RECONVIAGEN: TOWARDS ACCURATE MULTI-VIEW 3D OBJECT RECONSTRUCTION VIA GENERATION

Anonymous authors
Paper under double-blind review

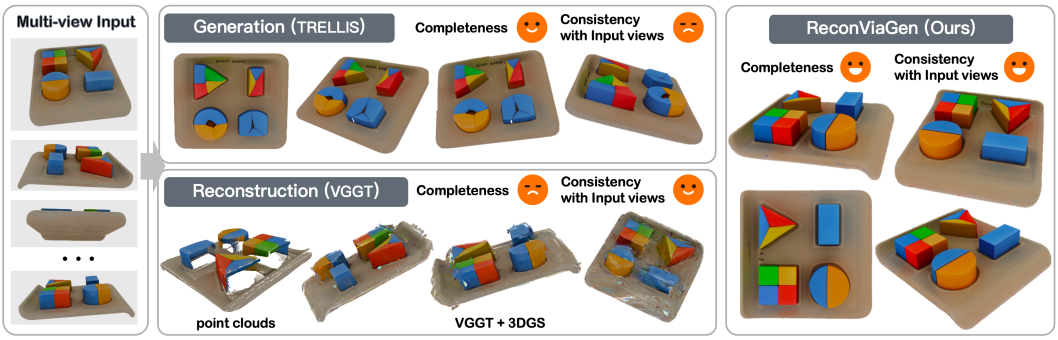

Figure 1: In the task of 3D object reconstruction from multi-view images, existing pure reconstruction methods can only produce incomplete results, while generation-based methods can get plausible complete results but with strong inconsistency with input images. Our ReconViaGen integrates 3D reconstruction and diffusion-based generation priors into one framework that leads to accurate reconstructions.

ABSTRACT

Existing multi-view 3D object reconstruction methods heavily rely on sufficient overlap between input views, where occlusions and sparse coverage in practice frequently yield severe reconstruction incompleteness. Recent advancements in diffusion-based 3D generative techniques offer the potential to address these limitations by leveraging learned generative priors to "hallucinate" invisible parts of objects, thereby generating plausible 3D structures. However, the stochastic nature of the inference process limits the accuracy and reliability of generation results, preventing existing reconstruction frameworks from integrating such 3D generative priors. In this work, we comprehensively analyze the reasons why diffusion-based 3D generative methods fail to achieve high consistency, including (a) the insufficiency in constructing and leveraging cross-view connections when extracting multi-view image features as conditions, and (b) the poor controllability of iterative denoising during local detail generation, which easily leads to plausible but inconsistent fine geometric and texture details with inputs. Accordingly, we propose ReconViaGen to innovatively integrate reconstruction priors into the generative framework and devise several strategies that effectively address these issues. Extensive experiments demonstrate that our ReconViaGen can reconstruct complete and accurate 3D models consistent with input views in both global structure and local details.

1 INTRODUCTION

In the field of 3D computer vision, multiview 3D object reconstruction has long been a fundamental yet challenging task, with numerous applications in areas such as VR, AR, and 3D modeling. Existing multiview reconstruction methods typically rely on enough visual cues and learned correspondences between views to estimate the 3D structure and appearance of the object Mildenhall et al. (2020); Kerbl et al. (2023); Leroy et al. (2024); Wang et al. (2024a; 2025a). However, these methods often face significant limitations when dealing with weak-texture objects or incomplete image captures due to occlusions or the presence of support surfaces. As a result, reconstructed 3D

models tend to have holes, artifacts, and missing/blurred geometric details, which severely restrict the reconstruction completeness He et al. (2024); Xu et al. (2024c).

Recent advances in diffusion-based 3D generative techniques have shown great promise in addressing these limitations. These techniques leverage 3D generative priors learned from large-scale 3D data to generate complete 3D outputs from sparse- or even single-view images Li et al. (2024b); Zhao et al. (2025); Li et al. (2025a); Zhang et al. (2024b); Ye et al. (2025). Such strong generative priors can effectively "hallucinate" the invisible portions of objects with plausible high-quality geometry and appearance, thereby showing great potential in 3D reconstruction by filling in the missing details and improving the completeness. However, the stochastic nature of the diffusion-based inference process introduces significant uncertainty and variability in the generated results, making it challenging to achieve high accuracy and reliability, especially the pixel-level alignment required in accurate reconstruction. This stochasticity has largely hindered the effective integration of diffusion-based 3D generative priors into existing multi-view reconstruction frameworks.

Pioneering explorations have been made in the field of 3D diffusion-based generation from multi-view images Xiang et al. (2024); Zhao et al. (2025). However, their predictions still suffer from inaccurate global structures and inconsistent local details. The inherent key reasons of the failure include (i) the insufficiency in constructing cross-view correlations when extracting multi-view image features as conditions, resulting in inaccurate estimation in both object geometry and texture, at the global and local level, (ii) the poor controllability and stability of the denoising process during inference, which easily results in inconsistency with input views especially in detailed geometry and texture estimation. To address these issues, we present ReconViaGen that innovatively integrates multi-view stereo priors into the diffusion-based generative framework for object reconstruction. Our solution includes three stages: (i) a pre-trained strong reconstructor Wang et al. (2025a) is developed to build a multi-view stereo understanding of the object geometry and texture, aggregated into a single global token list and a set of local token lists, for representing the global geometry and the detailed per-view appearance, respectively; (ii) a coarse-to-fine 3D generator Xiang et al. (2024) first estimates the coarse structure and then produces the fine textured mesh, under the conditioning of global and local tokens from the first stage, respectively; (iii) refining the estimated poses from the reconstructor using the generation from the second stage, and encouraging the pixel-wise alignment with input views using a novel rendering-aware velocity compensation mechanism, where input images coupled with estimated camera poses are used to explicitly guide the denoising trajectory of local latent representations.

Extensive experiments on the Dora-bench Chen et al. (2024) and OminiObject3D Wu et al. (2023b) datasets validate that our ReconViaGen can achieve state-of-the-art (SOTA) reconstruction performance in both global shape accuracy and completeness and local details in geometry and textures. Our contributions are summarized as follows:

- We propose a novel framework called ReconViaGen, which is the first to integrate strong reconstruction priors into a diffusion-based 3D generator for accurate and complete multi-view object reconstruction. A key design is to aggregate image features rich in reconstruction priors as multi-view-aware diffusion conditions.
- The generation adopts a coarse-to-fine paradigm, which leverages global and local reconstruction-based conditions to generate accurate coarse and then fine results in both geometry and texture. Additionally, a novel rendering-aware velocity compensation mechanism is proposed that constrains the denoising trajectory of local latent representations for detailed pixel-level alignment.
- Extensive experiments on the Dora-bench and OminiObject3D datasets are conducted that validate the effectiveness and superiority of the proposed ReconViaGen, which achieves SOTA performance.

2 Related Work

Single-view 3D Generation. Great developments have been made in single-view 3D Object Generation. Recent methods can be divided into two groups: 2D prior-based and 3D native generative methods. DreamFusion Poole et al. (2022) and its following successors Tang et al. (2024); Qiu et al. (2024); Wang et al. (2024b); Lin et al. (2023); Tang et al. (2023a) distill the 3D knowledge from pre-trained 2D models. Another line of work develops multi-view diffusion based on pre-trained image

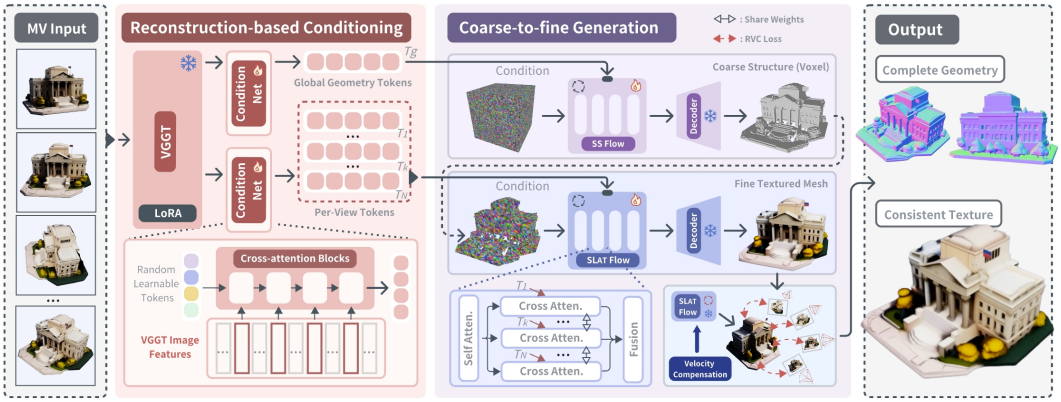

Figure 2: An overview illustration of the proposed ReconViaGen framework, which integrates strong reconstruction priors with 3D diffusion-based generation priors for accurate reconstruction at both the global and local level.

or video generators and conducts view fusion for 3D outputs Li et al. (2023); Xu et al. (2024b); Tang et al. (2025); Xu et al. (2024d); Wang et al. (2025b); Wei et al. (2024); Liu et al. (2023a; 2024); Xu et al. (2024e); Li et al. (2024a); Zuo et al. (2024); Wu et al. (2024a). Differently, 3D native generative methods employ diffusion based on different 3D representations like point clouds Luo & Hu (2021); Zhou et al. (2021); Nichol et al. (2022), voxel grids Hui et al. (2022); Tang et al. (2023b); Müller et al. (2023), Triplanes Chen et al. (2023); Wang et al. (2023b); Shue et al. (2023), 3D Gaussians Zhang et al. (2024a). More recently, 3D latent diffusion has been explored to directly learn the mapping between the image and 3D geometry Zhang et al. (2023); Zhao et al. (2024); Li et al. (2024b); Zhang et al. (2024b); Wu et al. (2024b); Li et al. (2025a); Zhao et al. (2025); Ye et al. (2025), which greatly improves the generation quality. However, multi-view 3D generation is still under-explored, suffering from high variations in generation, easy inconsistency with input images, or strong reliance on input viewpoints Xiang et al. (2024); Zhao et al. (2025), which hinders their direct application in accurate 3D object reconstruction.

