Vista4D: Video Reshooting with 4D Point Clouds

Kuan Heng Lin[1,3†] Zhizheng Liu[1,4†] Pablo Salamanca[1,2] Yash Kant[1,2]
Ryan Burgert[1,2,5†] Yuancheng Xu[1,2] Koichi Namekata[1,2,6†] Yiwei Zhao[2]
Bolei Zhou[4] Micah Goldblum[3] Paul Debevec[1,2] Ning Yu[1,2]

[1]Eyeline Labs [2]Netflix [3]Columbia University [4]UCLA [5]Stony Brook University [6]University of Oxford

https://eyeline-labs.github.io/Vista4D

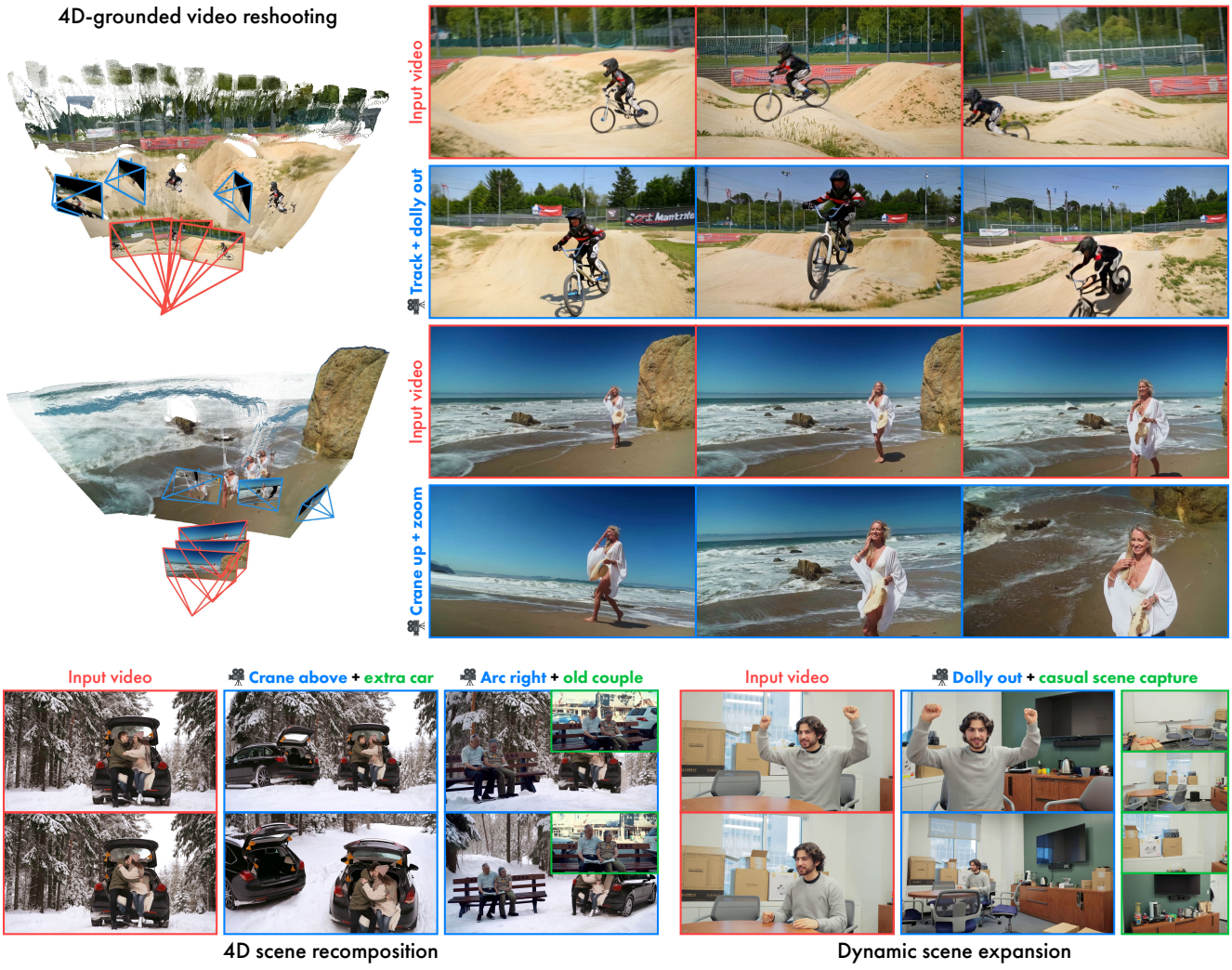

Figure 1. **4D-grounded video reshooting.** Given an input video, Vista4D re-synthesizes the scene with the same dynamics from different camera trajectories and viewpoints by grounding the input video and target cameras in a 4D point cloud. Vista4D is robust to point cloud artifacts and generalizes to real-world applications such as 4D scene recomposition and dynamic scene expansion.

[†]Work done during an internship at Eyeline Labs.

Abstract

We present Vista4D, a robust and flexible video reshooting framework that grounds the input video and target cameras in a 4D point cloud. Specifically, given an input video, our method re-synthesizes the scene with the same dynamics from a different camera trajectory and viewpoint. Existing video reshooting methods often struggle with depth estimation artifacts of real-world dynamic videos, while also failing to preserve content appearance and failing to maintain precise camera control for challenging new trajectories. We build a 4D-grounded point cloud representation with static pixel segmentation and 4D reconstruction to explicitly preserve seen content and provide rich camera signals, and we train with reconstructed multiview dynamic data for robustness against point cloud artifacts during real-world inference. Our results demonstrate improved 4D consistency, camera control, and visual quality compared to state-of-the-art baselines under a variety of videos and camera paths. Moreover, our method generalizes to real-world applications such as dynamic scene expansion and 4D scene recomposition.

1. Introduction

The camera is the visual portal to the filmmaker's world, guiding the audience's gaze as the story unfolds and constructing the narrative's visual language. While traditional visual effects can dramatically transform a raw film set into an immersive spectacle, the ability to manipulate the camera during post-production introduces another dimension of control over visual storytelling.

To this end, we synthesize or 'render' the dynamic scene specified by an input source video from novel camera trajectories and viewpoints, which we call *video reshooting*. Importantly, we must achieve faithful reconstruction of seen content in the source video and photorealistically plausible generation of unseen content, all while maintaining precise, user-definable camera control.

