4DIRECTOR: CONTROLLING VIDEO WORLD MODELS WITH RIGID 3D GEOMETRY

Anonymous authors
Paper under double-blind review

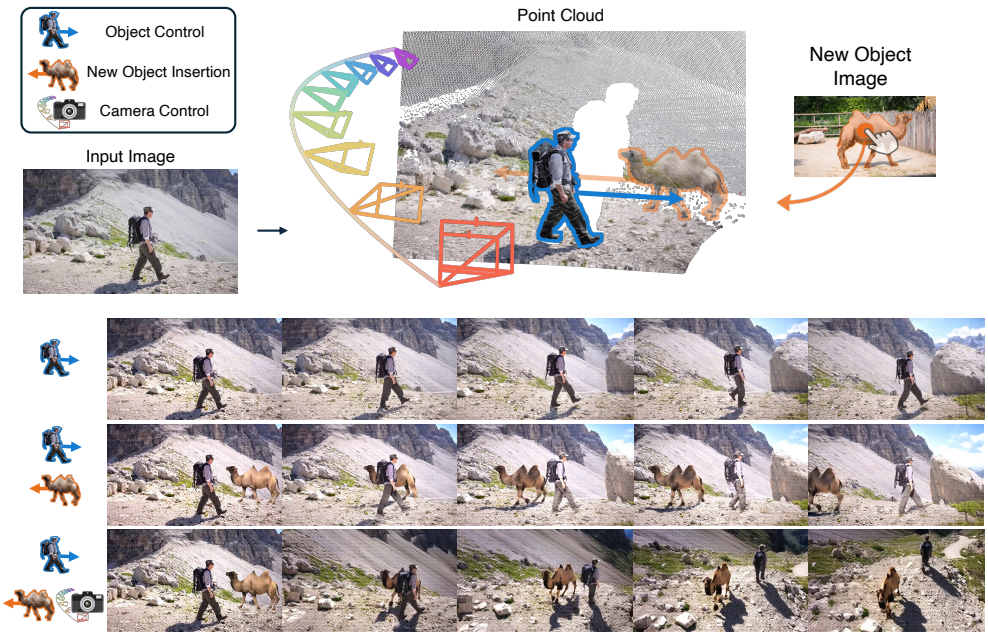
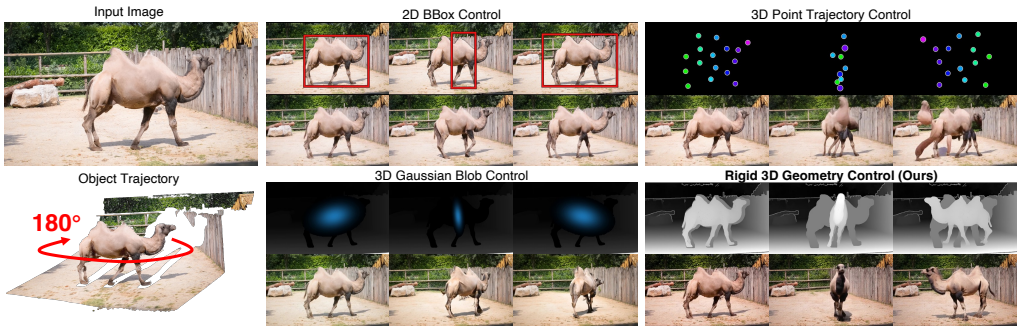


Figure 1: **Camera and object control from a single image.** *Top:* the image is lifted into a background point cloud and complete rigid 3D geometry for each object, in which the user draws the object and camera trajectories and, optionally, inserts a new object from a reference image. *Bottom:* 4Director generates videos that follow both trajectories, with plausible object motion, consistent appearance and illumination, and the background revealed by the camera filled in.

ABSTRACT

Precise control over camera and object motion is essential for professional video production. Existing methods control objects only coarsely, through image-plane cues that are ambiguous in depth and rotation or through 3D tracks and blobs that lack complete geometry and lose consistency across viewpoint changes. We introduce **4Director**, a video world model conditioned on an explicit 4D scene representation: each object is reconstructed once from the input image as a canonical mesh and moved by one prescribed rigid transformation per frame. This representation provides an intuitive 3D control interface and prevents unobserved geometry from being regenerated independently in every frame. We render the controlled scene as a depth video and introduce a *Motion Adapter* that transforms this geometric scaffold into video while synthesizing view-consistent appearance, illumination, and non-rigid dynamics. For training, we construct *RealCOD-Rigid*, a new dataset of 20,774 clips annotated with rigid 3D scenes by our automatic pipeline. We further introduce *Identity-Gated IoU* (IG-IoU), which jointly evaluates adherence to prescribed object motion and preservation of object identity. Experiments demonstrate that 4Director consistently outperforms prior methods in visual quality and in camera and object control. Anonymous Page: https://anonymous-author-page.github.io/4Director/.

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.



Method	Camera trajectory	3D object position	Object orientation	Complete geometry	New object insertion
MotionCtrl (Wang et al., 2024d)	✓	✓ᵖ	✗	✗	–
Perception-as-Control (Chen et al., 2025b)	✓	✓	✗	✗	–
SymphoMotion (Zhang et al., 2026)	✓	✓	✓ᵖ	✗	–
VerseCrafter (Zheng et al., 2026)	✓	✓	✓ᵖ	✗	–
4Director (ours)	✓	✓	✓	✓	✓

Figure 2: **Control representations.** *Top:* given an input image and an object trajectory (a 180° turn in place), we compare four control signals: a 2D bounding box, a 3D Gaussian blob, 3D point trajectories, and our rigid 3D geometry. *Bottom:* capabilities of the compared representations (✓ supported, ✓ᵖ partial, ✗ not supported, – not stated in the paper).

1 Introduction

A director stages a shot by designing the trajectories of the camera and every actor, and leaves how an actor walks or how the light falls to the cast and the crew. We develop **4Director**, a video world model (Bruce et al., 2024; Decart et al., 2024; Alonso et al., 2024) with the same principle: the user prescribes the trajectory of each object and the camera in 3D (Fig. 1), and the generator supplies the non-rigid dynamics, appearance and illumination along them.

Existing controls fall short of such trajectories in one of two ways: image-plane cues are ambiguous in depth and rotation, and 3D controls carry incomplete object geometry. Image-plane cues such as dragged paths, bounding boxes, masks, or point tracks (Yin et al., 2023; Wang et al., 2024b; Zhang et al., 2025c; Zhou et al., 2025a; Geng et al., 2025; Wang et al., 2024d) state where an object appears in each frame, but one projection is consistent with many 3D motions: a box does not determine the orientation of the object, a shrinking box does not distinguish an object moving away from one becoming smaller, and a 2D path conflates camera parallax with object motion (Fig. 2). 3D-aware controls avoid this ambiguity with rigid trajectories or 3D boxes (Fu et al., 2025; Shuai et al., 2025; Wang et al., 2025d), or with proxies lifted from the input, such as tracked 3D points (Zhang et al., 2026), spheres on tracked parts (Chen et al., 2025b) and one 3D Gaussian blob per object (Zheng et al., 2026). Their geometry, however, is incomplete: a trajectory or box carries no surface, and a lifted proxy captures only the visible shell of the object and has no geometry for regions that a turn or a new viewpoint brings into view (Fig. 2, Gaussian and point panels); point-based proxies are, moreover, driven by per-point trajectories, which are cumbersome to author one by one.

