REAL2SAM2REAL: GENERATIVE 3D CACHES AS COMPLEMENTARY CONTEXT FOR VIDEO DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video Diffusion Models (VDMs) synthesize high-quality videos, but precise camera and motion control remains challenging when large viewpoint shifts, changing occlusions, or complex dynamics reveal unseen regions. Diffusion priors alone can yield structural inconsistencies, while warped guidance can contain incomplete geometry and misleading view-dependent appearance cues. Real2SAM2Real complements VDM priors with a coarse 3D foreground cache of topologically complete meshes for intuitive editing and animation. Generative 3D lifting estimates these object and human meshes, including unobserved surfaces, from a single reference image. Normal maps rendered under specified camera trajectories and foreground motions guide video synthesis in the VDM, while the reference image provides appearance conditioning. Soft spatial-aligned injection and minimally invasive fine-tuning incorporate this guidance while largely preserving pretrained synthesis priors. Decoupled conditioning enables independent appearance and geometry editing. We train on around 2,000 monocular clips using instance-masked normal sequences with instance-level spatial perturbations to reduce sensitivity to differences between video-estimated training normals and normals rendered from coarse 3D caches at inference. Training requires neither explicit 3D assets nor paired multi-trajectory recordings. Real2SAM2Real supports decoupled camera and multi-instance motion control, including articulated humans. Across these challenging settings, experiments show strong spatiotemporal consistency and better overall control accuracy and video fidelity than evaluated state-of-the-art methods. Preliminary demonstrations suggest potential robotics applications.

1 INTRODUCTION

Video Diffusion Models (VDMs) (Wan et al., 2025; Yang et al., 2024b; HaCohen et al., 2024; Kong et al., 2024; Blattmann et al., 2023) synthesize high-fidelity videos using spatiotemporal priors learned from large-scale datasets. Recent work adapts these models through post-training (Zhang et al., 2023; Hu et al., 2022; Mou et al., 2024) for fine-grained camera and entity motion control. However, their training observations are 2D videos. While an ideal world model should simulate complete 4D spacetime, videos intrinsically capture dynamics solely on a projected 2D plane. This dimensional loss can limit geometric fidelity, spatiotemporal consistency, and explicit 3D-aware control in complex dynamic scenes.

Existing methods largely follow two paradigms. The first fine-tunes VDMs with parameterized camera embeddings (He et al., 2024; Wang et al., 2024c; Bai et al., 2025; Van Hoorick et al., 2024; Luo et al., 2025; Bai et al., 2024). This can require additional adaptation and risk degrading pretrained priors, while scale ambiguity may remain. The second uses pixel-wise 3D reconstructions, such as point clouds, in a "warp-and-inpaint" pipeline (Yu et al., 2024; 2025; Ren et al., 2025; Song et al., 2025; You et al., 2024; Liu et al., 2024). Although this reduces the modality gap, incomplete geometry and misleading cues in reflective or refractive regions can produce perspective ambiguities and artifacts. When motion exposes unobserved regions, insufficient geometric guidance can make structural consistency difficult to maintain.

We introduce **Real2SAM2Real**, which complements VDM priors with an editable, instance-complete foreground geometry cache (Fig. 1). From a single reference image, generative 3D lifting (Chen et al., 2025; Xiang et al., 2025; Hunyuan3D et al., 2025; Wu et al., 2025; Ye et al., 2025) builds

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

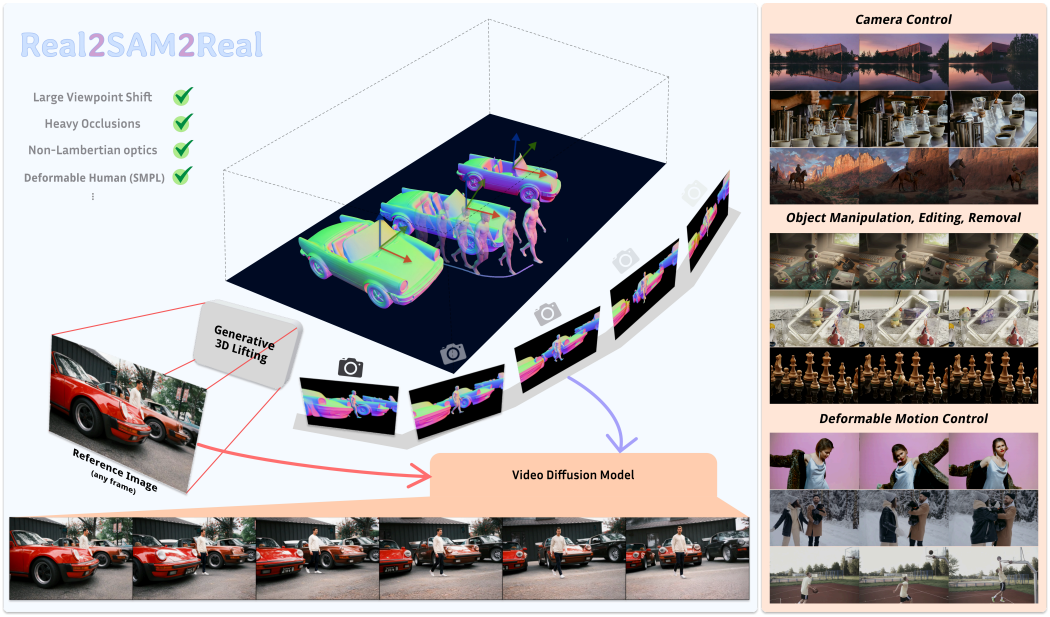

Figure 1: **Versatile 3D-Aware Video Generation via Real2SAM2Real. (Left)** From a single reference image at any frame, generative 3D lifting constructs an editable, *instance-complete* cache of objects and SMPL-based humans. The VDM uses coarse foreground geometry rendered according to the specified camera trajectory and foreground motion as guidance. **(Right)** Applications include camera control, entity manipulation and removal, and deformable motion control, with examples involving occlusions, viewpoint ambiguity, human motion, reflections, refractions, and flames.

a coarse cache that provides geometric priors and an intuitive interface for camera control and editing scene and character layouts. Background synthesis is left to the VDM. Coarse foreground geometry provides scene-anchored visual guidance, while minimally invasive fine-tuning aims to retain pretrained capabilities. Experiments show geometrically coherent generation with precise control in evaluated scenes involving large motions, heavy occlusions, and non-Lambertian effects (Fig. 2).