We will employ video diffusion models since they are powerful priors for generating dynamic content which is geometrically and temporally coherent [1–6]. We will further combine the diffusion models with 4D reconstruction which lifts the monocular source video into a 4D point cloud, providing spatiotemporal grounding for reconstruction and a rich signal for camera control. We present *Vista4D*, a video reshooting framework that grounds the source video and target cameras in a 4D point cloud with temporally-persistent static pixels, while leveraging the generative priors of video diffusion models.

Existing works for video reshooting [7–9] condition video diffusion models on per-frame depth-lifted point clouds rendered in the target cameras. However, they often struggle with geometry artifacts and/or temporal flickering due to imprecise 4D reconstruction of real-world dynamic videos as they are often trained on point cloud renders from precise depth maps. Moreover, they also struggle with accurate camera control and content preservation with challenging target camera trajectories and viewpoints.

Vista4D introduces the following key designs that not only show state-of-the-art visual quality and robustness to a wide variety of source videos and target cameras but also extend our method with capabilities beyond vanilla video reshooting. First, we build a 4D-grounded point cloud representation where static pixels are visible from any frame via segmentation and 4D reconstruction, as opposed to the per-frame 3D point cloud of baselines. Conditioning with static pixel temporal persistence establishes both explicit preservation of seen content and provides rich camera signals even when the target cameras have little per-frame overlap with the source video. Second, we augment model training with dynamic, 4D-reconstructed multiview video pairs that contain depth estimation artifacts from non-frontal views. Thus, Vista4D is significantly more robust to the quality of real-world point cloud renders while allowing us to additionally condition on the source video to utilize video model priors for geometric coherence. This further enables us to manipulate the 4D point cloud during inference for real-world applications beyond video reshooting.

Our contributions are as follows:

- We present Vista4D, a video reshooting framework that maintains geometric and physical plausibility with real-world inference, while explicitly preserving seen content by grounding generation in a 4D point cloud.
- Through extensive quantitative and qualitative comparisons, including a user study, we validate the improved content preservation, camera controllability, and visual quality of Vista4D over state-of-the-art baselines for a wide variety of videos and cameras.
- We show that our training extends Vista4D with capabilities that generalize to real-world applications such as dynamic scene expansion, 4D scene recomposition, and long video inference with memory.

2. Related work

Video reshooting with *explicit priors*. For video reshooting, and more broadly novel view synthesis of static scenes, 3D/4D point clouds provide an explicit and rich spatial prior. To this end, video reshooting methods with *explicit priors* [7–9, 12–14] use video depth estimators [15–17] to render per-frame camera-space point clouds as conditioning signals for video diffusion models. Depth estimation priors have also been widely used for static scene novel view synthesis (NVS) [18, 19] and video motion control [8, 20]. However, many of these methods often train on precise depth maps which inhibits generalization to imperfect real-world depth estimation, and their per-frame point cloud conditioning can struggle to preserve seen content and maintain accurate camera control with challenging camera trajectories.

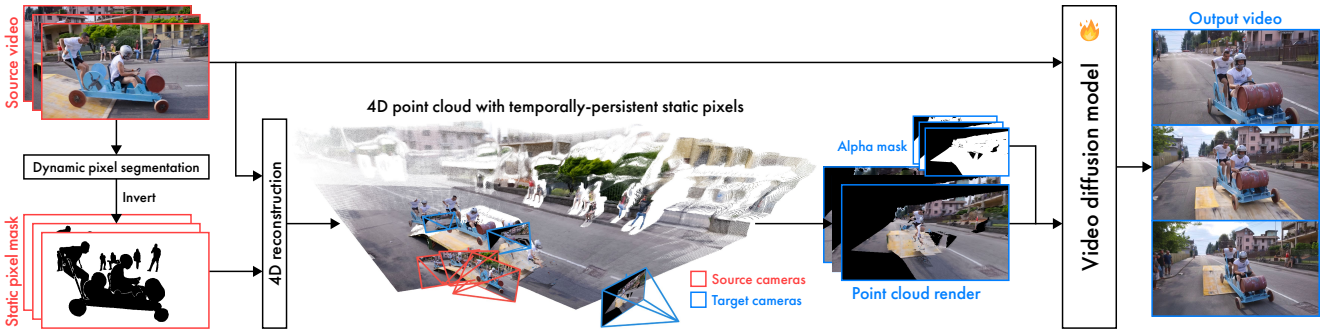

Figure 2. **Overview of Vista4D.** Given an input source video, we build a 4D point cloud where static pixels are temporally persistent via segmentation and 4D reconstruction. We then render the point cloud in the target cameras which users define. Lastly, the source video and point cloud render & alpha mask are jointly processed by the finetuned video diffusion model to generate a video of the same dynamic scene in the target cameras. We provide model architecture details in Supplementary B.

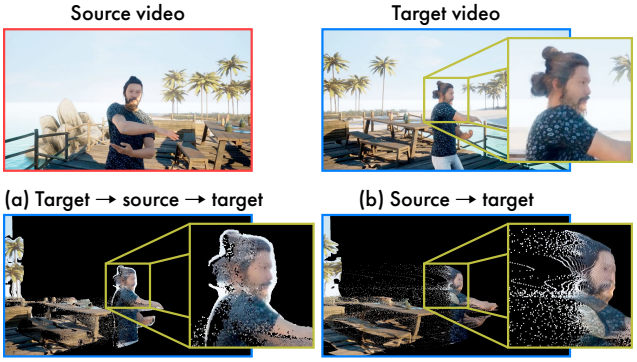

Figure 3. **Multiview 4D reconstruction artifacts.** (a) Double reprojection [7] first renders the target video point cloud in the source camera, then rerendering it in the target camera to create occluded regions for paired training, thus viewing the target video depth map from its frontal, artifact-free view. (b) In contrast, rendering the source video point cloud from the target camera with dynamic multiview data exposes non-frontal-view artifacts that better match real-world inference. The above source-target video pair is from MultiCamVideo [10] with 4D reconstruction by STream3R [11].

Video reshooting with *implicit priors*. Alternatively, video shooting methods can also use *implicit priors* for camera control such as camera embeddings [10, 21] or video references [22] by finetuning video diffusion models on time-synchronized synthetic multiview data. Image- and camera-conditioned diffusion models have also been used for static scene NVS [23–27] and camera-controlled video generation [28–30]. However, due to the inherent depth scale ambiguity of monocular videos, camera control from implicit-prior methods is often imprecise and unable to be explicitly 'previewed' unlike point clouds.