To address both shortcomings, we propose **4Director**, a video world model driven by an explicit 4D scene representation. Every object is a complete canonical mesh, reconstructed once and moved by one rigid transformation per frame. The meshes, a background point cloud and the camera share one coordinate frame. Explicit 3D space removes the ambiguity of image-plane cues and gives the user a direct interface: a turn is a rotation, motion in depth is a displacement rather than a change of size. Complete geometry removes the limitation of lifted proxies: the mesh has a surface on every side, so the geometry that a turn or new viewpoint reveals is rendered rather than regenerated (Fig. 2). And one rigid transformation moves the whole object without per-point trajectories.

We render the scene to a depth video in which every object moves as a single rigid body. We design a *Motion Adapter*, a trainable branch that injects this rigid rendering into a pretrained video

generator (Wang et al., 2025a). To train it, we build an automatic annotation pipeline that recovers the rigid 3D scene of a monocular clip, and run it on RealCOD-25K (Zhang et al., 2026) to construct *RealCOD-Rigid*, a dataset of 20,774 annotated clips. The adapter is trained to generate each clip from its rigid rendering, and so learns to supply what the rendering lacks: appearance, illumination, non-rigid dynamics and the background that a camera move reveals.

Since 4Director is capable of generating videos with explicit control on the objects and camera, we evaluate 4Director on visual quality, object control, and camera control. For visual quality, we employ standard metrics as well as a user study. Camera control is evaluated with the standard trajectory error. Object control lacks a comparable metric: mask IoU alone rewards correct placement regardless of identity. We therefore introduce *Identity-Gated IoU* (IG-IoU), which accumulates mask IoU only over frames in which the object's identity is preserved. From our evaluations, we find that 4Director outperforms prior methods on all three criteria. Our contributions are as follows:

- We introduce an ***explicit 4D scene representation*** with complete object geometry and per-frame rigid transformations that holds the camera and every object in one coordinate frame, resolving the ambiguity of image-plane cues and the incomplete geometry of lifted proxies.
- We present the ***Motion Adapter***, a trainable branch that drives a pretrained video generator with the rendered rigid scene, so that the geometry fixes the camera and object motion while the generator supplies appearance, illumination and non-rigid dynamics.
- We introduce ***RealCOD-Rigid***, a dataset of 20,774 clips annotated with complete object meshes, per-frame rigid transformations and camera trajectories by an automatic pipeline, and ***Identity-Gated IoU***, a metric that scores object control only where identity is preserved.
- Experiments and a user study show that 4Director outperforms state-of-the-art methods in visual quality, camera control, and object control.

2 Related Work

Video world models. World models roll out future observations (Ha & Schmidhuber, 2018; LeCun, 2022; Hafner et al., 2023; Cao et al., 2025b); diffusion and transformer backbones now produce realistic rollouts conditioned on actions, text or a camera path (Blattmann et al., 2023; Bruce et al., 2024; Parker-Holder et al., 2024; 2025; Decart et al., 2024; Alonso et al., 2024; Agarwal et al., 2025; Alhaija et al., 2025; Hu et al., 2023; Che et al., 2025; Yu et al., 2025b; He et al., 2025b; Li et al., 2025a), and extend the horizon with memory (Henschel et al., 2025; Qiu et al., 2023; Po et al., 2025; Yu et al., 2025a; Xiao et al., 2025; Li et al., 2025c; Wu et al., 2025; Chen et al., 2026; Duan et al., 2026; Sun et al., 2025; Hong et al., 2025). A geometry-aware line, including DeepVerse (Chen et al., 2025a), Voyager (Huang et al., 2025), Yume (Mao et al., 2025) and Aether (Zhu et al., 2025), adds reconstructed 3D structure for consistent exploration (Sun et al., 2024; Team et al., 2025; Yang et al., 2025). Their control, however, is expressed as text, actions or camera tokens, through which the motion of individual objects cannot be prescribed. 4Director instead exposes the scene itself: an explicit 4D scene representation edited before generation, with complete object meshes, one rigid transformation per frame and the camera in one coordinate frame. We therefore use the term *world model*, as VerseCrafter (Zheng et al., 2026) does, for a generator whose output follows an explicit, editable scene state rather than an interactive rollout.

Camera control. Pose conditioning, as Plücker rays (He et al., 2024; Sitzmann et al., 2021; Zhou et al., 2025b; Feng et al., 2024a; Li et al., 2025d; Wang et al., 2025f), epipolar attention (Xu et al., 2024), conditioning at chosen layers (Bahmani et al., 2025; 2024) or multi-view and reference-driven control (Kuang et al., 2024; Bai et al., 2024; He et al., 2025a; Luo et al., 2025; Zheng et al., 2024), says nothing about what the new view should contain. A second line renders lifted geometry along the target path: a colored point cloud (Yu et al., 2024), a cached 3D scene (Ren et al., 2025), warped geometry with its holes repaired (Hou et al., 2024; Yu et al., 2025c; Song et al., 2026; Popov et al., 2025; Hu et al., 2025), a source video re-rendered along a new path (Van Hoorick et al., 2024; Zhang et al., 2024; Bai et al., 2025a; Cao et al., 2026), a point cloud with a human body (Cao et al., 2025a), or depth maps for a world model (Alhaija et al., 2025), injected through a branch such as ControlNet or VACE (Zhang et al., 2023; Jiang et al., 2025). Lifted geometry covers only what the reference view saw, so what a camera move reveals is hallucinated anew in each frame, although complete shapes are available from image or 4D generation and reconstruction (Li et al., 2026; Voleti et al., 2024; Xie et al., 2025; Yao et al., 2025; Zhang et al., 2025b; Cao et al., 2024; Tang et al., 2026).

In contrast, 4Director reconstructs every object as a complete mesh, so the object geometry that a camera move reveals is rendered rather than regenerated, and its Motion Adapter learns to supply appearance, illumination and non-rigid dynamics on top of fixed geometry.

Object control. Image-plane cues, from dragged paths to tracked points and boxes, state where an object should appear (Wang et al., 2023; Yin et al., 2023; Ma et al., 2024; Zhang et al., 2025c; Zhou et al., 2025a; Wang et al., 2024b; Geng et al., 2025; Shi et al., 2024; Niu et al., 2024; Jain et al., 2024; Qiu et al., 2024; Li et al., 2025b; Wu et al., 2024; Xing et al., 2025; Burgert et al., 2025; Wang et al., 2025b; Chu et al., 2025); several systems add a camera module (Yang et al., 2024; Feng et al., 2024b; Lei et al., 2024; Li et al., 2025e; Zheng et al., 2025), MotionCtrl among them, with camera extrinsics and 2D object trajectories (Wang et al., 2024d). Moving the control into 3D, as a depth-augmented trajectory (Wang et al., 2025c; 2024c), a 6-DoF pose (Fu et al., 2025; Shuai et al., 2025; Liang et al., 2025), a 3D box (Wang et al., 2025d) or a coarse reconstruction in a 3D engine (Zhang et al., 2025d), fixes the 3D position, and all but the trajectory also fix orientation. The closest methods lift a proxy of the object from the source and render it with the camera: Perception-as-Control places spheres on tracked parts (Chen et al., 2025b), SymphoMotion (Zhang et al., 2026) and Diffusion as Shader (Gu et al., 2025) render 3D point tracks over a point cloud, and VerseCrafter moves one 3D Gaussian per object beside a background cloud (Zheng et al., 2026). Image-plane cues do not fix depth and rotation, 3D poses and boxes carry no surface, lifted proxies carry only the visible shell, and point-based ones are driven by per-point trajectories. In contrast, 4Director moves a complete canonical mesh by one rigid transformation per frame, so that position, orientation and the revealed surface are fixed before generation and one trajectory moves the whole object.

