World Reconstruction From Inconsistent Views

Lukas Höllein and Matthias Nießner

Technical University of Munich, Germany
https://lukashoel.github.io/video_to_world

Fig. 1: Our method reconstructs 3D worlds from video diffusion models. We propose a tailored non-rigid deformation of predicted pointcloud geometry (mid) that resolves the 3D inconsistencies inherent in generated video sequences. Then, we utilize this improved alignment to optimize a Gaussian Splatting [33] scene. Our worlds can be explored from novel views at high visual fidelity (right).

Abstract. Video diffusion models generate high-quality and diverse worlds; however, individual frames often lack 3D consistency across the output sequence, which makes the reconstruction of 3D worlds difficult. To this end, we propose a new method that handles these inconsistencies by non-rigidly aligning the video frames into a globally-consistent coordinate frame that produces sharp and detailed pointcloud reconstructions. First, a geometric foundation model lifts each frame into a pixel-wise 3D pointcloud, which contains unaligned surfaces due to these inconsistencies. We then propose a tailored non-rigid iterative frame-to-model ICP to obtain an initial alignment across all frames, followed by a global optimization that further sharpens the pointcloud. Finally, we leverage this pointcloud as initialization for 3D reconstruction and propose a novel inverse deformation rendering loss to create high quality and explorable 3D environments from inconsistent views. We demonstrate that our 3D scenes achieve higher quality than baselines, effectively turning video models into 3D-consistent world generators.

Keywords: video diffusion · world generation · 3D reconstruction

1 Introduction

World models build and simulate photorealistic environments. The diversity and high fidelity of video diffusion models [5, 76, 86] has made the generation of 3D

worlds a highly anticipated research goal. Automating this creation process has many applications such as more diverse robotics training or faster ideation and production in movies, gaming, and VR. Their vast prior knowledge makes video models a compelling foundation for world generation, where the goal is to lift these models into 3D generators. A common approach employs a generation-reconstruction loop, that autoregressively generates multiple video sequences and reconstructs 3D scenes from these frames [10, 61, 70].

The core challenge in leveraging video diffusion models for world generation is their lack of 3D consistency. While generated frame sequences look compelling, using them directly for 3D reconstruction leads to misaligned geometry and rendering artifacts (Figure 1). To this end, existing approaches add explicit camera trajectory control to the frame generation. This is achieved by conditioning the network on extrinsic parameters [21, 41, 91], or by progressively building up a 3D cache from all previous frames and rendering it into novel perspectives [56, 88]. More recent methods generate aligned RGB-D sequences [4, 26] or finetune with consistency constraints [15, 37]. Despite being pretrained and finetuned on huge amounts of data, the models still suffer from *generative drift*: generated frames inconsistently warp object geometry to novel viewpoints and do not follow input cameras [88]. We argue these generative models cannot be perfectly consistent and instead propose a lightweight alternative on the reconstruction side.

To this end, we propose a tailored non-rigid alignment of the scene geometry, that creates thin and sharp surfaces from an inconsistent initialization (Figure 1 mid). We then leverage this geometry as initialization in a *non-rigid aware* reconstruction, which yields high-quality, explorable 3D worlds (Figure 1 right). First, we utilize geometric foundation models [42] to obtain a dense pointcloud from all generated videos. This reveals the inconsistencies of the input frames as multiple unaligned surfaces, which we now can correct. Our tailored iterative frame-to-model ICP [6, 30] non-rigidly deforms these surfaces to align in a canonical space. We additionally employ a sparse correspondence term to resolve larger misalignments, followed by a global optimization that further sharpens the pointcloud. We use the aligned pointcloud as initialization for a *non-rigid aware* Gaussian Splatting [24, 33] optimization. Crucially, it factors out the misalignments by backward deforming the scene into the input frame spaces before rendering, which allows us to perform photometric optimization from an inconsistent image set. Our resulting 3D worlds are consistent and can be rendered from novel perspectives at high visual fidelity.
To summarize, our contributions are:

- We introduce a tailored non-rigid alignment of a scene pointcloud based on iterative frame-to-model ICP, that resolves the inconsistencies in the corresponding generated videos.
- We devise a *non-rigid aware* 3D reconstruction that leverages this alignment to optimize consistent 3D worlds from inconsistent views.
- We demonstrate that our lightweight reconstruction method can be employed with many state-of-the-art video diffusion models, effectively turning them into 3D world generators.

2 Related Work

2.1 Diffusion Models For Scene Generation

Image and video diffusion models (VDM) generate diverse, high-quality 2D content [38,55,57,60,76,86]. A highly anticipated goal is to leverage these models to generate 3D (e.g., NeRF [48], 3DGS [33], or meshes [12,65]). Early works leverage score distillation to create assets [54,66,79] or scene chunks [3,9,85], which remains a costly per-scene optimization. Recently, 2D models are finetuned with camera-control [21,22,44,56,73,88,91] to jointly generate multi-view images and then reconstruct them in 3D [10,23,41,61,63,72]. However, the generated frames remain geometrically inconsistent [88], which leads to floating artifacts when rendering. To this end, recent methods remove such areas, generate more images, or fix artifacts at render-time [20,45,81,82,84]. They primarily focus on real-world scenarios with reliable (but sparse) camera observations, whereas we focus on the generative setting, improving from dense (but inconsistent) generated views. Recent approaches finetune models on consistency objectives [4,15,37,40], which is a costly per-model training. In contrast, we propose a lightweight per-scene reconstruction that can turn any VDM into a 3D-consistent world generator. Instead of explicit 3D, recent works explore minute-long rollout of consistent videos [5, 28, 68], which needs vast compute at inference and can hit context limits. Our approach reconstructs consistent 3D worlds with high quality that can be rendered in real-time.

2.2 Geometric Foundation Models

Geometric foundation models (GFMs) predict 3D geometry from multiple images as input [32, 39, 42, 43, 78], enabling online reconstruction [2, 46, 67]. When combined with rigid alignment akin to bundle adjustment, recent methods obtain high-quality 3D scenes from uncalibrated images [17,27,43]. However, since GFMs are trained on real-world or synthetic images, they cannot correct for the generative drift inherent in VDMs. Instead, inconsistent images lead to non-overlapping surfaces, that reveal the geometric inaccuracy of generations. We exploit this insight to correct for these artifacts in a 3D alignment stage.

2.3 Pointcloud Alignment

The alignment of 3D geometry has been extensively studied. Foundational methods like ICP [6,58,64] or volumetric fusion [14] enable large-scale integration of depth measurements [11,13,30,51,93]. They were further extended to reconstruct textures [25, 52, 92] and dynamic reconstruction by optimizing for non-rigid deformations [8, 29, 50, 89, 94]. Inspired by these approaches we propose a tailored pointcloud alignment, that is suitable for inconsistent video generations. Specifically, we devise a non-rigid alignment energy akin to these approaches and utilize the reconstruction along with the optimized deformations in a *non-rigid aware* Gaussian Splatting [33] optimization to produce consistent 3D worlds.

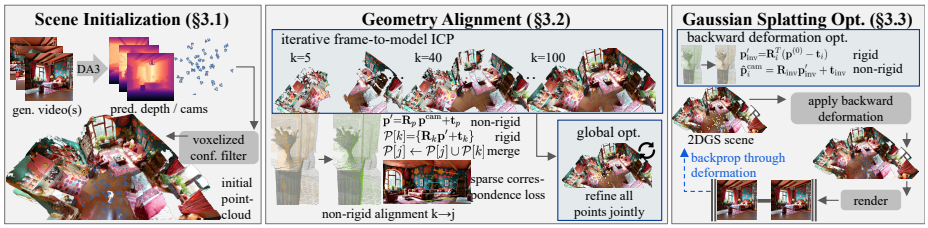

Fig. 2: Method overview. We propose a three stage method that reconstructs a 2DGS [24] scene from generated videos. First, we estimate multi-view depth and cameras with a geometric foundation model [42]. The resulting dense scene initialization is unaligned (multiple non-overlapping surfaces) due to the inconsistent input frames. We propose a tailored non-rigid geometry alignment that leverages iterative frame-to-model ICP [6, 30] and sparse correspondences, followed by global optimization, to create thin surfaces with detailed textures. Then, we leverage the alignment in a novel *non-rigid aware* 2DGS [24] optimization to obtain high-quality, consistent 3D worlds.

3 Method

Our method creates a 3D scene reconstruction from generated video sequences (see Figure 2). First, we initialize the scene by estimating dense per-pixel depth and poses and unproject them into a pointcloud (Section 3.1). Since the input frames are not 3D consistent, this geometry is unaligned and distorted. To this end, we improve the geometry in our non-rigid alignment stage (Section 3.2). The output geometry accurately depicts unified surfaces and textures. Finally, we use this refined pointcloud as initialization for our *non-rigid aware* Gaussian Splatting optimization (Section 3.3). We obtain consistent and detailed 3D worlds that can be explored in real-time from arbitrary positions.

3.1 Scene Initialization

We leverage video diffusion models (VDMs) [7, 76, 86] to generate input frames that depict static scenes from various perspectives. VDMs model the conditional probability distribution $p_\theta(\mathbf{V}|\mathbf{c})$, where $\mathbf{V}=\{\mathbf{I}_i\}_{i=0}^{N}$ is a continuous sequence of individual images $\mathbf{I}\in\mathbb{R}^{3\times H\times W}$. The condition $\mathbf{c}$ can be a text prompt and/or previous frames $\mathbf{V}_{\text{src}}$. This enables the generation of large-scale worlds via autoregressive roll-out of multiple sequences $\{\mathbf{V}\}_{i=0}^{N}$ [5, 28, 61]. VDMs can be conditioned on camera trajectories $\{\mathbf{R}_i\in\mathbb{R}^{3\times3}, \mathbf{t}_i\in\mathbb{R}^3, \mathbf{K}_i\in\mathbb{R}^{3\times3}\}_{i=1}^{N}$ to allow fine-grained per-frame spatial control. This is typically achieved through dense Plücker embeddings [41, 91] or 3D cache reprojection [56, 88].

Despite explicitly modeling multi-view control, the frames suffer from *generative drift*, making them inconsistent in 3D. The generated sequences are thus hard to reconstruct, with floating artifacts or distortions when rendering (Figure 1 right). We observe that geometric foundation models (GFMs) [32, 42, 43, 78] reveal these shortcomings and exploit this in a non-rigid alignment to obtain sharp and consistent surfaces (Figure 1 mid). Concretely, GFMs predict dense

per-pixel depth $\mathbf{d}\in\mathbb{R}^{H\times W}$ and confidence $\mathbf{c}\in\mathbb{R}^{H\times W}$ as well as per-frame pinhole cameras $(\mathbf{R},\mathbf{t},\mathbf{K})$:

$$\{\mathbf{d}_i,\mathbf{c}_i,\mathbf{R}_i,\mathbf{t}_i,\mathbf{K}_i\}_{i=0}^{N}=\mathcal{F}_{\text{GFM}}(\mathbf{V}), \tag{1}$$

where $\mathcal{F}_{\text{GFM}}$ denotes the pre-trained metric DepthAnything-3 [42] model. In the case of camera-controlled video diffusion models, we condition the GFM on the trajectory to obtain $\{\mathbf{d}_i,\mathbf{c}_i\}_{i=0}^{N}=\mathcal{F}_{\text{GFM}}(\mathbf{V},\mathbf{R}_i,\mathbf{t},\mathbf{K})$.