Multi-view 3D Reconstruction. Traditional methods conduct multi-view stereo (MVS) to reconstruct the visible surface of objects by triangulating correspondences across multiple calibrated images Furukawa et al. (2015); Galliani et al. (2015); Schönberger et al. (2016); Xu & Tao (2019). Learning-based MVS methods Yao et al. (2018; 2019); Chen et al. (2019); Cheng et al. (2020); Gu et al. (2020); Yang et al. (2020); Wang et al. (2021a) employ deep neural networks to enhance both reconstruction quality and computational efficiency. Scene-specific NeRF methods Lin et al. (2021); Wang et al. (2021b) adopt bundle adjustment from conventional SfM pipelines to jointly optimize camera parameters along with radiance field from dense views. Recently, DUSt3R Wang et al. (2024a) and its follow-up works Smart et al. (2024); Leroy et al. (2024); Wang et al. (2025a) together estimate point clouds and camera poses from paired or more views, which releases the reliance on camera parameters, but suffers from incomplete reconstruction results caused by the point cloud representation. Focusing on object reconstructions, large reconstruction models Hong et al. (2023) are explored to produce complete reconstructions via regressing a more compact or structured 3D representation (*e.g.* 3D Gaussians Kerbl et al. (2023) and Triplane) from multi-view inputs Wei et al. (2024); Xu et al. (2024b); Tang et al. (2025); Xu et al. (2024d), but requiring view inputs from certain camera poses. Follow-up methods further support pose-free reconstructions Wu et al. (2023a); Wang et al. (2023a); Jiang et al. (2023); He et al. (2024), while they tend to predict smooth and blurred details, especially in invisible regions. Differently, our method introduces diffusion-based 3D generation priors to advance pose-free object reconstruction in fidelity and completeness.

Generative Priors in 3D Object Reconstruction. Generative priors are introduced into 3D reconstruction frameworks to assist in predicting plausible geometries or textures in invisible portions of objects. Existing methods mainly introduce two kinds of priors: (i) diffusion-based 2D generative prior and (ii) regression-based 3D generative prior. The former is often used in single-view 3D reconstruction by generating plausible multi-view images first and conducting reconstruction Li et al. (2023; 2024a); Xu et al. (2024b); Tang et al. (2025); Xu et al. (2024d); Wang et al. (2025b); Wei et al. (2024); Liu et al. (2023a; 2024); Wu et al. (2024a). For pose-free sparse-view reconstruction, iFusion Wu et al. (2023a) leverages Zero123 Liu et al. (2023b) predictions within an optimization

pipeline to align poses and generate novel views for reconstruction. However, the inconsistency between views still limits the performance of this pipeline. Regression-based 3D generative priors are introduced to regress a unified compact 3D representation, avoiding this issue, for example 3D neural volume Jiang et al. (2023), Triplane Hong et al. (2023); Wei et al. (2024); Wang et al. (2023a), and 3D Gaussians He et al. (2024); Xu et al. (2024c); Smart et al. (2024). Diffusion-based generative priors prove superior to regressive ones in generating detailed results in both geometry and texture Li et al. (2025a); Zhao et al. (2025); Zhang et al. (2024b); Ye et al. (2025). One2345++ and its follow-up work Liu et al. (2024); Xu et al. (2024a) develop 3D volume diffusion conditioned by multi-view inputs. However, 3D volume suffer from poor compactness, so a trade-off between the diffusion learning difficulty and representation capability limits their performance. Differently, our method builds upon strong diffusion-based 3D generative priors Xiang et al. (2024), with powerful reconstruction priors Wang et al. (2025a) constraining the denoising process for accurate 3D outputs of high-fidelity details.

3 METHODOLOGY

3.1 PRELIMINARY

Given a set of N uncalibrated multi-view images of an object $I = \{I_i\}_{i=1}^{N}$, the task of pose-free multi-view reconstruction aims to obtain the complete 3D object O. Our framework leverages two kinds of strong priors to achieve complete and accurate reconstruction results: the reconstruction prior from VGGT Wang et al. (2025a) and the generation prior from TRELLIS Xiang et al. (2024). In this section, we first introduce these two priors as preliminaries.

Reconstruction prior of VGGT VGGT Wang et al. (2025a) achieves SOTA results in pose-free multi-view 3D reconstruction, providing a powerful reconstruction prior. It adopts a feed-forward transformer architecture designed for efficient and unified 3D scene reconstruction from single/multiple images. Multi-view images I are first fed into a DINO-based ViT Oquab et al. (2024) simultaneously for tokenization and feature extraction into ϕ_{dino}. Then, 24 self-attention layers further address ϕ_{dino} into 3D-aware features $\{\phi_i\}_{i=1}^{24}$ with an alternating attention strategy, switching between frame-wise and global self-attention to balance local and global information and enhance multi-view consistency. Finally, four prediction heads decode the output of 4 layers (4-th, 11-th, 17-th, and 23-rd), *i.e.*, $\phi_{\text{vggt}}(I) = \{\phi_4, \phi_{11}, \phi_{17}, \phi_{24}\}$, into camera parameters, depth map, point map, and tracking feature predictions. To adapt to object reconstruction, we fine-tune VGGT on an object-reconstruction dataset (see Sec. 4.1 for details). A LoRA fine-tuning on the VGGT aggregator is employed to preserve the pre-trained 3D geometric priors, with a multi-task objective:

$$\mathcal{L}_{\text{VGGT}}(\theta) = \mathcal{L}_{\text{camera}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{nmap}}, \tag{1}$$

where θ is the LoRA parameters, $\mathcal{L}_{\text{camera}}$, $\mathcal{L}_{\text{depth}}$ and $\mathcal{L}_{\text{nmap}}$ denote the camera pose loss, the depth loss, and the point map loss, respectively. In the following text, we simply use "VGGT" to refer to this fine-tuned VGGT.

Generation prior of TRELLIS TRELLIS Xiang et al. (2024) is a SOTA 3D generative model that provides a strong generation prior. It proposes a novel representation called Structured LATent (SLAT) that combines a sparse 3D grid with dense visual features extracted from a powerful vision foundation model, which captures both geometric (structure) and textural (appearance) information and enables decoding into multiple 3D representations. We choose TRELLIS as the 3D generator in our framework because it has shown great potential in 3D object generation He et al. (2025); Li et al. (2025b) and inspired many works in downstream applications Yang et al. (2025); Cao et al. (2025); Wu et al. (2024c). It employs a coarse-to-fine two-stage generation pipeline: generating the sparse structure (SS), represented as sparse voxels $\{p_i\}_i^L$, via SS Flow and then predicting structured latents (SLAT) for active SS voxels, represented as $X = \{(p_i, x_i)\}_i^V$, via SLAT Flow, where p_i, x_i, and V denotes the voxel position, the latent vector, and the number of voxels, respectively. The generation in both stages adopts rectified flow transformers Liu et al. (2022) with DINO-encoded image features as conditions. The result of SLAT Flow is then decoded into 3D outputs represented by radiance fields (RF), 3D Gaussians (3DGS), or meshes, *i.e.*, $O = \text{Dec}(x)$. Modeling the backward process as a time-dependent vector field $\boldsymbol{v}(x, t) = \nabla_t(x)$, the transformers $\boldsymbol{v}_\theta$ in both stages are trained by minimizing the conditional flow matching (CFM) objective Lipman et al. (2023):

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,x_0,\epsilon}\|\boldsymbol{v}_\theta(x, t) - (\epsilon - x_0)\|_2^2. \tag{2}$$

Overview Our ReconViaGen framework conducts reconstruction and generation simultaneously and utilizes the two priors in a complementary fashion. It builds upon TRELLIS to generate complete 3D outputs with strong generation priors to plausibly hallucinate invisible portions to compensate for the limitation of reconstruction. The proposed ReconViaGen adopts a coarse-to-fine reconstruction pipeline. As shown in Fig. 2, in the first stage, we use a pre-trained VGGT to provide reconstruction-based multi-view conditions at both the global and local levels. In the next stage, we respectively feed the global geometry and local per-view conditions into the SS and SLAT Flow transformers, for multi-view-aware generation. Finally, we further refine the estimated camera poses from VGGT using the generation and introduce pixel-level alignment constraints only in the inference stage for reconstructions highly consistent with input views in detailed geometry and textures.

3.2 Reconstruction-based Conditioning

We first introduce reconstruction priors in VGGT to provide strong multi-view-aware conditions for the coarse and detailed shape and texture generation of TRELLIS.

Global Geometry Condition VGGT learns a strong reconstruction prior to encode explicit 3D lifting information into multi-view image features. Therefore, we first aggregate VGGT features ϕ_{vggt} into a global geometry representation, serving as SS Flow conditions to generate more accurate coarse structures. Note that we did not use explicit reconstruction results like point clouds because VGGT features convey richer information including camera poses, depth, point maps, and tracking. A fixed-length token list T_g is aggregated from ϕ_{vggt} via a proposed Condition Net design shown in Fig. 2. Starting from a randomly initialized learnable token list T_{init}, four transformer cross-attention blocks progressively fuse layer-wise features of ϕ_{vggt} with the initial token list and produce T_g. Formulated as:

$$T^{i+1} = \text{CrossAttn}\big(Q(T^i), K(\phi_{\text{vggt}}), V(\phi_{\text{vggt}})\big), \quad i \in \{0, 1, 2, 3\}, \tag{3}$$

where T^0 is initilized with T_{init}, T^3 is the final output T_g, $Q(\cdot)$, $K(\cdot)$, and $V(\cdot)$ are linear layers respectively for query, key, and value projection, and ϕ_{vggt} is the VGGT features that concatenate all views on the token dimension. At the training stage of SS Flow, we freeze the VGGT layers and train the Condition Net together with DiT.

Local Per-View Condition A single token list condition can provide limited fine-grained information for geometry and texture generation in detail. We further adopt the Condition Net design to provide local per-view tokens as SLAT Flow conditions for fine-grained generation in both geometric and texture details. A random token list is initialized for each view and fed into the Condition Net to produce a view-specific token list $P_k, k \in [1, N]$:

$$P_k^{i+1} = \text{CrossAttn}\big(Q(P_k^i), K(\phi_k^{\text{vggt}}), V(\phi_k^{\text{vggt}})\big), \quad i \in \{0, 1, 2, 3\} \text{ and } k \in \{n\}_{n=1}^{N}, \tag{4}$$

where ϕ_k^{vggt} is VGGT features of the k-th view. The set of $\{P_k\}_{k=1}^{N}$ is sent into SLAT diffusion transformers offering per-view object appearance guidance for fine-grained generation.

3.3 Coarse-to-Fine Generation

The overall generation process consists of three stages: (i) coarse structure generation via SS Flow with global geometry condition; (ii) fine detail generation via SLAT Flow with local per-view condition; (iii) rendering-aware pixel-aligned refinement at the inference stage only.

Reconstruction-conditioned Flow To integrate the reconstruction prior into generation, the two stages of SS and SLAT Flow in TRELLIS take the global geometry condition T_g and local per-view conditions $\{P_k\}_{k=1}^{N}$ for coarse and fine diffusion guidance, respectively. In the first stage, we simply compute the cross-attention between the condition T_g and the noisy SS latent in each SS DiT block. In the second stage, as illustrated in Fig. 2, we encourage the cross-attention between the noisy SLAT and each view's condition P_k and conduct a weighted fusion in each SLAT DiT block, which can be formulated as:

$$y_{j+1} = \sum_{k=1}^{N} \text{CrossAttn}\big(Q(y'_j), K(P_k), V(P_k)\big) \cdot w_k, \quad j \in \{m\}_{m=1}^{M}, \tag{5}$$

where M is the number of SLAT DiT blocks, y'_j is the self-attention layer output of the noisy SLAT input y_j, and $w_k \in (0, 1)$ is the fusion weight computed via an MLP taking the cross-attention result

as input. After the first two stages, the 3D generator can generate multi-view-aware geometry and texture at both the global and local level.

