ remain aligned with the edit description. One way of reformulating the problem is by first choosing an anchor image A_{src} that is to be edited independently: $A_{edited} = f(A_{src}, T)$, where f is a preferred image editing model. Then, editing each of the remaining images, which we call query images and denote as $\{Q^i_{src}\}$,

Fig. 3: Monocular depth estimation vs our joint depth estimation approach. For the regions where geometry was not changed after the edit (zoomed-in) the depths of the unedited and edited scene from the same viewpoint should coincide. Monocular depths do not achieve as accurate alignment with respect to the unedited scene as in our approach where the depth maps of all images are estimated simultaneously.

can be conditioned on A_{edited} to ensure consistency. In conclusion, we simplify the task to processing triplets consecutively — given $\{Q_{src}, A_{src}, A_{edited}\}$, the goal is to produce Q_{edited} such that it is consistent with A_{edited}. As for now, existing image editing models do not guarantee that independently edited images, that is, $Q_{edited} = f(Q_{src}, T)$, will retain such consistency.

3.2 Overview

Our editing pipeline should be able to handle significant changes in the scene. For appearance changes, since at least two unedited views are available, one could use the geometry of the unedited scene obtained with a 3D reconstruction pipeline. For challenging geometric changes, however, we cannot utilize the geometry acquired from the unedited images. Instead, we propose to use the geometry of the edited scene. Moreover, we need to understand how the geometry of the edited scene relates to that of the unedited scene. In other words, the two 3D representations need to be aligned. However, solving the alignment accurately and robustly is challenging, especially if the majority of the scene geometry is changed. We sidestep the alignment problem altogether and instead aim to jointly estimate the 3D representations of both edited and unedited scene as realizations of one dynamic scene, which automatically results in a global alignment. This becomes possible with the recent advances in dynamic scene reconstruction [33]. The 3D representation of the globally aligned edited scene can now be used to render an image at a query viewpoint. Naturally, at this viewpoint, some regions need to be filled-in as they were not seen from the anchor viewpoint. A source query image provides an additional information about what should be depicted in these areas. A multi-reference image generation model [6, 7] can therefore be used to fulfill the task of final edit generation given an unedited image and its partially edited version obtained by warping. Fig. 2 presents an overview of our approach. Since the overall pipeline only requires one full forward pass and utilizes efficient foundation models, it only takes a few seconds to edit a pair of images consistently.

Having a method that can edit two images consistently, we extend the approach to editing sets of images by repeating the process for all the remaining

query images. We can then pass the resulting set of consistently edited images through a feedforward method such as AnySplat [26] to obtain a 3DGS representation of the edited scene.

Edited scene geometry estimation We first pass A_{src} and Q_{src} through the reconstruction model Depth Anything 3 [33] to obtain camera intrinsics, K_A and K_Q, respectively, and extrinsics, P_A and P_Q^0. The pose P_Q^0 is then refined into P_Q with the classic relative pose estimation from two views using RoMa [17] matches. The refinement affects the quality of the subsequent joint reconstruction discussed below. Further, we show in Sec. 4.5 that it improves the quality of the final edits.

Next, the three images $\{A_{src}, A_{edited}, Q_{src}\}$ and the corresponding camera intrinsics and extrinsics $\{(K_A, P_A), (K_A, P_A), (K_Q, P_Q)\}$ are passed through Depth Anything 3, constraining the output cameras to be equal to the ones provided. Using the same camera parameters for A_{edited} as for A_{src} is based on the assumption that the edit does not change the global motion. For example, if the edit is: "Change the view to the other side of the corridor" or "Turn to the right and show what is there", it violates our assumption. However, if the edit is: "Turn [depicted object] around", it aligns well with our assumption as it keeps the global scene pose unchanged. See also an example with the edit of the similar form, namely "Have the wooden dinosaur stand next to the stump", in Fig. 1.

Depth Anything 3 is trained to tackle dynamic scene reconstruction, and we leverage this ability for challenging nonrigid edits. The model views $\{A_{src}$, A_{edited}, $Q_{src}\}$ as images of a single dynamic scene, where possible changes in scene geometry and photometry over time appear at A_{edited}.

Edit initialization with warping A partially-filled edited image is rendered at a query viewpoint from projecting the point cloud of the edited scene onto (K_Q, P_Q), which can also be seen as warping A_{edited}, giving Q_{warp}. The warped edit provides a dense guidance on how the edit should look like from the query viewpoint, to ensure consistency with edited anchor view.

In this work, we use depth map / point cloud representation for both estimation and rendering. In theory, any other representation, such as NeRFs or 3D Gaussians, could work as well, as long as such a representation can be faithfully obtained from only a few images—in our case, as few as three images—and the underlying estimation method supports dynamic scenes. Another option is to upgrade to a desired representation starting from a point cloud. A challenge with this approach is having access to very limited data, namely a single edited image with the corresponding depth map. At this point, we found that manipulating point clouds directly works best.

Why not using monocular depth? A slightly simpler approach could be to use monocular depths estimated from an edited anchor image A_{edited} in order to warp this image from camera (K_A, P_A) to camera (K_Q, P_Q). In Fig. 3, we show

Fig. 4: Qualitative comparison of query-anchor consistency with Qwen anchor backbone. Our method GeM-NR preserves photometric and geometric details of the anchor edit.

qualitative examples of obtaining a monocular depth map from an edited image as compared to our alignment approach. Both estimators are provided with intrinsics. Ideally, the depth values of A_{src} and A_{edited} should coincide in the regions where the geometry was not changed. In the demonstrated examples, we zoom-in to the rightmost regions of the images, where the areas (leaves to the left, walls to the right) remain unchanged. Monocular depths (even in the calibrated setting shown in Fig. 3) do not follow the multi-view depths of the unedited scene as closely as in our approach, where the depth maps of all images are estimated simultaneously.