3 METHOD

We study camera and object control from a single image: given an image, a text prompt, a camera trajectory $\{\mathbf{E}^t\}_{t=1}^F$, and a rigid trajectory $\{\mathbf{T}_o^t\}_{t=1}^F$ for each object o that the user marks in the image, our goal is to synthesize a video of F frames that starts from the image, follows the camera trajectory and moves each object along its trajectory, with its identity preserved and plausible non-rigid dynamics. Here $\mathbf{E}^t$ is the world-to-camera matrix of frame t and $\mathbf{T}_o^t = (\mathbf{R}_o^t, \mathbf{t}_o^t) \in \mathrm{SE}(3)$ the rigid transformation of object o from its placement in the image to frame t (Fig. 3).

3.1 RIGID 3D GEOMETRY CONTROL

Rigid 3D geometry. We first lift the input image into 3D. MoGe-2 (Wang et al., 2025e) estimates its depth and intrinsic matrix $\mathbf{K}$, and SAM 2 (Ravi et al., 2025) turns the user's clicks on each object into a mask. The camera of the image defines the reference frame ($\mathbf{E}^1$ is the identity), in which two kinds of geometry are built. The background, the image outside the masks, is back-projected into a static point cloud $\mathcal{P}$. Each marked object is reconstructed by Pixal3D (Li et al., 2026) into a complete canonical mesh $\mathcal{M}_o$, with the sides the image does not show, and aligned to its mask and depth. Background, camera, and objects thus share one coordinate frame.

Camera and object trajectories. The user prescribes the trajectories in this frame by keyframes in a 3D viewer. Object o moves as a whole by $\mathbf{T}_o^t$, with $\mathbf{T}_o^1$ the identity, so that at frame t it is the placed mesh

$$\mathcal{M}_o^t = \{\mathbf{R}_o^t\,\mathbf{v} + \mathbf{t}_o^t \mid \mathbf{v} \in \mathcal{M}_o\}, \tag{1}$$

and the camera moves by $\mathbf{E}^t$. The scene thus carries only rigid motion. A new object can be inserted as a mesh reconstructed from a separate image and given a trajectory in the same way; its textured mesh is rendered into the input image, which then serves as the first frame. Altogether, the rigid 3D scene is the tuple

$$\mathcal{S} = \big(\mathcal{P},\ \{\mathcal{M}_o\},\ \{\mathbf{T}_o^t\},\ \mathbf{K},\ \{\mathbf{E}^t\}\big), \tag{2}$$

in which $\mathcal{P}$, the meshes and $\mathbf{K}$ are constant and only $\{\mathbf{T}_o^t\}$ and $\{\mathbf{E}^t\}$ vary in time.

Depth rendering. For each frame, $\mathcal{P}$ and the placed meshes $\{\mathcal{M}_o^t\}$ are projected with $(\mathbf{K}, \mathbf{E}^t)$ into a depth map, in which each pixel takes the depth of the nearest surface and pixels without geometry are empty. The resulting depth video $\mathbf{C}$ is the control. It carries only the viewpoint, the rigid transformation of every object and their occlusion; appearance, illumination, non-rigid dynamics such as the movement of a walker's limbs, and the background beyond what the image shows are absent, and supplying them is the task of the Motion Adapter.

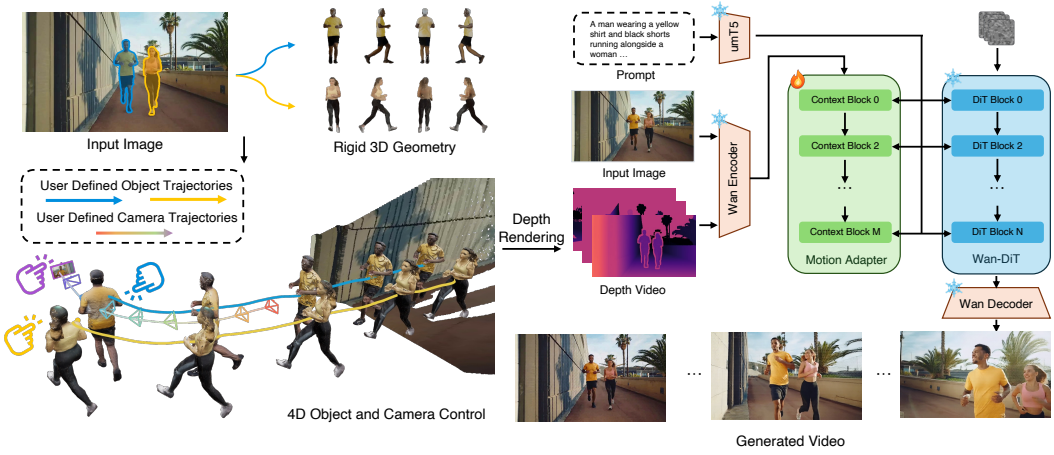


Figure 3: **Overview of 4Director.** *Left:* each object is reconstructed once as a complete canonical mesh and moved by one rigid transformation per frame; the meshes, a background point cloud and the camera share one coordinate frame, in which the user prescribes the trajectory of each object and the camera, and the scene is rendered to a depth video (Sec. 3.1). *Right:* the *Motion Adapter*, a trainable branch, injects this rigid rendering into a pretrained video generator, which supplies appearance, illumination and non-rigid dynamics (Sec. 3.2); it is trained on RealCOD-Rigid (Sec. 4).

3.2 Motion Adapter

Architecture. We build on Wan2.1-VACE-14B (Wang et al., 2025a; Jiang et al., 2025), a latent video diffusion transformer whose VAE, umT5 (Chung et al., 2023) text encoder and DiT serve as a generic video prior and are left unchanged. The *Motion Adapter* is a DiT-style branch that follows the context-branch design of VACE: as shown in Fig. 3 (right), the VAE encodes the depth video $\mathbf{C}$ and the input image into a context stream, which eight context blocks, one for every fifth DiT block, propagate with cross-attention to the prompt and inject into the DiT as linearly projected hints.

Training. Each training pair is built from one original clip (Sec. 4): the control video rendered from the clip's recovered rigid 3D scene is the condition, and the clip is the target to be generated; its first frame is the input image and its text description the prompt. The control video contains the rigid part of each object's motion but not the non-rigid part. A surface point i of object o is observed at

$$\mathbf{X}_i^t = \mathbf{R}_o^t \, \mathbf{Y}_i + \mathbf{t}_o^t + \mathbf{r}_i^t, \tag{3}$$

where $\mathbf{Y}_i$ is its position on the canonical mesh, the first two terms are the rigid motion of the object, and the residual $\mathbf{r}_i^t$ is its non-rigid motion, such as a swinging limb; only the first two terms enter the control video. To generate the clip, the adapter must therefore follow the rigid geometry and let the generator supply what the control omits: the non-rigid motion, with appearance and illumination.

The adapter is trained with the standard flow-matching objective. Let $\mathbf{z}$ be the latent of the clip and $\mathbf{z}_k = (1 - \sigma_k)\,\mathbf{z} + \sigma_k\,\boldsymbol{\epsilon}$ its noised version at timestep k, with Gaussian noise $\boldsymbol{\epsilon}$. The model predicts the velocity $\boldsymbol{\nu}(\mathbf{z}_k, k; \mathbf{C})$ from $\mathbf{z}_k$, $\mathbf{C}$, the image and the prompt, and the adapter minimizes

$$\mathbb{E}_{\mathbf{z},k,\boldsymbol{\epsilon}}\Big[\lambda_k \, \big\|\boldsymbol{\nu}(\mathbf{z}_k, k; \mathbf{C}) - (\boldsymbol{\epsilon} - \mathbf{z})\big\|_2^2\Big], \tag{4}$$

where λ_k weights the timesteps.