Rendering-aware Velocity Compensation To further encourage pixel-aligned consistency between generation results and input views, we develop a rendering-aware velocity compensation to constrain the diffusion trajectory according to inputs. In doing so, we first estimate camera pose with VGGT using the generation results from the second stage, with detailed implementation details included in the appendix. Inspired by the explicit normal regularization used in Hi3DGen Ye et al. (2025) to improve the input-output consistency, when $t < 0.5$, we decode the SLAT into O_t (*e.g.* a textured mesh) and conduct rendering for alignment. The SLAT Flow process initializes and updates a large number of noisy latents for all voxels simultaneously, which results in a challenging collaborative optimization problem. To solve this issue, we novelly propose a mechanism called Rendering-aware Velocity Compensation (RVC) to correct the predicted v for a more accurate generation consistent with input views. Specifically, we render images for O_t from the refined camera pose estimations C and calculate the difference between the rendered images and input images as:

$$\mathcal{L}_{\text{RVC}}(v_t) = \mathcal{L}_{\text{SSIM}} + \mathcal{L}_{\text{LPIPS}} + \mathcal{L}_{\text{DreamSim}}, \tag{6}$$

where $\mathcal{L}_{\text{SSIM}}$, $\mathcal{L}_{\text{LPIPS}}$, and $\mathcal{L}_{\text{DreamSim}}$ are SSIM Wang et al. (2004), LPIPS Zhang et al. (2018), and DreamSim losses Fu et al. (2023) (inspired by the practice in V2M4 Chen et al. (2025)), responsible for measuring the structural, perceptual, and semantic similarity, respectively. To exclude the influence of inaccurate pose estimation, we discard the losses corresponding to some images if their corresponding losses are higher than 0.8. By minimizing $\mathcal{L}_{\text{RVC}}$, we iteratively correct the predicted velocity in each SLAT denoising step with a compensation term Δv, derived as:

$$\Delta v_t = \frac{\partial \mathcal{L}}{\partial \hat{x_0}} \frac{\partial \hat{x_0}}{\partial v_t} = -t \frac{\partial \mathcal{L}}{\partial \hat{x_0}}, \tag{7}$$

where $\mathcal{L}$ represents $\mathcal{L}_{\text{RVC}}$ for simplicity and $\hat{x_0}$ is the predicted target SLAT at current timestep t, computed as $\hat{x_0} = x_t - t \cdot v_t$. The noisy SLAT of next step $x_{t_{\text{prev}}}$ can be updated as:

$$x_{t_{\text{prev}}} = x_t - (t - t_{\text{prev}})(v + \alpha \cdot \Delta v), \tag{8}$$

where α is a pre-defined hyperparameter that controls the extent of the compensation. In this way, the input images serve as a strong explicit guidance to find a denoising trajectory for each local SLAT vector, which leads to more accurate 3D results consistent with all input images in detail.

4 Experiments

4.1 Experiment Setup

Datasets For LoRA fine-tuning of the VGGT aggregator and Trellis sparse structure transformer, we employ 390k 3D data from the Objaverse dataset Deitke et al. (2024), a large-scale 3D object dataset that provides a rich variety of shapes and textures, with 60 views rendered per object for fine-tuning. For each object mesh, we render 150 view images in a resolution of 512×512 under uniform lighting conditions following TRELLIS Xiang et al. (2024). For evaluation, we selected two benchmark datasets to thoroughly assess the performance of our model: (i) Dora-Bench Chen et al. (2024), a benchmark organized based on 4 levels of complexity, combining 3D data selected from the Obajverse Deitke et al. (2023), ABO Collins et al. (2022), and GSO Downs et al. (2022) dataset; and (ii) OmniObject3D, a large-vocabulary 3D object dataset containing 6,000 high-quality textured meshes scanned from real-world objects, covering 190 daily categories. We randomly sample 300 objects from Dora-Bench and 200 objects covering 20 categories from OmniObject3D. We follow He et al. (2024) to render 24 views at different elevations, and randomly chose 4 of them as multi-view input for evaluation on OmniObject3D. On Dora-Bench, we follow the camera trajectory of TRELLIS Xiang et al. (2024) to render 40 views and choose 4 views (No.0, 9, 19, and 29) with a uniform interval to adapt to the setting of some baseline methods (LGM Tang et al. (2025) and InstantMesh Xu et al. (2024b)).

Evaluation Metrics We employ PSNR, SSIM, and LPIPS to evaluate the accuracy of synthesized novel views from 3D outputs, Chamfer Distance (CD) and F-score to evaluate the generated geometry accuracy and completeness. PSNR, SSIM, and LPIPS are evaluated on novel views of images rendered at the resolution of 512×512. CD and F-score are evaluated by sampling 100k points from the 3D outputs (using the center positions for 3D Gaussian outputs), with all object points normalized to the range of $[-1, 1]^3$. When calculating the F-score, the radius r is set to 0.1.

Table 1: Evaluation on the Dora-bench and OminiObject3D dataset. Best results are in **bold**.

Method	Dora-bench					OminiObject3D				
	PSNR↑	SSIM↑	LPIPS↓	CD↓	F-score↑	PSNR↑	SSIM↑	LPIPS↓	CD↓	F-score↑
VGGT Wang et al. (2025a)	-	-	-	0.112	0.921	-	-	-	0.091	0.900
TRELLIS-S Xiang et al. (2024)	16.562	0.876	0.103	0.176	0.807	16.021	0.771	0.264	0.102	0.906
TRELLIS-M Xiang et al. (2024)	16.706	0.882	0.111	0.144	0.843	16.861	0.790	0.242	0.072	0.932
Hunyuan3D-2.0-mv Zhao et al. (2025)	20.221	0.896	0.093	0.094	0.937	16.665	0.813	0.165	0.124	0.871
LGM Tang et al. (2025)	17.877	0.869	0.186	0.121	0.839	16.361	0.791	0.193	0.136	0.842
InstantMesh Xu et al. (2024b)	18.922	0.870	0.120	0.110	0.865	17.499	0.818	0.145	0.094	0.907
LucidFusion He et al. (2024)	16.509	0.835	0.144	0.131	0.831	16.254	0.771	0.144	0.114	0.868
ReconViaGen (Ours)	**22.632**	**0.911**	**0.090**	**0.090**	**0.953**	**19.767**	**0.847**	**0.141**	**0.059**	**0.959**

Baseline Methods Baseline methods for comparisons include (i) TRELLIS-S Xiang et al. (2024): generates 3D meshes from multi-view images using TRELLIS in the stochastic mode, which randomly chooses one input view to condition each step of denoising; (ii) TRELLIS-M Xiang et al. (2024): TRELLIS in the multidiffusion mode, which computes the average denoised results conditioned on all input views; (iii) Hunyuan3D-2.0-mv Zhao et al. (2025): concatenate DINO features of input images from fixed viewpoints as conditions to generate meshes[1]; (iv) InstantMesh Xu et al. (2024b): predicts Triplane for mesh outputs from multiple images with fixed viewpoints; (v) LGM Tang et al. (2025): predicts pixel-aligned 3D Gaussians from multiple images with fixed viewpoints; (vi) LucidFusion He et al. (2024): predicts relative coordinate maps for 3D Gaussian outputs; (vii) VGGT Wang et al. (2025a): reconstructs the point cloud from multi-view inputs in a feed-forward manner. We compare our methods with a wide range of existing SOTA baseline methods: (a) 3D generation models {i, ii, iii}; (b) large reconstruction models with known camera poses {iv, v}; (c) pose-free large reconstruction models with 3DGS or point cloud outputs {vi, vii}. For 3D generation models {i, ii, iii}, we use the same approach as Camera Pose Estimation in fine detail reconstruction to align the generated 3D models to the ground-truth models. Besides, we also compare with closed-source commercial 3D generation models like Hunyuan3D-2.5 and Meshy-5 on in-the-wild testing.

Implementation Details For LoRA fine-tuning of VGGT aggregator and TRELLIS transformer, we set the rank as 64, the alpha parameter for LoRA scaling as 128, and the dropout probability for LoRA layers as 0. We only apply the adapter to qkv mapping layer and the projectors of each attention layer. During fine-tuning VGGT aggregator, we randomly sample $1 \sim 4$ views from 150 images and use the AdamW optimizer with a fixed learning rate of 1×10^{-4}. For the fine-tuning of SS Flow and Slat-Flow transformer, we build upon TRELLIS Xiang et al. (2024), incorporating classifier-free guidance (CFG) with a drop rate of 0.3 and an AdamW optimizer with a fixed learning rate of 1×10^{-4}. We fine-tune the SS-Flow transformer using 8 NVIDIA A800 GPUs (80GB memory) for 40k steps with a batch size of 192. Differently, we finetune the SLat-Flow transformer with a batch size of 128. During inference, we set the CFG strengths in SS generation and SLAT generation to 7.5 and 3.0, and use 30 and 12 sampling steps to achieve optimal results. The α in rendering-aware velocity compensation is set to 0.1 in our practice.

4.2 Experiment Results

Quantitative Results We present the quantitative comparisons between our ReconViaGen and other baseline methods in Tab. 1 for evaluation on the Dora-bench and OminiObject3D dataset. The proposed method achieves consistently superior performance to other methods on both image-reconstruction consistency (PSNR, SSIM, and LPIPS), geometry accuracy (CD), and shape completeness (F-score). Impressively, our ReconViaGen seamlessly integrate the generation and reconstruction priors from TRELLIS Xiang et al. (2024) and VGGT Wang et al. (2025a), whose performance surpasses both of them. Note that VGGT performs better on Dora-bench than on OminiObject3D because uniformly-distributed views can capture richer visual cues than random views. Besides, our method also gets better results than previous SOTA pose-free multi-view reconstruction methods that integrate regression-based generation priors by a large margin, especially on PSNR, CD, and F-score, which validates the superiority of ReconViaGen. On the settings of more input views, we separately evaluate VGGT that can accept an arbitrary number of inputs for comparison, which is included in the appendix. We also present the camera pose estimation accuracy in the appendix.

[1]The fresh version, Hunyuan3D-2.5, has not been open-sourced, which is unsuitable for large-scale evaluation on benchmarks, so we use the open-sourced version, Hunyuan3D-2.0.