We utilize this information to construct a scene initialization as colored pointcloud via unprojection:

$$\mathbf{p}_i^{\text{cam}}(u,v)=\mathbf{K}_i^{-1}[u,v,1]^T\cdot\mathbf{d}_i(u,v),\bar{\mathcal{P}}=\{\mathbf{R}_i\mathbf{p}_i^{\text{cam}}+\mathbf{t}_i\}_{i=0}^{N},\mathcal{C}=\{\mathbf{I}_i(u,v)\}_{i=0}^{N}, \tag{2}$$

where (u,v) are pixel coordinates. We devise a filtering scheme to retain reliable points per scene chunk. This ensures a dense initialization that does not require hallucination, but only alignment, to create detailed surfaces of entire scenes. Each 3D point $\mathbf{p}_k\in\bar{\mathcal{P}}$ is assigned to a voxel $\mathbf{v}_k=\lfloor\mathbf{p}_k\,/\,s_{\text{vox}}\rfloor$ and then we compute the per-voxel confidence threshold: $\tau_{\text{loc}}[v]=\text{perc}\big(\{\mathbf{c}_k{:}\mathbf{v}_k{=}v\},\ \theta_{\text{loc}}\big)$, where $\mathbf{c}_k$ is the point confidence. We additionally compute the per-voxel occupancy threshold as $\tau_{\text{cnt}}=\text{perc}(\{\text{count}[v]\},\ \theta_{\text{cnt}})$ and use the two thresholds to obtain a filtered pointcloud: $\mathcal{P}=\{\mathbf{p}_k{:}\mathbf{c}_k\geq\tau_{\text{loc}}[\mathbf{v}_k]\big)\wedge\text{count}[\mathbf{v}_k]\geq\tau_{\text{cnt}}\}$.

The resulting scene initialization shows misaligned points, i.e., corresponding surfaces across frames to not perfectly overlap (Figure 1 mid-top). Next, we correct these artifacts in a 3D geometric alignment stage.

3.2 Geometry Alignment

We propose a two-stage non-rigid alignment of the pointcloud $\mathcal{P}$. First, a tailored non-rigid iterative frame-to-model ICP [1,6,30] obtains an initial alignment across all frames. Second, a global optimization further sharpens this alignment.

Non-rigid iterative frame-to-model ICP The goal of this stage is to iteratively align two point subsets $\mathcal{P}[j]$, $\mathcal{P}[k]$ where j,k denote image indices. In other words, the subsets refer to the confidence-filtered per-frame point predictions of the GFM. We optimize the camera extrinsics $(\mathbf{R}_k,\mathbf{t}_k)$ and the weights θ_k of a deformation neural field, implemented as a hashgrid MLP [49]. Given a 3D position $\mathbf{p}\in\mathcal{P}[k]$, the network predicts exponential coordinates $\xi=\mathcal{F}_\theta(\mathbf{p})\in\mathbb{R}^6$, which are interpreted as the coordinate representation of a twist $\boldsymbol{\xi}\in\mathfrak{se}(3)$. The corresponding per-point rigid transformation $(\mathbf{R}_p,\mathbf{t}_p)$ is obtained via the exponential map $T(\mathbf{p})=\exp(\boldsymbol{\xi}^\wedge)\in\text{SE}(3)$. Then, we non-rigidly transform $\mathcal{P}[k]$ as

$$\mathbf{p}'=\mathbf{R}_p\,\mathbf{p}^{\text{cam}}+\mathbf{t}_p,\quad\bar{\mathcal{P}}[k]=\{\mathbf{R}_k\mathbf{p}'+\mathbf{t}_k|\mathbf{p}\in\mathcal{P}[k]\}. \tag{3}$$

This aligns the per-frame pointcloud with $\mathcal{P}[j]$ by first transforming each point independently (i.e., non-rigidly) in the k-th camera space, followed by a rigid transformation from k to j. We apply this principle on all frames $\{\mathbf{I}_i\}_{i=1}^{N}$ by

fixing j=0 and iterating on $k\in\{1,...,N\}$. Each iteration optimizes $\mathbf{R}_k, \mathbf{t}_k, \theta_k$ with a point-to-plane dense data term:

$$\mathcal{L}_{\text{data}} = \frac{1}{\sum_{\mathbf{p}\in\mathcal{P}[k]} w(\mathbf{p})} \sum_{\mathbf{p}\in\mathcal{P}[k]} w(\mathbf{p}) \big((\mathbf{p}' - \mathbf{q}_{\text{nn}(j)}) \cdot \mathbf{n}_{\text{nn}(j)}\big)^2, \tag{4}$$

where $\mathbf{q}_{\text{nn}(j)}$ is the nearest neighbor of $\mathbf{p}'$ in $\mathcal{P}[j]$, $\mathbf{n}_{\text{nn}(j)}$ its normal vector and we classify inliers via

$$w(\mathbf{p}) = \begin{cases} 1, & \text{if } \|\mathbf{p}' - \mathbf{q}_{\text{nn}(i)}\|^2 < d_{\text{max}}^2, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

We also adopt the color ICP term $\mathcal{L}_{\text{color}}$ of Park *et al.* [52], which aligns points by minimizing differences in local tangent-plane color gradients. Additionally, we obtain 2D correspondence pairs via image matching of $\mathbf{I}_k$ and a subset of previous frames $\{\mathbf{I}_i\}_{i<k}$. We utilize the RoMa matcher [18, 19] to obtain pairs (s_i, t_i) and a certainty weight w_i, where s_i is the pixel-coordinate in a source image and t_i in $\mathbf{I}_k$. Then, we compute the weighted sparse correspondence loss

$$\mathcal{L}_{\text{corr}} = \frac{\sum_i w_i \|\mathcal{P}_j[\mathbf{s}_i] - \mathcal{P}_k[\mathbf{t}_i]\|^2}{\sum_i w_i}. \tag{6}$$

In practice we found $\mathcal{L}_{\text{corr}}$ is highly beneficial, as it allows to align points that exceed the nearest neighbor distance d_{max}. We further regularize the non-rigid deformations to be small and locally similar, akin to an ARAP constraint [69]:

$$\mathcal{L}_{\text{tv}} = \frac{1}{|\mathcal{P}[k]|} \sum_{p\in\mathcal{P}[k]} \|\mathcal{F}_{\theta_k}(\mathbf{p}) - \mathcal{F}_{\theta_k}(\mathbf{p} \pm \mathbf{s}_{\text{vox}})\|^2, \tag{7}$$

where we compare the twist predictions of each point with its 6 axis-aligned neighbors at distance s_{vox}. The final objective then becomes $E(\mathbf{R}_k, \mathbf{t}_k, \theta_k)$=$\mathcal{L}_{\text{data}}$+ $\lambda_{\text{color}}\,\mathcal{L}_{\text{color}} + \lambda_{\text{corr}}\,\mathcal{L}_{\text{corr}} + \lambda_{\text{tv}}\,\mathcal{L}_{\text{tv}}$. We optimize this energy with ADAM [35] in a coarse-to-fine fashion with progressively smaller $d_{\text{max}}, s_{\text{vox}}$, where we predict non-rigid deformations with $\mathcal{F}_\theta$ only in the finest scale.

After each iteration we merge the aligned points into the current model as $\mathcal{P}[j] \leftarrow \mathcal{P}[j] \cup \bar{\mathcal{P}}[k]$ using an adaptive outlier removal scheme. It discards points, whose ICP losses are larger than the median absolute deviation (MAD) across all previous frames. We compute the k-th thresholds g_k^d=perc$\big(\mathcal{L}_{\text{data}}, \theta_d\big)$ and g_k^c=perc$\big(\mathcal{L}_{\text{color}}, \theta_c\big)$ and accumulate them in $\tau_d = \text{median}(\{g^d\}_{i<k}) + \sigma_{\text{d}} \cdot 1.4826 \cdot \text{MAD}(\{g^d\}_{i<k})$ and τ_c, respectively. A point is then only merged if both its data and color ICP terms are smaller than τ_d and τ_c, respectively.

Global non-rigid optimization After the iterative ICP stage, the pointcloud $\mathcal{P}[0]$ now represents an aligned 3D scene, where different per-frame points make up a unified surface. We further improve this alignment by *jointly* optimizing all cameras and deformation networks. We find that this refinement leads to thinner surfaces and sharpens the overall texture quality of the pointcloud (e.g., compare

the fence in Figure 5 bottom). Concretely, the frame-to-model optimization of the previous stage can drift over time towards aligned but thicker surfaces, since only the k-th frame is optimized at a time.

To this end, we optimize the global energy

$$E(\{\mathbf{R}_i, \mathbf{t}_i, \theta_i\}_{i=1}^N) = \mathcal{L}_{\text{data}}^{\text{global}} + \lambda_{\text{color}} \mathcal{L}_{\text{color}}^{\text{global}} + \lambda_{\text{anchor}} \mathcal{L}_{\text{anchor}}, \tag{8}$$

that modifies the correspondence search in the ICP losses. For each point $\mathbf{p} \in \mathcal{P}[0]$ we find its k=5 nearest neighbors in *other* frames and use them to calculate the respective objectives. As gradients now flow to both $\mathbf{p}'$ and $\mathbf{q}_{\text{nn}(j)}$, this drives the non-rigid deformations towards a thinner surface across all frames.

Additionally, we regularize the updates to the cameras and deformation networks to be as small as possible, which avoids degenerate solutions:

$$\mathcal{L}_{\text{anchor}} = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{M} \sum_{k=1}^{M} \| \mathcal{F}_{\theta_i}(\mathbf{a}_{k,i}) - \xi_{k,i}^{(0)} \|^2 + \| \xi_{g,i} - \xi_{g,i}^{(0)} \|^2 \right], \tag{9}$$

where we subsample points from each frame $\{\mathbf{a}_{k,i}\}_{k=1}^M \subset \mathcal{P}[i]$ and encourage similarity of their twists to the state before global optimization $\xi_{k,i}^{(0)}$. We similarly regularize the cameras by comparing their twist parameterizations $\xi_{g,i}$ and $\xi_{g,i}^{(0)}$.

3.3 Non-rigid Gaussian Splatting Optimization

After the geometry alignment we have obtained $\mathcal{P}[0]$, which represents an aligned 3D scene, as well as the non-rigid *forward* deformations $\{\mathbf{R}_i, \mathbf{t}_i, \mathcal{F}_{\theta_i}\}_{i=1}^N$ from each frame to the first frame. This is similar to a 4D scene reconstruction that is obtained in the dynamic reconstruction literature [50, 53, 83, 89]. In contrast, our goal is not to render a dynamic scene from a time-dependent state, but to unify the N inconsistent geometry predictions into a single canonical state. Thus, we propose a novel *non-rigid aware* Gaussian Splatting [33] optimization. The output 3D reconstruction then renders photorealistic novel views of consistent 3D worlds, while being trained from the inconsistent image set $\{\mathbf{I}_i\}_{i=0}^N$.