Edit refinement with multi-reference editing To fill-in the gaps in Q_{warp}, one could use an image inpainting model. The problem with this approach is that it does not guarantee the consistency of the filled-in areas with those in Q_{src}. We need to be able to constraint the image generation process both by forcing the edit to closely follow Q_{warp} where possible, and also by letting the model infer how to fill the remaining areas based on the semantics of the corresponding

Fig. 5: Multi-view editing of 10 images with Qwen anchor backbone. GeM-NR successfully edits the full sequence while maintaining multi-view consistency.

regions in Q_{src} and the overall content in Q_{warp}. Hence the model should be able to accept multiple visual inputs and understand the relation between them from the text description, which we refer to as multi-reference image editing.

Recent image editing methods [6,7,60] allow for multi-reference inputs: a list of images $I_{list} = [I_1, I_2, ...]$ and a text prompt $\hat{T}$ specifying how the images should be combined to create a final image $I_{edited} = f_{multi}\left(I_{list}, \hat{T}\right)$. We propose to include the warping of the initial edited view into the target view as conditioning when editing. This provides the constraint directly in the image frame of the view that should be edited, making it easier for the model to create an image that respects the warping from the initial edited view, encouraging consistency, while also respecting the additional information given in the unedited target view,

$$Q_{edited} = f_{multi}\left([Q_{src}, Q_{warp}], \text{concat}(T, T_+)\right), \tag{1}$$

Fig. 6: Application of our method in interior design.

"Remove the clutter and add a large stylish brown leather armchair standing in the corner but a bit out and facing the room center"

where our additional instruction T_+ is: "The suggested appearance is in the second image. Stick to this change, but refine it to keep consistency with respect to the first image".

4 Experiments

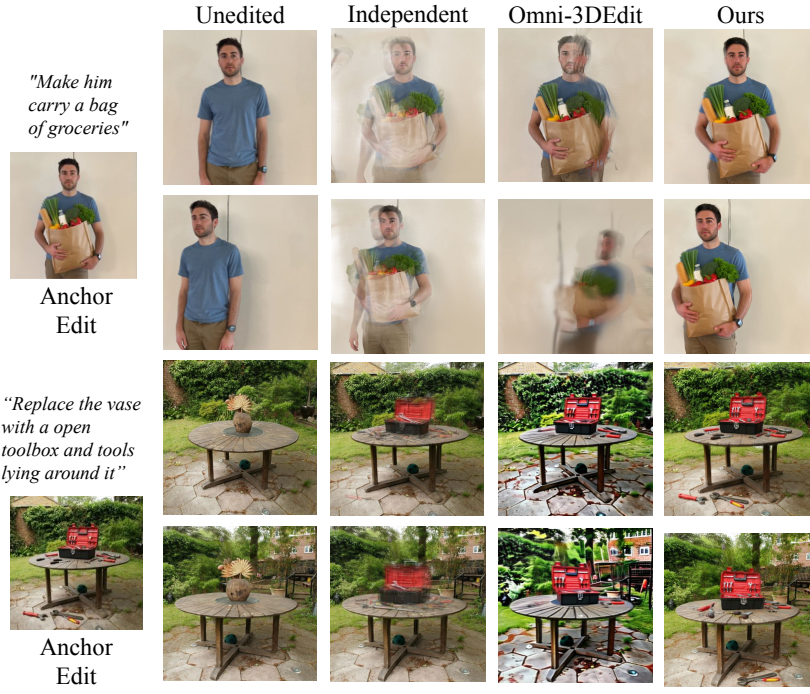
We divide our evaluation into two main tasks: (1) general multi-view editing (Figs. 5—6), where we also evaluate 3DGS representations generated from the edited multi-view images (Fig. 7) and show the 3D reconstruction capabilities on the edited images (Fig. 8), and (2) image pair editing, where an object is masked out and replaced with another one in a pair of images (Fig. 9). We evaluate view consistency for both tasks.

4.1 Experimental setup

We evaluate our method on general multi-view editing across a variety of different edit types. For evaluation, we use 38 prompts over 4 editing categories (general nonrigid edits, object addition, object removal, and appearance change). We use a combination of 15 test scenes taken from the following datasets: SPIn-NeRF [41], IN2N [22], Mip-NeRF360 [3] and BlendedMVS [66]. For this task, we compare GeM-NR with Omni-3DEdit [37] which performs multi-view editing by directly editing a set of up to 10 images given one edited anchor image. The approach of Omni-3DEdit is a data generation pipeline followed by training a model on a created dataset. The generated paired edits are from specific categories such as object removal, object addition, and appearance change. Omni-3DEdit cannot directly perform a nonrigid or any complex edit, but if the task can be split into first removing an object and then adding a new object, Omni-3DEdit can be run twice to achieve a similar result.

For evaluating consistent image pair editing, we use image pairs taken from DreamBooth [50] and Mip-NeRF360 [3], using a total of 38 pairs in the test set. For this task, we compare GeM-NR with Edicho [2] which performs consistent image editing by computing explicit correspondences from the unedited images

Fig. 7: Renders from 3DGS of edited scene. Our approach gives sharp renders that preserve the details from the anchor edit.



that are used to guide the denoising process. Edicho [2] presents results with two different backbone 2D editors: ControlNet [71, 73] for global editing and BrushNet [27] for inpainting-based editing. The code release for Edicho only includes the inpainting-based approach. Hence, we only use BrushNet [27] for comparison.

Implementation details. We use FLUX.2 [klein] [7] as our backbone multi-reference image editing model. For geometry estimation, we use Depth Anything 3 [33]. For the general multi-view editing, initial anchor edits are generated using Qwen [60] (version Qwen-Image-Edit-2509) which is the editing method used when training Omni-3DEdit. When comparing image pair editing performance to Edicho, we use BrushNet [27] for the initial anchor edits, since this is the method used by Edicho. Since BrushNet requires masks and inpaints the masked object area, we adapt our method to work with this type of input by providing a masked anchor edit as a reference and estimating the object geometry instead of the full scene geometry. Finally, for all tasks, we also edit anchors using FLUX.2 [klein] [7] and include the results in the comparisons. To generate 3DGS representation from the edited multi-view images, we use AnySplat [26], a feedforward method that can generate 3D Gaussians from a set of sparse unposed images in seconds.