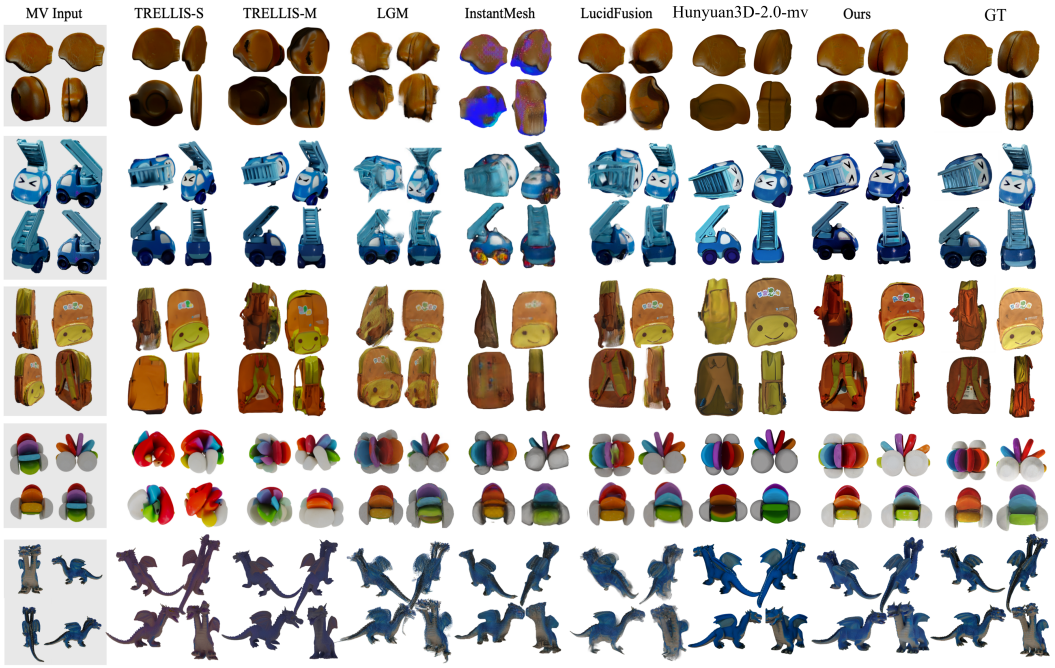

Figure 3: Reconstruction result comparisons between our ReconViaGen and other baseline methods on samples from the Dora-bench and OminiObject3D datasets. Zoom in for better visualization.

Figure 4: Reconstruction results on in-the-wild samples. Note that commercial 3D generators require input images from orthogonal viewpoints, while ours can accept views from arbitrary camera poses for robust outputs. Zoom in for better visualization in detail.

Qualitative Results We further present extensive qualitative comparisons to demonstrate the superiority of our ReconViaGen. We first select some examples from the OminiObject3D and Dora-bench dataset for visualization, as shown in Fig. 3. The reconstruction results of ReconViaGen have the most accurate geometry and textures compared to other methods. We further evaluate several baseline methods on in-the-wild multi-view images. As shown in Fig. 4, our ReconViaGen exhibits strong robustness even in comparison with the multi-view version of closed-source commercial 3D

Figure 5: Qualitative comparisons for different variants of ReconViaGen for ablative study. Zoom in for better visualization in detail.

generation models like Hunyuan3D-2.5 and Meshy-5. More qualitative results are included in the appendix.

4.3 Ablation Study

The proposed ReconViaGen framework comprises three novel designs to integrate reconstruction priors into the diffusion-based 3D generation: (i) the global geometry condition (GGC); (ii) the per-view condition (PVC); and (iii) the rendering-aware velocity compensation (RVC). We conduct ablation studies to validate the individual effectiveness of each component. On the Dora-bench dataset, we start from a basic TRELLIS-M baseline (ReconViaGen without all designs, Tab. 2a) and progressively add one component, leading to 3 variants (b,c,d). As shown in Tab. 2, integrating GGC, which strongly improves the prediction accuracy of coarse structure, brings a large performance gain on almost all metrics. Further integrating PVC can lead to extra improvement, especially on PSNR, which proves the effectiveness in improving local per-view alignment. Finally, adopting RVC, though in the inference stage only, brings additional increments in both shape completeness and fine-grained accuracy in geometry and texture. Qualitative comparisons in Fig. 5 visualize the positive effect of each component: global geometry conditioning greatly corrects the global shape, per-view conditioning produces local details in geometry and texture of high consistency with each view, and rendering-aware velocity compensating impressively refines the fine-grained appearance, leading to high-quality results. More ablation results on the detailed designs, including the number of image inputs and the choice of condition form for SS and SLAT Flow, can be seen in the appendix.

Table 2: Quantitative ablation results on the Dora-bench dataset.

	GGC	PVC	RVC	PSNR↑	SSIM↑	LPIPS↓	CD↓	F-score↑
(a)	✗	✗	✗	16.706	0.882	0.111	0.144	0.843
(b)	✓	✗	✗	20.462	0.894	0.102	0.093	0.941
(c)	✓	✓	✗	21.045	0.905	0.093	0.093	0.937
(d)	✓	✓	✓	**22.632**	**0.911**	**0.090**	**0.090**	**0.953**

5 Conclusion

In this paper, we have presented ReconViaGen, a novel coarse-to-fine framework that effectively integrates strong reconstruction priors with diffusion-based 3D generative priors for accurate and complete multi-view 3D object reconstruction. We first analyze the inherent reasons leading to the challenge of leveraging diffusion-based 3D generative priors into reconstruction: insufficient cross-view correlation modeling and stochastic denoising process with weak constraint from input images. Therefore, we effectively use powerful reconstruction priors with three novelly designed mechanisms to enhance the multi-view correlation awareness in 3D diffusion learning and establish strong constraints for a reliable denoising process. Extensive experiments have demonstrated that ReconViaGen achieves SOTA performance in both global shape accuracy and completeness as well as local details in geometry and textures. As future work, with the development of 3D reconstruction and 3D generation, stronger reconstruction or generation priors can be integrated into our framework to further improve reconstruction quality via generation.

REFERENCES

Ziang Cao, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Physx-3d: Physical-grounded 3d asset generation. *arXiv preprint arXiv:2507.12465*, 2025.

Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *ICCV*, pp. 2416–2425, 2023.

Jianqi Chen, Biao Zhang, Xiangjun Tang, and Peter Wonka. V2m4: 4d mesh animation reconstruction from a single monocular video. *arXiv preprint arXiv:2503.09631*, 2025.

Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *ICCV*, pp. 1538–1547, 2019.

Rui Chen, Jianfeng Zhang, Yixun Liang, Guan Luo, Weiyu Li, Jiarui Liu, Xiu Li, Xiaoxiao Long, Jiashi Feng, and Ping Tan. Dora: Sampling and benchmarking for 3d shape variational auto-encoders. *arXiv preprint arXiv:2412.17808*, 2024.

Shuo Cheng, Zexiang Xu, Shilin Zhu, Zhuwen Li, Li Erran Li, Ravi Ramamoorthi, and Hao Su. Deep stereo using adaptive thin volume representation with uncertainty awareness. In *CVPR*, pp. 2524–2534, 2020.

Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *CVPR*, pp. 21126–21136, 2022.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *ICCV*, pp. 13142–13153, 2023.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *NeurIPS*, 36, 2024.

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *ICRA*, pp. 2553–2560, 2022.

Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981. doi: 10.1145/358669.358692.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. volume 36, pp. 50742–50768, 2023.

Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015.

Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Massively parallel multiview stereopsis by surface normal diffusion. In *ICCV*, pp. 873–881, 2015.

Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *CVPR*, pp. 2495–2504, 2020.

Hao He, Yixun Liang, Luozhou Wang, Yuanhao Cai, Xinli Xu, Hao-Xiang Guo, Xiang Wen, and Yingcong Chen. Lucidfusion: Generating 3d gaussians with arbitrary unposed images. *arXiv preprint arXiv:2410.15636*, 2024.

Xianglong He, Zi-Xin Zou, Chia-Hao Chen, Yuan-Chen Guo, Ding Liang, Chun Yuan, Wanli Ouyang, Yan-Pei Cao, and Yangguang Li. Sparseflex: High-resolution and arbitrary-topology 3d shape modeling. *arXiv preprint arXiv:2503.21732*, 2025.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. In *ICLR*, 2023.

Ka-Hei Hui, Ruihui Li, Jingyu Hu, and Chi-Wing Fu. Neural wavelet-domain diffusion for 3d shape generation. In *SIGGRAPH Aisa*, pp. 1–9, 2022.

Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 406–413. IEEE, 2014.

Hanwen Jiang, Zhenyu Jiang, Yue Zhao, and Qixing Huang. Leap: Liberate sparse-view 3d modeling from camera poses. *arXiv preprint arXiv:2310.01410*, 2023.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *TOG*, 42(4):139–1, 2023.

Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o(n) solution to the pnp problem. *IJCV*, 81(2):155–166, 2009. doi: 10.1007/s11263-008-0152-6.

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

Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023.

Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *arXiv preprint arXiv:2405.11616*, 2024a.

Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024b.

Yangguang Li, Zi-Xin Zou, Zexiang Liu, Dehu Wang, Yuan Liang, Zhipeng Yu, Xingchao Liu, Yuan-Chen Guo, Ding Liang, Wanli Ouyang, et al. Triposg: High-fidelity 3d shape synthesis using large-scale rectified flow models. *arXiv preprint arXiv:2502.06608*, 2025a.

Zhihao Li, Yufei Wang, Heliang Zheng, Yihao Luo, and Bihan Wen. Sparc3d: Sparse representation and construction for high-resolution 3d shapes modeling. *arXiv preprint arXiv:2505.14521*, 2025b.

Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *ICCV*, pp. 5741–5751, 2021.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, pp. 300–309, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023.

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *NeurIPS*, 36, 2023a.

Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *CVPR*, pp. 10072–10083, 2024.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, pp. 9298–9309, 2023b.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*, pp. 2837–2845, 2021.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pp. 405–421, 2020.

Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulo, Peter Kontschieder, and Matthias Nießner. Diffrf: Rendering-guided 3d radiance field diffusion. In *CVPR*, pp. 4328–4338, 2023.

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pp. 1–31, 2024.

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

Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *CVPR*, pp. 9914–9925, 2024.

Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, pp. 501–518. Springer, 2016.

J Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *CVPR*, pp. 20875–20886, 2023.

Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. *arXiv preprint arXiv:2408.13912*, 2024.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. In *ICLR*, 2024.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *ECCV*, pp. 1–18. Springer, 2025.

Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *ICCV*, pp. 22819–22829, 2023a.

Zhicong Tang, Shuyang Gu, Chunyu Wang, Ting Zhang, Jianmin Bao, Dong Chen, and Baining Guo. Volumediffusion: Flexible text-to-3d generation with efficient volumetric encoder. *arXiv preprint arXiv:2312.11459*, 2023b.