Backward deformation optimization First, we obtain the inverse of the per-frame deformations $\{\mathbf{R}_i, \mathbf{t}_i, \mathcal{F}_{\theta_i}\}_{i=1}^N$, which transforms $\mathcal{P}[0]$ into all other frames. Obtaining $\{\mathbf{R}_i^{-1}, \mathbf{t}_i^{-1}\}_{i=1}^N$ is straightforward and we optimize a single deformation neural field $\mathcal{F}_{\theta_{\text{inv}}}^{-1}$, that predicts the inverse exponential coordinates $\xi^{\text{inv}} = \mathcal{F}_{\theta_{\text{inv}}}^{-1}(\mathbf{p}, i)$ for any point $\mathbf{p} \in \mathcal{P}[0]$ to any view i. We similarly implement $\mathcal{F}_{\theta_{\text{inv}}}^{-1}$ as a hashgrid MLP [49], but additionally condition the network on a learnable view embedding. We generate training pairs $\{\mathbf{p}_i^{\text{cam}}, \mathbf{p}^{(0)}\}_{i=1}^M$ by sampling a subset of points from all per-frame predictions in their respective camera space and transforming them to $\mathcal{P}[0]$ via Equation (3). We apply the *backward* deformation on these pairs to obtain the predictions $\hat{\mathbf{p}}_i^{\text{cam}}$:

$$\mathbf{p}'_{\text{inv}} = \mathbf{R}_i^T (\mathbf{p}^{(0)} - \mathbf{t}_i), \quad \hat{\mathbf{p}}_i^{\text{cam}} = \mathbf{R}_{\text{inv}} \mathbf{p}'_{\text{inv}} + \mathbf{t}_{\text{inv}}, \tag{10}$$

where $(\mathbf{R}_{\text{inv}}, \mathbf{t}_{\text{inv}})$ is the corresponding per-point inverse deformation that we obtain from ξ^{inv}. Then, we define the total objective as

$$\mathcal{L}_{\text{inverse}} = \frac{1}{M}\sum_{i=1}^{M} \|\hat{\mathbf{p}}_i^{\text{cam}} - \mathbf{p}_i^{\text{cam}}\|^2, \quad E(\theta_{\text{inv}}) = \mathcal{L}_{\text{inverse}} + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}}. \tag{11}$$

3D Scene Consolidation We leverage $\mathcal{F}_{\theta_{\text{inv}}}^{-1}$ to train a Gaussian Splatting [33] scene representation from the inconsistent image set $\{\mathbf{I}_i\}_{i=0}^N$. Concretely, we convert $(\mathcal{P}[0], \mathcal{C})$ into the attributes of 2DGS [24], which parameterizes 2D Gaussian disks with position, rotation, scales, opacity, and color attributes. We directly assign $\mathcal{P}[0]$ as the positions and use their estimated normals to orient the disks along the surface. We calculate the average euclidean distance of each point to its k=10 nearest neighbors to determine the Gaussian scale. We uniformly initialize the opacities to 0.1 and convert the pointcloud colors into the degree-0 coefficients of Spherical Harmonics.

Since the generated image set $\{\mathbf{I}_i\}_{i=0}^N$ is inconsistent, we cannot directly render our scene from the corresponding cameras to train the 2D Gaussians. Doing so would effectively *undo* the geometric alignment to explain the generative drift in the image observations (see Figure 6). To this end, we propose a *non-rigid aware* rendering objective, that first transforms the Gaussians into a per-frame camera space. Concretely, we apply Equation (10) on the Gaussian positions and rotations, which non-rigidly deforms them into the inconsistent frame spaces, and transform the scales using $(\mathbf{R}_i^{-1}, \mathbf{t}_i^{-1})$. Then, we follow 2DGS [24] to rasterize the Gaussians and perform volumetric alpha blending to render the image set $\{\hat{\mathbf{I}}_i\}_{i=0}^N$. We jointly optimize the 2D Gaussians $\mathcal{G}$ and cameras $\{\mathbf{R}_i, \mathbf{t}_i\}_{i=0}^N$ with standard rendering objectives:

$$\mathcal{L}_{\text{rend}} = \frac{1}{N}\sum_{i=0}^{N}\big(\lambda_1\|\hat{\mathbf{I}}_i - \mathbf{I}_i\|^1 + \lambda_2 \text{LPIPS}(\hat{\mathbf{I}}_i, \mathbf{I}_i)\big), \tag{12}$$

where LPIPS denotes the perceptual loss [31]. Since Equation (10) is differentiable, this loss optimizes the 2D Gaussians in their canonical state to explain the image observations. Crucially, the non-rigid deformation *factors out* the geometric inconsistencies, which keeps the scene representation aligned while optimizing for sharp and detailed appearance. We additionally adopt the depth and normal regularizers of 2DGS [24] and optimize the final objective

$$E(\mathcal{G}, \{\mathbf{R}_i, \mathbf{t}_i\}_{i=0}^N) = \mathcal{L}_{\text{rend}} + \lambda_d\mathcal{L}_d + \lambda_n\mathcal{L}_n. \tag{13}$$

Since $\mathcal{P}[0]$ is a dense and accurate surface initialization, we observe fast convergence in only a few thousand iterations, similar to EDGS [36]. For specular scenes we additionally optimize higher degree Spherical Harmonics coefficients in a second training phase. In practice, we find it beneficial to freeze the cameras and positions in this stage to avoid degenerate solutions that model the generative drift of the input images into the view-dependent effects.

Method	Consistency		Fidelity	
	3D	Photometric	CLIP-IQA+	CLIP Aesthetic
DA3 [42]	69.53	71.62	31.64	36.64
3DGS-MCMC [34]	67.34	69.58	35.56	35.06
VGGT-X [43]	65.73	65.58	38.21	37.24
VGGT-X†	69.66	67.05	41.44	37.53
Ours	**79.29**	**86.59**	**46.56**	**37.61**

Table 1: Quantitative comparison. We compare against baselines on the single video 3D reconstruction task and report averaged results across all source video models. Our method achieves the highest 3D and photometric consistency, which underlines its ability to reconstruct worlds from inconsistent frames. We also improve in terms of image fidelity, which signals our images are of higher sharpness and texture detail.

4 Results

4.1 Implementation Details

We select uniform hyperparameters for all results. The coarse-to-fine ICP optimizes with s_{vox}=[4cm, 2cm] and $d_{\max}$=[5cm, 3cm] for 50 and 150 iterations, respectively, with a learning rate of 1e−3 and λ_{color}=0.05, λ_{corr}=1.0, λ_{tv}=10.0. We set the percentile thresholds as θ_{loc}=15.0, θ_{cnt}=50.0, θ_{d}=θ_{g}=75.0 and the maximum standard deviations for outlier removal to σ_d=2.5, σ_c=1.5. We detect up to 5,000 correspondences in up to 20 image pairs for $\mathcal{L}_{\text{corr}}$. The global optimization runs for another 100 iterations with λ_{anchor}=50.0. Both stages take a total of 25 minutes / 20GB on average for N=50 images on a single A6000 GPU. We subsample $\mathcal{P}[0]$ to roughly 1.5M Gaussians to optimize a 3D scene in 10 minutes / 8GB for 5,000 iterations following the setup of 2DGS [24], but without densification. We optionally optimize the SH for another 10,000 iterations.

4.2 Single Video 3D Reconstruction

We compare our method against recent 3D reconstruction methods on the generated frames from multiple state-of-the-art video diffusion models. Concretely, we generate single video sequences depicting various indoor/outdoor scenes and camera motions from text with Wan-2.2 [76] and with camera-control using ViewCrafter [88], Gen3C [56], Seva [91], and Voyager [26]. We additionally adopt the recent autoregressive world generators Genie3 [5] and HY-WorldPlay [71].

We reconstruct 3D scenes from N=50 images sampled from these videos with various baselines. First, with 3DGS-MCMC [34] from a random initialization and optimizable cameras from DA3 [42]. Second, we utilize the DA3 [42] points as 3DGS [33] initialization and regularize their positions with a depth loss [90] against the DA3 depth. We refer to this as "DA3" [42] and refer to the suppl. material for more details. We compare VGGT-X [43], which *rigidly* aligns correspondences, akin to bundle adjustment, before optimizing 3DGS [33]. We refer to this as "VGGT-X†" when using DA3 [42] instead of VGGT [43] predictions.

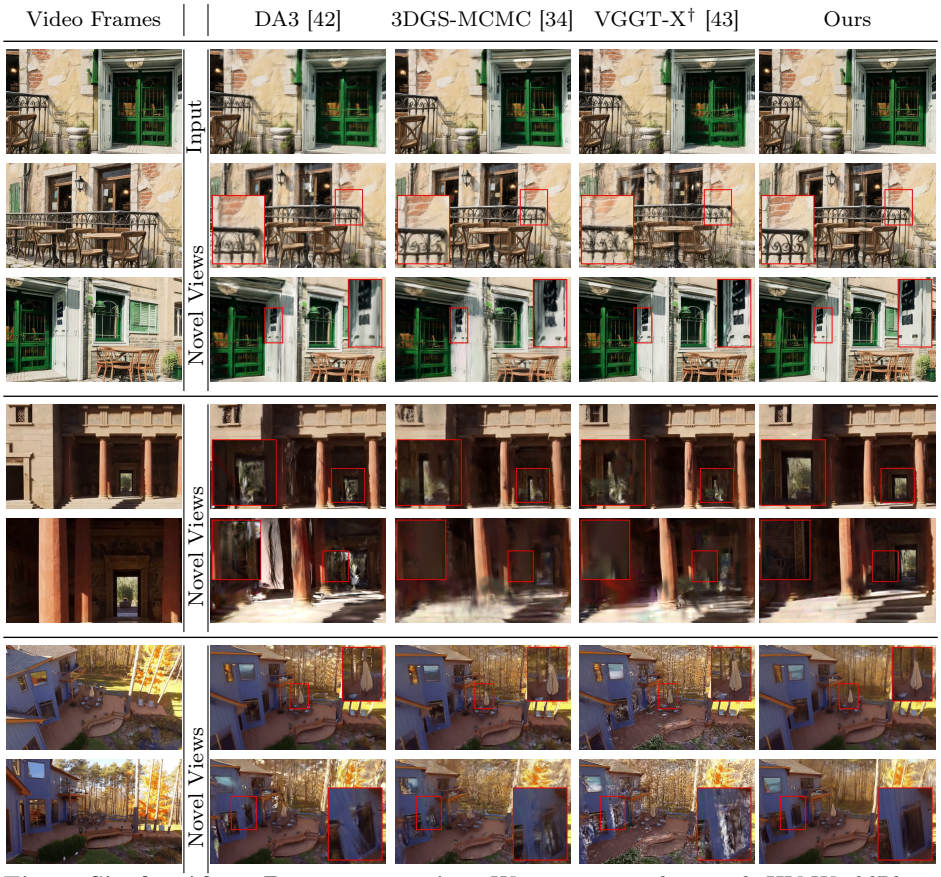

Fig. 3: Single video 3D reconstruction. We generate videos with HY-WorldPlay [71] (top), Genie3 [5] (mid), ViewCrafter [88] (bottom) and 3D reconstruct them. Our method optimizes consistent worlds from inconsistent generated frames. Compared to baselines, the renderings are of higher visual fidelity from both input and novel views.