Fig. 8: 3D reconstructions of the edited images, obtained using VGGT [56].

Evaluation metrics. For consistent image pair editing, we follow [2] and use CLIP [44] to evaluate text alignment (TA) and edit consistency (EC). For general multi-view editing, we evaluate both consistency and text alignment across the edited views. Multi-view consistency is measured with MEt3R [1], which compares DINO [9] embeddings at matches obtained by Dust3R [58]. To verify that the edited images preserve the relative poses as the unedited images, we also report mean average accuracies (mAA) computed by thresholding symmetrized epipolar distances and confidences of RoMa matches, where the ground truth poses are obtained by running COLMAP on sets of unedited images. Finally, we evaluate the resulting 3D Gaussians by measuring how well their renderings align with the text prompt.

4.2 General multi-view editing

Multi-view consistency. We evaluate multi-view consistency on sets of 10 images. Tab. 1 reports the consistency metrics, both across all edit types and for the each edit category separately. In general, our method achieves the highest consistency, with the largest improvements for general nonrigid editing involving significant geometry changes. TA measures how well the edit instruction is preserved after multi-view editing, for which we observe that Omni-3DEdit is limited in scope and struggles to perform an edit for appearance change and general non-rigid editing, while GeM-NR handles a broad range of edits. Examples of this can be seen in Fig. 5, where our method succeeds at performing multi-view editing, while Omni-3DEDit fails, and independently edited images suffer from significant inconsistencies. The top two rows of Fig. 4 present examples, where Omni-3DEdit performs the edit successfully, but GeM-NR preserves consistent details and appearance with respect to the anchor edit even better. Additional qualitative results can be found in Sec. A in the supplementary material. We also show in

Table 1: Consistency evaluation of edited multi-view images. GeM-NR gives improved multi-view consistency and preserves edit instructions, cf. high TA scores.

		Edit Consistency		3D Rec. Consistency			Text Alignment
		MEt3R↓	mAA, % ↑	PSNR↑	SSIM↑	LPIPS↓	TA / TA dir ↑
All Edit Types							
Qwen	Independent	0.248	31.47	20.70	0.677	0.213	**0.258** / **0.202**
Qwen	Omni-3DEdit	0.236	33.99	20.06	**0.690**	0.164	0.230 / 0.127
Qwen	**Ours**	0.194	36.71	21.63	0.653	0.167	0.252 / 0.198
FLUX.2	Independent	0.231	34.03	21.12	0.670	0.203	0.254 / 0.190
FLUX.2	Omni-3DEdit	0.238	34.27	19.57	0.677	0.171	0.233 / 0.122
FLUX.2	**Ours**	**0.190**	**38.58**	**21.96**	0.659	**0.162**	0.254 / 0.193
	Unedited	0.186	47.71	25.76	0.769	0.116	0.204 / -
General Nonrigid							
Qwen	Independent	0.303	23.15	18.62	0.613	0.284	0.265 / 0.240
Qwen	Omni-3DEdit	0.288	23.78	18.21	**0.662**	0.205	0.251 / 0.184
Qwen	**Ours**	0.215	31.56	20.68	0.612	0.194	0.268 / **0.252**
FLUX.2	Independent	0.291	24.91	18.93	0.598	0.275	**0.269** / 0.234
FLUX.2	Omni-3DEdit	0.288	21.56	17.60	0.628	0.216	0.248 / 0.172
FLUX.2	**Ours**	**0.216**	**32.35**	**20.91**	0.607	**0.188**	0.268 / 0.247
	Unedited	0.216	47.31	25.23	0.745	0.128	0.202 / -
Object Addition							
Qwen	Independent	0.217	41.22	20.72	0.738	0.220	**0.242** / **0.190**
Qwen	Omni-3DEdit	0.125	46.60	23.46	**0.780**	**0.100**	0.221 / 0.120
Qwen	**Ours**	**0.120**	46.61	24.40	0.766	0.105	0.226 / 0.151
FLUX.2	Independent	0.165	45.97	22.82	0.769	0.158	0.240 / 0.165
FLUX.2	Omni-3DEdit	0.134	46.99	22.92	0.779	0.107	0.232 / 0.121
FLUX.2	**Ours**	0.121	**47.17**	**24.78**	0.777	0.111	0.238 / 0.144
	Unedited	0.097	53.33	29.03	0.860	0.070	0.191 / -
Object Removal							
Qwen	Independent	0.166	41.79	24.06	0.734	0.143	**0.207** / **0.156**
Qwen	Omni-3DEdit	0.203	38.27	20.83	0.690	0.174	.201 / 0.114
Qwen	**Ours**	0.153	45.66	24.04	0.712	0.135	0.201 / 0.155
FLUX.2	Independent	0.165	44.62	**24.66**	**0.741**	0.142	0.204 / **0.156**
FLUX.2	Omni-3DEdit	0.202	40.90	20.81	0.698	0.170	0.199 / 0.104
FLUX.2	**Ours**	**0.151**	**46.43**	24.44	0.719	**0.133**	0.204 / 0.154
	Unedited	0.152	48.15	25.95	0.790	0.101	0.203 / -
Appearance Change							
Qwen	Independent	0.256	29.15	**20.74**	0.673	0.193	**0.278** / **0.200**
Qwen	Omni-3DEdit	0.257	34.03	19.67	**0.674**	**0.158**	0.232 / 0.099
Qwen	**Ours**	0.224	32.87	20.29	0.614	0.187	0.271 / 0.199
FLUX.2	Independent	0.244	31.27	20.52	0.652	0.199	0.270 / 0.185
FLUX.2	Omni-3DEdit	0.260	34.95	19.09	0.663	0.168	0.237 / 0.098
FLUX.2	**Ours**	**0.214**	**36.31**	20.63	0.626	0.174	0.270 / 0.193
	Unedited	0.214	46.17	24.80	0.741	0.132	0.210 / -

Table 2: Text alignment (TA↑ / TA dir↑) of the renders from 3DGS of edited scene with Qwen anchor backbone. GeM-NR improves over Omni-3DEdit. It especially improves for general nonrigid edits and object addition, where independent and Omni-3DEdit struggle.