The core contributions of this work can be summarized as follows:

1. **We introduce a topologically complete 3D geometry cache for intuitive editing and animation.** These spatial anchors support camera and entity control while helping mitigate artifacts from incomplete geometry (e.g., holes and erroneous facades) and misleading cues in reflective or refractive regions.
2. **We use instance-masked normal sequences as a cross-modal bridge between the 3D cache and the VDM.** Normal maps provide dense, scale-invariant geometry without texture details, complementing the appearance condition. We derive RGB-normal training pairs directly from monocular videos, without paired multi-trajectory recordings or 3D assets, and apply instance-level random perturbations to the normal maps to help reduce the training-inference gap.
3. **We design a Soft Spatial-Aligned Injection mechanism and a minimally invasive fine-tuning strategy to decouple appearance and geometry, enabling independent editing of both.** This design adapts the VDM by training approximately 1.8% of its parameters on around 2,000 video clips, while spatial perturbations help accommodate imperfect geometric guidance.

2 Background & Related Works

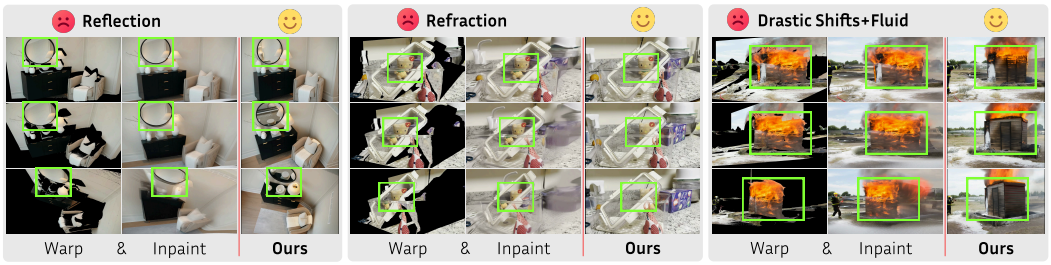

Figure 2: **Addressing the limitations of warp-and-inpaint pipelines.** Point-cloud warping can entangle geometry with view-dependent appearance: (a) reflections and (b) refractions introduce misleading spatial cues. (c) Incomplete geometry, background depth errors, and the geometric encoding of unstructured dynamics (e.g., fire) can produce stretched textures and erroneous facades under large viewpoint shifts. Our appearance-decoupled, *instance-complete* cache reduces these artifacts while maintaining spatiotemporal coherence.

2.1 Camera Control

Current methodologies for camera control can be broadly classified into two dominant paradigms. The first paradigm attempts to adapt foundational video models to parameterized camera embeddings. Methods (He et al., 2024; Wang et al., 2024c; Bai et al., 2025; Van Hoorick et al., 2024; Luo et al., 2025; Bai et al., 2024) directly inject camera trajectories (e.g., $[\mathbf{R}|\mathbf{t}]$ extrinsic matrices or Plücker rays) into the generative backbone through various attention mechanisms, feature concatenation, or similar conditioning modules. While integrating camera-centric parameterized control signals enables explicit trajectory controllability, this approach is limited by scale ambiguity and a cross-modal gap. Aligning these signals with visual features requires additional adaptation.

Instead of numerical poses, the second paradigm relies on reprojecting explicit pixel-wise 3D reconstructions as dense spatial guidance. Methods (Yu et al., 2024; 2025; Ren et al., 2025; Song et al., 2025; Hu et al., 2025; Ma et al., 2025; You et al., 2024; Xiao et al., 2024; Cao et al., 2025; Liu et al., 2024; Zhang et al., 2025) lift monocular depth into 3D point clouds or tracks (Wang et al., 2025a; Piccinelli et al., 2024; 2025; Wang et al., 2025b; Bochkovskii et al., 2024; Lin et al., 2025; Huang et al., 2025), warp them to novel views, and condition the VDM on these proxy videos. While this "warp-and-inpaint" strategy mitigates the cross-modal gap, the imperfections of monocular reconstructions expose the model to geometric errors. Specifically, the unclosed nature of these visible shells can introduce perspective ambiguities (via structural holes and erroneous facades), while non-Lambertian regions (e.g., reflections and refractions) yield misleading geometric cues. Together, these flawed spatial proxies can trigger cascading artifacts and structural degradation. Both paradigms provide limited explicit 3D context for unobserved regions. When large camera movements unveil previously hidden surfaces, this limited geometric guidance leaves the model to infer missing geometry from its generative priors, which can lead to structural degradation.

2.2 Motion Control

While camera control defines the global observation window, motion control guides the motion and interactions of objects and articulated subjects. This domain exhibits significant diversity, predominantly relying on a wide spectrum of abstract object representations or pixel-level cues as conditioning signals. For instance, 3DTrajMaster (Fu et al., 2024) explicitly conditions the generation on the numerical embeddings of 6-DoF object pose trajectories. For articulated motion, Motion-2-to-3 (Pi et al., 2024) uses 2D motion data for 3D human motion generation. Motion Prompting (Geng et al., 2025) uses point trajectories, while Diffusion as Shader (Gu et al., 2025) conditions generation on 3D tracking videos. Uni3C (Cao et al., 2025) combines SMPL-X human meshes with monocular scene point clouds. For non-human foreground objects, these conditions provide limited geometric coverage, leaving the VDM to infer unobserved surfaces from its spatiotemporal priors. This reliance can lead to structural degradation under large motions and complex occlusions.

To address these limitations, our **Real2SAM2Real** framework introduces an explicitly editable, *instance-complete* 3D proxy for objects and SMPL-based humans via instance-masked normal maps. This appearance-decoupled geometric guidance provides priors for unobserved regions, mitigating the "inpainting trap" and cross-modal alignment issues of prior methods. Furthermore, by integrating this 3D context in a minimally invasive and fault-tolerant manner, our approach largely preserves foundational VDM priors while enabling 3D control and easy editing of diverse foreground instances.

3 METHODOLOGY

3.1 ENTITIES AS ANCHORS: INSTANCE-COMPLETE 3D GEOMETRY CACHE

To synthesize a spatiotemporally consistent video $V \in \mathbb{R}^{T\times H\times W\times 3}$ of length T, we take as input a single reference image $I_{ref} \in \mathbb{R}^{H\times W\times 3}$, a user-specified camera trajectory $\mathbf{C} = \{c_t\}_{t=1}^{T}$, and a set of instance-specific spatial manipulation sequences $\mathbf{M} = \{\mathbf{M}_k\}_{k=1}^{K}$. We decouple spatial layout from appearance through an editable 3D staging environment. Full-scene reconstruction from a single image is ambiguous, particularly for distant backgrounds. We therefore model foreground instances explicitly and use the VDM's pretrained priors for background synthesis. Let Φ_{3D} denote a layout-aware, multi-instance 3D generation framework (e.g., SAM3D (Chen et al., 2025)). The initial static 3D representation is formulated as the union of decoupled 3D foreground entities:

$$\mathcal{S}_{\text{3D}}^{(0)} = \Phi_{\text{3D}}(I_{ref}) = \bigcup_{k=1}^{K} O_k \tag{1}$$

where O_k denotes the 3D mesh of the k-th instance, including objects and SMPL-based humans. To enable intuitive motion control, we apply the user-specified temporal editing parameters $m_{k,t} \in \mathbf{M}_k$ (e.g., 6-DoF rigid transformations or non-rigid deformations) to each decoupled instance. The dynamic scene configuration at any temporal frame t is formulated as:

$$\mathcal{S}_{\text{3D}}^{(t)} = \bigcup_{k=1}^{K} \mathcal{T}(O_k, m_{k,t}) \tag{2}$$

where $\mathcal{T}(\cdot)$ denotes geometric manipulation. Here, topological completeness refers to full instance surfaces, including estimated unobserved regions. These independently animatable meshes determine depth ordering and occlusions during rendering, reducing structural holes and cardboard effects caused by incomplete proxies.