We showcase the reconstruction results in Figures 3 and 4 and refer to the suppl. material for more samples and animated results. Our method renders sharper and more detailed textures that are of comparable visual fidelity as the input video frames, just in a 3D consistent space. In contrast, the baselines do not correct for the generative drift inherent in the generated frames. This can lead to blurrier renderings from input poses (e.g., the brick textures in Figure 3top). It becomes especially noticeable from novel perspectives: the inconsistent image observations lead to floating artifacts in the 3D scenes that limit their explorability from arbitrary positions. Our method optimizes the scenes in a *non-rigid aware* fashion which leads to greatly improved viewpoint stability, i.e., we can explore the 3D worlds from novel perspectives while retaining the high visual fidelity of the video frames. Our method effectively turns any video diffusion model into a 3D world generator that allows persistent, high quality, and real-time rendering.

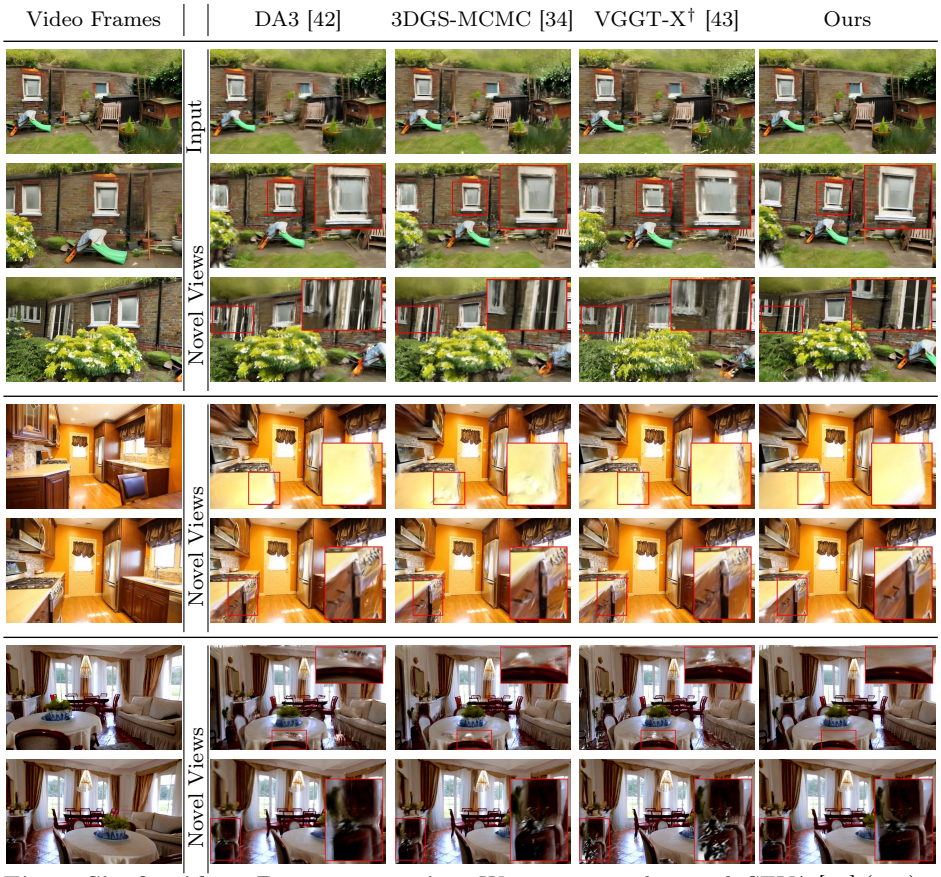

Fig. 4: Single video 3D reconstruction. We generate videos with SEVA [91] (top), Gen3C [56] (mid), Wan [76] (bottom) and 3D reconstruct these frames. Inconsistencies in the generations lead to blurry textures for the baselines compared to the corresponding video, and to severe floating artifacts from novel views. In contrast, our method creates 3D consistent worlds with high fidelity beyond the generated perspectives.

We confirm this in Table 1 by calculating consistency and fidelity metrics following the established WorldScore benchmark [16], averaged across all VDMs per reconstruction method. Our method obtains the highest consistency scores, which underlines its ability to reconstruct 3D worlds from inconsistent views. The rendering quality is the highest among baselines and comparable to that of the input videos (CLIP-IQA+: 47.39, CLIP Aesthetic: 39.04). Additionally, we compare the 3D pointclouds that are used as initialization for all applicable baselines in Figure 5. In comparison to the DA3 [42] predictions, our results greatly improve the alignment of individual surfaces (e.g., unifies multiple windows and slides in the garden scene). VGGT-X [43] obtains sparser pointclouds and less-precise alignment since they do not model non-rigid deformations.

Fig. 5: Pointcloud reconstructions. We compare the quality of the reconstructed pointclouds that are used as initialization for Gaussian Splatting [33] optimization in the subsequent stages for each method. Our approach achieves the highest alignment and compelling textures for individual objects with no overlap of multiple surfaces.

4.3 Ablations

The key components of our method are the non-rigid alignment (Eqs. (3) and (10)) and its usage in the 3D reconstruction. We ablate these decisions in Figure 6 and Table 2 (averaged across four SEVA [91] scenes) and additionally compare dropping the correspondence loss, global optimization, and point filtering.

This highlights that both only rigid alignment and vanilla 3DGS optimization ("no inv") are not enough to obtain aligned surfaces and consistent renderings. Dropping the correspondences fails to align surfaces beyond the matching distance $d_{\max}$, which depending on the amount of generative drift can be noticeable in the pointcloud and renderings. No point filtering or global optimization adversely impact the texture quality of reconstructed and rendered surfaces. In contrast, our full approach obtains the highest rendering quality and alignment.

Method	Consistency		Fidelity	
	3D	Photometric	CLIP-IQA+	CLIP Aesthetic
only rigid	76.98	64.93	11.43	49.12
no inv	78.28	55.16	15.81	46.39
no filt	72.58	64.88	55.17	51.05
no corr	65.62	74.76	44.81	46.27
no global	63.90	77.85	49.69	51.61
Ours	**79.79**	**86.88**	**55.30**	**52.40**

Table 2: Ablation study. 3D reconstruction from only rigid alignment, without correspondences, or without backward deformation ("no inv") negatively impacts 3D consistency and image quality. No point filter or global optimization decrease image fidelity and consistency. Our full method achieves the highest consistency and fidelity.

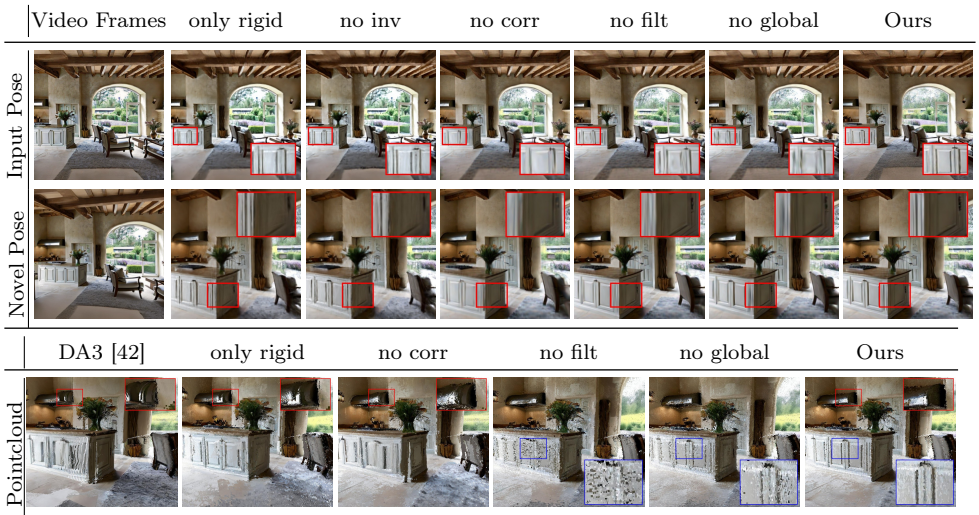

Fig. 6: Ablation. We compare our full method on rendering quality of input views (top) and novel views (mid), as well as on pointcloud reconstruction quality (bottom). The DA3 [42] prediction contains misalignments that cannot be resolved by only rigid alignment or without correspondences. Removing the color filtering and global optimization steps adversely impact texture quality. Removing our novel inverse deformation loss reduces sharpness. In contrast, our method obtains flat/sharp surfaces which better constrain the 3D reconstruction, yielding high quality rendering results.

4.4 Large-Scale World Generation

A single generated video is limited in the amount of scene exploration it can show. Recent works exploit VDMs autoregressively to generate multiple sequences that depict entire 360 degree scenes [5, 10, 28, 61]. Figures 1 and 7 compare against WorldExplorer [61] and VGGT-X [43] by generating and then reconstructing up to 32 video sequences with SEVA [91] via the progressive scene expansion strategy of [61]. The generative drift grows larger across multiple videos and thus the 3D reconstructions suffer from increased consistency problems. While the baselines can depict complete and diverse worlds, the amount of exploration they enable is still limited (i.e., novel views far beyond the training poses suffer from extensive floating artifacts). In contrast, our worlds remain consistent and retain a high rendering fidelity, even from extreme novel perspectives.