	All Edit Types	General Nonrigid	Object Addition	Object Removal	Appearance Change
Independent	0.250 / 0.191	0.249 / 0.214	0.225 / 0.134	**0.212** / 0.170	**0.275** / **0.206**
Omni-3DEdit	0.229 / 0.130	0.244 / 0.194	0.232 / 0.130	0.207 / 0.126	0.226 / 0.092
Ours	**0.258** / **0.212**	**0.269** / **0.273**	**0.246** / **0.188**	0.211 / **0.173**	0.274 / 0.198

Fig. 9: Qualitative image pair editing examples for Edicho and our GeM-NR.

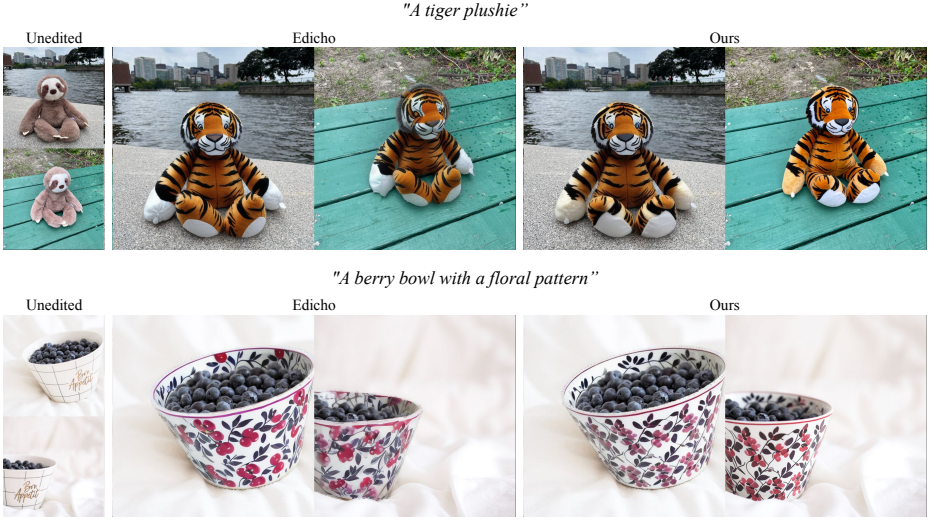


Fig. 6 an example of how GeM-NR can be used for consistent multi-view editing in interior design applications.

Edited 3D representations. The multi-view edits can also be used to generate a 3D Gaussian model representing the edited scene. Tab. 2 demonstrates that GeM-NR gives edited 3D Gaussians that best align with the desired edit instruction, with largest improvements in the categories of object addition and general nonrigid edit where significant changes in geometry occur. Comparable results are achieved when using FLUX.2 [klein] backbone for the anchor image, as can be seen in Sec. A in the supplementary material. In Fig. 7 we observe that our method gives renders that are sharp and clearly preserve the details from the edited anchor view. We also show in Fig. 8 that a 3D reconstruction of the edited scene can be obtained by running VGGT [56] on the set of edited images.

Table 3: Consistent image pair editing (inpainting) evaluation.

	mAA, % ↑	MEt3R (object) ↓	TA / TA dir (object) ↑	EC (object) ↑
DreamBooth images				
BrushNet (uses masks)				
Independent	-	0.646	0.279 / 0.218	0.834
Edicho	-	0.324	**0.287** / 0.244	0.887
Ours	-	**0.258**	0.280 / **0.245**	**0.899**
FLUX.2 [klein] (with masks)				
Independent	-	0.528	**0.271** / 0.210	0.871
Ours	-	**0.244**	0.268 / **0.224**	**0.902**
Mip-NeRF 360 test scenes				
BrushNet (uses masks)				
Independent	23.68	0.500	**0.252** / 0.177	0.885
Edicho	28.13	0.319	0.248 / 0.178	**0.941**
Ours	28.04	0.285	0.248 / 0.179	0.920
FLUX.2 [klein] (with masks)				
Independent	30.22	0.271	0.242 / 0.193	0.905
Ours	**33.89**	**0.217**	0.238 / **0.198**	0.920

4.3 Image pair editing

Tab. 3 shows evaluation results for image pair editing on DreamBooth images and Mip-NeRF 360 test scenes. MEt3R, TA, EC are computed on the masked images where only the object is kept (since the background varies a lot between the images, see top row in Fig. 9). GeM-NR often achieves as good or better consistency as well as alignment with respect to the edit description.

4.4 Runtimes

In Table 4, we compare runtimes of per-image editing using different methods. Our method is significantly faster than Omni-3DEdit, especially for complex edits where Omni-3DEdit has to perform two forward passes to complete the edit, and it is comparable to Edicho (while achieving higher quality images and better consistency as shown in Fig. 9).

Table 4: Runtime comparison. Per-image editing runtime in seconds.

Method	Edicho	Omni3DEdit	Omni3DEdit (complex edits)	**Ours**
Time, s	3.5	8.3	16.6	**3.4**

4.5 Ablation Studies

Tab. 5 shows the ablation configurations and results on the held-out validation dataset. Here, we used 4 separate validation scenes extracted from MipNeRF360 [3], IN2N [22] and SPIn-NeRF [41] datasets, with 11 prompts. We evaluate several modifications to the proposed method: (1) Reference text (Input edit) — whether to provide the original edit description; if not, provide "Edit the first image in the same way as shown in the second image"; (2) Reference image(s) — which of the images to provide: original query Q_{src}, edited anchor A_{edited}, and/or depth-warp Q_{warp}. We also compare our method to the baseline approach (first row) that concatenates the two inputs into one image and asks the model to edit two images consistently. Finally, we evaluate our method without pose refinement (last row). Our method is highlighted in gray. It achieves a trade-off across all performance evaluation metrics, as indicated in the last metrics column that computes a balanced score over all metrics. The detailed prompts used for different configurations can be found in Sec. C in the supplementary material.