Inference. At inference, the scene built from the input image and directed by the user is rendered to its control video $\mathbf{C}$, and the adapter generates the video from $\mathbf{C}$, the image and the prompt. Since the adapter sees the scene only as a depth video, a directed scene is treated exactly like a recovered one: new trajectories and inserted objects require no additional training (Fig. 1, Appendix I).

4 The RealCOD-Rigid Dataset

Training the Motion Adapter requires clips paired with their rigid 3D scenes, which no dataset provides and which cannot be annotated by hand at scale. We therefore recover the rigid 3D scene

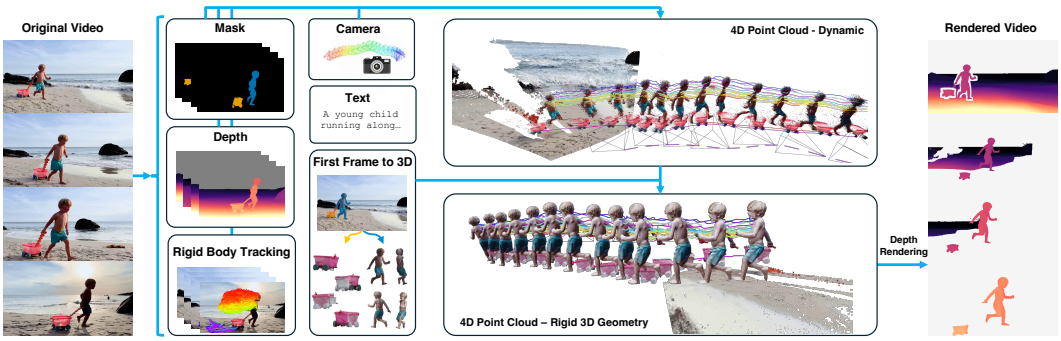


Figure 4: **RealCOD-Rigid annotation pipeline.** From each clip we estimate object masks, depth and the camera trajectory, which together give a dynamic 4D point cloud of the scene (*4D Point Cloud – Dynamic*); we reconstruct a canonical mesh for every masked object from the first frame (*First Frame to 3D*) and track 3D points on it as a rigid body (*Rigid Body Tracking*), which yields one rigid transformation per frame and aligns the canonical mesh to every frame (*4D Point Cloud – Rigid 3D Geometry*), giving the rigid 3D scene. Rendering it to a depth video yields the control; the original clip is the training target, with its text description as the prompt.

of each clip automatically. Starting from the clips of RealCOD-25K (Zhang et al., 2026), each with a text prompt and SAM3 masks for one or two objects, the pipeline lifts the first frame into 3D as in Sec. 3.1 and then tracks the camera and every object through the clip; the 20,774 clips that pass every stage and are not used for evaluation form RealCOD-Rigid. It has four stages (Fig. 4).

(1) Camera and depth estimation. MegaSaM (Li et al., 2025f), with MoGe-2 (Wang et al., 2025e) as depth prior and UniDepthV2 (Piccinelli et al., 2025) as metric branch, estimates per-frame depth, the intrinsic matrix $\mathbf{K}$ and the camera trajectory $\{\mathbf{E}^t\}$; its coordinate frame, of arbitrary scale, is the scene frame of the clip. The background point cloud $\mathcal{P}$ is the first frame back-projected into this frame with the object masks removed.

(2) First frame to 3D. Pixal3D (Li et al., 2026) reconstructs the canonical mesh $\mathcal{M}_o$ from the object's crop in the first frame. Pixal3D is pixel-aligned, so every visible mesh point comes with the first-frame pixel it was reconstructed from, and the depth of stage (1) places that pixel in the scene. A robust similarity fit (rotation, translation and scale) on these pairs moves the mesh onto the object at $t = 1$, which defines $\mathbf{T}_o^1$ as the identity.

(3) Rigid body tracking. TAPIP3D (Zhang et al., 2025a) tracks 3D points on each object through the clip in the scene frame, seeded inside the object's mask; these tracks are the observations $\mathbf{X}_i^t$ of Eq. (3). Every frame needs the rigid transformation that moves the object from its first-frame placement to where the tracks observe it. Not every track moves rigidly: a limb swings and some tracks drift, and a plain fit would follow them. We therefore find the object's *stable core* Ω_o, the tracks that move together rigidly across the clip, by alternating a robust fit with the selection of its inliers, and fit each frame on the core alone,

$$(\mathbf{R}_o^t, \mathbf{t}_o^t) = \arg\min_{\mathbf{R},\,\mathbf{t}} \sum_{i \in \Omega_o} \rho\big(\|\mathbf{R}\,\mathbf{Y}_i + \mathbf{t} - \mathbf{X}_i^t\|\big), \tag{5}$$

with $\mathbf{Y}_i$ the position of track i at $t = 1$ and a robust loss ρ.

(4) Depth rendering. Placing each mesh by its rigid transformation completes the rigid 3D scene of the clip. Where the per-frame depth of stage (1) still shows a limb swinging, this scene moves each object as one rigid body. Rendering it gives the control video (Sec. 3.2).

5 Experiments

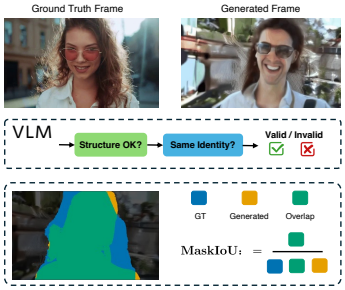
5.1 Setup

Implementation details. We initialize the Motion Adapter (3.0 B parameters) from the released VACE branch, perturb its attention and feed-forward matrices once, and train all of it on the 20,774

Table 1: **Quantitative comparison on joint camera and object control.** (a) Visual quality and text alignment (FID, FVD, CLIP-SIM), camera control (RotErr, TransErr) and object control (recognition rate, IG-IoU). Best **bold**, second best underlined. (b) Identity-Gated IoU on one frame: a vision–language judge decides whether the generated object is still the same object; a valid frame contributes its mask IoU and an invalid frame zero (Eq. 6).

Method	Visual quality / text			Camera control		Object control	
	FID ↓	FVD ↓	CLIP ↑	Rot. (°) ↓	Trans. ↓	Rec. ↑	IG-IoU ↑
MotionCtrl	115.6	1653.3	27.5	4.74	0.231	32.6	18.8
PaC	55.7	1454.6	30.2	7.00	0.228	80.5	40.8
VerseCrafter	55.0	484.7	31.7	6.16	0.174	94.8	54.8
SymphoMotion	45.8	405.4	30.9	3.89	0.132	90.2	52.9
4Director	**44.1**	**370.4**	**31.8**	**3.65**	**0.122**	**94.9**	**60.4**

(a) Quantitative comparison



(b) Identity-Gated IoU

Table 2: **VBench-I2V and user study.** VBench-I2V scores on the eight applicable dimensions (percent) and mean user ratings of visual quality, control accuracy and consistency (1 to 5, Sec. 5.3). 4Director leads on every dimension; VBench margins over the next best stay within 2 points, whereas the user study separates them clearly. Best **bold**, second best underlined.