Tencent Hunyuan3D Team. Hunyuan3d 2.0: High-resolution 3d asset generation. https://3d-models.hunyuan.tencent.com, 2025.

Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *CVPR*, pp. 14194–14203, 2021a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025a.

Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023a.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, pp. 20697–20709, 2024a.

Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *CVPR*, pp. 4563–4573, 2023b.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *NeurIPS*, 36, 2024b.

Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. In *ECCV*, pp. 57–74. Springer, 2025b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *TIP*, 13(4):600–612, 2004. doi: 10.1109/TIP.2004.827628.

Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. 2021b.

Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Meshlrm: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385*, 2024.

Chin-Hsuan Wu, Yen-Chun Chen, Bolivar Solarte, Lu Yuan, and Min Sun. ifusion: Inverting diffusion for pose-free reconstruction from sparse views. *arXiv preprint arXiv:2312.17250*, 2023a.

Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *arXiv preprint arXiv:2405.20343*, 2024a.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *arXiv preprint arXiv:2405.14832*, 2024b.

Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Xun Cao, Philip Torr, et al. Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. *arXiv preprint arXiv:2505.17412*, 2025.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *CVPR*, pp. 803–814, 2023b.

Zhennan Wu, Yang Li, Han Yan, Taizhang Shang, Weixuan Sun, Senbo Wang, Ruikai Cui, Weizhe Liu, Hiroyuki Sato, Hongdong Li, et al. Blockfusion: Expandable 3d scene generation using latent tri-plane extrapolation. *TOG*, 43(4):1–17, 2024c.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024.

Chao Xu, Ang Li, Linghao Chen, Yulin Liu, Ruoxi Shi, Hao Su, and Minghua Liu. Sparp: Fast 3d object reconstruction and pose estimation from sparse views. In *ECCV*, pp. 143–163. Springer, 2024a.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024b.

Jiale Xu, Shenghua Gao, and Ying Shan. Freesplatter: Pose-free gaussian splatting for sparse-view 3d reconstruction. *arXiv preprint arXiv:2412.09573*, 2024c.

Qingshan Xu and Wenbing Tao. Multi-scale geometric consistency guided multi-view stereo. In *CVPR*, pp. 5483–5492, 2019.

Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024d.

Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. Dmv3d: Denoising multi-view diffusion using 3d large reconstruction model. In *ICLR*, 2024e.

Jiayu Yang, Wei Mao, Jose M Alvarez, and Miaomiao Liu. Cost volume pyramid based depth inference for multi-view stereo. In *CVPR*, pp. 4877–4886, 2020.

Xianghui Yang, Huiwen Shi, Bowen Zhang, Fan Yang, Jiacheng Wang, Hongxu Zhao, Xinhai Liu, Xinzhou Wang, Qingxiang Lin, Jiaao Yu, et al. Hunyuan3d 1.0: A unified framework for text-to-3d and image-to-3d generation. *arXiv preprint arXiv:2411.02293*, 2024.

Yunhan Yang, Yufan Zhou, Yuan-Chen Guo, Zi-Xin Zou, Yukun Huang, Ying-Tian Liu, Hao Xu, Ding Liang, Yan-Pei Cao, and Xihui Liu. Omnipart: Part-aware 3d generation with semantic decoupling and structural cohesion. *arXiv preprint arXiv:2507.06165*, 2025.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *ECCV*, pp. 767–783, 2018.

Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *CVPR*, pp. 5525–5534, 2019.

Chongjie Ye, Yushuang Wu, Ziteng Lu, Jiahao Chang, Xiaoyang Guo, Jiaqing Zhou, Hao Zhao, and Xiaoguang Han. Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. *arXiv preprint arXiv:2503.22236*, 3, 2025.

Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *TOG*, 42(4):1–16, 2023.

Bowen Zhang, Yiji Cheng, Jiaolong Yang, Chunyu Wang, Feng Zhao, Yansong Tang, Dong Chen, and Baining Guo. Gaussiancube: Structuring gaussian splatting using optimal transport for 3d generative modeling. *arXiv e-prints*, pp. arXiv–2403, 2024a.

Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *TOG*, 43(4):1–20, 2024b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *NeurIPS*, 36, 2024.

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *ICCV*, pp. 5826–5835, 2021.

Qi Zuo, Xiaodong Gu, Lingteng Qiu, Yuan Dong, Zhengyi Zhao, Weihao Yuan, Rui Peng, Siyu Zhu, Zilong Dong, Liefeng Bo, et al. Videomv: Consistent multi-view generation based on large video generative model. *arXiv preprint arXiv:2403.12010*, 2024.

A Appendix

A.1 Details on camera pose estimation

For better alignment with the input images, we register them into the TRELLIS generation space. Specifically, we first render 30 images from randomly sampled camera views on a sphere, concatenate them with the input images, and feed them into VGGT for pose estimation. Since the camera poses of the rendered views are known, we can recover coarse camera poses for the input images in the TRELLIS space. While VGGT provides robust pose predictions, they remain insufficiently accurate for constructing pixel-level rendering constraints.

To refine the results, we render images and depth maps using the coarse poses, then apply an image matching method to establish 2D-2D correspondences between rendered and input images. Leveraging the depth maps and camera parameters of rendered views, we further obtain 2D-3D correspondences between each input image and the generated object. By aggregating multi-view correspondences, we solve for refined camera poses C using a PnP Lepetit et al. (2009) solver with RANSAC Fischler & Bolles (1981). This image-matching-based refinement effectively corrects the initial pose predictions from TRELLIS's generative priors, yielding higher accuracy. The refined poses enable pixel-wise constraints from the input views, thereby supporting finer detail alignment in generation.

A.2 Evaluation with more input images

Table 3: Evaluation with more input images on the Dora-bench dataset. Best results are in **bold**.

Method	**Uniform** (PSNR↑ / LPIPS↓) 6 views	8 views	10 views	**Limited View** (PSNR↑ / LPIPS↓) 6 views	8 views	10 views
Object VGGT + 3DGS	18.476/0.123	19.890/0.109	21.363/0.102	16.498/0.139	16.774/0.135	17.121/0.133
ReconViaGen (Ours)	**22.823/0.089**	**23.067/0.090**	**23.193/0.087**	**21.427/0.098**	**21.782/0.099**	**21.866/0.103**

To thoroughly evaluate and validate the effectiveness of ReconViaGen, we compare it against 3DGS reconstruction initialized with point clouds and camera poses from object VGGT (denoted as object VGGT + 3DGS) on the Dora-Bench dataset. We conduct experiments under two input scenarios: uniformly and limited-view sampled views. As shown in Tab. 3, ReconViaGen consistently outperforms object VGGT+3DGS at 6/8/10 input views, regardless of the sampling strategy. This advantage arises because the generative prior in ReconViaGen plays a crucial role in completing invisible regions of the object.

A.3 Evaluation of Camera Pose Estimation

Table 4: Evaluation of camera pose estimation on the Dora-bench dataset. Best results are in bold.

Method	RRE↓	Acc.@15° ↑	Acc.@30° ↑	TE↓
VGGT Wang et al. (2025a)	8.575	90.67	92.00	0.066
Object VGGT	**7.257**	93.44	94.11	0.055
Ours	7.925	**93.89**	**96.11**	**0.046**

To assess the performance of our finetuned object VGGT and the effectiveness of our proposed camera pose estimation strategy, we evaluate pose prediction quality on the Dora-Bench dataset. We adopt both rotation and translation metrics: relative rotation error (RRE, in degrees), the proportion of RRE values below 15° and 30°, and translation error (TE), measured as the distance between predicted and ground-truth camera centers. For evaluation, we use four input images and transform both predicted and ground-truth poses into the coordinate system of the first image, which is excluded from the metric computation. To address translation scale ambiguity, we compute relative translations between views for both predictions and ground truth and normalize them by their respective mean L2-norm. As reported in Tab. 4, the finetuned object VGGT achieves clear improvements over the original VGGT. Our method further delivers the best overall performance, as the generative prior effectively 'densifies' sparse views. However, our RRE is slightly higher than that of object VGGT, likely due to minor discrepancies between the generated 3D model and the ground-truth geometry.

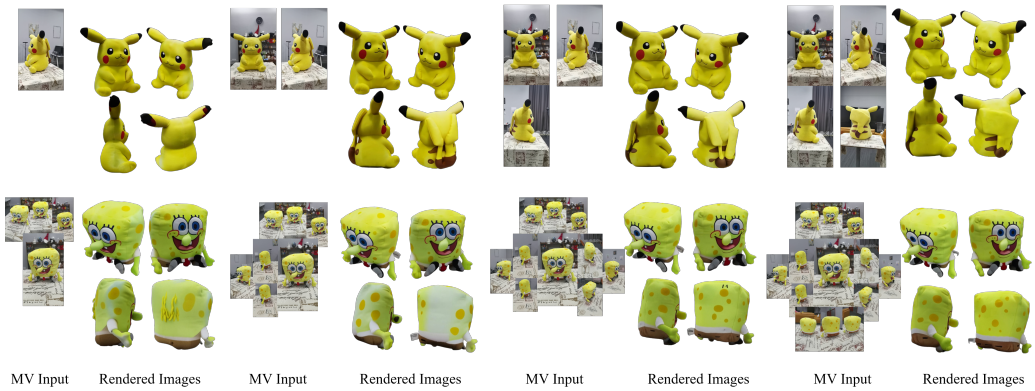

Figure 6: Qualitative comparisons for different numbers of input images with ReconViaGen. Zoom in for better visualization in detail.

A.4 Ablation Study on the Number of Input Images

Table 5: Quantitative ablation results of the number of input images on the Dora-bench dataset.

Number of Images	PSNR↑	SSIM↑	LPIPS↓	CD↓	F-score↑
1	18.438	0.887	0.106	0.135	0.838
2	19.568	0.894	0.099	0.131	0.867
4	22.632	0.911	0.090	0.090	0.953
6	22.823	0.912	**0.089**	0.084	0.958
8	**23.067**	**0.914**	0.090	**0.081**	**0.961**

Since ReconViaGen can take an arbitrary number of input images, a natural question is how reconstruction quality scales with the number of images. To investigate this, we conducted an ablation study varying the number of input views on Dora-Bench, with results summarized in Tab. 5. We observe that reconstruction performance consistently improves as more images are provided. However, the marginal gains gradually diminish, indicating a saturation effect when the number of views becomes large. The visualization results are shown in Fig. 6, which also shows that our ReconViaGen can process any number of input images from any viewpoint.

A.5 Ablation Study on the Form of Condition

Table 6: Quantitative ablation results of condition at SS Flow on the Dora-bench dataset.