Table 5: Ablation study of consistent image pair editing. Our method is highlighted in gray. It achieves a trade-off across all performance evaluation axes.

Ref. text	Ref. image(s)			Metrics				
Input edit	Original query	Edited anchor	Depth-warp	MEt3R ↓	mAA, % ↑	TA ↑	EC ↑	Balanced ↑
Simultaneous pair editing								
Concatenate both inputs								
✓	N/A	N/A	N/A	0.229	35.40	0.271	0.888	8.08
Editing one conditioned on another								
Original edit text prompt preserved; no warp								
✓	✓	✗	✗	0.238	31.70	0.214	0.924	7.43
✓	✓	✓	✗	0.069	13.53	0.271	0.888	3.53
Original edit text prompt ignored; with warp								
✗	✗	✗	✓	0.219	39.83	0.268	0.827	8.47
✗	✓	✗	✓	0.220	40.74	0.268	0.844	8.87
✗	✗	✓	✓	0.231	38.74	0.266	0.823	8.25
✗	✓	✓	✓	0.185	26.11	0.271	0.863	5.92
Original edit text prompt preserved; with warp								
✓	✓	✓	✓	0.173	23.96	0.275	0.913	5.86
✓	✓	✗	✓	0.199	40.63	0.274	0.885	**9.50**
No relative pose refinement				0.158	39.88	0.278	0.894	9.28
			Unedited	0.190	49.72	0.214	0.924	-

5 Conclusion

We present GeM-NR, a method for multi-view consistent edits that substantially change scene geometry and appearance. Our approach is flexible, handling diverse editing tasks, and fast, requiring only seconds per image. Extensive evaluations on different edit categories show that our method shows improved multi-view consistency and that the edited images can be used to generate 3D representations that align well with the specified edit instruction.

Currently, our multi-view consistent editing is conditioned on a single anchor edit, which does not ensure consistency across all views. We observe that this approach works well for scenes with limited view point variations, but it limits ability to handle larger scenes with more extreme viewpoint variation. A potential solution is to condition the editing not only on one anchor view, but also other previously edited images.

References

1. Asim, M., Wewer, C., Wimmer, T., Schiele, B., Lenssen, J.E.: Met3r: Measuring multi-view consistency in generated images. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6034–6044 (2025)
2. Bai, Q., Ouyang, H., Xu, Y., Wang, Q., Yang, C., Cheng, K.L., Shen, Y., Chen, Q.: Edicho: Consistent image editing in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15277–15287 (October 2025)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. Proceedings of the Computer Vision and Pattern Recognition Conference (2022)
4. Bengtson, J., Nilsson, D., Lee, D.I., Lochman, Y., Kahl, F.: 3d-consistent multi-view editing by correspondence guidance (2026), https://arxiv.org/abs/2511.22228
5. Bengtson, J., Nilsson, D., Lin, C.T., Büsching, M., Kahl, F.: Adjustable visual appearance for generalizable novel view synthesis. In: Wallraven, C., Liu, C.L., Ross, A. (eds.) Pattern Recognition and Artificial Intelligence. pp. 157–171. Springer Nature Singapore, Singapore (2025)
6. Black Forest Labs: FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2 (2025)
7. Black Forest Labs: FLUX.2 [klein]: Towards Interactive Visual Intelligence. https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence (2025)
8. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2023)
9. Caron, M., Touvron, H., Misra, I., J'egou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 9630–9640 (2021), https://api.semanticscholar.org/CorpusID:233444273
10. Chen, D.Y., Tennent, H., Hsu, C.W.: Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8619–8628 (June 2024)
11. Chen, L., Li, R., Zhang, G., Wang, P., Zhang, L.: Fast multi-view consistent 3d editing with video priors. Proceedings of the AAAI Conference on Artificial Intelligence **40**(4), 2948–2956 (Mar 2026). https://doi.org/10.1609/aaai.v40i4.37286, https://ojs.aaai.org/index.php/AAAI/article/view/37286
12. Chen, M., Laina, I., Vedaldi, A.: Dge: Direct gaussian 3d editing by consistent multi-view editing. In: European Conference on Computer Vision. pp. 74–92. Springer (2024)
13. Chen, Y., Chen, Z., Zhang, C., Wang, F., Yang, X., Wang, Y., Cai, Z., Yang, L., Liu, H., Lin, G.: Gaussianeditor: Swift and controllable 3d editing with gaussian splatting (2023)
14. Chung, J., Hyun, S., Heo, J.P.: Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8795–8805 (June 2024)
15. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., et. al., E.R.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025), https://arxiv.org/abs/2507.06261
16. Dong, J., Wang, Y.X.: Vica-nerf: View-consistency-aware 3d editing of neural radiance fields. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
17. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: Roma: Robust dense feature matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19790–19800 (2024)
18. Erkoç, Z., Dai, A., Nießner, M.: Worldagents: Can foundation image models be agents for 3d world models? arXiv preprint arXiv:2603.19708 (2026)
19. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 12606–12633. PMLR (21–27 Jul 2024), https://proceedings.mlr.press/v235/esser24a.html
20. Fu, T.J., Hu, W., Du, X., Wang, W., Yang, Y., Gan, Z.: Guiding instruction-based image editing via multimodal large language models. In: ICLR (2024), https://arxiv.org/abs/2309.17102
21. Gomel, E., Wolf, L.: Diffusion-based attention warping for consistent 3d scene editing (2024), https://arxiv.org/abs/2412.07984
22. Haque, A., Tancik, M., Efros, A.A., Holynski, A., Kanazawa, A.: Instruct-nerf2nerf: Editing 3d scenes with instructions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 19740–19750 (2023)
23. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross-attention control. In: International Conference on Learning Representations (2023)
24. Hertz, A., Voynov, A., Fruchter, S., Cohen-Or, D.: Style aligned image generation via shared attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4775–4785 (June 2024)
25. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)

26. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., et al.: Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. ACM Transactions on Graphics (TOG) **44**(6), 1–16 (2025)
27. Ju, X., Liu, X., Wang, X., Bian, Y., Shan, Y., Xu, Q.: Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 150–168. Springer Nature Switzerland, Cham (2025)
28. Koh, E., Hyun, S., Lee, M., Chung, J., Seo, K., Heo, J.P.: Diffusion feature field for text-based 3d editing with gaussian splatting. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2026), https://openreview.net/forum?id=Kf9eNbp4wy
29. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742
30. Lee, D.I., Doh, H., Chi, S., Duan, R., Kim, S., Ramani, K.: Dynamic-editor: Training-free text-driven 4d scene editing with multimodal diffusion transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026)
31. Lee, D.I., Park, H., Seo, J., Park, E., Park, H., Baek, H.D., Shin, S., Kim, S., Kim, S.: Editsplat: Multi-view fusion and attention-guided optimization for view-consistent 3d scene editing with 3d gaussian splatting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11135–11145 (2025)
32. Li, L., Huang, Z., Feng, H., Zhuang, G., Chen, R., Guo, C., Sheng, L.: Voxhammer: Training-free precise and coherent 3d editing in native 3d space. arXiv preprint arXiv:2508.19247 (2025)
33. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Zhao, Y., Peng, S., Guo, H., Zhou, X., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=yirunib8l8
34. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=PqvMRDCJT9t
35. Liu, A., Tucker, R., Jampani, V., Makadia, A., Snavely, N., Kanazawa, A.: Infinite nature: Perpetual view generation of natural scenes from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2021)
36. Liu, X., Gong, C., qiang liu: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=XVjTT1nw5z
37. Liyi, C., Pengfei, W., Guowen, Z., Zhiyuan, M., Lei, Z.: Omni-3dedit: Generalized versatile 3d editing in one-pass. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026)
38. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11461–11471 (June 2022)
39. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2022)

40. Mirzaei, A., Aumentado-Armstrong, T., Brubaker, M.A., Kelly, J., Levinshtein, A., Derpanis, K.G., Gilitschenski, I.: Watch your steps: Local image and scene editing by text instructions. In: ECCV (2024)
41. Mirzaei, A., Aumentado-Armstrong, T., Derpanis, K.G., Kelly, J., Brubaker, M.A., Gilitschenski, I., Levinshtein, A.: SPIn-NeRF: Multiview segmentation and perceptual inpainting with neural radiance fields. In: CVPR (2023)
42. Müller, N., Schwarz, K., Rössle, B., Porzi, L., Bulò, S.R., Nießner, M., Kontschieder, P.: Multidiff: Consistent novel view synthesis from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10258–10268 (June 2024)
43. Park, J., Choi, T.E., Jun, Y., Hwang, S.J.: Wave: Warp-based view guidance for consistent novel view synthesis using a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11906–11915 (October 2025)
44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/radford21a.html
45. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents (2022), https://arxiv.org/abs/2204.06125
46. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12179–12188 (October 2021)
47. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(3) (2022)
48. Rojas, S., Philip, J., Zhang, K., Bi, S., Luan, F., Ghanem, B., Sunkavalli, K.: Datenerf: Depth-aware text-based editing of nerfs. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XI. p. 267–284. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-73247-8_16, https://doi.org/10.1007/978-3-031-73247-8_16
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
50. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22500–22510 (June 2023)
51. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 36479–36494. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf

52. Seo, J., Fukuda, K., Shibuya, T., Narihira, T., Murata, N., Hu, S., Lai, C.H., Kim, S., Mitsufuji, Y.: Genwarp: Single image to novel views with semantic-preserving generative warping. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 80220–80243. Curran Associates, Inc. (2024). https://doi.org/10.52202/079017-2550, https://proceedings.neurips.cc/paper_files/paper/2024/file/92e886487a8354b03d8bf4416eae6d7d-Paper-Conference.pdf
53. Song, L., Cao, L., Gu, J., Jiang, Y., Yuan, J., Tang, H.: Efficient-nerf2nerf: Streamlining text-driven 3d editing with multiview correspondence-enhanced diffusion models. arXiv preprint arXiv:2312.08563 (2023)
54. Tung, J., Chou, G., Cai, R., Yang, G., Zhang, K., Wetzstein, G., Hariharan, B., Snavely, N.: Megascenes: Scene-level view synthesis at scale. In: ECCV (2024)
55. Wang, B., Dutt, N.S., Mitra, N.J.: Proteusnerf: Fast lightweight nerf editing using 3d-aware image context. Proc. ACM Comput. Graph. Interact. Tech. **7**(1) (may 2024). https://doi.org/10.1145/3651290, https://doi.org/10.1145/3651290
56. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
57. Wang, J., Lin, C., Sun, L., Cao, Z., Yin, Y., Nie, L., Yuan, Z., Chu, X., Wei, Y., Liao, K., et al.: Geometry-guided reinforcement learning for multi-view consistent 3d scene editing. arXiv preprint arXiv:2603.03143 (2026)
58. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20697–20709 (2024)
59. Wang, Y., Yi, X., Wu, Z., Zhao, N., Chen, L., Zhang, H.: View-consistent 3d editing with gaussian splatting. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29 – October 4, 2024, Proceedings, Part XXXV. p. 404–420. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-72761-0_23, https://doi.org/10.1007/978-3-031-72761-0_23
60. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., ming Yin, S., Bai, S., Xu, X., Chen, Y., Chen, Y., Tang, Z., Zhang, Z., Wang, Z., Yang, A., Yu, B., Cheng, C., Liu, D., Li, D., Zhang, H., Meng, H., Wei, H., Ni, J., Chen, K., Cao, K., Peng, L., Qu, L., Wu, M., Wang, P., Yu, S., Wen, T., Feng, W., Xu, X., Wang, Y., Zhang, Y., Zhu, Y., Wu, Y., Cai, Y., Liu, Z.: Qwen-image technical report (2025), https://arxiv.org/abs/2508.02324
61. Wu, J., Bian, J.W., Li, X., Wang, G., Reid, I., Torr, P., Prisacariu, V.: GaussCtrl: Multi-View Consistent Text-Driven 3D Gaussian Splatting Editing. ECCV (2024)
62. Xia, R., Tang, Y., Zhou, P.: Towards scalable and consistent 3d editing (2025), https://arxiv.org/abs/2510.02994
63. Xiao, S., Wang, Y., Zhou, J., Yuan, H., Xing, X., Yan, R., Li, C., Wang, S., Huang, T., Liu, Z.: Omnigen: Unified image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13294–13304 (June 2025)
64. Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., et. al., F.H.: Qwen2 technical report (2024), https://arxiv.org/abs/2407.10671
65. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)
66. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. Proceedings of the Computer Vision and Pattern Recognition Conference (2020)