Method	VBench-I2V								User study (1–5)		
	Subject consistency	Background consistency	Motion smoothness	Dynamic degree	Aesthetic quality	Imaging quality	I2V subject	I2V background	Quality ↑	Control ↑	Consistency ↑
MotionCtrl	70.8	84.3	92.9	94.0	43.8	45.7	81.6	84.2	1.49	1.44	1.79
PaC	90.0	91.6	95.1	93.0	55.0	61.6	92.4	92.5	1.98	2.01	2.41
VerseCrafter	88.9	91.9	98.7	91.0	56.7	64.2	96.0	96.8	3.69	2.39	3.88
SymphoMotion	89.5	91.4	98.5	91.0	54.8	59.1	95.4	96.7	3.35	2.33	3.69
4Director	**90.3**	**92.0**	**98.8**	**96.0**	**57.0**	**64.6**	**96.3**	**97.0**	**4.62**	**4.82**	**4.69**

pairs of RealCOD-Rigid at 832×480 and 81 frames for three epochs on 24 GPUs with a global batch of 24, using AdamW with a peak learning rate of 5×10^{-5} and 25 warmup steps. At inference we use 20 sampling steps and a classifier-free guidance scale of 5.0 for every case.

Baselines. We compare against four public methods for joint camera and object control: MotionCtrl (Wang et al., 2024d), Perception-as-Control (PaC) (Chen et al., 2025b), SymphoMotion (Zhang et al., 2026) and VerseCrafter (Zheng et al., 2026), each with its released weights, resolution and clip length. To compare fairly, we give every method the same motion to follow: the camera trajectory and object motion recovered from each evaluation clip are converted into whatever control each baseline expects, from 2D trajectories to 3D Gaussians.

Metrics. All methods are evaluated on the same 100 clips of RealCOD-25K that are excluded from training, with joint camera and object motion, using four groups of metrics. (1) *Visual quality*: FID (Heusel et al., 2017), FVD (Unterthiner et al., 2018) and the eight dimensions of VBench-I2V (Huang et al., 2024) that apply to our videos. (2) *Text alignment*: CLIP-SIM (Radford et al., 2021). (3) *Camera control*: following CameraCtrl (He et al., 2024), the rotation error (RotErr) and relative translation error (TransErr) between the target trajectory and the one re-estimated from the generated video with the pipeline of Sec. 4. (4) *Object control*: Identity-Gated IoU, defined next.

5.2 Identity-Gated IoU

Mask IoU against the source clip's object masks measures whether the generated object is placed where the trajectory prescribes, but it also rewards an object that is correctly placed and no longer the same object. Identity-Gated IoU (IG-IoU) credits placement only where identity is preserved. Each video is scored on the same $N = 16$ frames, spaced uniformly over the frames in which the object is visible in the source clip. Let $\mathrm{MaskIoU}_f$ denote the intersection over union of the generated and the reference mask in frame f. The identity gate $v_f \in \{0, 1\}$ is set by a vision–language judge that

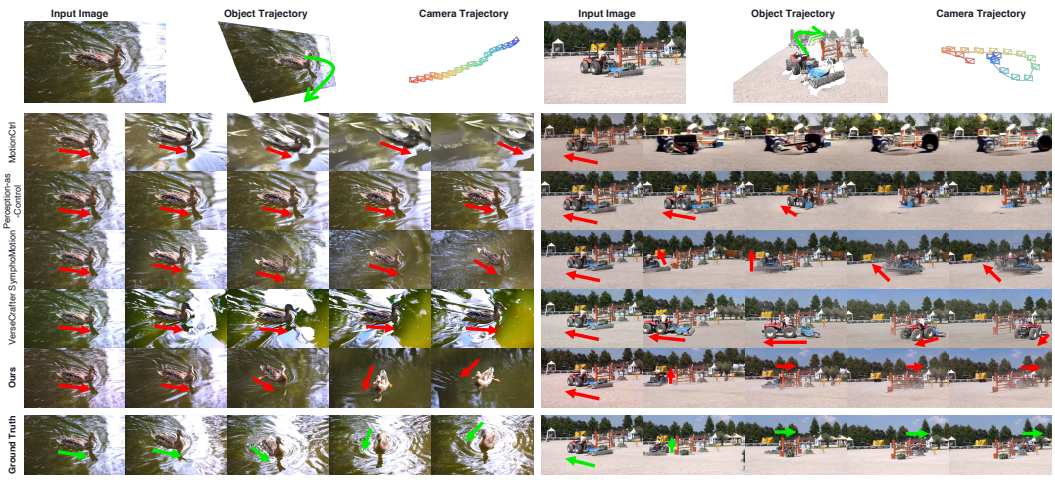


Figure 5: **Qualitative comparison on joint camera and object control.** For each clip, the top row shows the input image and the prescribed trajectories, and the rows below show temporally aligned frames from the four baselines, 4Director and the source video. Arrows mark the object heading, green in the source and red in the generated videos. Only 4Director completes the turn of the source video; the baselines keep the original heading, turn only partway, blur or lose the subject.

compares the generated crop with the reference crop: $v_f = 1$ if the object is intact and still the same object, and $v_f = 0$ otherwise. We report

$$\text{Recognition rate} = \frac{1}{N}\sum_{f=1}^{N} v_f, \qquad \text{IG-IoU} = \frac{1}{N}\sum_{f=1}^{N} v_f \,\text{MaskIoU}_f, \tag{6}$$

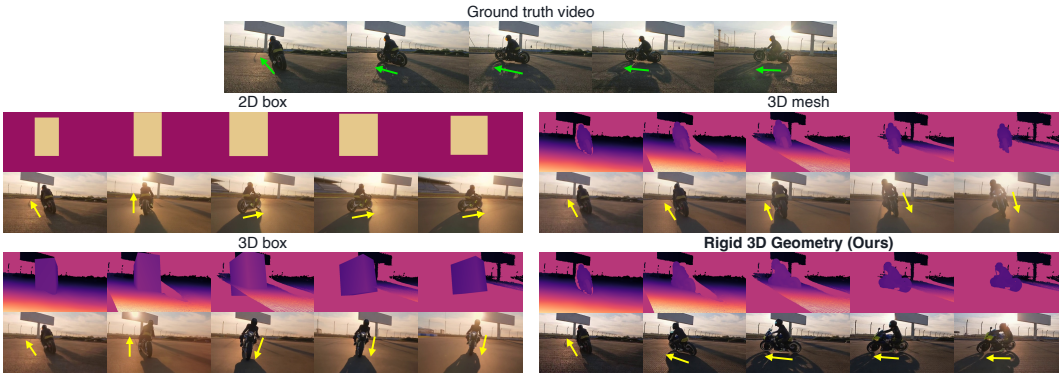
averaged over the evaluation clips: a frame that fails the gate contributes zero but stays in the denominator. Masks come from SAM3 (Carion et al., 2025) and the judge is Qwen3-VL (Bai et al., 2025b) (Table 1b; protocol in the appendix). Because the reference mask is the object's silhouette at its true pose, MaskIoU also penalizes an orientation error that changes the silhouette; a 180° turn of a front–back symmetric object is not penalized and is shown qualitatively (Figs. 2 and 15).

5.3 Joint Camera and Object Control

Quantitative comparison. As shown in Table 1a, 4Director achieves the best score on every metric. (1) *Visual quality and text alignment*: it lowers FID by 3.7% and FVD by 8.6% relative to SymphoMotion, the next best on both, and matches VerseCrafter on CLIP-SIM. (2) *Camera control*: it reduces the rotation error from 3.89° for SymphoMotion, the next best, to 3.65° and the translation error by 7.6%, with the other three baselines clearly behind. (3) *Object control*, where the margins are widest: IG-IoU rises from 54.8 for VerseCrafter, the next best, to 60.4, a 10.2% relative gain at the same recognition rate (94.9 against 94.8). 4Director thus places the object more accurately rather than keeping it recognizable more often, which we attribute to the complete mesh fixing its extent and orientation, as a Gaussian blob or tracked points do not (Fig. 2); MotionCtrl, driven by 2D trajectories, keeps the object recognizable in only 32.6% of frames, so its IG-IoU drops to 18.8.