4D reconstruction. To provide explicit geometric priors for video reshooting, we lift the input video into a world-space point cloud with 4D reconstruction. Traditional structure from motion [31–33] rely on multiview geometry constraints but are not robust to dynamic scenes. With the strong performance of learning-based video depth estimation models [15–17, 34–36], recent works [37–39] combine these depth priors and camera optimization with SLAM [40] to obtain robust and coherent dynamic scene reconstruction. Followed by recent success in end-to-end 3D reconstruction methods [41–43], end-to-end 4D reconstruction models [11, 44–47] have also emerged as more efficient alternatives. Some recent methods also predict 4D Gaussians from monocular videos [48–50], enabling novel view synthesis at small viewpoint deviations from the input videos.

3. 4D-grounded video reshooting

Given an input source video $\mathbf{X}^{\text{src}}$, we first build a 4D point cloud via 4D reconstruction with temporally-persistent static pixels defined by static pixel masks from segmentation. We then render the point cloud from the target cameras and jointly condition the finetuned video diffusion model on the source video and point cloud render, producing the output video. Section 3.1 builds the temporally-persistent point cloud; Section 3.2 explains the importance of training with noisily-reconstructed multiview data; Section 3.3 discusses joint conditioning of source videos and point cloud renders; and Section 3.4 describes data and training details. Our method is illustrated in Figure 2.

3.1. Building a temporally-persistent 4D point cloud

To explicitly preserve seen content in the source video and provide more accurate camera control especially when target cameras have little per-frame overlap with the source video, we build a temporally-persistent 4D point cloud. We first use 4D reconstruction [11, 44] to obtain depths $\mathbf{D}^{\text{src}}$, camera extrinsics $\mathbf{T}^{\text{src}}$, and camera intrinsics $\mathbf{K}^{\text{src}}$, and we use segmentation [51–53] to obtain a static pixel mask $\mathbf{M}^{\text{stc}}$. We then lift the source video into a world-space per-frame 3D point cloud

$$\mathbf{P} = \Omega\left(\Phi^{-1}\left([\mathbf{X}^{\text{src}}, \mathbf{D}^{\text{src}}], \mathbf{K}^{\text{src}}\right), \mathbf{T}^{\text{src}}\right), \quad (1)$$

where Φ^{-1} and Ω are the inverse perspective projection and world-space transformation. Since the per-frame point

cloud $\mathbf{P}$ is grounded in world space, we use $\mathbf{M}^{\text{stc}}$ to make static pixels persistent across all frames to incorporate explicit 4D context in our point cloud rendering, obtaining the temporally-persistent point cloud $\overline{\mathbf{P}}$. Then, we render $\overline{\mathbf{P}}$ from the target cameras, obtaining the point cloud render $\mathbf{X}^{\text{src}\to\text{tgt}}$ and its alpha mask $\mathbf{M}^{\text{src}\to\text{tgt}}$ as temporally persistent, 4D-grounded priors for the video diffusion model.

3.2. Training with noisy multiview data

So far, generating our 4D point cloud requires source-target video pairs: The source video builds the temporally-persistent point cloud, and the target video defines the target cameras. Because 4D reconstruction methods are imperfect, the point cloud render during inference often contain *geometric artifacts* when the target cameras deviate far from the frontal view of the lifted point cloud. This is especially true for dynamic pixels where depth estimators cannot leverage multiview geometry constraints from moving cameras. Existing methods [7–9] instead train with artifact-free point clouds, which essentially simplify video reshooting to inpainting. For example, as illustrated in Figure 3 (a), TrajectoryCrafter [7] applies double-reprojection to monocular videos to obtain paired data of point cloud renders and target videos, which always views the depth maps from their frontal, artifact-free view. In contrast, we train with multiview dynamic-scene videos with 4D-reconstructed depths and cameras, which results in spatially mismatching point clouds artifacts compared to the target video as shown in Figure 3 (b). Thus, our method moves beyond inpainting and instead corrects imperfect point cloud geometry.

As real-world multiview video datasets with dynamic scenes are rare and small in scale, we use synthetic multiview dynamic videos to train our model as in [10]. Moreover, to ensure the model is generalizable to real-world video inputs while being robust to noisy 4D reconstruction, we train with a mix of multiview synthetic and real-world monocular data. For monocular data, following TrajectoryCrafter [7], we first render the point cloud of the target video from heuristic-generated source cameras to produce $\mathbf{X}^{\text{tgt}\to\text{src}}$. Then, we render $\mathbf{X}^{\text{tgt}\to\text{src}}$ back to the original target cameras to produce the double-reprojected point cloud render.

3.3. Conditioning on source videos and point clouds

Point cloud artifacts during real-world inference obfuscate not only geometry but also appearance information from the source video. Thus, while some existing methods only condition on point cloud renders [8, 9], we also condition on source videos to utilize video diffusion model priors for transferring geometric and appearance information like implicit-prior methods do [10, 22]. Unlike TrajectoryCrafter's cross-attention injection of source videos [7], we concatenate the patchified latent tokens of the source video and point cloud render with the noisy target latent tokens along the frame dimension. We find that in-context conditioning best preserves source video content and is thus more robust to point cloud artifacts, which we ablate in Supplementary F.

Table 1. **Camera control accuracy and 3D consistency.** Vista4D consistently shows the most accurate camera control compared to baselines with superior rotation, translation, and intrinsics errors. Our method also significantly outperforms baselines in per-frame 3D consistency with the lowest reprojection error under SuperGlue (RE@SG) [57–59]. **Bold** indicates best results.

Method	Translation error ⇓	Rotation error ⇓	Intrinsics error ⇓	RE@SG ⇓
ReCamMaster [10]	1.574	12.79	11.16	23.66
CamCloneMaster [22]	2.132	23.77	6.422	23.38
TrajectoryCrafter [7]	1.434	6.838	6.671	120.5
EX-4D [9]	1.325	5.941	5.182	13.11
GEN3C [8]	1.309	4.751	5.085	12.99
Vista4D (ours)	**1.251**	**4.647**	**4.927**	**7.504**

Table 2. **Novel-view video synthesis.** Vista4D shows comparable to superior noel-view video synthesis performance on the `iphone` dataset [60]. EPE (endpoint error) measures optical flow error between the generated and ground truth videos and indicates scene motion reconstruction. **Bold** indicates best results.