3.2 NORMAL AS A BRIDGE: 3D-FREE DATA CURATION & RANDOM PERTURBATION

To translate these coupled explicit 3D priors (both camera and entity dynamics) into a visual domain easily digestible by the VDM, we employ a graphics renderer $\mathcal{R}$. The dynamic 3D cache $\mathcal{S}_{\text{3D}}^{(t)}$ is projected along the target camera pose c_t to yield a sequence of instance-masked normal maps $\mathbf{N} \in \mathbb{R}^{T\times H\times W\times 3}$:

$$\mathbf{N} = \{n_t\}_{t=1}^{T}, \quad \text{where } n_t = \mathcal{R}(\mathcal{S}_{\text{3D}}^{(t)}, c_t) \tag{3}$$

Normal maps encode local surface orientations independently of scene scale, and their three-channel format is compatible with the pretrained VAE. We construct RGB-normal pairs directly from monocular videos, without paired multi-trajectory recordings, explicit 3D assets, or reprojected conditioning videos. NormalCrafter (Bin et al., 2025) estimates pseudo-normal sequences; SAM 3 (Carion et al., 2025) and vision-language models identify foreground instances, whose masks are applied to the estimated normals to obtain $\mathbf{N}$. To reduce the training-inference gap caused by noisy geometry and coarse boundaries in rendered 3D caches, we randomly perturb each instance's normal sequence with translation, rotation, non-uniform scaling, and boundary displacement. Each instance shares one perturbation parameter set across all frames, while RGB targets remain unchanged. This reduces sensitivity to imperfect geometric guidance.

3.3 VDM AS A NEURAL COMPOSITOR: SOFT SPATIAL-ALIGNED INJECTION AND MINIMALLY INVASIVE FINE-TUNING

Our generative backbone is Wan2.2-Fun-A14B-InP, with high- and low-noise DiT experts. Let $z_0 = \mathcal{E}(V)$ denote the video latent and $z_s = (1-\sigma_s)z_0 + \sigma_s\epsilon$, where σ_s is the noise level at timestep

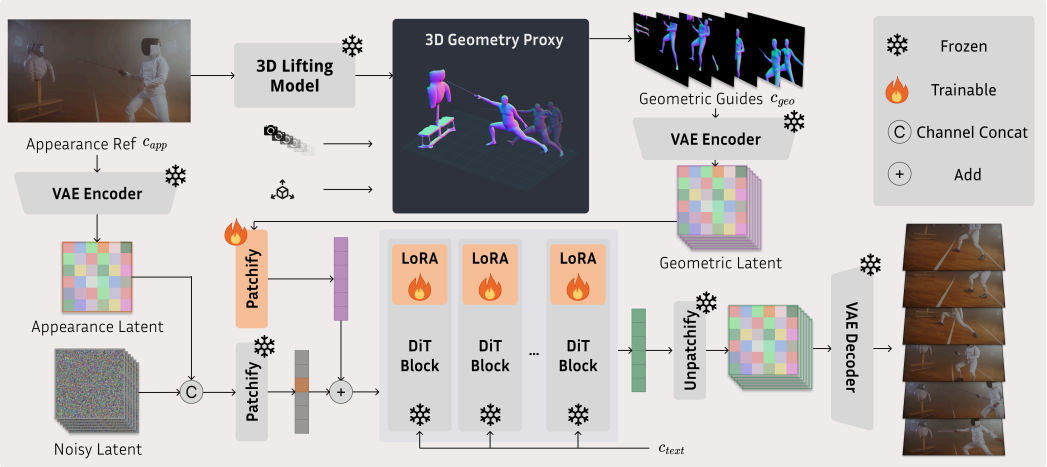

Figure 3: **Pipeline Architecture and Decoupled Injection Mechanism.** The reference frame enters through the backbone's native inpainting condition. A frozen VAE encodes instance-masked normal sequences, and a zero-initialized 3D convolution maps these latents to additive video features. A frozen T5 encoder supplies optional text conditioning. For each noise expert, only the geometry embedding layer and LoRA modules are trained.

s and $\epsilon \sim \mathcal{N}(0, I)$. We train a conditional velocity predictor v_θ using flow matching:

$$\mathcal{L} = \mathbb{E}_{z_0,\epsilon,s}\left[\rho(v_\theta(z_s, s, \mathbf{c}_{app}, \mathbf{c}_{geo}, \mathbf{c}_{text}) - (\epsilon - z_0))\right], \tag{4}$$

where ρ averages squared residuals with outlier masking. The appearance condition $\mathbf{c}_{app}$ combines a binary mask with VAE latents of a video containing the reference image at its designated temporal position. It enters the backbone's native inpainting pathway through channel concatenation with z_s. During training, the reference frame is sampled from the target video. Optional text guidance $\mathbf{c}_{text}$ is encoded by a frozen T5 encoder.

For geometric guidance, the frozen VAE encodes $\mathbf{N}$, and a zero-initialized, trainable 3D convolution with a $1 \times 2 \times 2$ kernel maps these latents to geometric features $\mathbf{c}_{geo}$. These features are added to the frozen input patch embedding's output across all video frames. This *soft spatial-aligned injection* uses a separate additive branch for geometry while retaining the pretrained inpainting pathway for appearance. The two conditions can be edited independently; motion transfer is performed by editing the reference image while keeping the normal-map sequence fixed.

We use *minimally invasive fine-tuning*, training only the geometry patch embedding and LoRA modules in attention and feed-forward projections, totaling approximately 1.8% of the base parameters per expert. The two noise experts are fine-tuned separately on their corresponding noise ranges, while the pretrained DiT weights, VAE, and T5 encoder remain frozen. Combined with instance-level normal perturbations, these limited updates reduce overfitting to imperfect geometry while largely preserving the backbone's appearance and background synthesis priors. The native text-conditioning pathway supports optional editing of visual effects and styles alongside 3D control.

4 Experimental Results

4.1 Datasets and Experimental Setup

We implement our framework based on Wan2.2-Fun-A14B-InP and fine-tune it using around 2,000 monocular video clips, from which our 3D-free pipeline automatically constructs training pairs of RGB videos and randomly perturbed instance-masked normal sequences. We fine-tune the high- and low-noise experts separately at 720p (81 frames), using LoRA and trainable geometry embeddings. Details of the training data curation and training setup are provided in Appendix B, and the test data, baseline conditioning, and evaluation metrics of each experiment in Appendix C.