67. Ye, J., Xie, S., Zhao, R., Wang, Z., Yan, H., Zu, W., Ma, L., Zhu, J.: Nano3d: A training-free approach for efficient 3d editing without masks (2025), https://arxiv.org/abs/2510.15019
68. You, M., Zhu, Z., Liu, H., Hou, J.: Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. In: International Conference on Learning Representations (2025)
69. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: ViewCrafter: Taming Video Diffusion Models for High-fidelity Novel View Synthesis . IEEE Transactions on Pattern Analysis & Machine Intelligence (01), 1–18 (Sep 5555). https://doi.org/10.1109/TPAMI.2025.3613256, https://doi.ieeecomputersociety.org/10.1109/TPAMI.2025.3613256
70. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 31428–31449. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/64008fa30cba9b4d1ab1bd3bd3d57d61-Paper-Datasets_and_Benchmarks.pdf
71. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
72. Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10146–10156 (June 2023)
73. Zhao, S., Chen, D., Chen, Y.C., Bao, J., Hao, S., Yuan, L., Wong, K.Y.K.: Uni-controlnet: All-in-one control to text-to-image diffusion models. Advances in Neural Information Processing Systems (2023)
74. Zhu, Z., Chen, H., Li, P., Wei, M.: Coreeditor: Correspondence-constrained diffusion for consistent 3d editing. IEEE Transactions on Visualization and Computer Graphics **32**(3), 2838–2851 (2026). https://doi.org/10.1109/TVCG.2026.3657658
75. Zhuang, J., Zeng, Y., Liu, W., Yuan, C., Chen, K.: A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 195–211. Springer Nature Switzerland, Cham (2025)

Supplementary Material

Overview

In this supplementary material, we show additional experimental results (Sec. A), hyperparameter tuning results (Sec. B), and details for the prompts used in the ablation study (Sec. C).

A Additional Experimental Results

Tab. B.1 presents text alignment metrics for renderings from edited 3DGS when using FLUX.2 [klein] to edit the anchor image. The results are comparable to those with the Qwen backbone for anchor images presented in Tab. 2 in the main paper. Additional qualitative examples of the multi-view consistent editing can be found in Fig. B.1 and B.2.

B Hyperparameter Tuning

In Fig. B.3 we show the results of hyperparameter tuning done on the held-out validation set. We use the same validation set as in the ablation study. The hyperparameters in our method are split into two groups. The first group, shown along the y-axis in Fig. B.3, consists of the hyperparameters related to relative pose refinement, namely: (1) whether to perform relative pose refinement, (2) maximum epipolar error in robust estimation, (3) certainty threshold for RoMa matches, and (4) minimum percentage of inliers to consider the refinement successful (or otherwise revert to the initial pose). The second group, shown along the x-axis in Fig. B.3, consists of the hyperparameters related to masking and depth estimation: (1) whether to provide an additional input to the multi-reference model, where the input is a query image with masked out area that is to be edited (the area is computed automatically based on the image difference between the anchor image and its edited version and warping), (2) certainty threshold as a percentile for the predicted depths, (3) strength of mask erosion if the query image is to be masked.

C Ablation configurations detailed

Table C.2 shows the detailed text decriptions of the prompts used in the ablation study.

Table B.1: Text alignment (TA↑ / TA dir↑) of the renders from 3DGS of edited scene with FLUX.2 [klein] anchor backbone. GeM-NR clearly improves over Omni-3DEdit. It especially improves for the general nonrigid edits and object addition, where both Independent and Omni-3DEdit struggle.

	All Types	General Nonrigid	Object Addition	Object Removal	Appearance Change
Unedited	0.201 / -	0.194 / -	0.201 / -	0.200 / -	0.206 / -
Independent	0.253 / 0.192	0.254 / 0.227	0.243 / 0.169	**0.213** / **0.170**	0.271 / 0.187
Omni-3DEdit	0.233 / 0.131	0.244 / 0.185	0.240 / 0.140	0.206 / 0.114	0.233 / 0.101
Ours	**0.259** / **0.205**	**0.270** / **0.269**	**0.253** / **0.176**	0.211 / 0.163	**0.272** / **0.190**

Fig. B.1: Additional qualitative results for multi-view editing with Qwen.