Method	mPSNR ⇑	mSSIM ⇑	mLPIPS ⇓	PSNR ⇑	SSIM ⇑	LPIPS ⇓	EPE ⇓
ReCamMaster [10]	10.84	0.444	0.692	10.96	0.262	0.755	4.681
CamCloneMaster [22]	11.14	0.444	0.651	11.17	0.260	0.713	4.318
TrajectoryCrafter [7]	13.82	**0.492**	0.569	13.06	**0.320**	0.656	2.375
EX-4D [9]	12.85	0.479	0.596	12.64	0.305	0.669	4.269
GEN3C [8]	12.19	0.447	0.608	12.06	0.260	0.679	3.019
Vista4D (ours)	**14.09**	0.480	**0.461**	**14.14**	0.310	**0.514**	**1.142**

Thus, given the source video $\mathbf{X}^{\text{src}}$, point cloud render $\mathbf{X}^{\text{src}\to\text{tgt}}$ and its alpha mask $\mathbf{M}^{\text{src}\to\text{tgt}}$, and target cameras $\mathbf{C}^{\text{tgt}} = (\mathbf{K}^{\text{tgt}}, \mathbf{T}^{\text{tgt}})$, we finetune a video diffusion transformer $\boldsymbol{\epsilon}_\theta$ to generate the target video $\mathbf{X}^{\text{tgt}}$ with the flow matching objective

$$\mathcal{L} = \left\| \boldsymbol{\epsilon}_\theta \left(\mathbf{X}_t^{\text{tgt}}, \mathbf{X}^{\text{src}\to\text{tgt}}, \mathbf{M}^{\text{src}\to\text{tgt}}, \mathbf{X}^{\text{src}}, \mathbf{C}^{\text{tgt}}, t \right) - \mathbf{V} \right\|, \tag{2}$$

where $\mathbf{V} = \mathbf{X}^{\text{tgt}} - \boldsymbol{\epsilon}$ and $\mathbf{X}_t^{\text{tgt}}$ is the noisy target video at timestep t by sampled Gaussian $\boldsymbol{\epsilon}$. We inject the target cameras $\mathbf{C}^{\text{tgt}}$ as Plücker embeddings [54–56] via zero-initialized linear projections, with an identity-initialized projection after self-attention, inspired by ReCamMaster [10]. We provide model architecture details in Supplementary B.

Conditioning the model with both the source video and point cloud render allows the model to learn to propagate geometry and appearance information from the source to the output video. For monocular training videos without a ground-truth $\mathbf{X}^{\text{src}}$, we condition the model on $\mathbf{X}^{\text{tgt}\to\text{src}}$ as an occluded source video with its alpha mask to still learn this propagation.

3.4. Training details and datasets

For the base video generation model, we build off of `Wan2.1-T2V-14B` [2], a pretrained text-to-video flow

Table 3. **Video fidelity.** Vista4D consistently outperform point-cloud-conditioned (explicit-prior) baselines for the video fidelity metrics FID, FVD, CLIP-T, and metrics from VBench [61] and VBench-2.0 [62]. Implicit-prior methods (ReCamMaster and CamCloneMaster) outperform our method in some metrics due to their low camera control accuracy (Table 1) that result in output videos with similar, usually more static, cameras to the input video which produces better FID, FVD, and VBench consistency metrics. **Bold** indicates best results.

Method	Camera control	FID ⇓	FVD $\times 10^3$ ⇓	CLIP-T ⇑	VBench [61] & VBench-2.0 [62]					
					Aesthetic quality ⇑	Imaging quality ⇑	Subject consistency ⇑	Background consistency ⇑	Temporal style ⇑	Human anatomy ⇑
ReCamMaster [10]	Extrinsics	**94.15**	**1.203**	0.319	0.552	0.701	**0.913**	**0.934**	0.243	0.759
CamCloneMaster [22]	Ref. video	101.4	1.406	0.321	0.560	0.709	0.886	0.915	0.247	0.711
TrajectoryCrafter [7]	Point cloud	125.6	1.640	0.305	0.509	0.650	0.854	0.906	0.241	0.790
EX-4D [9]	Point cloud	124.6	1.481	0.296	0.480	0.660	0.849	0.894	0.226	0.687
GEN3C [8]	Point cloud	113.5	1.441	0.318	0.519	0.660	0.857	0.913	0.245	0.775
Vista4D (ours)	Point cloud	105.4	1.418	**0.326**	**0.567**	**0.716**	0.883	0.916	**0.253**	**0.857**

Table 4. **User study.** Participants consistently perfer Vista4D over baselines on source video content preservation, camera control accuracy, and overall video fidelity. **Bold** indicates best results.

Method	Source preservation ⇑	Camera accuracy ⇑	Overall fidelity ⇑
ReCamMaster [10]	9.921%	1.905%	4.365%
CamCloneMaster [22]	15.63%	6.429%	11.03%
TrajectoryCrafter [7]	0.952%	5.952%	0.476%
EX-4D [9]	1.587%	6.508%	0.794%
GEN3C [8]	4.841%	11.03%	5.952%
Vista4D (ours)	**67.06%**	**68.17%**	**77.38%**

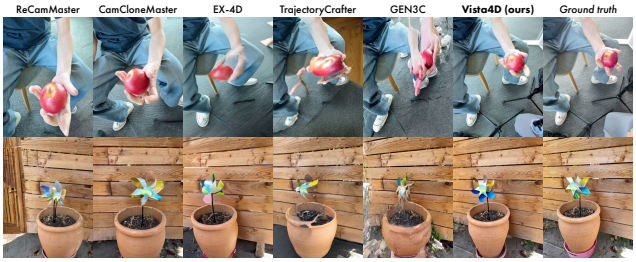

Figure 4. **Qualitative comparison on novel-view synthesis.** We show two samples of Vista4D compared our baselines on the `iphone` dataset [60].

matching [63] video diffusion transformer [64]. We fine-tune the model at a resolution of 672×384 for 30,000 steps, then at 1280×720 for 300 steps, both with 49-frame videos, a global batch size of 8, and the AdamW optimizer with a learning rate of 1×10^{-5}. We train the patchify layers for $\mathbf{X}^{\text{src}}$ and $\mathbf{X}^{\text{src}\to\text{tgt}}$, self-attention layers, camera encoders, and projectors, while freezing all other parameters.