4.2 Camera Control

We evaluate on 200 test cases constructed from 100 in-the-wild images (Unsplash), each paired with 2 randomly sampled GLB camera trajectories, with ground-truth poses directly recorded from the GLB environment. We compare against ReCamMaster (Bai et al., 2025), ViewCrafter (Yu et al., 2024), TrajectoryCrafter (Yu et al., 2025), WorldForge (Song et al., 2025), and GEN3C (Ren et al., 2025). Warp-based baselines are conditioned on an identical WorldForge-generated proxy for fairness, and GEN3C renders its own 3D cache along the same trajectory; ReCamMaster receives ground-truth poses converted to its required format. Camera fidelity is measured by ATE, RPE-T, and RPE-R via GLOMAP (Pan et al., 2024) and evo (Grupp, 2017), following (Song et al., 2025; Bai et al., 2025), with VBench (Huang et al., 2024) for video quality. As shown in Tab. 1 and Fig. 4, Baseline methods exhibit cascading artifacts under large viewpoint shifts, whereas Real2SAM2Real maintains faithful appearance and spatiotemporal coherence in the evaluated scenes.

Table 1: Quantitative comparison of camera control on 200 test cases. Camera accuracy is reported by ATE, RPE-T and RPE-R, computed with GLOMAP (Pan et al., 2024) and evo (Grupp, 2017) following WorldForge (Song et al., 2025) and ReCamMaster (Bai et al., 2025); video quality is reported by VBench (Huang et al., 2024). The best and second-best results are shaded in red and blue; ablations are not ranked.

Method	**VBench** (↑)				**Camera Accuracy** (↓)		
	Subject	Background	Aesthetic	Imaging	ATE	RPE-T	RPE-R
ReCamMaster (Bai et al., 2025)	90.29	91.39	55.54	68.98	0.0620	0.0182	**1.0649**
ViewCrafter (Yu et al., 2024)	89.95	91.34	62.03	74.20	0.0809	0.0712	3.6352
TrajectoryCrafter (Yu et al., 2025)	90.63	93.01	57.42	74.02	0.0443	0.0286	1.8461
WorldForge (Song et al., 2025)	89.92	92.73	56.87	70.87	0.1281	0.0605	2.2211
GEN3C (Ren et al., 2025)	**91.42**	93.32	57.85	73.13	**0.0210**	0.0143	1.1461
Ours (w/o Soft Injection)	86.26	89.80	59.60	71.67	0.0553	0.0268	1.1309
Ours (w/o Perturbation)	89.19	91.75	61.40	75.04	0.0280	0.0148	1.1553
Ours (w/o Both)	86.22	89.76	59.58	71.69	0.0645	0.0265	1.2557
Ours (Full)	91.33	**93.70**	**62.12**	**75.09**	0.0234	**0.0132**	1.0940

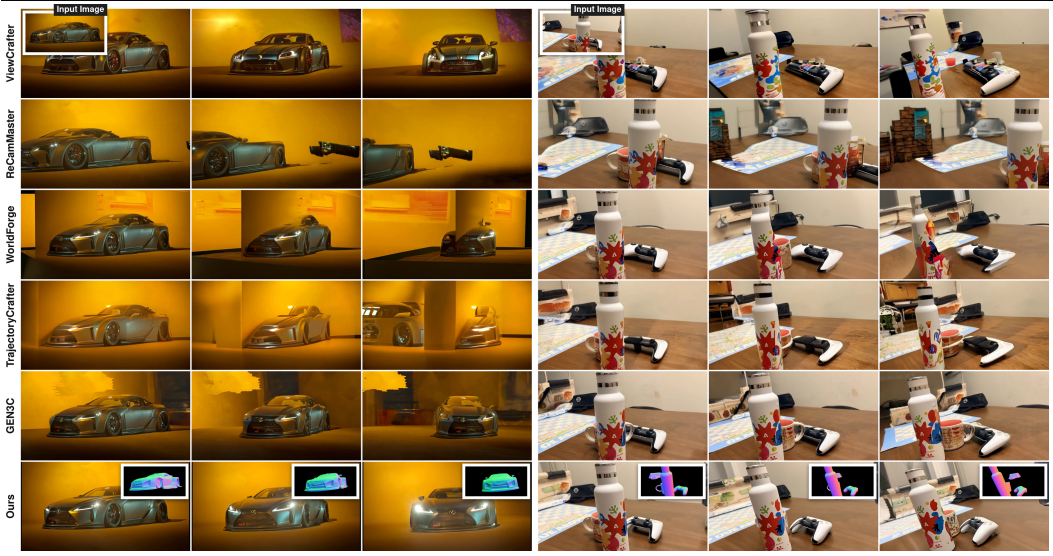

Figure 4: **Qualitative comparison of camera control.** Two scenes are shown side by side; within each, frames advance in time from left to right. Baseline methods exhibit cascading artifacts and structural breakdowns under large viewpoint shifts. Ours maintains faithful appearance and spatiotemporal coherence in the shown scenes. Videos are presented in the HTML supplementary material.

4.3 High-Dynamic Scenes

We evaluate dynamic scenes, in which non-rigid motion and camera motion occur simultaneously. With SAM 3D Body (Yang et al., 2026), we obtain a dynamic 3D cache by estimating a per-frame

3D human mesh from the source video, and render its instance-masked normal maps as our condition. The source video thereby serves as ground truth, and each method is tasked with reconstructing it from a single input frame. We evaluate on 300 real-world clips from HumanVid (Wang et al., 2024b) and OpenVid-HD (Nan et al., 2024), where each baseline additionally receives the strongest motion conditioning it supports, derived from the same clip. As shown in Tab. 2 and Fig. 5, the baselines exhibit structural collapse and appearance drift under large non-rigid motion, whereas Real2SAM2Real maintains spatiotemporal coherence. In practice, this setting doubles as a low-cost motion capture pipeline: a casually filmed performance directly yields a dynamic 3D cache, with no dedicated motion-capture equipment.

Table 2: Quantitative comparison of deformable human video reconstruction on 300 clips from HumanVid (Wang et al., 2024b) and OpenVid-HD (Nan et al., 2024). Video quality is reported by VBench (Huang et al., 2024), and human pose accuracy (PA-MPJPE (Kocabas et al., 2020), AKD-2D, MKR) by extending the 2D keypoint protocol of Monkey-Net (Siarohin et al., 2019) to 3D. Poses are estimated by SAM 3D Body (Yang et al., 2026). The best and second-best results are shaded in red and blue.