Qualitative comparison. Figure 5 shows two further clips in which the object turns while the camera moves; the arrows mark its heading. Only 4Director turns the object as far as the source video does. On the duck, Perception-as-Control and VerseCrafter keep the original heading, and MotionCtrl and SymphoMotion turn it only slightly; on the tractor, SymphoMotion and VerseCrafter turn it only halfway, and Perception-as-Control and MotionCtrl lose and blur it, respectively. The tractor also passes behind the fence in the source video; 4Director reproduces this occlusion, whereas SymphoMotion and VerseCrafter keep the tractor in front (see also Appendix F).

User study. To complement the automatic metrics, 20 participants rated the five methods on nine cases from 1 to 5 on *visual quality*, *control accuracy* and *consistency* (protocol in Appendix H.2).



Control	Control encodes: Depth	Rotation	Complete geometry	Visual quality / text: FID ↓	FVD ↓	CLIP ↑	Camera control: Rot. (°) ↓	Trans. ↓	Object control: Rec. ↑	IG-IoU ↑
2D box				58.3	608.8	31.2	9.61	0.233	**95.0**	48.5
3D box	✓	✓		54.7	556.0	31.1	5.00	<u>0.142</u>	93.3	50.3
3D mesh	✓		✓	<u>53.0</u>	<u>485.6</u>	<u>31.2</u>	<u>4.84</u>	0.147	91.9	<u>53.0</u>
Rigid 3D geometry (ours)	✓	✓	✓	**44.1**	**370.4**	**31.8**	**3.65**	**0.122**	<u>94.9</u>	**60.4**

Figure 6: **Ablation on the control representation.** *Top:* the source video and three alternative controls, each shown as the control video above the video generated from it; arrows mark the motorcycle heading, and only the rigid 3D geometry follows the prescribed turn. *Bottom:* what each control encodes and its scores; the three arms are trained with the recipe of 4Director, and the last row is 4Director as in Table 1a. Best **bold**, second best <u>underlined</u>.

As Table 2 shows, 4Director receives the highest rating on all three, 4.6 to 4.8 against at most 3.9 for the next best, with the largest margin on control accuracy (4.82 against 2.39).

5.4 Ablation Study

Figure 6 replaces the rigid 3D geometry by three alternative controls. The 3D box and the 3D mesh each drop one of its two ingredients: the 3D box keeps depth and orientation but has no surface, and the 3D mesh keeps the complete surface but does not rotate. Dropping either lowers IG-IoU from 60.4 to 50.3 without the surface and to 53.0 without rotation, while the recognition rate stays between 91.9 and 95.0 for every control; the loss is thus mainly in placement, not identity. In the top panel, only the rigid 3D geometry follows the prescribed turn: with the mesh that does not rotate, the motorcycle ends up facing the camera instead of turning sideways, an orientation error that also changes its silhouette. The 2D box, which encodes neither depth nor orientation, has the lowest IG-IoU (48.5), and the four controls rank in the same order on FID, FVD and IG-IoU.

6 Conclusion and Future Work

We present 4Director, a video world model controlled with rigid 3D geometry: an explicit 4D scene representation in which every object is a complete canonical mesh moved by one rigid transformation per frame, sharing one coordinate frame with the background and the camera. The Motion Adapter injects a depth rendering of this scene, which carries the camera and the rigid motion of every object, together with the input image into a pretrained video generator that supplies appearance, illumination and non-rigid dynamics. To train the adapter, we automatically annotated 20,774 clips of RealCOD-25K with rigid 3D scenes (RealCOD-Rigid); to evaluate object control, we introduced IG-IoU. Experiments, ablations and a user study show higher visual quality and more accurate camera and object control than four state-of-the-art baselines. The representation, however, controls motion only at the level of a rigid body: an articulated or deforming object moves as one whole, so finer motion such as running, jumping or the movement of the limbs cannot be prescribed and is decided by the generator, which may even keep a mostly articulated object in its first-frame pose (Appendix G). Extending the rigid 3D scene with articulated parts, so that such motion can be prescribed, is a natural next step.

AI USE STATEMENT

In this work, we used generative AI tools to draft sections of this paper and to aid and polish its writing. We have not used generative AI tools for retrieval and discovery, for research ideation or execution, for generating synthetic datasets, or for proving mathematical claims: the representation, the annotation pipeline, the training runs and the evaluation protocol were designed and executed by the authors, and every reported number was checked by the authors against the recorded experimental artifacts. Pretrained models used as components of the method and of the evaluation, such as the vision–language judge of Sec. 5.2, are described in the paper. We have reviewed all AI-assisted text for accuracy and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, et al. Cosmos-transfer1: Conditional world generation with adaptive multimodal control. *arXiv preprint arXiv:2503.14492*, 2025.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. In *Advances in Neural Information Processing Systems*, 2024.

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. *arXiv preprint arXiv:2407.12781*, 2024.

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. AC3D: Analyzing and improving 3D camera control in video diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22875–22889, 2025.

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*, 2025a.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

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.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Proceedings of the International Conference on Machine Learning*, 2024.

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13–23, 2025.

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 *ACM SIGGRAPH Asia 2025 Conference Papers*, 2025a.

Wei Cao, Chang Luo, Biao Zhang, Matthias Nießner, and Jiapeng Tang. Motion2VecSets: 4D latent vector set diffusion for non-rigid shape reconstruction and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20496–20506, 2024.

Wei Cao, Marcel Hallgarten, Tianyu Li, Daniel Dauner, Xunjiang Gu, Caojun Wang, Yakov Miron, Marco Aiello, Hongyang Li, Igor Gilitschenski, et al. Pseudo-simulation for autonomous driving. *arXiv preprint arXiv:2506.04218*, 2025b.

Wei Cao, Hao Zhang, Fengrui Tian, Yulun Wu, Yingying Li, Shenlong Wang, Ning Yu, and Yaoyao Liu. FreeOrbit4D: Training-free arbitrary camera redirection for monocular videos via foreground-complete 4D reconstruction. In *ACM SIGGRAPH Conference Papers*, pp. 1–12, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. In *International Conference on Learning Representations*, 2025.

Junyi Chen, Haoyi Zhu, Xianglong He, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Zhoujie Fu, Jiangmiao Pang, et al. Deepverse: 4d autoregressive video generation as a world model. *arXiv preprint arXiv:2506.01103*, 2025a.

Kaijin Chen, Dingkang Liang, Xin Zhou, Yikang Ding, Xiaoqiang Liu, Pengfei Wan, and Xiang Bai. Out of sight but not out of mind: Hybrid memory for dynamic video world models. *arXiv preprint arXiv:2603.25716*, 2026.

Yingjie Chen, Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. Perception-as-control: Fine-grained controllable image animation with 3d-aware motion representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b.

Ruihang Chu, Yefei He, Zhekai Chen, Shiwei Zhang, Xiaogang Xu, Bin Xia, Dingdong Wang, Hongwei Yi, Xihui Liu, Hengshuang Zhao, et al. Wan-move: Motion-controllable video generation via latent trajectory guidance. *arXiv preprint arXiv:2512.08765*, 2025.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. UniMax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *International Conference on Learning Representations*, 2023.

Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. Project page, 2024. URL https://oasis-model.github.io/.

Zicheng Duan, Jiatong Xia, Zeyu Zhang, Wenbo Zhang, Gengze Zhou, Chenhui Gou, Yefei He, Feng Chen, Xinyu Zhang, and Lingqiao Liu. LiveWorld: Simulating out-of-sight dynamics in generative video world models. *arXiv preprint arXiv:2603.07145*, 2026.