Datasets. For multiview time-synchronized videos, we adopt the synthetic MultiCamVideo dataset from ReCamMaster [10], and we run 4D reconstruction across all views with STream3R [11]. For real-world monocular videos, we adopt a 60K subset from `OpenVidHD-0.4M` [65] and run 4D reconstruction with π^3 [44]. For segmenting static pixels, inspired by Uni4D [38], we obtain semantic classes with RAM [66], filter for dynamic subjects/nouns with `Llama-3.1-8B-Instruct` [67], and segment per-frame dynamic pixels with Grounded SAM 2 [51–53] and invert the resulting masks.

4. Experiments

Baselines. We compare Vista4D to state-of-the-art video reshooting methods. For explicit-prior methods, TrajectoryCrafter [7] introduces the double-reprojection technique to generate training pairs from monocular dynamic videos, EX-4D [9] proposes the Depth Watertight Mesh during inference to train on tracking-based inpainting, and GEN3C [8] builds a 3D cache with pooling-based fusion for sparse-view novel-view synthesis. For implicit-prior methods, ReCamMaster [10] constructs a synthetic multiview time-synchronized video dataset to train a camera-conditioned model, and CamCloneMaster [22] replicates camera trajectories from reference videos. We use our 672×384 checkpoint for all quantitative evaluations and the user study.

Evaluation dataset. For quantitative evaluation, we create an evaluation dataset of high quality, diverse 110 video-camera pairs: We select 51 videos from DAVIS [68] and the royalty-free stock video website Pexels [69]. Then, we run 4D reconstruction with π^3 [44] and segmentation with Grounded SAM 2 [51–53], and we design two to three camera trajectories for each video with our camera design UI, which we show examples of in Supplementary D.

4.1. Quantitative comparisons

We quantitatively compare Vista4D to baselines and show our method's superiority on three dimensions: Camera control and 3D consistency, novel-view video synthesis, and video fidelity. We include details of each quantitative evaluation metric in Supplementary E.

Camera control accuracy and 3D consistency. We compare camera control accuracy and 3D consistency Vista4D to baselines in Table 1 on our 110 video-camera pair dataset. For camera control accuracy, we measure translation, rotation, and intrinsics error between target cameras from the evaluation dataset and 4D-reconstruction-predicted cameras

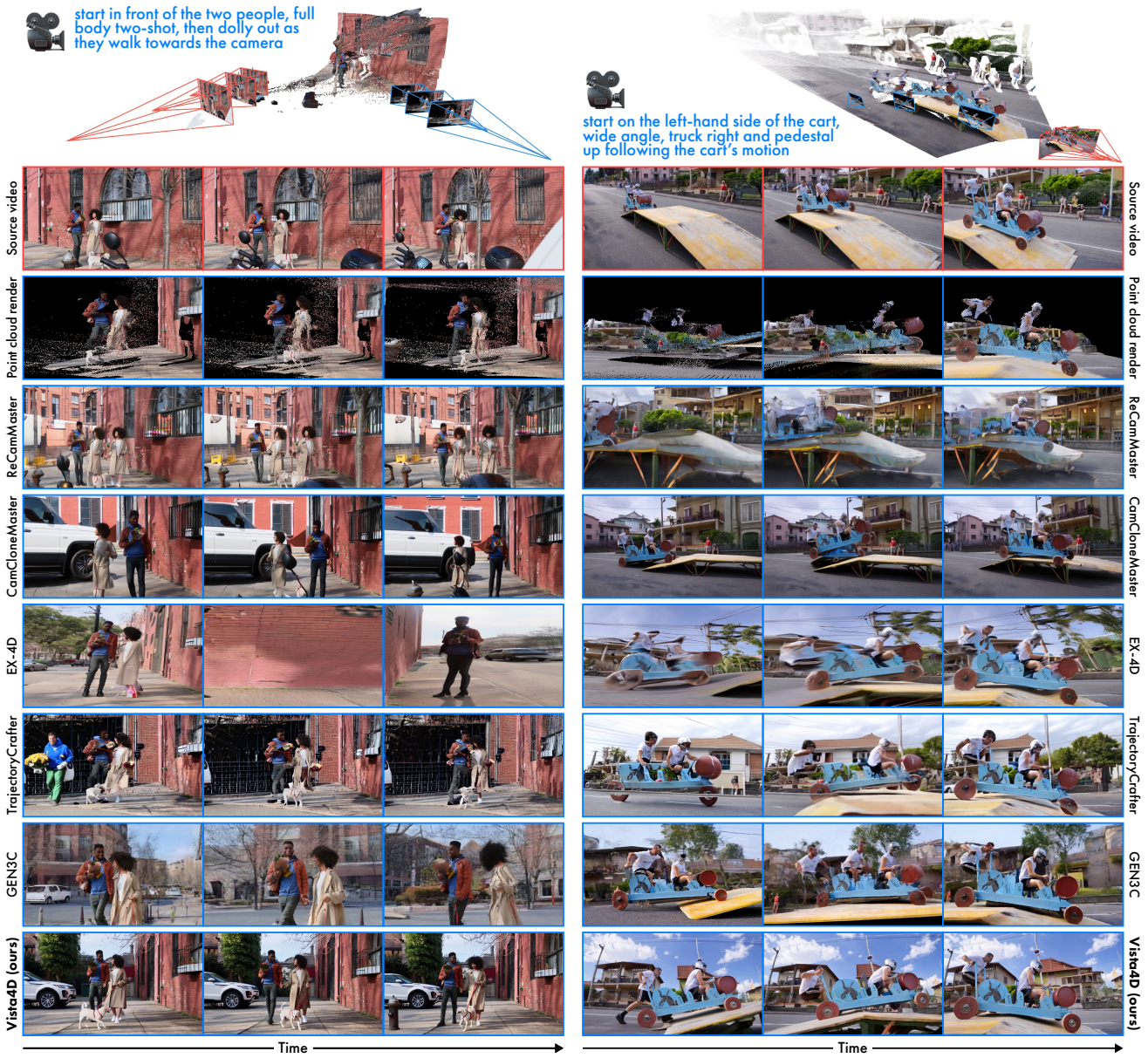

Figure 5. **Qualitative comparison on real-life monocular videos.** We show two video reshooting examples of Vista4D compared to our baselines, TrajectoryCrafter [7], GEN3C [8], EX-4D [9], ReCamMaster [10], and CamCloneMaster [22].

of generated videos from each method [10, 56]. For 3D consistency between the source and output videos, following Pippo [59], we adopt the per-frame reprojection error of SuperPoint [58] landmarks under SuperGlue (RE@SG) [57]. Our method consistently exhibits more accurate camera control compared to baselines, especially against implicit-prior methods. Moreover, our method significantly outperforms baselines in 3D consistency, showcasing its output geometric plausibility despite noisy real-world 4D reconstruction.