	Form of Condition	PSNR↑	SSIM↑	LPIPS↓	CD↓	F-score↑
(i)	Feature Volume	16.229	0.858	0.126	0.172	0.814
(ii)	Concatenation	19.749	0.871	0.137	0.121	0.873
(iii)	PVC	19.878	0.882	0.135	0.120	0.870
(iv)	GGC	**20.462**	**0.894**	**0.102**	**0.093**	**0.941**

For SS Flow, as described in the method section, we explored several strategies to leverage VGGT features for sparse structure generation on Dora-Bench: (i) Downsampling the point cloud from VGGT to a 64^3 resolution occupancy volume, projecting DINO features from each view into the volume, and averaging them to form a feature-volume condition; (ii) Fusing VGGT features with DINO features for each view through several linear layers, then concatenating all input-view tokens as condition; (iii) adopting the same local per-view condition (PVC) used in our SLAT Flow; (iv) employing the proposed global geometry condition (GGC). For fair comparison, we use the original SLat Flow in TRELLIS and train all models for 40k steps. As shown in Tab. 6, our GGC achieves the best performance among all strategies. We attribute this to the limitations of the alternative designs: for (i), inaccurate predicted poses or point clouds lead to erroneous projections, introducing noise into the condition and harming generation; for (ii) and (iii), view-level features are not effectively aggregated, resulting in redundancy and making the model overly dependent on the accuracy of VGGT outputs.

Table 7: Quantitative ablation results of condition at SLAT Flow on the Dora-bench dataset.

	Form of Condition	PSNR↑	SSIM↑	LPIPS↓	CD↓	F-score↑
(i)	GGC	17.784	0.858	0.120	0.097	0.939
(ii)	PVC	**22.632**	**0.911**	**0.090**	**0.090**	**0.953**

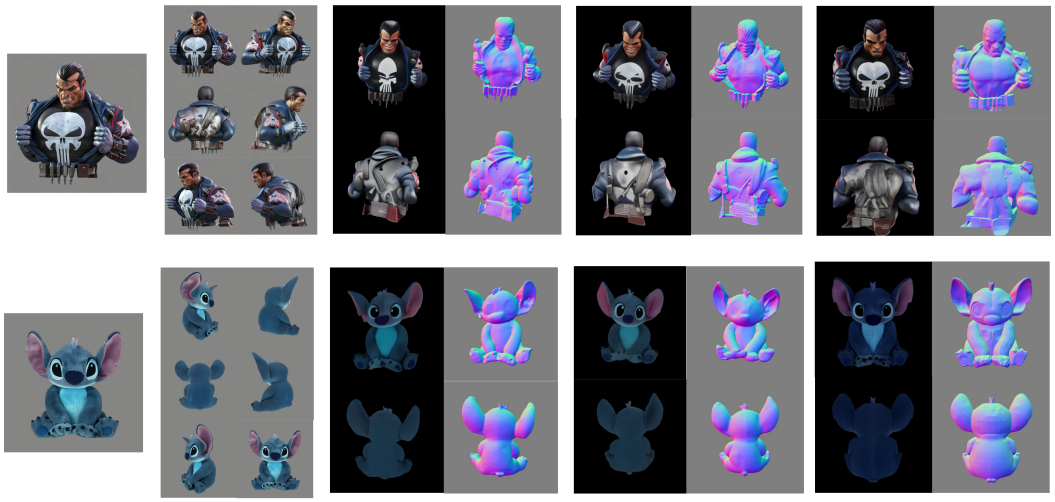

Figure 7: Reconstruction result comparisons between TRELLIS-M, TRELLIS-S, and our ReconViaGen on samples produced by the multi-view image generator.

For SLat Flow, we conduct an ablation study on Dora-Bench with two conditioning strategies: (i) the same global geometry condition (GGC) used in SS Flow, and (ii) the local per-view condition (PVC). For fairness, we pair both variants with SS Flow conditioned on GGC and train all models for 40k steps. As shown in Tab. 7, PVC substantially outperforms GGC in SLat Flow. We attribute this to the information compression in GGC, which leads to a loss of fine-grained details in the condition and degrades performance. This observation also explains why we adopt PVC instead of GGC for SLat Flow.

The following description provides an interpretability of GGC and PVC. GGC performs global multi-view aggregation, which strengthens overall structural consistency—this is exactly what SS needs, since SS focuses on coarse geometry where global cues dominate and multi-view compression is acceptable. In contrast, SLAT requires fine-grained appearance and local geometry refinement on a fixed coarse shape. PVC's per-view feature interaction preserves high-frequency cues that GGC would smooth out. Therefore, GGC naturally aligns with global structure formation, while PVC is better suited for detail-aware refinement.

A.6 Reconstruction on generated multi-view images or videos

Table 8: Quantitative comparison of generated multi-view images on the Dora-bench dataset.

Method	PSNR↑	SSIM↑	LPIPS↓	CD↓	F-score↑
ours (generated 6-view)	14.379	0.808	0.226	0.190	0.723
ours (single-view)	18.438	0.887	0.106	0.135	0.838
ours (real 6-view)	**22.823**	**0.912**	**0.089**	**0.084**	**0.958**

Given the growing interest in multi-view image generation using large image and video generative models, we further evaluate the robustness of our approach on such generated data. These multi-view images are hallucinated from a single view and often suffer from cross-view inconsistencies in fine details. Specifically, we generate 6-view samples using the open-sourced multi-view generator Hunyuan3D-1.0 Yang et al. (2024). The quantitative comparisons on Dora-Bench are summarized in Tab. 8. In our observations, the results of generated 6 views as input are much worse than those

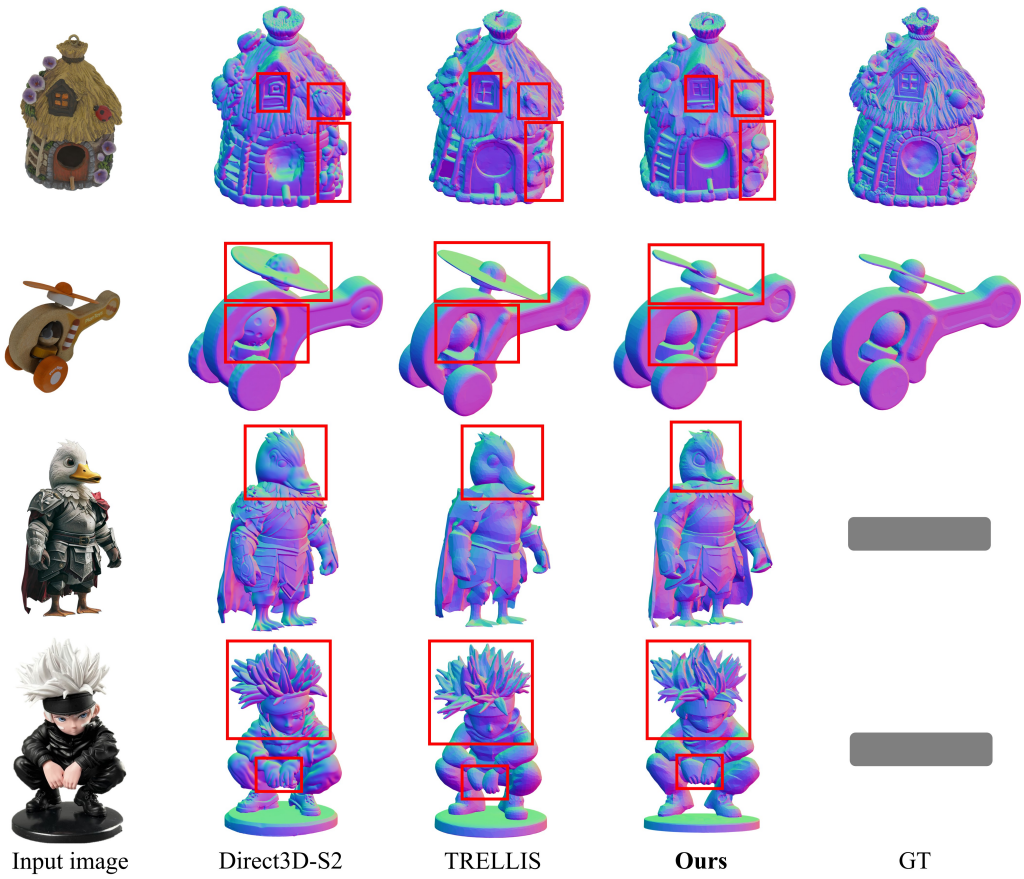

Figure 8: Qualitative result of our ReconViaGen and other baseline methods with single-view input on samples from the Dora-bench and in-the-wild scenarios. Since in-the-wild cases lack corresponding ground truth meshes, these ground truth meshes are not displayed in the figure. Zoom in for better visualization.

of single-view and real 6 view input due to much severe inconsistency in multi-view generation. However, when the inconsistency in generated multi-view images is not severe, visualizations in Fig. 7 show that ReconViaGen exhibits strong robustness to moderate cross-view inconsistency. Please refer to the supplementary video for additional results on generated videos.

A.7 More reconstruction results

We further showcase our method on in-the-wild data, including not only multiple objects but also scenes, even from generated dynamic object videos. For scene reconstruction, we segment individual objects, reconstruct them separately, and then register the reconstructed 3D objects back into the scene using our predicted camera poses. Please refer to the supplementary video for qualitative results.

A.8 Comparison experiment of single-view input

Table 9: Quantitative results of single view input on the Dora-bench dataset.

Method	PSNR↑	SSIM↑	LPIPS↓	CD↓	F-score↑
TRELLIS Xiang et al. (2024)	15.264	0.858	0.182	0.162	0.781
Direct3D-S2 Wu et al. (2025)	-	-	-	0.165	0.805
Ours	**18.438**	**0.887**	**0.106**	**0.135**	**0.838**

Figure 9: More qualitative comparisons for (c) vs (d) ablation study. Each case is labeled with a corresponding PSNR value. Zoom in for better visualization in detail.

To further demonstrate the superiority of our ReconViaGen, we conduct comparison experiments with single-view input. We select TRELLIS Xiang et al. (2024) and Direct3D-S2 Wu et al. (2025) as representative single-view generation baselines for comparison. For evaluation on the Dora-Bench dataset, we use a single randomly selected view as input. Quantitative comparisons between our ReconViaGen and the baseline methods are summarized in Tab. 9. Note that since Direct3D-S2 Wu et al. (2025) outputs geometry without color, we do not report its visual metrics. The quantitative data demonstrate that ReconViaGen consistently achieves superior performance across both visual and geometric evaluation criteria. We further provide extensive qualitative comparisons on both Dora-Bench and in-the-wild scenarios, as shown in Fig. 8. The visual evidence confirms that our proposed method consistently outperforms the baseline approaches, yielding reconstructed geometry and textures that exhibit the strongest fidelity and alignment with the input view.