4.5 Limitations

Our method reconstructs 3D scenes from inconsistent generated views, however, some limitations remain (see supplementary material). First, we tackle the problem of *generative drift* in VDMs, i.e., existing objects move geometrically inconsistent. However, VDMs can also suffer from hallucinations (e.g., revisiting previous scene areas creates novel or removes existing objects, changes their texture, etc.). Adopting robust reconstruction mechanisms to detect such outlier

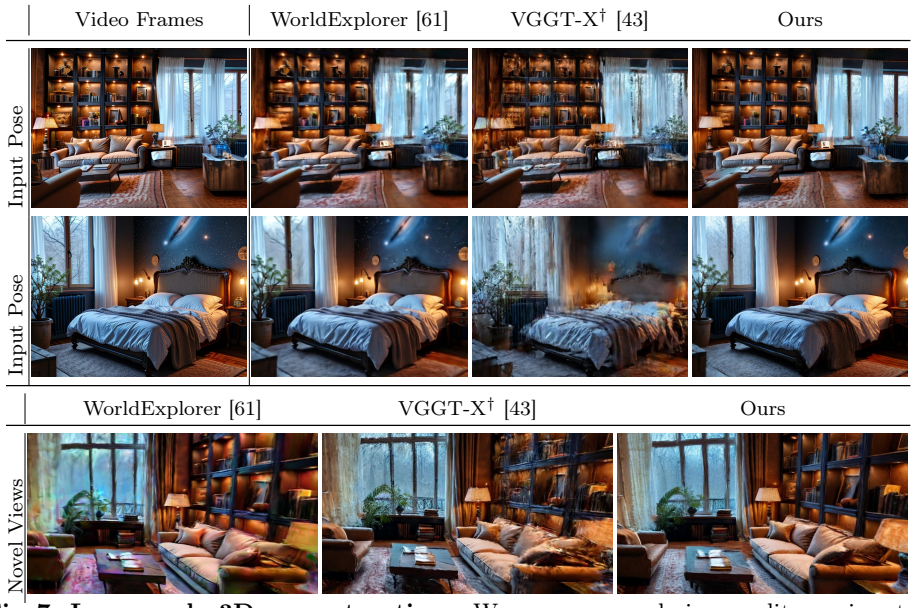

Fig. 7: Large-scale 3D reconstructions. We compare rendering quality on input poses and novel views for entire 360 degree explorable scenes. Our method creates 3D consistent worlds with a high rendering fidelity far beyond the training views. Please see the supplementary material for animated flythroughs.

frames could resolve these problems [59]. Second, our method proposes a per-scene alignment stage for generated frames. While this lightweight approach can turn any video diffusion model into a reliable world generator, it still entails additional computational cost before reconstruction. One could explore finetuning VDMs with our alignments acting as gradient signals for the generators.

5 Conclusion

We have presented a method to reconstruct 3D worlds from the inconsistent generated views of video diffusion models [5, 26, 56, 76, 88, 91]. We exploit geometric foundation models [42] to reveal the generative drift inherent in the generations (Section 3.1). Our tailored non-rigid alignment creates sharp and flat surfaces that resolve these inconsistencies (Section 3.2). Finally, we propose a novel *non-rigid aware* Gaussian Splatting [33] optimization that optimizes a canonical 3D scene representation from the distorted frame observations (Section 3.3). Overall, this leads to sharp and high-quality scene renderings from novel poses, without floating artifacts that otherwise occur due to the inaccurate frames. We believe this demonstrates the exciting ability to turn any video diffusion model into a reliable 3D world generator, which will open up further research avenues and make world generation more practical across many real-world applications.

Acknowledgements

This project was funded by the ERC Consolidator Grant Gen3D (101171131). We also thank Angela Dai for the video voice-over.

References

1. Amberg, B., Romdhani, S., Vetter, T.: Optimal step nonrigid icp algorithms for surface registration. In: 2007 IEEE conference on computer vision and pattern recognition. pp. 1–8. IEEE (2007)
2. Antsfeld, L., Chidlovskii, B., Cabon, Y., Leroy, V., Revaud, J.: S-must3r: Sliding multi-view 3d reconstruction (2026), https://arxiv.org/abs/2602.04517
3. Bahmani, S., Shen, T., Ren, J., Huang, J., Jiang, Y., Turki, H., Tagliasacchi, A., Lindell, D.B., Gojcic, Z., Fidler, S., Ling, H., Gao, J., Ren, X.: Lyra: Generative 3d scene reconstruction via video diffusion model self-distillation. In: arXiv preprint arXiv:2509.19296 (2025)
4. Bai, Y., Fang, S., Yu, C., Wang, F., Huang, Q.: Geovideo: Introducing geometric regularization into video generation model (2025), https://arxiv.org/abs/2512.03453
5. Ball, P.J., Bauer, J., Belletti, F., Brownfield, B., Ephrat, A., Fruchter, S., Gupta, A., Holsheimer, K., Holynski, A., Hron, J., Kaplanis, C., Limont, M., McGill, M., Oliveira, Y., Parker-Holder, J., Perbet, F., Scully, G., Shar, J., Spencer, S., Tov, O., Villegas, R., Wang, E., Yung, J., Baetu, C., Berbel, J., Bridson, D., Bruce, J., Buttimore, G., Chakera, S., Chandra, B., Collins, P., Cullum, A., Damoc, B., Dasagi, V., Gazeau, M., Gbadamosi, C., Han, W., Hirst, E., Kachra, A., Kerley, L., Kjems, K., Knoepfel, E., Koriakin, V., Lo, J., Lu, C., Mehring, Z., Moufarek, A., Nandwani, H., Oliveira, V., Pardo, F., Park, J., Pierson, A., Poole, B., Ran, H., Salimans, T., Sanchez, M., Saprykin, I., Shen, A., Sidhwani, S., Smith, D., Stanton, J., Tomlinson, H., Vijaykumar, D., Wang, L., Wingfield, P., Wong, N., Xu, K., Yew, C., Young, N., Zubov, V., Eck, D., Erhan, D., Kavukcuoglu, K., Hassabis, D., Gharamani, Z., Hadsell, R., van den Oord, A., Mosseri, I., Bolton, A., Singh, S., Rocktäschel, T.: Genie 3: A new frontier for world models (2025)
6. Besl, P.J., McKay, N.D.: Method for registration of 3-d shapes. In: Sensor fusion IV: control paradigms and data structures. vol. 1611, pp. 586–606. Spie (1992)
7. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
8. Bozic, A., Palafox, P., Zollhöfer, M., Dai, A., Thies, J., Nießner, M.: Neural non-rigid tracking. Advances in Neural Information Processing Systems **33**, 18727–18737 (2020)
9. Chen, D.Z., Li, H., Lee, H., Tulyakov, S., Nießner, M.: Scenetex: High-quality texture synthesis for indoor scenes via diffusion priors. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 21081–21091. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.01992, https://doi.org/10.1109/CVPR52733.2024.01992
10. Chen, L., Zhou, Z., Zhao, M., Wang, Y., Zhang, G., Huang, W., Sun, H., Wen, J.R., Li, C.: Flexworld: Progressively expanding 3d scenes for flexiable-view synthesis. arXiv preprint arXiv:2503.13265 (2025)

11. Chen, Y., Medioni, G.: Object modelling by registration of multiple range images. Image and vision computing **10**(3), 145–155 (1992)
12. Chen, Y., He, T., Huang, D., Ye, W., Chen, S., Tang, J., Chen, X., Cai, Z., Yang, L., Yu, G., et al.: Meshanything: Artist-created mesh generation with autoregressive transformers. arXiv preprint arXiv:2406.10163 (2024)
13. Choi, S., Zhou, Q.Y., Koltun, V.: Robust reconstruction of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5556–5565 (2015)
14. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 303–312 (1996)
15. Danier, D., Gao, G., McDonagh, S., Li, C., Bilen, H., Aodha, O.M.: View-consistent diffusion representations for 3d-consistent video generation (2025), https://arxiv.org/abs/2511.18991
16. Duan, H., Yu, H.X., Chen, S., Fei-Fei, L., Wu, J.: Worldscore: A unified evaluation benchmark for world generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27713–27724 (2025)
17. Duisterhof, B., Zust, L., Weinzaepfel, P., Leroy, V., Cabon, Y., Revaud, J.: Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion (2024), https://arxiv.org/abs/2409.19152
18. Edstedt, J., Nordström, D., Zhang, Y., Bökman, G., Astermark, J., Larsson, V., Heyden, A., Kahl, F., Wadenbäck, M., Felsberg, M.: Roma v2: Harder better faster denser feature matching. arXiv preprint arXiv:2511.15706 (2025)
19. 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)
20. Fischer, T., Bulò, S.R., Yang, Y.H., Keetha, N., Porzi, L., Müller, N., Schwarz, K., Luiten, J., Pollefeys, M., Kontschieder, P.: Flowr: Flowing from sparse to dense 3d reconstructions (2025), https://arxiv.org/abs/2504.01647
21. Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P.P., Barron, J.T., Poole, B.: Cat3d: Create anything in 3d with multi-view diffusion models. Advances in Neural Information Processing Systems (2024)
22. Höllein, L., Božič, A., Müller, N., Novotny, D., Tseng, H.Y., Richardt, C., Zollhöfer, M., Nießner, M.: Viewdiff: 3d-consistent image generation with text-to-image models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5043–5052 (2024)
23. Höllein, L., Cao, A., Owens, A., Johnson, J., Nießner, M.: Text2room: Extracting textured 3d meshes from 2d text-to-image models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7909–7920 (2023)
24. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: ACM SIGGRAPH 2024 conference papers. pp. 1–11 (2024)
25. Huang, J., Dai, A., Guibas, L.J., Nießner, M.: 3dlite: towards commodity 3d scanning for content creation. ACM Trans. Graph. **36**(6), 203–1 (2017)
26. Huang, T., Zheng, W., Wang, T., Liu, Y., Wang, Z., Wu, J., Jiang, J., Li, H., Lau, R.W., Zuo, W., Guo, C.: Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. arXiv preprint arXiv:2506.04225 (2025)
27. Huang, Z., Wang, P., Zhang, J., Liu, Y., Li, X., Wang, W.: 3r-gs: Best practice in optimizing camera poses along with 3dgs (2025), https://arxiv.org/abs/2504.04294

28. HunyuanWorld, T.: Hy-world 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. arXiv preprint (2025)
29. Innmann, M., Zollhöfer, M., Nießner, M., Theobalt, C., Stamminger, M.: Volumedeform: Real-time volumetric non-rigid reconstruction. In: European conference on computer vision. pp. 362–379. Springer (2016)
30. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., et al.: Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In: Proceedings of the 24th annual ACM symposium on User interface software and technology. pp. 559–568 (2011)
31. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
32. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025)
33. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
34. Kheradmand, S., Rebain, D., Sharma, G., Sun, W., Tseng, Y.C., Isack, H., Kar, A., Tagliasacchi, A., Yi, K.M.: 3d gaussian splatting as markov chain monte carlo. Advances in Neural Information Processing Systems **37**, 80965–80986 (2024)
35. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
36. Kotovenko, D., Grebenkova, O., Ommer, B.: Edgs: Eliminating densification for efficient convergence of 3dgs. arXiv preprint arXiv:2504.13204 (2025)
37. Kupyn, O., Manhardt, F., Tombari, F., Rupprecht, C.: Epipolar geometry improves video generation models. arXiv preprint arXiv:2510.21615 (2025)
38. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2023)
39. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r (2024), `https://arxiv.org/abs/2406.09756`
40. Li, X., Wang, T., Gu, Z., Zhang, S., Guo, C., Cao, L.: Flashworld: High-quality 3d scene generation within seconds. arXiv preprint arXiv:2510.13678 (2025)
41. Liang, H., Cao, J., Goel, V., Qian, G., Korolev, S., Terzopoulos, D., Plataniotis, K.N., Tulyakov, S., Ren, J.: Wonderland: Navigating 3d scenes from a single image. arXiv preprint arXiv:2412.12091 (2024)
42. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
43. Liu, Y., Luo, C., Tang, Z., Peng, J., Zhang, Z.: Vggt-x: When vggt meets dense novel view synthesis. arXiv preprint arXiv:2509.25191 (2025)
44. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. arXiv preprint arXiv:2309.03453 (2023)
45. de Lutio, R., Fischer, T., Chang, Y.Y., Zhang, Y., Wu, J.Z., Ren, X., Shen, T., Tothova, K., Gojcic, Z., Turki, H.: Artifixer: Enhancing and extending 3d reconstruction with auto-regressive diffusion models (2026), `https://arxiv.org/abs/2603.00492`
46. Maggio, D., Lim, H., Carlone, L.: Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. arXiv preprint arXiv:2505.12549 (2025)