Novel-view video synthesis. We compare novel-view video synthesis quality of Vista4D to baselines in Table 2 on the real-world time-synchronized multiview dataset, `iphone` [60]. We measure masked (indicated by the prefix "m") and full PSNR, SSIM, and LPIPS for synthesis quality [60], along with optical flow endpoint error (EPE) for motion quality [70]. Our method outperforms baselines in PSNR and LPIPS, indicating our superior spatial reconstruction quality. We also significantly outperform baselines for EPE, indicating our method's ability to preserve source video motion. Note that even though we are behind TrajectoryCrafter [7] for SSIM, viewing the synthesized videos quickly reveal significant artifacts in the latter's outputs not caught by SSIM, examples of which we show in Figure 4.

Video fidelity. We evaluate video fidelity and quality of

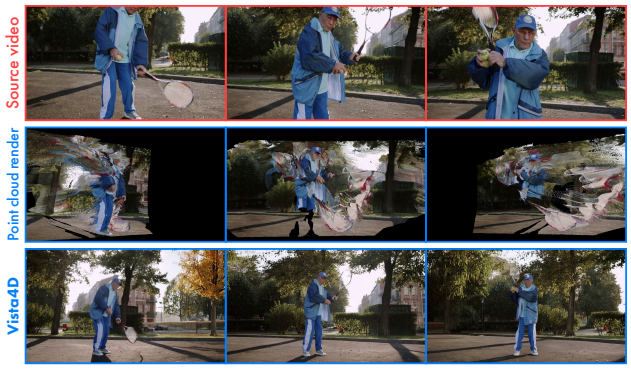

Figure 6. **Robustness to segmentation failure.** We simulate segmentation failure by not segmenting the tennis racket as dynamic. Vista4D is generally robust to these point cloud streaks as it utilizes the in-context-conditioned source video to correct the artifacts.

Figure 7. **Dynamic scene expansion.** With our 4D-grounded temporally-persistent point cloud, Vista4D can do video reshooting with additional scene information from casual scene captures or alternate angles by doing joint 4D reconstruction of these frames with the source video. Doing so reduces video model hallucinations and provides stronger control beyond the source video.

Vista4D to baselines in Table 3 on our 110 video-camera pair dataset. We use FID [71], FVD [72], VBench (aesthetic quality, imaging quality, subject consistency, background consistency, and temporal style) [61], and VBench-2.0 (human anatomy) [62] to evaluate video fidelity, and CLIP-T [73] for prompt alignment. Our method consistently outperforms point-cloud-conditioned (explicit-prior) baselines, especially in aesthetic quality, imaging quality, and human anatomy due

Figure 8. **4D scene recomposition.** By directly editing the 4D point cloud, Vista4D can recompose 4D scenes from the source video or other inserted videos. Importantly, our method synthesizes physically plausible lighting when inserting a rhino lit by sunlight through leaves into an otherwise overcast scene.

to our robustness to point cloud artifacts. Implicit-prior methods (ReCamMaster and CamCloneMaster) perform better in FID, FVD, subject consistency, and background consistency because they often fail to follow the target cameras, resulting in relatively little camera change from the source video and thus seemingly better metrics. Qualitative comparisons are in Figure 5 and Supplementary A, along with the user study in Table 4, show our method's clear high video fidelity.

4.2. Qualitative comparisons

We qualitatively compare Vista4D to baselines in Figure 5 on two example real-life monocular videos, where we show the point cloud render to illustrate the intended cameras in addition to the written description. Explicit-prior methods (EX-4D, TrajectoryCrafter, and GEN3C) all struggle with point cloud artifacts from target cameras at non-frontal views of the depth maps, resulting in subject and background artifacts (*e.g.*, TrajectoryCrafter, left video; all three methods, right video) or camera control failure (*e.g.*, EX-4D and GEN3C, left video). Implicit-prior methods (ReCamMaster and CamCloneMaster) similarly struggle with precise camera control (ReCamMaster, left video; both methods, right video) and subject artifacts (both methods, left video).

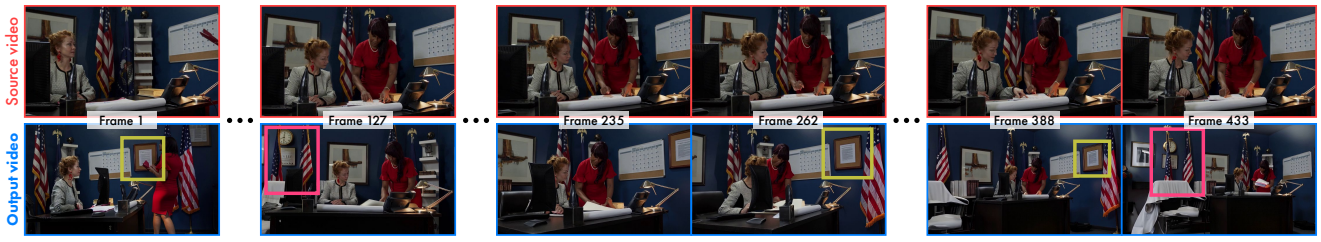

Figure 9. **Long video inference with memory.** Vista4D can reshoot long videos by doing inference in chunks. By registering static pixels of newly generated chunks back into the temporally-persistent 4D point cloud, Vista4D maintains an explicit, 4D-grounded memory of generated content. We showcase this above as the camera arcs around the scene, indicated with color-matched yellow and pink boxes.

In contrast, our method produces high-quality outputs that not only faithfully preserves source video content but also follows the target cameras. We include more comprehensive qualitative results and comparisons in Supplementary A.

User study. We show the results of our user study in Table 4, where we ask participants to compare our method to baselines on three dimensions: Source video content preservation, camera control accuracy, and overall video fidelity. We randomly select a subset of 30 video-camera pairs from the 110-pair evaluation dataset and invite 42 participants to select their preferred method for each pair and each dimension. Users consistently prefer our method by a wide margin over baselines on all dimensions, especially overall video fidelity due to our method's robustness to point cloud artifacts and challenging camera trajectories and viewpoints. We include details of our user study in Supplementary D.