A.9 More qualitative examples for the (c) vs (d) ablation study

We present more qualitative examples and label the PSNR of each case for the (c) vs (d) ablation study in Fig. 9. The qualitative analysis confirms that the 1.587 dB improvement is primarily due to RVC's ability to correct color and texture drift and enforcing high-frequency aligned with the input views.

A.10 Ablation study on RVC hyperparameters

The RVC stage in ReconViaGen involves three main hyperparameters: the extent of RVC α, timestep t, and outlier rejection threshold t_o. Then we conduct extensive ablation experiments on these three hyperparameters on the Dora-Bench dataset. As shown in Tab. 10, for α, varying α from 1 to 0.01, we observe stable performance across a wide range, with the best results around $\alpha = 0.1$. Extremely small values slightly reduce accuracy, but no instability was observed. Importantly, α

Table 10: Ablation study on the extent of RVC α on the Dora-bench dataset.

α	PSNR↑	SSIM↑	LPIPS↓	CD↓	F-score↑
1	22.475	0.911	0.091	0.091	0.941
0.5	22.151	**0.913**	0.093	**0.089**	0.949
0.1	**22.632**	0.911	**0.090**	0.090	**0.953**
0.05	21.321	0.910	0.089	0.091	**0.953**
0.01	21.079	0.906	0.095	0.092	0.941

Table 11: Ablation study on timestep t in RVC on the Dora-bench dataset.

t	PSNR↑	SSIM↑	LPIPS↓	CD↓	F-score↑	Runtime↓
0.7	22.149	0.907	0.090	0.094	0.938	9.2767
0.6	22.477	0.912	0.091	0.094	0.940	8.2999
0.5	**22.632**	0.911	0.090	0.090	**0.953**	6.8652
0.4	22.270	0.909	0.089	**0.089**	0.944	5.7155
0.3	22.057	**0.915**	**0.088**	0.091	0.937	**4.7825**

introduces no runtime overhead, as it only changes the update weight of a vector. As shown in Tab. 11, applying RVC at different timesteps shows that $t = 0.5$ offers the best trade-off between fidelity and cost. As expected, runtime scales linearly with the fraction of steps with RVC (due to decoding/rendering), ranging from 4.8s ($t = 0.3$) to 9.3s ($t = 0.7$). As shown in Tab. 12, for outlier rejection threshold t_o, results are insensitive across a broad range. We choose $t_o = 0.8$ as it yields strong PSNR/F-score. Therefore, $\alpha = 0.1$, $t = 0.5$, and $t_o = 0.8$ is the optimal combination of hyperparameters. Importantly, the best-performing configuration ($\alpha = 0.1, t = 0.5, t_o = 0.8$) is shared across all assets and datasets, and no per-asset tuning is required. This confirms that the final RVC step generalizes well and does not rely on instance-specific adjustments.

A.11 The detailed latency of each component in ReconViaGen

To demonstrate the efficiency of our method in detail, we report the inference time for each component on an NVIDIA L20 GPU (46GB memory) with different numbers of input images in Tab. 13. As shown in the table, the dominant runtime contributors are pose estimation and RVC. Nevertheless, the total inference time remains reasonable. The RVC is a justified trade-off for accuracy and efficiency. RVC provides a significant quality boost of approximately 1.587 dB PSNR compared to the pipeline without it. This gain is crucial for achieving high-fidelity 3D reconstruction and eliminating visual inconsistencies to input images. Importantly, even without pose estimation and RVC, the results still surpass current SOTA reconstruction methods, while RVC further improves fine geometry and texture alignment.

A.12 Theoretical explanation on how RVC interacts with rectified flow objectives

RVC does not modify the underlying rectified-flow training objective or the learned vector field v_θ. Instead, it is an inference-time correction mechanism that adjusts the denoising trajectory according to rendering errors from the input views. Concretely, the rectified flow predicts the direction of the straightened transport path as:

$$x_{t-\Delta t} = x_t - \Delta t\, v_\theta(x_t, t). \tag{9}$$

RVC introduces a small correction:

$$x_{t-\Delta t} = x_t - \Delta t(v_\theta + \alpha\, \Delta v), \tag{10}$$

where $\Delta v = -t\, \frac{\partial L_{\text{RVC}}}{\partial x_0}$. This adjustment is orthogonal to the rectified-flow objective. The flow model is fully trained under the standard CFM loss. And RVC only guides the denoising trajectory toward a solution that is better aligned with input images. The correction does not alter the learned velocity field or its theoretical properties. In summary, RVC provides input-conditioned guidance during inference, but does not change the theoretical formulation or training of rectified flow.

Table 12: Ablation study on the outlier rejection threshold t_o in RVC on the Dora-bench dataset.

t_o	PSNR↑	SSIM↑	LPIPS↓	CD↓	F-score↑
1	21.618	0.912	0.094	0.091	0.938
0.9	22.075	0.906	0.090	0.091	0.946
0.8	**22.632**	**0.911**	**0.090**	0.090	**0.953**
0.7	22.450	0.909	0.094	**0.089**	0.941
0.6	22.312	0.906	0.093	0.094	0.944

Table 13: Inference time (in seconds) of each component under varying numbers of input views.

Component \Number of images	1	3	5	7	9
DINO feature + VGGT feature	0.2157s	0.4393s	0.8789s	1.3788s	2.0745s
GGC	0.0169s	0.0175s	0.0184s	0.0199s	0.0224s
SS Flow + SS Decoder	4.4193s	4.4039s	4.4105s	4.4141s	4.4267s
PVC	0.0017s	0.0019s	0.0022s	0.0023s	0.0025s
SLat Flow + SLat Decoder	2.7178s	3.0591s	3.7154s	3.9062s	4.4952s
Pose estimation	20.6085s	24.0013s	27.4166s	30.8672s	35.3407s
RVC	5.9726s	6.4080s	7.5483s	10.3359s	12.2985s
Overall	33.9525s	38.3310s	43.9903s	50.9244s	58.6605s

A.13 Ablation study on decoding to different 3D representations in RVC

Table 14: Ablation study on decoding to different 3D representations in RVC on the Dora-bench dataset.

3d representation	PSNR↑	SSIM↑	LPIPS↓	CD↓	F-score↑
Radiance Field (RF)	21.899	0.908	0.0910	0.0917	0.937
Mesh	21.561	0.906	0.0924	**0.0879**	**0.954**
3DGS	**22.632**	**0.911**	**0.0901**	0.0895	0.953

To investigate the influence of 3D representation in RVC, we decode the output SLAT to different 3D representations and report their metrics on the Dora-Bench dataset. In Tab. 14, mesh shows slightly better geometric accuracy due to its structured nature. 3DGS provides superior rendering quality, leading to the best overall visual fidelity. Based on its superior rendering quality and overall balanced performance, choosing 3DGS as the representation is more appropriate.

A.14 Attention visualizations in GGC and cross-view token similarity in PVC

To fully understand the proposed GGC and PVC, we provide attention visualization in GGC and cross-view token similarity in PVC. Specifically, for attention visualization in GGC, we first average attention maps along the dimension of multi-head in each attention layer, and then max pool the attention maps along the dimension of learnable tokens. Finally, we resize the attention map as the same size as input images. For cross-view token similarity in PVC, we calculate the similarity between tokens of two different input images and select a token from one image to visualize its similarity to tokens of another image. As shown in Fig. 10, for GGC's global aggregation in SS, attention visualizations of the first three layers all show a strong correlation between the global learnable tokens and object features in images. In Fig. 11, for PVC's local correspondence in SLAT, we find there is high similarity between corresponding points across different viewpoints.

A.15 Ablation study on the length of learnable tokens in GGC

We conduct the ablation study on the length of learnable tokens in GGC. Since the length of SS latent is $16^3 = 4096$, we believe 4096 is enough for the length of learnable tokens in GGC to represent the visible structure. As shown in Tab. 15, the significant performance drop shows that 2048 tokens are insufficient to represent information of input. Although 8192 tokens offer a slight improvement in PSNR and CD, they introduce additional computational overhead. Therefore, 4096 learnable tokens provide the balance between performance and efficiency.

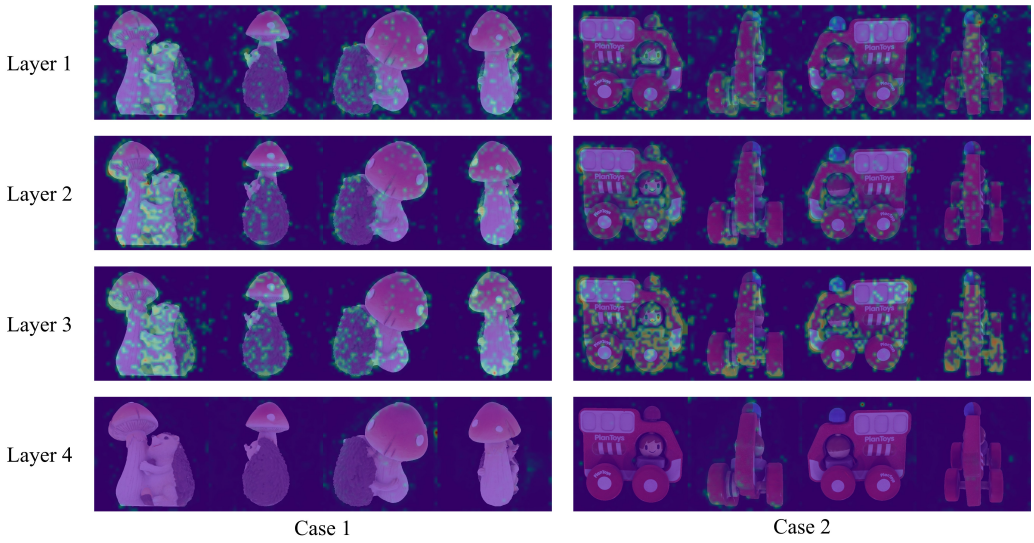

Figure 10: Attention visualization in GGC.

Table 15: Ablation study on the length of learnable tokens in GGC on the Dora-bench dataset.

Length of tokens	PSNR↑	SSIM↑	LPIPS↓	CD↓	F-score↑
2048	18.366	0.890	0.107	0.114	0.884
4096	20.462	**0.894**	0.102	0.093	**0.941**
8192	**20.527**	0.887	**0.105**	**0.092**	0.938

A.16 Evaluation on the DTU dataset

Table 16: Evaluation on the DTU dataset.