47. Mescheder, L., Dong, W., Li, S., Bai, X., Santos, M., Hu, P., Lecouat, B., Zhen, M., Delaunoy, A., Fang, T., Tsin, Y., Richter, S.R., Koltun, V.: Sharp monocular view synthesis in less than a second (2026), https://arxiv.org/abs/2512.10685
48. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
49. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics **41**(4), 1–15 (Jul 2022). https://doi.org/10.1145/3528223.3530127, http://dx.doi.org/10.1145/3528223.3530127
50. Newcombe, R.A., Fox, D., Seitz, S.M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 343–352 (2015)
51. Nießner, M., Zollhöfer, M., Izadi, S., Stamminger, M.: Real-time 3d reconstruction at scale using voxel hashing. ACM Transactions on Graphics (ToG) **32**(6), 1–11 (2013)
52. Park, J., Zhou, Q.Y., Koltun, V.: Colored point cloud registration revisited. In: Proceedings of the IEEE international conference on computer vision. pp. 143–152 (2017)
53. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields (2021), https://arxiv.org/abs/2011.12948
54. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net (2023), https://openreview.net/forum?id=FjNys5c7VyY
55. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022)
56. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. arXiv preprint arXiv:2503.03751 (2025)
57. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022), https://arxiv.org/abs/2112.10752
58. Rusinkiewicz, S., Levoy, M.: Efficient variants of the icp algorithm. In: Proceedings third international conference on 3-D digital imaging and modeling. pp. 145–152. IEEE (2001)
59. Sabour, S., Vora, S., Duckworth, D., Krasin, I., Fleet, D.J., Tagliasacchi, A.: Robustnerf: Ignoring distractors with robust losses. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20626–20636 (2023)
60. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022)
61. Schneider, M.A., Höllein, L., Nießner, M.: Worldexplorer: Towards generating fully navigable 3d scenes. In: SIGGRAPH Asia 2025 Conference Papers. pp. 1–11 (2025)
62. Schuhmann, C.: Improved aesthetic predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor (2022), gitHub repository

63. Schult, J., Tsai, S., Höllein, L., Wu, B., Wang, J., Ma, C.Y., Li, K., Wang, X., Wimbauer, F., He, Z., Zhang, P., Leibe, B., Vajda, P., Hou, J.: Controlroom3d: Room generation using semantic proxy rooms (2023), https://arxiv.org/abs/2312.05208
64. Segal, A., Haehnel, D., Thrun, S., et al.: Generalized-icp. In: Robotics: science and systems. vol. 2, p. 435. Seattle, WA (2009)
65. Siddiqui, Y., Alliegro, A., Artemov, A., Tommasi, T., Sirigatti, D., Rosov, V., Dai, A., Nießner, M.: Meshgpt: Generating triangle meshes with decoder-only transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19615–19625 (2024)
66. Siddiqui, Y., Monnier, T., Kokkinos, F., Kariya, M., Kleiman, Y., Garreau, E., Gafni, O., Neverova, N., Vedaldi, A., Shapovalov, R., Novotný, D.: Meta 3d assetgen: Text-to-mesh generation with high-quality geometry, texture, and PBR materials. CoRR **abs/2407.02445** (2024). https://doi.org/10.48550/ARXIV.2407.02445, https://doi.org/10.48550/arXiv.2407.02445
67. Smart, B., Zheng, C., Laina, I., Prisacariu, V.A.: Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs. arXiv preprint arXiv:2408.13912 (2024)
68. Song, K., Chen, B., Simchowitz, M., Du, Y., Tedrake, R., Sitzmann, V.: History-guided video diffusion. arXiv preprint arXiv:2502.06764 (2025)
69. Sorkine, O., Alexa, M.: As-rigid-as-possible surface modeling. In: Proceedings of EUROGRAPHICS/ACM SIGGRAPH Symposium on Geometry Processing. pp. 109–116 (2007)
70. Sun, W., Chen, S., Liu, F., Chen, Z., Duan, Y., Zhang, J., Wang, Y.: Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. In: International Conference on Computer Vision (ICCV) (2025)
71. Sun, W., Zhang, H., Wang, H., Wu, J., Wang, Z., Wang, Z., Wang, Y., Zhang, J., Wang, T., Guo, C.: Worldplay: Towards long-term geometric consistency for real-time interactive world model. arXiv preprint (2025)
72. Szymanowicz, S., Zhang, J.Y., Srinivasan, P., Gao, R., Brussee, A., Holynski, A., Martin-Brualla, R., Barron, J.T., Henzler, P.: Bolt3D: Generating 3D Scenes in Seconds. arXiv:2503.14445 (2025)
73. Tang, S., Zhang, F., Chen, J., Wang, P., Furukawa, Y.: Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. arXiv (2023)
74. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow (2020), https://arxiv.org/abs/2003.12039
75. Teed, Z., Deng, J.: Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras (2022), https://arxiv.org/abs/2108.10869
76. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models (2025), https://arxiv.org/abs/2503.20314
77. Wang, J., Chan, K.C.K., Loy, C.C.: Exploring clip for assessing the look and feel of images (2022), https://arxiv.org/abs/2207.12396

78. 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)
79. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. Advances in Neural Information Processing Systems **36**, 8406–8441 (2023)
80. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
81. Warburg, F., Weber, E., Tancik, M., Holynski, A., Kanazawa, A.: Nerfbusters: Removing ghostly artifacts from casually captured nerfs (2023), https://arxiv.org/abs/2304.10532
82. Wei, J., Leutenegger, S., Schaefer, S.: Gsfix3d: Diffusion-guided repair of novel views in gaussian splatting. arXiv preprint arXiv:2508.14717 (2025)
83. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20310–20320 (2024)
84. Wu, J.Z., Zhang, Y., Turki, H., Ren, X., Gao, J., Shou, M.Z., Fidler, S., Gojcic, Z., Ling, H.: Difix3d+: Improving 3d reconstructions with single-step diffusion models (2025), https://arxiv.org/abs/2503.01774
85. Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P.P., Verbin, D., Barron, J.T., Poole, B., et al.: Reconfusion: 3d reconstruction with diffusion priors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21551–21561 (2024)
86. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
87. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20331–20341 (2024)
88. 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. arXiv preprint arXiv:2409.02048 (2024)
89. Yunus, R., Lenssen, J.E., Niemeyer, M., Liao, Y., Rupprecht, C., Theobalt, C., Pons-Moll, G., Huang, J.B., Golyanik, V., Ilg, E.: Recent trends in 3d reconstruction of general non-rigid scenes. In: Computer Graphics Forum. vol. 43, p. e15062. Wiley Online Library (2024)
90. Zhang, B., Fang, C., Shrestha, R., Liang, Y., Long, X., Tan, P.: Rade-gs: Rasterizing depth in gaussian splatting. arXiv preprint arXiv:2406.01467 (2024)
91. Zhou, J.J., Gao, H., Voleti, V., Vasishta, A., Yao, C.H., Boss, M., Torr, P., Rupprecht, C., Jampani, V.: Stable virtual camera: Generative view synthesis with diffusion models. arXiv preprint arXiv:2503.14489 (2025)
92. Zhou, Q.Y., Koltun, V.: Color map optimization for 3d reconstruction with consumer depth cameras. ACM Transactions on Graphics (ToG) **33**(4), 1–10 (2014)
93. Zhou, Q.Y., Park, J., Koltun, V.: Fast global registration. In: European conference on computer vision. pp. 766–782. Springer (2016)
94. Zollhöfer, M., Nießner, M., Izadi, S., Rehmann, C., Zach, C., Fisher, M., Wu, C., Fitzgibbon, A., Loop, C., Theobalt, C., et al.: Real-time non-rigid reconstruction using an rgb-d camera. ACM Transactions on Graphics (ToG) **33**(4), 1–12 (2014)

A Supplemental Video

Please watch our attached video for a comprehensive evaluation of the proposed method. We include rendered videos of multiple generated scenes from novel trajectories, that showcase the quality of both single video 3D reconstructions and our large-scale scenes. This highlights our methods abilities to create consistent 3D worlds from inconsistent views with high quality renderings and fewer floating artifacts than the baseline methods.

B Extended Baseline Discussion

B.1 Comparison Against Dynamic Reconstruction

Fig. A1: Comparison against dynamic reconstruction. We compare against "Deformable 3DGS" [87] on reconstructing inconsistent generated video sequences. Since they do not model an aligned canonical space, their renderings both with deformation (dyn) and without (static) suffer from floating artifacts and decreased image quality.

After the geometry alignment we have obtained $\mathcal{P}[0]$, which represents an aligned 3D scene, as well as the non-rigid *forward* deformations $\{\mathbf{R}_i, \mathbf{t}_i, \mathcal{F}_{\theta_i}\}_{i=1}^{N}$ from each frame to the first frame (see Section 3.2). This is similar to a 4D scene reconstruction that is obtained in the dynamic reconstruction literature [50, 53, 83, 89]. We compare against "Deformable 3DGS" [87] as an exemplary method of that line of work. Concretely, they assign a timestep to each viewpoint and a learnable deformation network then transforms all canonical Gaussians to

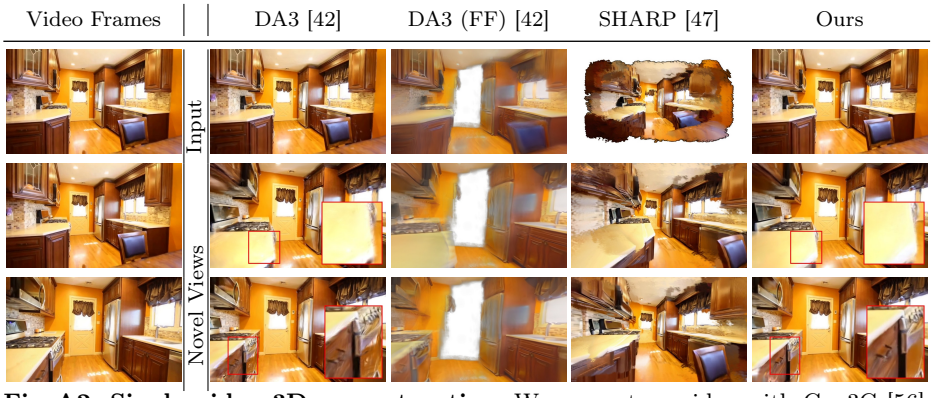

Fig. A2: Single video 3D reconstruction. We generate a video with Gen3C [56] and 3D reconstruct the frames. The feedforward 3DGS prediction from DA3 [42] suffers from decreased quality in comparison to optimizing 3DGS from scratch from the DA3 predicted depth/cams. Combining SHARP [47] and DA3 to obtain feedforward 3DGS predictions from multiple cameras suffers from the misaligned world space of the DA3 predictions (even though we utilize our optimized poses). In contrast, our method obtains high-quality renderings from inconsistent views by lifting the corresponding geometry into an aligned space.