Robustness to segmentation failure. Since Vista4D uses Grounded SAM 2 to segment dynamic pixels, segmentation failures can result in point cloud streaking artifacts. However, Vista4D is generally robust to them. For example, we simulate segmentation failure in Figure 6 by deliberately not segmenting the tennis racket, and Vista4D corrects the streaking just like it corrects imperfect point cloud geometry, that is, by utilizing the in-context-conditioned source video. Broadly, we observe that streaking artifacts during inference are rare or inconsequential, especially compared to the improved camera control and 4D consistency from static pixel temporal persistence.

4.3. Ablation study

We study the effects of our data and model conditioning on source video content preservation and robustness to imperfect 4D reconstruction. We perform ablations combinations of the following design choices: No depth artifacts (by always doing double reprojection), no source video, cross-attention source video injection, and no temporal persistence. We find that the combination of training with depth artifacts and the (in-context conditioned) source video enables our model's ability to be robust to 4D reconstruction artifacts, particularly both spatial artifacts (imprecise depths from non-frontal views) and temporal artifacts (jittering depths). We also find that removing temporal persistence reduces our model's ability to both preserve static content and maintain accurate camera control when the source video and target cameras have low per-frame overlap. We show examples of both findings in Supplementary F.

4.4. Applications

Dynamic scene expansion. Video reshooting requires the video diffusion models to hallucinate pixels not existent in the source video, even though we often have more visual information of the environment. For example, we may have casual captures of a scene or alternate camera angles on a film set. Vista4D's explicit context grounding with temporally-persistent 4D point clouds enables us to incorporate this information by doing joint 4D reconstruction of the source video and additional scene frames. Figure 7 shows an example of dynamic scene expansion, where the addition of temporally-persistent casual scene captures enables more faithful environment reproduction.

4D scene recomposition. As Vista4D is trained to be robust to point cloud artifacts, we can directly edit and recompose the 4D point cloud to manipulate, duplicate, delete, and even insert new subjects while maintaining their dynamics. Figure 8 shows examples of 4D scene recomposition. Notably, the Figure includes an example where we insert the point cloud of a rhino illuminated by sunlight through leaves into an overcast funeral procession scene. Our method naturally blends these differing lighting conditions, generating a region of dappled light around the rhino while keeping the procession in soft shadows under the trees.

Long-video inference with memory. For video reshooting on long videos beyond the video diffusion model's trained context window, our temporally-persistent 4D point cloud acts as an explicit, compressed context to retain generated static content across camera viewpoint changes. To do so, we autoregressively generate chunks of the video in target cameras that fit within our model's context window, where we train a variant of our model based on the first-frame-conditioned `Wan2.1-I2V-14B` to ensure visual consistency between chunks. Figure 9 shows an example of long-video inference that retains memory of generated content.

Acknowledgements. We would like to thank Aleksander Hołyński, Wenqi Xian, Dan Zheng, Mohsen Mousavi, Li Ma, and Lingxiao Li for their technical discussions; Ryan Tabrizi, Tianyi Lorena Yan, and Shreyas Havaldar for appearing in our demo videos; Lukas Lepicovsky, David Rhodes, Nhat Phong Tran, Dacklin Young, and Johnson Thomasson for their production support; Jeffrey Shapiro, Ritwik Kumar, and Hossein Taghavi for their executive support; Jennifer Lao and Lianette Alnaber for their operational support.

References

[1] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2

[2] Team Wan. Wan: Open and advanced large-scale video generative models, 2025. 4, 1, 3, 5, 8

[3] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners, 2025.

[4] Weijie Kong et al. Hunyuanvideo: A systematic framework for large video generative models, 2025.

[5] NVIDIA. World simulation with video foundation models for physical ai, 2025.

[6] Genmo Team. Mochi 1, 2024. 2

[7] Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. TrajectoryCrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *ICCV*, 2025. 2, 3, 4, 5, 6, 1, 10, 11

[8] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3d-informed world-consistent video generation with precise camera control. In *CVPR*, 2025. 2, 4, 5, 6, 1, 3

[9] Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. EX-4D: Extreme viewpoint 4d video synthesis via depth watertight mesh, 2025. 2, 4, 5, 6, 1, 3

[10] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. ReCamMaster: Camera-controlled generative rendering from a single video. In *ICCV*, 2025. 3, 4, 5, 6, 1, 2, 9

[11] Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. Stream3r: Scalable sequential 3d reconstruction with causal transformer, 2025. 3, 5, 1, 2

[12] Hyeonho Jeong, Suhyeon Lee, and Jong Chul Ye. Reangle-a-video: 4d video generation as video-to-video translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11164–11175, October 2025. 2

[13] Meng YOU, Zhiyu Zhu, Hui LIU, and Junhui Hou. NVS-solver: Video diffusion model as zero-shot novel view synthesizer. In *ICLR*, 2025.

[14] Zezhong Qian, Xiaowei Chi, Yuming Li, Shizun Wang, Zhiyuan Qin, Xiaozhu Ju, Sirui Han, and Shanghang Zhang. Wristworld: Generating wrist-views via 4d world models for robotic manipulation, 2025. 2

[15] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *CVPR*, pages 22831–22840, June 2025. 2, 3

[16] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *CVPR*, pages 2005–2015, June 2025.

[17] Tian-Xing Xu, Xiangjun Gao, Wenbo Hu, Xiaoyu Li, Song-Hai Zhang, and Ying Shan. Geometrycrafter: Consistent geometry estimation for open-world videos with diffusion priors. In *ICCV*, pages 6632–6644, October 2025. 2, 3

[18] Yash Kant, Aliaksandr Siarohin, Michael Vasilkovsky, Riza Alp Guler, Jian Ren, Sergey Tulyakov, and Igor Gilitschenski. invs : Repurposing diffusion inpainters for novel view synthesis. In *SIGGRAPH Asia*, 2023. 2

[19] Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Multidiff: Consistent novel view synthesis from a single image. In *CVPR*, pages 10258–10268, June 2024. 2

[20] Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. Trajectory attention for fine-grained video motion control. In *ICLR*, 2025. 2

[21] Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. *ECCV*, 2024. 3

[22] Yawen Luo, Jianhong Bai, Xiaoyu Shi, Menghan Xia, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Tianfan Xue. CamCloneMaster: Enabling reference-based camera control for video generation. In *SIGGRAPH Asia*, 2025. 3, 4, 5, 6, 1, 2

[23] Jensen (Jinghao) Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025. 3

[24] Yash Kant, Aliaksandr Siarohin, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, and Igor Gilitschenski. Spad: Spatially aware multi-view diffusers. In *CVPR*, 2024.