Method	PSNR↑	SSIM↑	LPIPS↓
TRELLIS-M Xiang et al. (2024)	15.8675	0.6010	0.3237
Hunyuan3D-2.0-mv Team (2025)	19.0951	0.6763	0.2654
Instantmesh Xu et al. (2024b)	18.8966	0.6712	0.2527
Ours	**21.7639**	**0.7576**	**0.2175**

DTU Jensen et al. (2014) is a real-world multi-view dataset. To further verify the generalizability of our method on real multi-view captures, we evaluate three SOTA multi-view reconstruction baselines (TRELLIS-M, InstantMesh, and Hunyuan3D-2.0-mv) and our method on the DTU dataset, and the results are summarized in Tab. 16. Our approach substantially improves reconstruction fidelity on real-world multi-view data (+2.7 PSNR over the strongest baseline). These results confirm that ReconViaGen generalizes well to real multi-view captures, without any dataset-specific tuning. Qualitative comparisons in Fig. 12 further demonstrate the superiority of our ReconViaGen in real multi-view captures.

A.17 Ablation study on RVC only baseline with our full-version model

We conduct experiments of TRELLIS-M with RVC on Dora-Bench. As shown in the Tab. 17, RVC alone consistently improves the baseline, demonstrating that RVC is indeed effective. However, its performance remains far below our full model, since RVC mainly provides local refinement and cannot replace the global structure aggregation (GGC) and per-view detail modeling (PVC). This confirms that RVC is beneficial but complements rather than substitutes the other components.

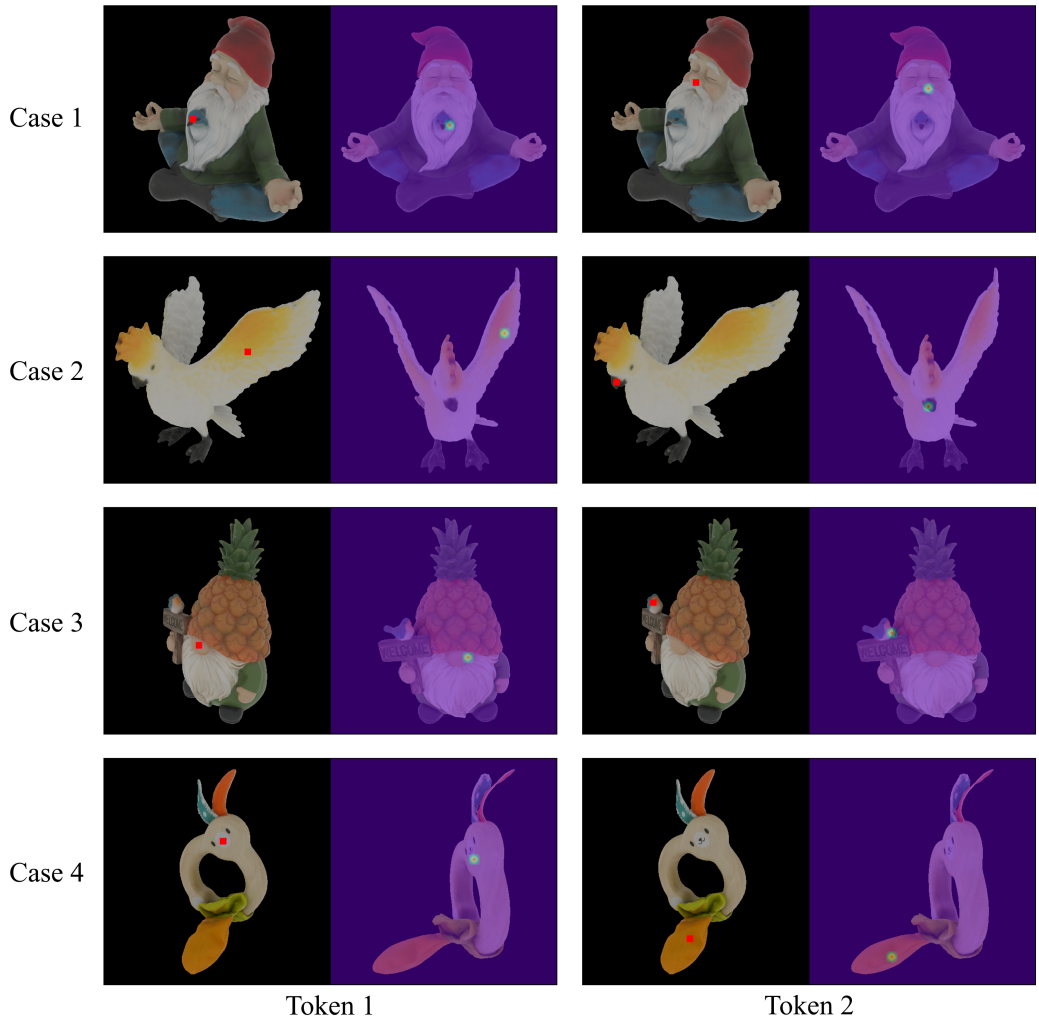

Figure 11: Cross-view token similarity in PVC.

Table 17: Ablation on RVC only baseline with our full-version model.

Method	PSNR↑	SSIM↑	LPIPS↓	CD↓	F-score↑
TRELLIS-M Xiang et al. (2024)	16.706	0.882	0.111	0.144	0.843
TRELLIS-M Xiang et al. (2024) w. RVC	17.728	0.857	0.101	0.139	0.851
Ours	**22.632**	**0.911**	**0.090**	**0.090**	**0.953**

A.18 Ablation study on alternative design for weighted fusion in SLat-Flow

The weighted fusion design aims to trade off multi-view specificity and computational efficiency. The SLAT latent already encodes global geometry and acts as a "volume memory". Therefore, the result of cross attention between SLAT latent and each individual view already reflects how much that view contributes to the current latent state. The MLP is then used to map this interaction to a scalar weight, enabling view-dependent importance estimation without requiring joint attention over all views. Another alternative design is to add all the cross attention results from all views together and then apply a MLP. We ablation on this alternative design with ours in Tab. 18. The alternative yields worse performance. We think direct summation reduces view-specific differences, confirming the need for a view-aware weighting scheme. A direct way is applying a cross-attention with the KV cache coming from all the views. We experimented with this design, but the KV cache grows linearly with the number of views and quickly exceeds GPU memory. Thus, per-view cross-attention with a learned weight provides the balance of accuracy, interpretability, and memory efficiency.

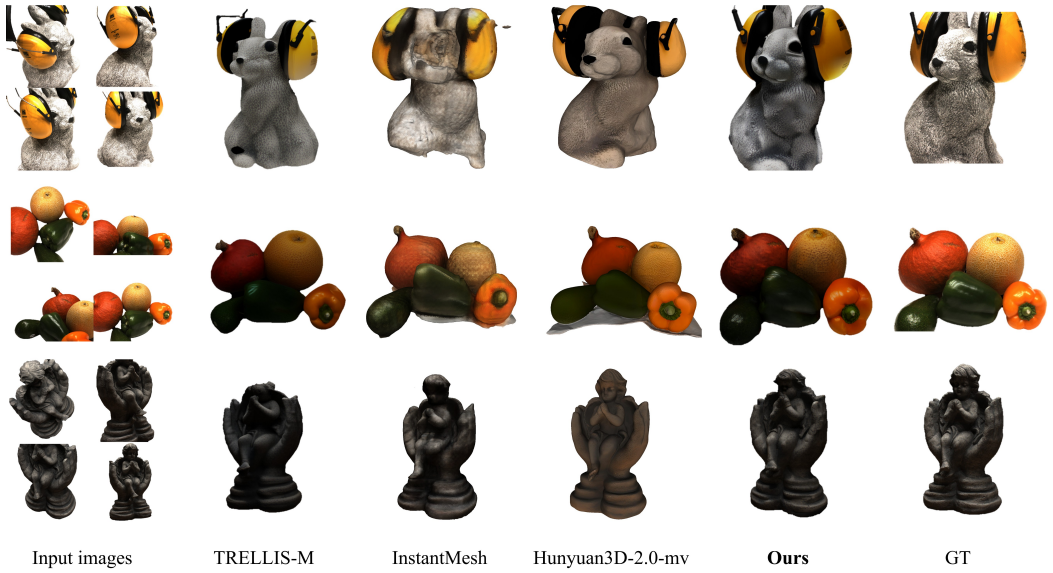

Figure 12: Qualitative comparisons on the DTU datset.

Table 18: Ablation on alternative design for weighted fusion in SLat-Flow on the Dora-bench dataset.

Method	PSNR↑	SSIM↑	LPIPS↓	CD↓	F-score↑
Add up + MLP	20.809	0.894	**0.091**	0.098	0.924
Our weighted fusion (c)	**21.045**	**0.905**	0.093	**0.093**	**0.937**

A.19 Clarification for omitting the visual metrics of VGGT in the Dora-Bench evaluation

Using point clouds from VGGT directly for image rendering does indeed produce poor results. For example, due to the discrete representation of point clouds, when rendering an object from the front, rays may pass through the front point clouds and hit the back point clouds instead. This is due to deficiencies in point cloud representation, therefore we believe that reporting visual metrics in this way is unfair to VGGT. To use the output of VGGT for image rendering, a direct way is to combine the VGGT with 3DGS optimization. We specifically compared ours with VGGT+3DGS in Sec.A.2. However, 3DGS optimization introduces many artifacts and floaters, which negatively impacts geometric metrics like CD and F-score, which is unfair to VGGT. 3D reconstruction often focuses more on geometry, so we choose to report only VGGT's geometric metrics and omit its visual metrics.

A.20 Sensitivity analysis of the quality to camera pose accuracy

Table 19: Ablation study on camera poses under different extent of rotational perturbation on Dora Bench.

Method	PSNR↑	SSIM↑	LPIPS↓	CD↓	F-score↑
Ours w/o RVC	21.045	0.905	0.093	0.093	0.937
Ours w. GT pose	**23.957**	**0.917**	0.092	**0.086**	0.955
Ours w. 3° perturbation	23.606	0.914	**0.089**	0.088	**0.957**
Ours w. 5° perturbation	22.788	0.916	0.093	0.092	0.951
Ours w. 10° perturbation	21.221	0.909	0.094	0.095	0.950
Ours w. 30° perturbation	20.939	0.903	0.094	0.093	0.941
Ours w. 50° perturbation	20.906	0.903	0.097	0.093	0.938
Ours	22.632	0.911	0.090	0.090	0.953

To investigate the robustness and sensitivity of RVC to camera pose accuracy, we intentionally introduce severe rotational perturbation to the ground-truth camera pose in RVC. As shown in Tab. 19, RVC improves reconstruction significantly when accurate poses are available. With moderate errors ($\leq$ 10), the degradation is minimal, showing high tolerance. Even with extreme perturbations 30° and 50° (far beyond practical estimation errors), the performance remains comparable to the baseline "Ours w/o RVC", demonstrating that RVC does not destabilize or corrupt the denoising trajectory. The key to this robustness lies in our outlier mechanism within the rendering-based loss.

A.21 THE USE OF LARGE LANGUAGE MODELS

We only utilize LLMs to refine the writing style and enhance the clarity of exposition. The LLMs are not involved in research ideation, experimental design or data analysis.