the dynamic scene state at that timestep. At inference, this allows to replay the dynamically moving scene from any timestep at any observed (or novel) view. These methods define a canonical scene state from which the Gaussians are warped either as the first timestep or as a learned embedding. However, when applying this principle to the dynamic reconstruction of inconsistent generated video sequences, we observe severe limitations (see Figure A1). The key observation is that these methods *do not model* an aligned canonical space. Instead, the gradients from the rendering loss incentify a strong deformation network to explain the training views from *any canonical scene constellation*. This is sufficient to *reproduce* the inconsistent views when applying the deformation network at inference time (marked as "dyn"). However, when rendering the scene at a fixed timestep (to simulate a static and unified 3D reconstruction from inconsistent views; marked as "static"), severe degradation becomes visible through floating artifacts. In contrast, our method first optimizes for a geometric alignment, that explicitly models the canonical space. Using this as a basis for deformable Gaussian Splatting optimization leads to consistent and high-quality 3D reconstructions.

B.2 Discussion Of DA3 Feedforward 3DGS Prediction

In our main results, we utilize the DA3 [42] points as 3DGS [33] initialization and regularize their positions with a depth loss [90] against the DA3 depth. We refer to this as "DA3" [42] and additionally compare against DA3 by utilizing its feed-forward 3DGS prediction head in Figure A2 (marked as "DA3 (FF)"). We

noticed severe quality degradation both in terms of color-balance and background density in comparison to re-optimizing the 3DGS scenes, across all available released model checkpoints. To ensure a fair comparison, we thus opt for the "DA3" comparison described above. It is conceptually similar to the capabilities of the feed-forward head, but without the quality degradations. Concretely, the FF head predicts the remaining Gaussian attributes and places them at the positions obtain from unprojecting the predicted depth maps. Similarly, the point-based initialization and depth loss in our "DA3" implementation ensures that the Gaussians are positioned in this way.

B.3 Comparison Against SHARP

Recently, the monocular Gaussian Splatting predictor SHARP [47] demonstrated impressive results on novel view synthesis from a single image. We propose a baseline comparison on our task by predicting and merging per-view 3DGS attributes from our generated video sequence with this method. Since the per-view predictions do not yet reside in a unified world space, we utilize our cameras $\{\mathbf{R}_i, \mathbf{t}_i\}_{i=1}^{N}$ that we optimize from the DA3 [42] predictions to align them. However, Figure A2 shows that this is insufficient to obtain aligned worlds, which highlights the need for our non-rigid alignment and reconstruction stages. While the per-frame predictions render high-quality textures, they remain geometrically unaligned, since the input generated frames are inconsistent. In contrast, our full approach is able to produce aligned and high-quality surfaces.

C Additional Implementation Details

C.1 Voxelized Confidence Filter

global 0 global 10 global 20 global 40 ours voxel-based

Fig. A3: Voxel-based confidence filtering. We propose a point pruning strategy of DA3 [42] predictions to retain reliable, but spatially dense geometry, based on the predicted confidence scores. Naive global filtering schemes can prune too few points (global 0; noticeable stretch artifacts), which makes it hard to obtain high quality surface alignments. Global filtering with too high thresholds (global 10-40) can remove important scene content (e.g., the lamp or windows), which requires additional densification in the reconstruction stage. In contrast, our voxel-based filtering obtains dense surfaces that can directly be aligned in the subsequent stages and require no further densification to model complete scenes.

After unprojecting the initial DA3 [42] predictions into the pointcloud $\bar{\mathcal{P}}$, we propose a voxelized confidence filtering scheme to obtain the scene initialization

$\mathcal{P}$ (see Section 3.1). We compare this filtering scheme against common global percentile thresholds that are similarly based on the predicted confidence scores in Figure A3. The challenge with this global filtering scheme is to find reliable thresholds across scenes that do not prune too few or too much points. Pruning too few points creates a scene initialization with many stretched-out artifacts between objects. This makes our subsequent geometry alignment stage harder, as the nearest neighbor correspondence search in the ICP terms is less reliable and struggles to create unified surfaces. In contrast, removing too much points creates hard to reconstruct background regions in the photometric Gaussian Splatting optimization. While this can be tackled with densification, the introduction of additional Gaussian parameters undermines the previous efforts of obtaining an aligned pointcloud geometry: they are harder to regularize onto a single surface and could re-introduce floating artifacts if the corresponding background regions are inconsistently generated in the input sequences. Our voxel-based strategy solves both problems by retaining only reliable points, but densely across the entire scenes. This makes it possible to optimize our final scene representation without densification and thus we obtain 3D consistent worlds that can render complete and high-quality images also for the background regions.

C.2 Details Of Quantitative Evaluation

We quantitatively evaluate our main results by calculating consistency and fidelity metrics of the established WorldScore [16] benchmark. Here, we provide additional details. Since our input video sequences are generated with video diffusion models, the individual frames are not 3D consistent (i.e., they suffer from generative drift; see Figure 1). Common reconstruction metrics like PSNR, SSIM [80], LPIPS [31] compare rendered and observed images. However, since the input frames are inconsistent, these metrics would only measure the degree to which the respective reconstruction methods can *reproduce the inconsistencies*. In contrast, we are interested to measure how geometrically consistent the 3D scenes are. This entails that renderings from input camera poses do not precisely (i.e., pixel-wise) reproduce the input frames, but instead small movements of individual objects become noticeable, that solve the distortions inherent in the input frames. To this end, we follow WorldScore [16] and calculate two types of consistency metrics. The "3D consistency" measures reprojection error from one frame to another based on a reconstruction of the rendered frames with DROID-SLAM [75]. Similarly, the "photometric consistency" estimates the optical flow between consecutive frames using RAFT [74] and calculates the warping error. We are also interested to reason about the methods abilities to render sharp textures without blurriness or floating artifacts. To quantify this image quality, we additionally adopt the fidelity scores CLIP-IQA+ [77] and CLIP Aesthetic [62].

We calculate these metrics on re-renderings of the input video sequence after 3D reconstruction that we uniformly downsample to 100 frames (between 3-5 second clips on average). This ensures that image fidelity is not negatively impacted by rendering novel views that show unoptimized areas like backgrounds. It also allows to compare the image fidelity against the video frames itself, which serve

as the upper bound in this category. Importantly, this still allows to quantify the 3D consistency, since the input frames themselves are inconsistent. Naively reproducing them thus yields lower scores in these metrics. We report average results across all source video diffusion models (VDM) for each reconstruction method, which amounts to 16 different scenes (2-3 per VDM).

D Limitations

Our method reconstructs 3D scenes from inconsistent generated views, however, some limitations remain. We tackle the problem of *generative drift* in video diffusion models (VDM), i.e., existing objects move geometrically inconsistent. However, VDMs can also suffer from hallucinations (e.g., revisiting previous scene areas creates novel or removes existing objects, changes their texture, etc.). We visualize this phenomenon in Figure A4, where a street lamp suddenly appears in the input video sequence upon revisiting that area. Our aligned pointcloud geometry thus includes this object (it is still aligned from multiple frames that show the object). Our optimized 3D scene renders the street lamp in both input perspectives, since the optimization is jointly performed from all views on a single static 3D world and thus averages out inconsistent viewpoints. Adopting robust reconstruction mechanisms to detect such outlier frames could resolve these problems [59].

Fig. A4: Limitations. Our method reconstructs 3D consistent worlds from inconsistent video sequences through geometry alignment and non-rigid Gaussian Splatting optimization. However, video diffusion models can also hallucinate novel content when revisiting areas (e.g., the street lamp appears suddenly). We can still align multiple frames that show such hallucinations, but they become a part of the 3D-consistent static geometry. Additionally, 3D reconstruction optimizes for the average explanation of all input views and as such, these inconsistencies stay visible (e.g., the street lamp is visible in both viewpoints when rendering from our 3D scene).

E Additional Results

E.1 Single Video 3D Reconstruction

We show additional results on the single video 3D reconstruction task in Figures A5 and A6. Our method renders sharper and more detailed textures that

are of comparable visual fidelity as the input video frames, just in a 3D consistent space. In contrast, the baselines do not correct for the generative drift inherent in the generated frames, which leads to blurrier renderings (especially noticeable from novel perspectives). Our method optimizes the scenes in a *non-rigid aware* fashion which leads to greatly improved viewpoint stability, i.e., we can explore the 3D worlds from novel perspectives while retaining the high visual fidelity of the video frames. Our method effectively turns any video diffusion model into a 3D world generator that allows persistent, high quality, and real-time rendering.

Fig. A5: Single video 3D reconstruction. We generate videos with ViewCrafter [88] (top), Voyager [26] (mid), Genie3 [5] (bottom) and 3D reconstruct these frames. Inconsistencies in the generations lead to blurry textures for the baselines compared to the corresponding video, and to floating artifacts from novel views. In contrast, our method creates 3D consistent worlds with high fidelity beyond generated perspectives.