[25] Chen Liu and et al. Zero-1-to-3: Zero-shot novel view synthesis from a single image. *arXiv:2303.11328*, 2023.

[26] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3D generation. In *Proc. ICLR*, 2024.

[27] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024. 3

[28] Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and

Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models, 2025. 3

[29] Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *Proc. ICLR*, 2025.

[30] Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470*, 2024. 3

[31] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 3

[32] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *CVPR*, pages 21686–21697, 2024.

[33] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision (ECCV)*, 2024. 3, 9

[34] Gene Chou, Wenqi Xian, Guandao Yang, Mohamed Abdelfattah, Bharath Hariharan, Noah Snavely, Ning Yu, and Paul Debevec. Flashdepth: Real-time streaming video depth estimation at 2k resolution. *arXiv preprint arXiv:2504.07093*, 2025. 3

[35] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025.

[36] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025. 3

[37] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *CVPR*, pages 10486–10496, June 2025. 3

[38] David Yifan Yao, Albert J. Zhai, and Shenlong Wang. Uni4d: Unifying visual foundation models for 4d modeling from a single video. In *CVPR*, pages 1116–1126, June 2025. 5, 3

[39] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. Vipe: Video pose engine for 3d geometric perception, 2025. 3, 9

[40] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 3

[41] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 3

[42] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025.

[43] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 3

[44] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning, 2025. 3, 5, 1, 9

[45] Q. Sun et al. Monst3r: A simple approach for estimating geometry in dynamic scenes. *arXiv:2410.03825*, 2024.

[46] D. Zhuo et al. Streaming 4d visual geometry transformer. *arXiv:2507.11539*, 2025.

[47] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4d: Leveraging video generators for geometric 4d scene reconstruction. *arXiv preprint arXiv:2504.07961*, 2025. 3

[48] Jiahui Lei, Yijia Weng, Adam W. Harley, Leonidas Guibas, and Kostas Daniilidis. MoSca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In *CVPR*, pages 6165–6177, June 2025. 3

[49] Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. In *ICCV*, pages 9660–9672, October 2025. 10

[50] Shizun Wang, Xingyi Yang, Qiuhong Shen, Zhenxiang Jiang, and Xinchao Wang. Gflow: Recovering 4d world from monocular video. In *AAAI*. AAAI Press, 2025. ISBN 978-1-57735-897-8. 3

[51] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos, 2024. 3, 5

[52] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, Yuda Xiong, Hao Zhang, Feng Li, Peijun Tang, Kent Yu, and Lei Zhang. Grounding DINO 1.5: Advance the "edge" of open-set object detection, 2024.

[53] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded SAM: Assembling open-world models for diverse visual tasks, 2024. 3, 5

[54] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Adv. Neural Inform. Process. Syst.*, volume 37, pages 16240–16271. Curran Associates, Inc., 2024. 4, 1

[55] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *ICLR*, 2025.

[56] Yuancheng Xu, Wenqi Xian, Li Ma, Julien Philip, Ahmet Levent Taşel, Yiwei Zhao, Ryan Burgert, Mingming He, Oliver

Hermann, Oliver Pilarski, Rahul Garg, Paul Debevec, and Ning Yu. Virtually being : Customizing camera-controllable video diffusion models with multi-view performance captures. In *SIGGRAPH Asia*, 2025. 4, 6, 1

[57] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, June 2020. 4, 6

[58] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR Workshops*, June 2018. 6, 10

[59] Yash Kant, Ethan Weber, Jin Kyu Kim, Rawal Khirodkar, Su Zhaoen, Julieta Martinez, Igor Gilitschenski, Shunsuke Saito, and Timur Bagautdinov. Pippo: High-resolution multi-view humans from a single image. In *CVPR*, 2025. 4, 6, 10

[60] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Adv. Neural Inform. Process. Syst.*, volume 35, pages 33768–33780. Curran Associates, Inc., 2022. 4, 5, 6, 10

[61] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, pages 21807–21818, June 2024. 5, 7, 10

[62] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness, 2025. 5, 7, 10

[63] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023. 5

[64] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, October 2023. 5

[65] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. OpenVid-1M: A large-scale high-quality dataset for text-to-video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. 5, 1

[66] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize Anything: A strong image tagging model. *arXiv preprint arXiv:2306.03514*, 2023. 5, 3

[67] Llama Team. The Llama 3 herd of models, 2024. 5, 3

[68] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, June 2016. 5

[69] Pexels. Pexels: Free stock photos & videos, 2025. 5

[70] Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael Ryoo, Paul Debevec, and Ning Yu. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *CVPR*, pages 13–23, June 2025. 6

[71] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 7, 10

[72] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2019. 7, 10

[73] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Int. Conf. Machine Learn.*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. 7, 10

[74] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 1

[75] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4):376–380, 2002. 1, 9

[76] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, Lei Zhao, Zhuoyi Yang, Xiaotao Gu, Xiaohan Zhang, Guanyu Feng, Da Yin, Zihan Wang, Ji Qi, Xixuan Song, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Yuxiao Dong, and Jie Tang. CogVLM2: Visual language models for image and video understanding, 2024. 3

[77] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. 3, 5

[78] Brandon Castellano. PySceneDetect, 2025. 3

[79] Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. 5

[80] Brent Yi, Chung Min Kim, Justin Kerr, Gina Wu, Rebecca Feng, Anthony Zhang, Jonas Kulhanek, Hongsuk Choi, Yi Ma, Matthew Tancik, and Angjoo Kanazawa. Viser: Imperative, web-based 3d visualization in python, 2025. 6, 8

[81] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 9

[82] Philipp Lindenberger, Paul-Edouard Sarlin, Marc Pollefeys, and Mihai Dusmanu. Lightglue: Local feature matching at light speed. In *ICCV*, 2023. 10

[83] 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. *Communications of the ACM*, 65(1):99–106, 2021. 10

[84] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 10

[85] Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *European Conference on Computer Vision*, pages 36–54. Springer, 2024. 10