Method	**VBench** (↑)				**Human Pose Accuracy** (↓)		
	Subject	Background	Aesthetic	Imaging	PA-MPJPE (mm)	AKD-2D (px)	MKR (%)
ReCamMaster (Bai et al., 2025)	90.47	90.80	54.41	65.81	34.81	78.09	9.29
Uni3C (Cao et al., 2025)	89.07	90.21	54.12	64.05	59.45	108.58	17.85
DaS (Gu et al., 2025)	**92.58**	**92.07**	55.04	67.22	55.33	57.09	3.48
TrajectoryCrafter (Yu et al., 2025)	90.12	90.59	52.94	63.76	37.30	28.37	1.91
WorldForge (Song et al., 2025)	89.81	90.80	51.91	61.74	50.59	48.19	6.12
Ours	91.83	91.12	**56.81**	**69.63**	**22.80**	**17.80**	**0.58**



Figure 5: **Qualitative comparison under high-dynamic scene motion.** Real2SAM2Real synthesizes coherent video under large non-rigid motion coupled with camera motion. This is a challenging regime in which diffusion priors alone can break down. All methods reconstruct the real video (last row) from the same single input frame (inset, top left); to let every baseline operate at its full capability, each is additionally given the strongest motion conditioning it supports, i.e. camera poses, point-cloud renderings, tracking videos, or a dense warped video. The insets in the *Ours* row are the instance-masked normal maps that condition our model. The baselines over-rely on diffusion priors and exhibit structural collapse and appearance drift, whereas Real2SAM2Real maintains spatiotemporal coherence and stays closest to the real video. Videos are presented in the HTML supplementary material.

4.4 Entity Motion Control & Manipulation

We construct a test set of 100 cases, each specifying randomized object translation and rotation trajectories within a GLB scene, with ground-truth 6-DoF poses directly recorded from the environment. We compare against DaS (Gu et al., 2025) and ViewCrafter (Yu et al., 2024). While ViewCrafter is not designed for entity manipulation, it can perform the task when conditioned on a warped video proxy, reaching pose accuracy comparable to DaS. Object pose accuracy is measured by Rotation Error (RotErr) and Translation Error (TransErr) against ground-truth GLB trajectories, with VBench metrics for video quality. As shown in Tab. 3 and Fig. 6, baseline methods suffer from structural breakdowns and degradation under large entity movements and occlusions, whereas Real2SAM2Real maintains precise control and spatiotemporal coherence.

Table 3: Quantitative comparison of object manipulation. Video quality is reported by VBench (Huang et al., 2024), and object 6-DoF accuracy by rotation and translation errors (RotErr, TransErr) of object poses tracked with FoundationPose (Wen et al., 2024). The best and second-best results are shaded in red and blue; ablations are not ranked.

Method	**VBench** (↑)				**Object 6-DoF Accuracy** (↓)	
	Subject	Background	Aesthetic	Imaging	RotErr	TransErr
DaS (Gu et al., 2025)	96.37	96.66	60.23	70.99	22.7625	48.1813
ViewCrafter (Yu et al., 2024)	94.12	94.50	60.24	71.32	21.1082	41.9168
Ours (w/o Soft Injection)	91.23	93.16	59.38	69.45	7.4432	17.3950
Ours (w/o Perturbation)	93.93	94.89	61.48	73.39	11.9049	22.6497
Ours (w/o Both)	91.92	93.12	59.91	71.20	9.9061	21.0134
Ours (Full)	**96.39**	**97.08**	**61.21**	**73.42**	**4.1932**	**12.0991**



Figure 6: **Qualitative comparison of object manipulation.** Two scenes are shown side by side; within each, frames advance in time from left to right. Baseline methods exhibit cascading artifacts and structural breakdowns under large entity movements and occlusions. Videos are presented in the HTML supplementary material.

4.5 Ablation Study



Figure 7: Ablation on Soft Spatial-Aligned Injection and Instance-Level Spatial Perturbation. Videos are presented in the HTML supplementary material.

Soft Spatial-Aligned Injection To evaluate our geometry injection strategy, we replace the additive geometry branch with channel concatenation ("w/o Soft Injection"). As shown in Fig. 7, this variant exhibits appearance distortion and color bleeding near coarse or misaligned normal boundaries.

Compared with the full model, it yields lower scores on all four VBench metrics and higher pose errors in both camera and object control (Tables 1 and 3). Our separate geometry branch works alongside the native inpainting pathway, better preserving reference appearance while incorporating 3D guidance.

Instance-level Spatial Perturbation The "w/o Perturbation" variant removes the instance-level random perturbations applied to normal maps during training. In Fig. 7, this variant exhibits appearance distortion around misaligned normal boundaries, indicating greater sensitivity to the gap between training normals and normals rendered from lifted 3D caches. Removing perturbations lowers subject and background consistency and increases pose errors in both tasks (Tables 1 and 3). These results support using perturbed normals to improve robustness to imperfect geometric guidance.

The "w/o Both" variant also exhibits appearance distortion (Fig. 7). Compared with all three ablation variants, the full model achieves higher subject and background consistency and lower pose errors in both tasks, supporting the combined use of soft injection and normal perturbations.

4.6 Robotics Application

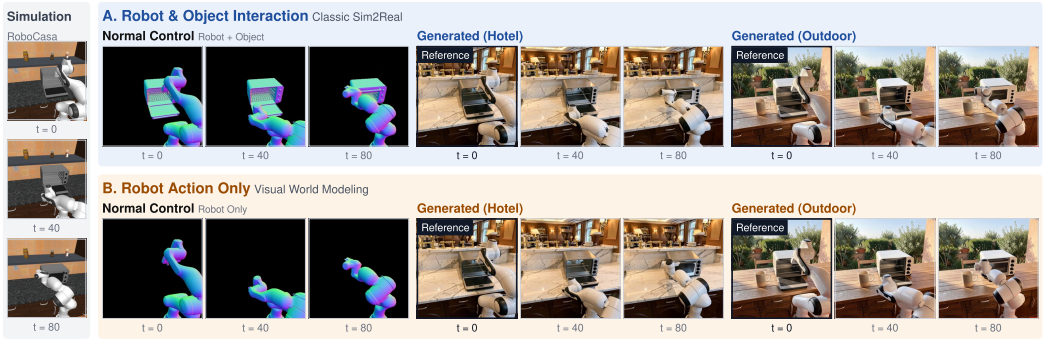

Figure 8: **Robotics application.** Left: a simulated RoboCasa (Nasiriany et al., 2026) episode ("Close the toaster oven door."), frames t=0, 40, 80 of 81. Rows: the normal-map control and the videos Real2SAM2Real generates from it in two scenes (hotel coffee station, outdoor terrace) at the same time steps. (A) the control contains the robot and the interacted object (Sim2Real); (B) it contains only the robot (visual world modeling). The frame tagged *reference* is the input reference image; all other frames are generated. Videos are presented in the HTML supplementary material.