Wanquan Feng, Jiawei Liu, Pengqi Tu, Tianhao Qi, Mingzhen Sun, Tianxiang Ma, Songtao Zhao, Siyu Zhou, and Qian He. I2vcontrol-camera: Precise video camera control with adjustable motion strength. *arXiv preprint arXiv:2411.06525*, 2024a.

Wanquan Feng, Tianhao Qi, Jiawei Liu, Mingzhen Sun, Pengqi Tu, Tianxiang Ma, Fei Dai, Songtao Zhao, Siyu Zhou, and Qian He. I2vcontrol: Disentangled and unified video motion synthesis control. *arXiv preprint arXiv:2411.17765*, 2024b.

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. In *International Conference on Learning Representations*, 2025.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

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.

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. *arXiv preprint arXiv:2503.10592*, 2025a.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025b.

Roberto Henschel, Levon Khachatryan, Hayk Poghosyan, Daniil Hayrapetyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. StreamingT2V: Consistent, dynamic, and extendable long video generation from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

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 *Advances in Neural Information Processing Systems*, 2017.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, et al. Relic: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

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.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson WH Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *arXiv preprint arXiv:2506.04225*, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8079–8088, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

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. *Advances in Neural Information Processing Systems*, 37:16240–16271, 2024.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview, 2022. Version 0.9.2.

Guojun Lei, Chi Wang, Hong Li, Rong Zhang, Yikai Wang, and Weiwei Xu. Animateanything: Consistent and controllable animation for video generation. *arXiv preprint arXiv:2411.10836*, 2024.

Dong-Yang Li, Wang Zhao, Yuxin Chen, Wenbo Hu, Meng-Hao Guo, Fang-Lue Zhang, Ying Shan, and Shi-Min Hu. Pixal3D: Pixel-aligned 3D generation from images. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2026.

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition. *arXiv preprint arXiv:2506.17201*, 2025a.

Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. VMem: Consistent interactive video scene generation with surfel-indexed view memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025c.

Teng Li, Guangcong Zheng, Rui Jiang, Tao Wu, Yehao Lu, Yining Lin, Xi Li, et al. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. *arXiv preprint arXiv:2502.10059*, 2025d.

Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Ying Shan, and Yuexian Zou. Image conductor: Precision control for interactive video synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025e.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025f.

Jingyun Liang, Jingkai Zhou, Shikai Li, Chenjie Cao, Lei Sun, Yichen Qian, Weihua Chen, and Fan Wang. Realismotion: Decomposed human motion control and video generation in the world space. *arXiv preprint arXiv:2508.08588*, 2025.

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 *ACM SIGGRAPH Asia 2025 Conference Papers*, 2025.

Wan-Duo Kurt Ma, John P Lewis, and W Bastiaan Kleijn. Trailblazer: Trajectory control for diffusion-based video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025.

Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. In *European Conference on Computer Vision*, pp. 111–128. Springer, 2024.

Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model. *Google DeepMind technical report*, 2024. URL https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/.

Jack Parker-Holder, Shlomi Fruchter, et al. Genie 3: A new frontier for world models. *Google DeepMind Blog*, 4, 2025.

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.

Ryan Po, Yotam Nitzan, Richard Zhang, Berlin Chen, Tri Dao, Eli Shechtman, Gordon Wetzstein, and Xun Huang. Long-context state-space video world models. *arXiv preprint arXiv:2505.20171*, 2025.

Stefan Popov, Amit Raj, Michael Krainin, Yuanzhen Li, William T Freeman, and Michael Rubinstein. Camctrl3d: Single-image scene exploration with precise 3d camera control. *arXiv preprint arXiv:2501.06006*, 2025.

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. FreeNoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023.

Haonan Qiu, Zhaoxi Chen, Zhouxia Wang, Yingqing He, Menghan Xia, and Ziwei Liu. Freetraj: Tuning-free trajectory control in video diffusion models. *arXiv preprint arXiv:2406.16863*, 2024.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

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. In *International Conference on Learning Representations*, 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6121–6132, 2025.

Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Xincheng Shuai, Henghui Ding, Zhenyuan Qin, Hao Luo, Xingjun Ma, and Dacheng Tao. Free-form motion control: Controlling the 6D poses of camera and objects in video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 2021.

Chenxi Song, Yanming Yang, Tong Zhao, Ruibo Li, and Chi Zhang. Taming video models for 3d and 4d generation via zero-shot camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40352–40363, 2026.

Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

Jiapeng Tang, Wei Cao, Biao Zhang, Chang Luo, Yaoyao Liu, and Matthias Nießner. Motion2VecSets: Non-rigid shape reconstruction and tracking with 4D latent set diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

HunyuanWorld Team, Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li, Sheng Zhang, et al. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint arXiv:2507.21809*, 2025.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

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.

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. In *European Conference on Computer Vision*, pp. 439–457. Springer, 2024.

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

Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. Ati: Any trajectory instruction for controllable video generation. *arXiv preprint arXiv:2505.22944*, 2025b.

Hanlin Wang, Hao Ouyang, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Qifeng Chen, Yujun Shen, and Limin Wang. LeviTor: 3D trajectory oriented image-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025c.

Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21686–21697. IEEE, 2024a.

Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*, 2024b.

Qinghe Wang, Yawen Luo, Xiaoyu Shi, Xu Jia, Huchuan Lu, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. CineMaster: A 3D-aware and controllable framework for cinematic text-to-video generation. In *ACM SIGGRAPH 2025 Conference Papers*, pp. 1–10, 2025d.

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*, 2025e.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36:7594–7611, 2023.

Zhouxia Wang, Yushi Lan, Shangchen Zhou, and Chen Change Loy. ObjCtrl-2.5D: Training-free object control with camera poses. *arXiv preprint arXiv:2412.07721*, 2024c.

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*, 2024d.

Zun Wang, Jaemin Cho, Jialu Li, Han Lin, Jaehong Yoon, Yue Zhang, and Mohit Bansal. Epic: Efficient video camera control learning with precise anchor-video guidance. *arXiv preprint arXiv:2505.21876*, 2025f.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284*, 2025.

Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, pp. 331–348. Springer, 2024.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WORLDMEM: Long-term consistent world simulation with memory. In *Advances in Neural Information Processing Systems*, 2025.

Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. In *International Conference on Learning Representations*, pp. 33421–33441, 2025.

Jinbo Xing, Long Mai, Cusuh Ham, Jiahui Huang, Aniruddha Mahapatra, Chi-Wing Fu, Tien-Tsin Wong, and Feng Liu. Motioncanvas: Cinematic shot design with controllable image-to-video generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025.

Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. CamCo: Camera-controllable 3D-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024.

Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

Zhongqi Yang, Wenhang Ge, Yuqi Li, Jiaqi Chen, Haoyuan Li, Mengyin An, Fei Kang, Hua Xue, Baixin Xu, Yuyang Yin, et al. Matrix-3d: Omnidirectional explorable 3d world generation. *arXiv preprint arXiv:2508.08086*, 2025.

Chun-Han Yao, Yiming Xie, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d 2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4d generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13248–13258. IEEE, 2025.

Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. DragNUWA: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. *arXiv preprint arXiv:2506.03141*, 2025a.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. *arXiv preprint arXiv:2501.08325*, 2025b.

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, 2025c.

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.

Bowei Zhang, Lei Ke, Adam Harley, and Katerina Fragkiadaki. Tapip3d: Tracking any point in persistent 3D geometry. *Advances in Neural Information Processing Systems*, 38:135284–135303, 2025a.

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. *arXiv preprint arXiv:2411.05003*, 2024.

Guiyu Zhang, Yabo Chen, Xunzhi Xiang, Junchao Huang, Zhongyu Wang, and Li Jiang. Symphomotion: Joint control of camera motion and object dynamics for coherent video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11127–11137, 2026.