We additionally compare the recent method Voyager [26] more closely in Figure A7. They propose a joint modeling of RGB-D sequences in the latent

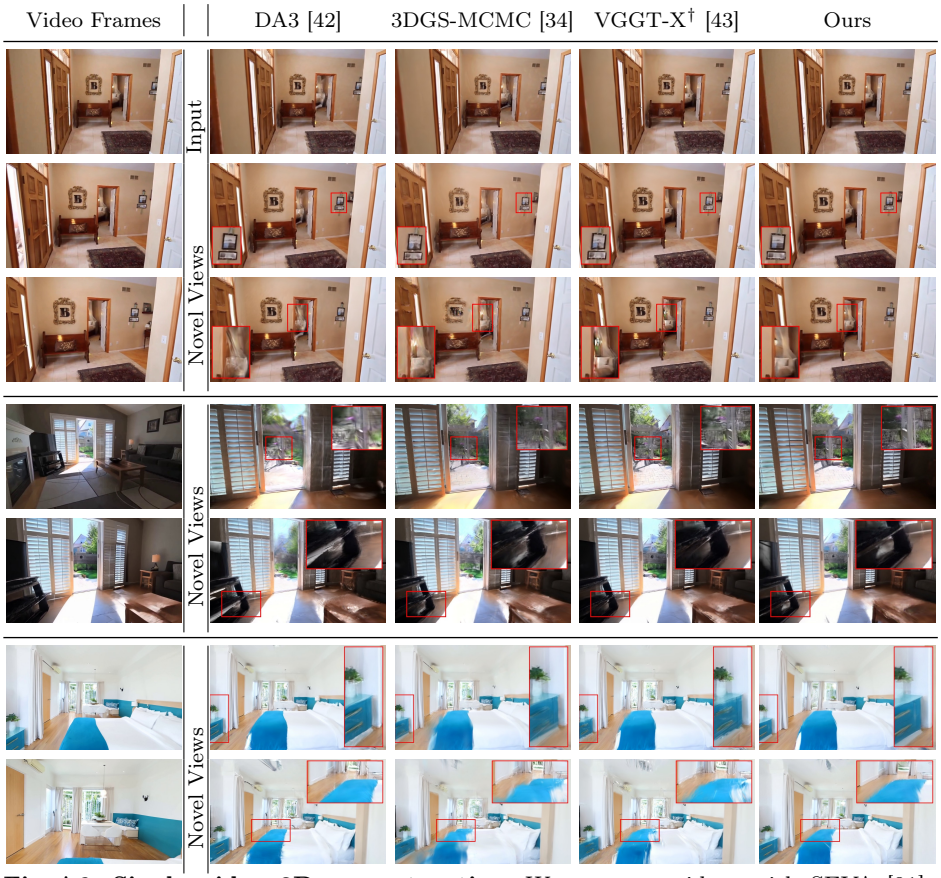

Fig. A6: Single video 3D reconstruction. We generate videos with SEVA [91] (top), Gen3C [56] (mid), Wan [76] (bottom) and 3D reconstruct these frames. Inconsistencies in the generations lead to blurry textures for the baselines compared to the corresponding video, and to floating artifacts from novel views. In contrast, our method creates 3D consistent worlds with high fidelity beyond generated perspectives.

Fig. A7: Single video 3D reconstruction. We generate videos with Voyager [26] and 3D reconstruct these frames. Our method optimizes consistent worlds from inconsistent generated frames. Compared to baselines, the renderings are of higher visual fidelity from both input and novel views. The pointcloud visualizations demonstrate that our method obtains aligned surfaces with compelling textures, whereas the baselines produce sparser or less aligned geometry.

space of a video diffusion model, which allows for explicit 3D consistency supervision during training and thus yields improved consistency in the generated output sequence. Nevertheless, generative drift still remains noticeable in the reconstructed and rendered scenes (Figure A7 top). Additionally, the generated RGB-D sequence still contains unaligned and overlapping surfaces with similar artifacts as the DA3 [42] predictions. This underlines the motivation of our approach, which proposes a lightweight alignment on the reconstruction side (instead of further finetuning of a generative model). Our pointcloud geometry contains single, aligned surfaces with compelling texture quality, which serves as strong initialization for photometric optimization with Gaussian Splatting [24, 33].

E.2 Large-Scale World Generation

We show additional results on the large-scale world generation task in Figures A8 to A10. Concretely, we generate and then reconstruct up to 32 video sequences with SEVA [91] via the progressive scene expansion strategy of [61]. The generative drift grows larger across multiple videos and thus the 3D reconstructions suffer from increased consistency problems. While the baselines can depict complete and diverse worlds, the amount of exploration they enable is still limited (i.e., novel views far beyond the training poses suffer from extensive floating artifacts). In contrast, our worlds remain consistent and retain a high rendering fidelity, even from extreme novel perspectives.

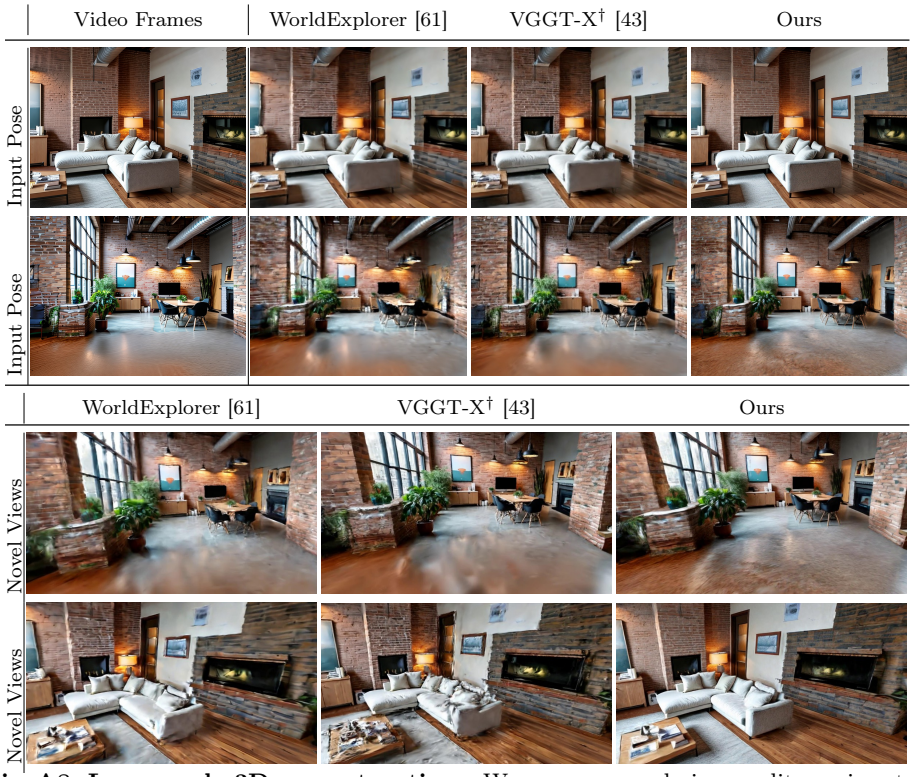

Fig. A8: Large-scale 3D reconstructions. We compare rendering quality on input poses and novel views for entire 360 degree explorable scenes. Our method creates 3D consistent worlds with a high rendering fidelity far beyond the training views. Please see the supplementary video for animated flythroughs.

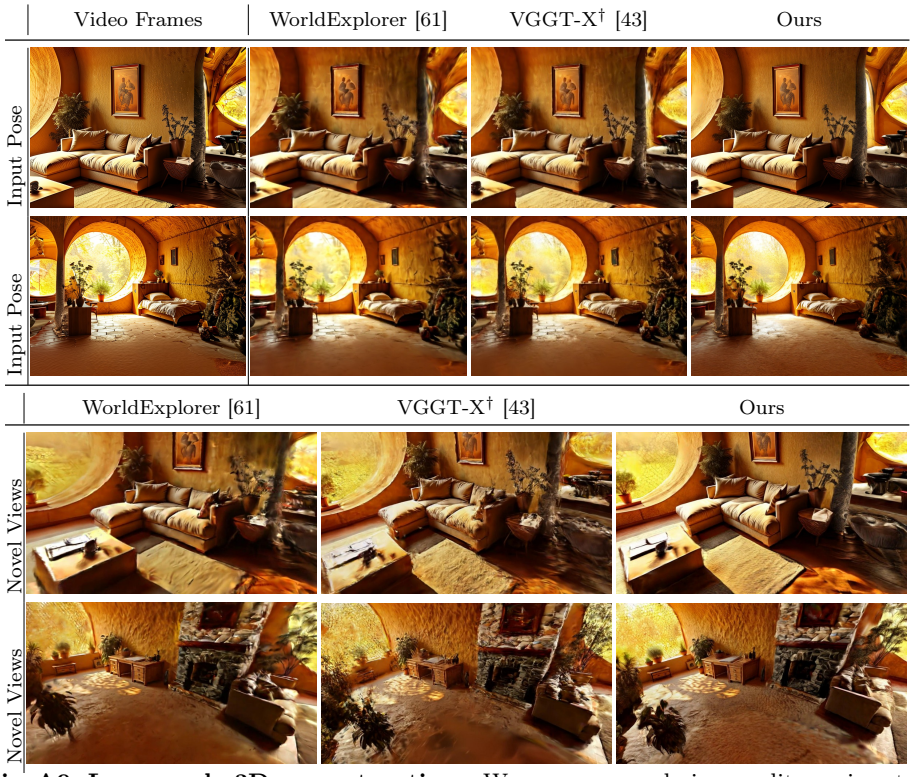

Fig. A9: Large-scale 3D reconstructions. We compare rendering quality on input poses and novel views for entire 360 degree explorable scenes. Our method creates 3D consistent worlds with a high rendering fidelity far beyond the training views. Please see the supplementary video for animated flythroughs.

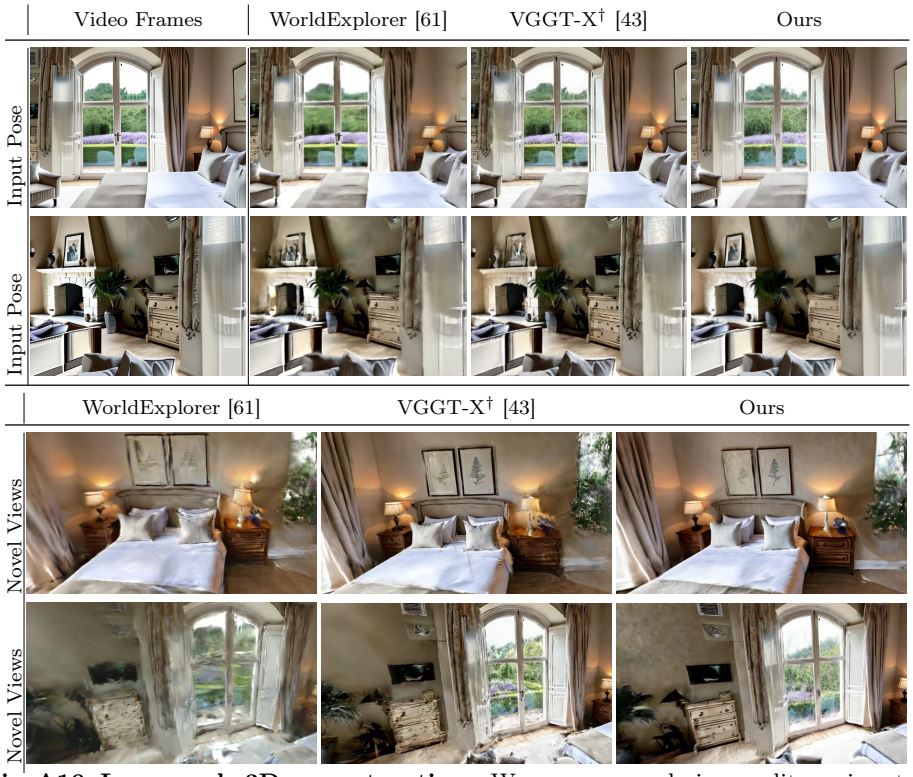

Fig. A10: Large-scale 3D reconstructions. We compare rendering quality on input poses and novel views for entire 360 degree explorable scenes. Our method creates 3D consistent worlds with a high rendering fidelity far beyond the training views. Please see the supplementary video for animated flythroughs.