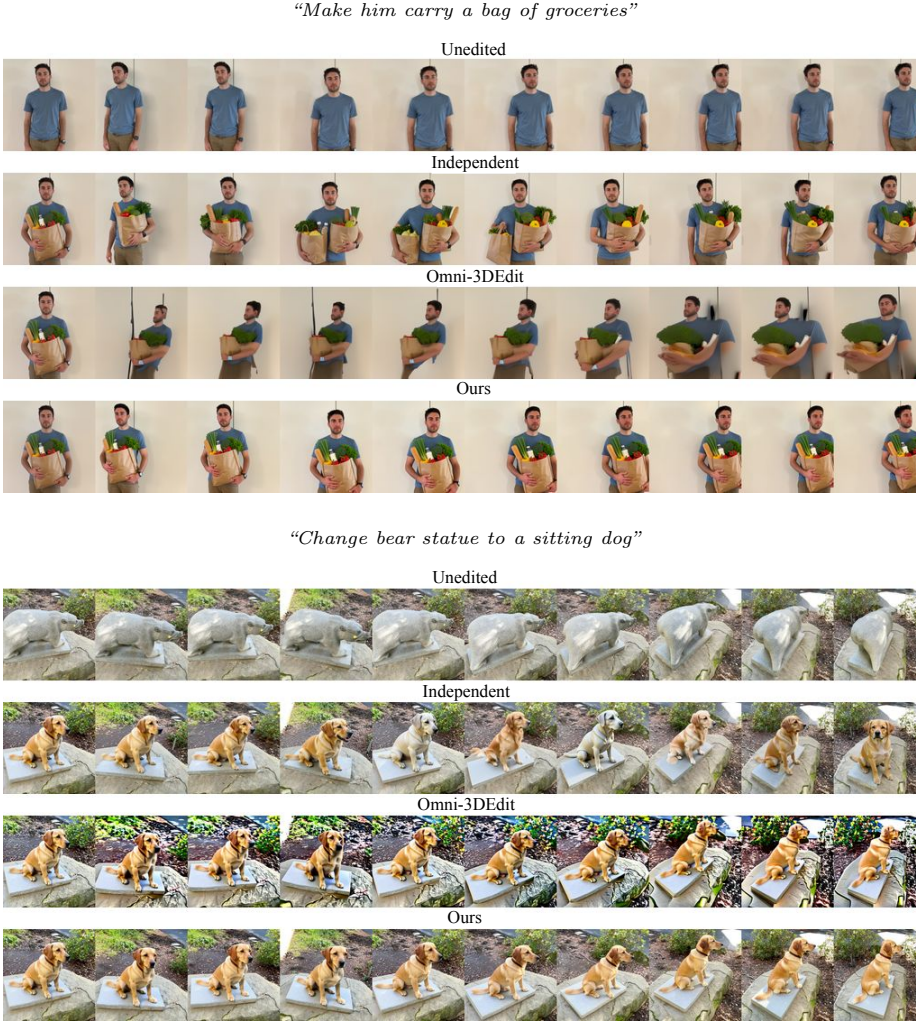


Fig. B.2: Edited multi-view images. We observe that our method gives consistent edits with preserved details, such as the patches on the jacket.

Fig. B.3: Quantitative results of hyperparameter tuning of our method on the held-out validation set. The best set of hyperparameters (outlined cell) is chosen such that it maximizes the `balanced` score: $\texttt{balanced} = \texttt{mAA} \cdot (1 - \texttt{MEt3R}/2)$.

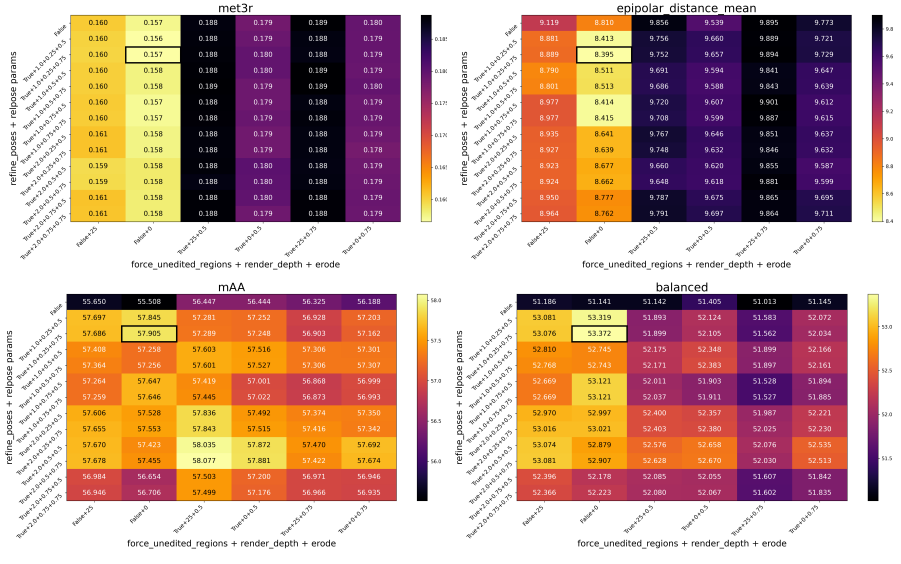


Table C.2: Ablation study configurations in consistent image pair editing.

Ref. text	Ref. image(s)			
Input edit	Original query	Edited anchor	Depth-warp	Prompt
Simultaneous pair editing				
Concatenate both inputs				
✓	N/A	N/A	N/A	"Edit these two concatenated images in a consistent way"
Editing one conditioned on another				
Original edit text prompt preserved; no warp				
✓	✓	✗	✗	T
✓	✓	✓	✗	"Edit the first image as shown in the second image"
Original edit text prompt ignored; with warp				
✗	✗	✗	✓	"Inpaint this image"
✗	✓	✗	✓	"Edit this image as shown in the second image. Inpaint the second image"
✗	✗	✓	✓	"Inpaint the first image. Use the second image as a guidance on how to inpaint"
✗	✓	✓	✓	"Edit the first image in the same way as shown in the second image. The suggested appearance is in the third image. Stick to this change, but refine it to keep consistency with respect to the first image"
Original edit text prompt preserved; with warp				
✓	✓	✓	✓	concat(T, "The suggested appearance is in the second image. Stick to this change, but refine it to keep consistency with respect to the first image. For reference, the same edit at a different viewpoint is provided in the third image")
✓	✓	✗	✓	concat(T, "The suggested appearance is in the second image. Stick to this change, but refine it to keep consistency with respect to the first image")