As a preliminary application, we use Real2SAM2Real for Sim2Real and visual world modeling without robot-specific training (Fig. 8). A RoboCasa (Nasiriany et al., 2026) episode is rendered to normal-map control, while FLUX.2-klein-9B generates reference images for different scenes from the corresponding simulated frame while preserving its layout.

For Sim2Real, the control contains both the robot and the interacted object (Fig. 8A), preserving the simulated interaction across reference-defined environments. For visual world modeling, only the robot is controlled (Fig. 8B), and the model infers the object's response from the reference image and arm motion; the toaster oven door closes consistently with the arm in both scenes. These qualitative examples suggest potential for realistic manipulation video generation and limited visual world modeling; systematic evaluation remains future work.

5 Conclusion

We presented Real2SAM2Real, a highly controllable 3D-aware video generation framework that addresses the structural collapse and generative hallucinations of video diffusion models (VDMs) under high-dynamic motion and complex occlusions. Our coarse 3D foreground cache uses topologically complete meshes for intuitive animation. Normal maps rendered under specified camera trajectories and foreground motions guide the VDM, which synthesizes appearance while largely retaining its pretrained priors. Adapted from around 2,000 training clips, Real2SAM2Real maintains spatiotemporal coherence across camera control, entity manipulation, and high-dynamic scenes, even under large viewpoint shifts, severe occlusions, and non-Lambertian effects.

AI USE STATEMENT

We used generative AI tools for manuscript translation, language editing, LaTeX formatting, and figure refinement. The data pipeline also uses pretrained vision-language, segmentation, and normal-estimation models to construct instance-masked geometric conditions, as described in Appendix B.1. FLUX.2-klein-9B generates reference images for the robotics examples (Section 4.6). The RGB training targets are taken from existing monocular videos. The authors take responsibility for the final text, figures, methodology, and reported results.

ETHICS STATEMENT

Our study uses only existing video datasets and online images; we did not recruit participants or collect new recordings of individuals. Controllable video generation may be misused to create misleading depictions of people or events. Generated content should be identified as synthetic, and applications should respect privacy, consent, and applicable dataset and model licenses. The system may inherit biases from pretrained models and source data. The robotics examples demonstrate visual generation, and physical deployment would require separate validation.

REPRODUCIBILITY STATEMENT

Section 3 describes the geometry representation, conditioning mechanism, and training objective. Appendix B documents the data processing pipeline and implementation settings, while Section 4 reports the evaluation datasets, baselines, metrics, and ablations. We will publicly release our code and model weights to support reproduction and further evaluation.

REFERENCES

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Xiao Fu, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. *arXiv preprint arXiv:2412.07760*, 2024.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14834–14844, 2025.

Yanrui Bin, Wenbo Hu, Haoyuan Wang, Xinya Chen, and Bing Wang. Normalcrafter: Learning temporally consistent normals from video diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8330–8339, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Aleksei Bochkovskii, AmaãĢl Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024.

Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16901–16911, 2024.

Xiao Fu, Xian Liu, Xintao Wang, Sida Peng, Menghan Xia, Xiaoyu Shi, Ziyang Yuan, Pengfei Wan, Di Zhang, and Dahua Lin. 3dtrajmaster: Mastering 3d trajectory for multi-entity motion in video generation. *arXiv preprint arXiv:2412.07759*, 2024.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, et al. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1–12, 2025.

Michael Grupp. evo: Python package for the evaluation of odometry and slam. `https://github.com/MichaelGrupp/evo`, 2017.

Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2025.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Adam W Harley, Yang You, Xinglong Sun, Yang Zheng, Nikhil Raghuraman, Yunqi Gu, Sheldon Liang, Wen-Hsuan Chu, Achal Dave, Suya You, et al. Alltracker: Efficient dense point tracking at high resolution. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5253–5262. IEEE, 2025.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh. *arXiv preprint arXiv:2506.05554*, 2025.

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. In *NVIDIA Research Whitepapers arXiv:2508.10934*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Team Hunyuan3D, Shuhui Yang, Mingxin Yang, Yifei Feng, Xin Huang, Sheng Zhang, Zebin He, Di Luo, Haolin Liu, Yunfei Zhao, et al. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material. *arXiv preprint arXiv:2506.15442*, 2025.

Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5252–5262. IEEE, 2020.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Kunhao Liu, Ling Shao, and Shijian Lu. Novel view extrapolation with video diffusion priors. *arXiv preprint arXiv:2411.14208*, 2024.

Yawen Luo, Xiaoyu Shi, Jianhong Bai, Menghan Xia, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Camclonemaster: Enabling reference-based camera control for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–10, 2025.

Baorui Ma, Huachen Gao, Haoge Deng, Zhengxiong Luo, Tiejun Huang, Lulu Tang, and Xinlong Wang. You see it, you got it: Learning 3d creation on pose-free videos at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2016–2029, 2025.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 4296–4304, 2024.

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. *arXiv preprint arXiv:2407.02371*, 2024.

Soroush Nasiriany, Sepehr Nasiriany, Abhiram Maddukuri, and Yuke Zhu. Robocasa365: A large-scale simulation framework for training and benchmarking generalist robots. In *International Conference on Learning Representations (ICLR)*, 2026.

Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L Schönberger. Global structure-from-motion revisited. In *European Conference on Computer Vision*, pp. 58–77. Springer, 2024.

Huaijin Pi, Ruoxi Guo, Zehong Shen, Qing Shuai, Zechen Hu, Zhumei Wang, Yajiao Dong, Ruizhen Hu, Taku Komura, Sida Peng, et al. Motion-2-to-3: Leveraging 2d motion data to boost 3d motion generation. *arXiv preprint arXiv:2412.13111*, 2024.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10106–10116, 2024.

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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132. IEEE, 2025.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2372–2381. IEEE, 2019.

Chenxi Song, Yanming Yang, Tong Zhao, Ruibo Li, and Chi Zhang. Worldforge: Unlocking emergent 3d/4d generation in video diffusion model via training-free guidance. *arXiv preprint arXiv:2509.15130*, 2025.

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. In *European Conference on Computer Vision*, pp. 313–331. Springer, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

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*, 2025b.

Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, and Dahua Lin. HumanVid: Demystifying training data for camera-controllable human image animation. In *NeurIPS*, 2024b.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024c.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17868–17879, 2024.

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

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21469–21480, 2025.

Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan. Trajectory attention for fine-grained video motion control. *arXiv preprint arXiv:2411.19324*, 2024.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024a.

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollar, and Kris Kitani. SAM 3D Body: Robust full-body human mesh recovery. *arXiv preprint arXiv:2602.15989*, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024b.

Chongjie Ye, Yushuang Wu, Ziteng Lu, Jiahao Chang, Xiaoyang Guo, Jiaqing Zhou, Hao Zhao, and Xiaoguang Han. Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25050–25061, 2025.

Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. *arXiv preprint arXiv:2405.15364*, 2024.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 100–111, 2025.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2050–2062, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

Supplementary Material for: "Real2SAM2Real: Generative 3D Caches as Complementary Context for Video Diffusion"

A Video Results

Since static images are inherently limited in conveying the spatiotemporal coherence and perceptual quality of generated videos, we highly encourage readers to refer to the static HTML webpage (`index.html`) provided in the supplementary material, or directly navigate the video folders, to view the comprehensive video synthesis results.

We provide nearly 100 qualitative video comparisons. All video results corresponding to the experiments in the main paper can be accessed via the provided HTML file, covering the following scenarios and applications:

- **Camera Control under Extreme Trajectories:** Camera control along long-range trajectories with large rotations and translations.
- **Camera Control:** Large viewpoint changes with occlusions, compared with the baselines.
- **Reflection, Refraction, and Fluid Dynamics:** Scenes with non-Lambertian surfaces and time-varying effects (e.g., smoke and fire), whose appearance is synthesized by the video diffusion model rather than encoded in the 3D cache.
- **Dynamic Scenes:** Reconstruction of real videos with large non-rigid motion and camera motion from a single frame, compared side by side with all baselines.
- **Motion Transfer:** Editing the reference image while keeping the normal-map sequence fixed to transfer motion across different foreground appearances, e.g., retargeting a captured human performance to a new character.
- **Object Manipulation:** 3D editing of individual foreground entities, such as 6-DoF rigid transformations.
- **Joint Control via 3D Guidance and Text:** Joint control by 3D geometric guidance and text prompts that specify stylistic or environmental effects.
- **Ablation Studies:** Visual comparisons of the effects of soft spatial-aligned injection and instance-level perturbation.
- **Robotics Application:** Videos generated from normal-map controls of robot manipulation episodes (e.g., closing a toaster oven door), illustrating potential applications in Sim2Real and visual world modeling.

B Implementation Details

B.1 Training Data Curation

To construct paired RGB–normal training data, we apply the 3D-free pipeline described in Sec. 3.2 to in-the-wild monocular videos from DL3DV (Ling et al., 2024), HumanVid (Wang et al., 2024b), and OpenVid-HD (Nan et al., 2024). Specifically, given a video, objects are identified by Qwen2-VL (Wang et al., 2024a), and YOLO-World (Cheng et al., 2024) localizes each instance via text-conditioned bounding boxes, from which SAM 3 (Carion et al., 2025) generates pixel-level instance masks. NormalCrafter (Bin et al., 2025) estimates temporally consistent normals, to which we apply the instance masks to form RGB–normal training pairs. In total, we use around 2,000 video clips, with approximately 13% from DL3DV, 17% from HumanVid, and 70% from OpenVid-HD. We uniformly sample 81 frames from each clip.

B.2 Training Setup

We implement our framework based on Wan2.2-Fun-A14B-InP and fine-tune it on two NVIDIA H200 GPUs (batch size of 1, BF16 mixed precision) for approximately two days. We report results using the checkpoint at training step 6,000. The high- and low-noise experts are fine-tuned separately, each with its own LoRA modules ($r = 64$, $\alpha = 32$) and geometry patch embedding. We use AdamW

with a cosine learning-rate schedule and $\epsilon = 10^{-10}$. LoRA uses $lr = 10^{-4}$ and weight decay 0.03, while the geometry embedding uses $lr = 10^{-5}$ and no weight decay. Training uses 720p videos of 81 frames, with independent 10% dropout for the normal, reference-frame, and text conditions.

The flow-matching loss masks extreme velocity residuals: $\rho(e) = D^{-1} \sum_{i=1}^{D} e_i^2 \mathbf{1}(|e_i| \leq 50)$, where D is the number of scalar residuals. The mean is taken over all elements, including masked entries.

C Experimental Setup

Each experiment is described in the same order: test data, baselines and their conditioning, and evaluation metrics.

C.1 Camera Control

Test Data We source 100 in-the-wild images from Unsplash, spanning indoor scenes, outdoor environments, and object-centric compositions. For each image, we reconstruct the foreground instances as posed meshes with SAM 3D (Chen et al., 2025) and place them in a GLB scene, with the camera at the source viewpoint. We sample two camera trajectories per image, yielding 200 test cases. Each trajectory orbits a pivot on the optical axis at the depth of the foreground meshes; the first trajectory moves in a random in-plane direction, and the second moves horizontally to the left or right. Each trajectory contains 81 camera poses recorded directly in the GLB scene, so ground-truth poses are available without post-hoc reconstruction.

Baselines and Conditioning All methods start from the same source image and follow the same ground-truth trajectory, using their released checkpoints and recommended inference settings. ReCamMaster (Bai et al., 2025) takes the source image repeated as a static 81-frame video, together with the ground-truth poses converted to its camera format. ViewCrafter (Yu et al., 2024), TrajectoryCrafter (Yu et al., 2025), and WorldForge (Song et al., 2025) share an identical warped video: following the warping procedure of WorldForge, the source image is lifted to a point cloud with the depth and intrinsics estimated during SAM 3D reconstruction and rendered along the ground-truth trajectory. Each of these methods samples the shared warp at its native length (25, 49, and 49 frames, respectively). GEN3C (Ren et al., 2025) builds its own 3D cache from the source image and renders it along the ground-truth trajectory at its native length of 121 frames. Ours renders the instance-masked normal maps of the same GLB scene along the ground-truth trajectory (81 frames).

Evaluation Metrics Following (Song et al., 2025; Bai et al., 2025), we measure camera accuracy with Absolute Trajectory Error (ATE) and Relative Pose Error in translation and rotation (RPE-T, RPE-R). For each generated video, we estimate the camera trajectory from all frames with GLOMAP (Pan et al., 2024) and compare it against the ground truth with evo (Grupp, 2017) after Sim(3) alignment; RPE is computed between consecutive frames. Video quality is evaluated with VBench (Huang et al., 2024), following (Bai et al., 2025): we report Subject Consistency, Background Consistency, Aesthetic Quality, and Imaging Quality.

C.2 High-Dynamic Scenes

Test Data We collect 300 real-world clips, 100 from HumanVid (Wang et al., 2024b) and 200 from OpenVid-HD (Nan et al., 2024), selected for large non-rigid motion coupled with large camera motion. The source videos are at 1280×720 and 16 fps, and all methods are evaluated within source frames 0 to 80. For our method, we obtain the dynamic 3D cache by estimating a per-frame human mesh with SAM 3D Body (Yang et al., 2026), and render it as instance-masked normal maps. Since the cache is recovered from the source video, the source video serves as ground truth, and each method reconstructs it from a single frame.

Baselines and Conditioning All methods start from a single frame of the source video, and each baseline additionally receives the strongest motion conditioning it supports, derived from the same source video. ReCamMaster (Bai et al., 2025) receives the camera trajectory estimated from the source video by ViPE (Huang et al., 2025). Uni3C (Cao et al., 2025) receives the camera trajectory and scene point cloud produced by its own pipeline from the source video. DaS (Gu et al., 2025)

receives the 3D tracking video estimated from the source video. TrajectoryCrafter (Yu et al., 2025) and WorldForge (Song et al., 2025) share a dense warped video: using AllTracker (Harley et al., 2025) with frame 0 as the query frame, we track every pixel through the source video and forward-splat the pixels of frame 0 to each subsequent frame, keeping the most confident pixel when several land on the same location and marking uncovered regions as holes. All pixels of the warped video come from frame 0, so no appearance of the ground-truth video is leaked. Ours receives the normal-map sequence rendered from the SAM 3D Body cache.

Temporal and Spatial Alignment Since the methods sample frames differently, we compare them only on the source frames covered by all methods, taking the nearest output frame where sampling points differ. All videos are evaluated on a common grid, obtained by cropping the central 1080×720 region of the source frame and resizing it to 720×480. The native cropping and resizing of each method is determined by matching its output frames against the source video.

Evaluation Metrics For human pose accuracy, we run SAM 3D Body on each generated video and on the ground-truth video with identical crops and box prompts, and compare the two sets of pose estimates. AKD-2D follows the keypoint protocol of Monkey-Net (Siarohin et al., 2019) and measures the average 2D distance of 23 body and foot joints on the 720×480 grid. PA-MPJPE (Kocabas et al., 2020) extends this protocol to 3D: it is computed over the same 23 joints after root removal and similarity (Procrustes) alignment, in millimeters. MKR is the fraction of frames in which the main subject cannot be detected without a box prompt. Video quality is evaluated with VBench (Huang et al., 2024): we report Subject Consistency, Background Consistency, Aesthetic Quality, and Imaging Quality.

C.3 Entity Motion Control

Test Data We construct 100 test cases, each assigning randomized 6-DoF trajectories (translation and rotation) to one or more foreground objects in a GLB scene, while the camera remains static. Each trajectory spans 81 frames, and ground-truth object poses are recorded directly in the GLB scene without post-hoc reconstruction.

Baselines and Conditioning All methods start from the same source image and follow the same object trajectories. DaS (Gu et al., 2025) receives a 3D tracking video: we lift the source image to a point cloud with monocular geometry estimation, move the points of the manipulated objects along the ground-truth trajectories with the camera fixed, and render the result. ViewCrafter (Yu et al., 2024) is not designed for entity manipulation; we instead provide a warped video proxy, obtained by forward-warping the source image with the same moving point cloud, which enables it to perform the task. Ours moves the SAM 3D meshes of the objects along the same trajectories and renders their instance-masked normal maps.

Evaluation Metrics We track the 6-DoF pose of each manipulated object in the generated videos with FoundationPose (Wen et al., 2024), using the SAM 3D mesh of the object as its model; the object is registered in the first frame with its instance mask and tracked in subsequent frames. Since generated videos have no depth, we estimate per-frame relative depth with Depth Anything V2 (Yang et al., 2024a) and calibrate its scale and shift on the first frame against the depth of the mesh rendered at the ground-truth pose, yielding metric depth. All methods share the camera intrinsics estimated during SAM 3D reconstruction, which are also used to render the ground truth. Tracked poses are converted to motion relative to the first frame and compared with the ground-truth relative motion, excluding the first frame, so that registration errors in the first frame cancel out and the metrics reflect whether the objects move correctly.

D Robustness to 3D Cache Errors

Single-image 3D reconstruction is inherently under-constrained, and the 3D cache produced by SAM 3D may contain geometric and placement errors. Our design treats the cache as a coarse spatial anchor rather than an exact rendering constraint: it enters through soft spatial-aligned injection, and training-time instance-level perturbation teaches the model to tolerate misaligned normals (Sec. 3.2).

To quantify this tolerance, we inject controlled errors into the estimated cache on the camera-control benchmark.

Setup Each object in the cache is displaced in a random direction by a magnitude δ, given as a percentage of its own bounding-box diagonal. We evaluate eight levels, $\delta \in \{3, 6, 9, 12, 16, 19, 22, 25\}\%$, on a subset of 30 test cases, and group them into three bands in Tab. 4; all methods, including the baselines, are re-evaluated on this subset, so the numbers differ from those in Tab. 1. At the strongest level, objects may float above their supporting surfaces and intersect their neighbors. Besides camera accuracy and VBench, we report PSNR, SSIM, and LPIPS of the first generated frame against the source image, which reveal whether cache errors propagate into the generated pixels.

Results Video quality is largely insensitive to cache errors: VBench stays at 0.84, and first-frame PSNR, SSIM, and LPIPS remain flat, although the displacement grows about fivefold from the lowest to the highest band. Camera accuracy degrades gradually: ATE rises monotonically from 0.015 to 0.027, indicating that the cache does anchor the spatial layout, while RPE-T and RPE-R remain nearly unchanged. Even at the strongest corruption, our ATE and RPE-T remain lower than those of all warp-based baselines evaluated on the same cases.

Table 4: Robustness to 3D cache errors on 30 camera-control test cases. δ: random displacement of each cached object, as a percentage of its bounding-box diagonal. PSNR, SSIM, and LPIPS compare the first generated frame with the source image. Since all methods are evaluated on this subset, the numbers are not directly comparable to Tab. 1.

Method	δ (%)	PSNR ↑	SSIM ↑	LPIPS ↓	VBench ↑	ATE ↓	RPE-T ↓	RPE-R ↓
Ours	0	30.63	0.8998	0.0669	0.838	0.015	0.014	1.087
Ours	3–6	30.71	0.8998	0.0661	0.840	0.017	0.014	1.066
Ours	9–16	30.69	0.8998	0.0660	0.840	0.023	0.015	1.058
Ours	19–25	30.61	0.8994	0.0662	0.840	0.027	0.015	1.038
TrajectoryCrafter (Yu et al., 2025)	–	–	–	–	0.822	0.034	0.029	1.811
ViewCrafter (Yu et al., 2024)	–	–	–	–	0.829	0.054	0.062	3.756
WorldForge (Song et al., 2025)	–	–	–	–	0.816	0.076	0.039	1.920

E LIMITATION

Our framework supports human motion and deformable objects, but the assets produced by most 3D lifting models are not directly simulation-ready. Accurately simulating complex interactions between objects, contact-induced deformation, or fracture remains challenging. In these settings, a coarse 3D cache may provide misleading geometric guidance, limiting the physical plausibility and controllability of the generated interactions.

Appearance is directly anchored only at the reference frame. For other frames, the VDM must propagate reference appearance using its spatiotemporal priors and coarse geometric guidance, without an equally direct appearance constraint. Fine textures and instance identity may therefore drift under large viewpoint changes or prolonged occlusions, while the appearance of newly revealed surfaces depends on generative inference.