Hao Zhang, Chun-Han Yao, Simon Donné, Narendra Ahuja, and Varun Jampani. Stable part diffusion 4d: Multi-view rgb and kinematic parts video generation. *Advances in Neural Information Processing Systems*, 38:114172–114195, 2025b.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025c.

Zhiyuan Zhang, Dongdong Chen, and Jing Liao. I2v3d: Controllable image-to-video generation with 3d guidance. *arXiv preprint arXiv:2503.09733*, 2025d.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

Sixiao Zheng, Zimian Peng, Yanpeng Zhou, Yi Zhu, Hang Xu, Xiangru Huang, and Yanwei Fu. Vidcraft3: Camera, object, and lighting control for image-to-video generation. *arXiv preprint arXiv:2502.07531*, 2025.

Sixiao Zheng, Minghao Yin, Wenbo Hu, Xiaoyu Li, Ying Shan, and Yanwei Fu. VerseCrafter: Dynamic realistic video world model with 4D geometric control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40277–40290, 2026.

Haitao Zhou, Chuang Wang, Rui Nie, Jinlin Liu, Dongdong Yu, Qian Yu, and Changhu Wang. TrackGo: A flexible and efficient method for controllable video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025a.

Jensen 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. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12405–12414. IEEE, 2025b.

Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8535–8546, 2025.

A Implementation Details

A.1 Rigid 3D Scene Construction

Camera and depth. The scene coordinate frame is produced by MegaSaM (Li et al., 2025f) camera tracking, bundle adjustment and consistent-video-depth refinement, with MoGe-2 (Wang et al., 2025e) as the monocular depth prior and UniDepthV2 (Piccinelli et al., 2025) as the metric-depth

branch. MegaSaM returns one intrinsic matrix per clip and one world-to-camera matrix per frame. Metric cues enter only through MegaSaM's per-frame alignment of the monocular disparity, and no later stage rescales cameras, depth or tracks; the result is a shared coordinate system, not a metric reconstruction.

Object tracks. TAPIP3D (Zhang et al., 2025a) is run with its released checkpoint (16-frame window, stride 8) in a forward and a reversed pass so that tracks from later anchors extend back to the first frame. Per object, eight anchor frames are spread over the clip and 128 farthest-point queries are drawn inside the mask at each anchor, giving 1,024 queries per object. An observation is valid if the tracker's visibility score is at least 0.9, the point has positive camera-axis depth, and it projects inside the object's mask dilated by 2 px; a track is retained if at least 20% of its frames are valid.

Canonical mesh and alignment. Pixal3D (Li et al., 2026) receives the first frame cropped to a square of side $1.1\times$ the longer side of the mask's bounding box, with the field of view derived from the MegaSaM intrinsics, and returns a textured mesh (100k faces). Rendering the mesh in its reconstruction camera pairs visible mesh points with first-frame pixels; one correspondence is kept per pixel inside the mask, the mask is eroded by 1 px, and 20% of the 8×8 pixel cells are withheld from the fit. The similarity transform is a closed-form Umeyama fit with a median $+\,3\,\mathrm{MAD}$ inlier threshold. Residuals are normalized by the object diameter d, the extent of the correspondence cloud. An object passes the 3D gate if the residual on the withheld correspondences has median ≤ 0.15, p90 ≤ 0.30 and p95 ≤ 0.50, and the silhouette gate if the placed mesh rasterized into the first-frame camera has IoU ≥ 0.5 with the mask. A clip is complete only if every object passes both gates; otherwise it is excluded from the training set.

Stable core. Per frame, Eq. (5) is solved with the Tukey loss as ρ and with track weights that combine observation support, a body-center prior and a spatial-balancing factor, so that a densely tracked region cannot dominate the pose. Per frame, a seeded RANSAC (threshold $0.035\,d$) followed by Tukey-reweighted iterations (Tukey constant $0.05\,d$) updates the rotation by weighted Kabsch and the translation by a robust geometric median. A track is stable if its weighted inlier fraction at $0.05\,d$ is at least 0.70 over frames with a valid fit and it is observed in at least 30% of them; the core is re-fit until the stable set repeats and must contain at least 30 tracks. If the core is too small or degenerate, tracks are clustered by the invariance of their pairwise distances over time and the most rigid cluster seeds the core.

A.2 Rendering and Encoding

The background point cloud is the first-frame depth back-projected after removing the union of all object masks dilated by 9 px. The background is rasterized as a splatted point cloud and each placed mesh with one face per pixel, both with the same intrinsics and camera-axis depth Z. The front-most entity at each pixel defines the depth Z_t, and pixels covered by no entity are invalid. The 1st and 99th percentiles of $1/Z_t$ are computed per entity over all frames, and $[\alpha, \beta]$ is their envelope over entities, so that the range of a small near object is never clipped by the background. Because the range is fixed per clip, a given depth maps to the same gray value in every frame. A valid pixel is encoded as

$$\mathbf{C}_t = \Big\lfloor 255\ \mathrm{clip}\Big(\frac{1/Z_t - \alpha}{\beta - \alpha},\, 0,\, 1\Big)\Big\rceil, \tag{7}$$

and stored as 8-bit gray (near $= 255$, far $= 0$); invalid pixels receive the constant 128.

A.3 Motion Adapter

Context stream. The adapter follows the conditioning branch of Wan2.1-VACE-14B (Wang et al., 2025a; Jiang et al., 2025) in architecture and is initialized from its released weights. The VAE encoding of the control video is paired with an all-ones mask that marks every pixel for synthesis, and the encoding of the input image is prepended along time; the eight context blocks share the architecture of the Wan-DiT blocks.

Training. Table 3 lists the hyper-parameters of the two variants. Both train on the 20,774 clips of RealCOD-Rigid at $81 \times 832 \times 480$ in bf16 with a learning rate that warms up linearly and then

stays constant. The timestep is sampled uniformly over the 1,000-step Wan flow schedule (shift 5), and the timestep weight λ_k of Eq. (4) is the scheduler's default. Before training the full adapter, its attention and feed-forward matrices are perturbed once with uniform noise of magnitude 3×10^{-4} times each tensor's standard deviation, in the manner of NoisyTune; the LoRA variant starts from the unperturbed released weights.

Table 3: **Training hyper-parameters of the two variants.**

	4Director (full)	4Director-LoRA
Trainable parameters	entire adapter, 3.049 B	rank-16 LoRA on the context blocks' attention and feed-forward projections, $\approx$15.3 M
Initialization	official VACE + NoisyTune-style noise	official VACE
GPUs / global batch	3×8 / 24	3×8 / 24
Epochs	3	3
Peak learning rate	5×10^{-5}	5×10^{-5}
Warmup steps	25	25
Weight decay	0.01	0.01
Seed	42	42

Inference. Both variants use 20 flow-matching sampling steps, classifier-free guidance 5.0 with the pipeline's default negative prompt, sigma shift 5.0, seed 42, 832×480, 81 frames at 16 fps, tiled VAE decoding and hint scale 1.0. The same checkpoints and settings were used for every result in this paper.

B Dataset Details

The annotation pipeline takes the 25,318 source clips of RealCOD-25K (Zhang et al., 2026), monocular clips of 81 frames at 832×480 and 16 fps, each with one text prompt and SAM3 instance masks for one or two objects (29,338 objects in total). The clips are stock and open-source web videos; we reuse the videos, text prompts and instance masks, and derive every rigid 3D scene, trajectory and control video of RealCOD-Rigid with our pipeline. An object is allowed to vanish in some frames so that genuine occlusion is represented rather than suppressed. Clips that fail a stage, because of too few correspondences or missing outputs, and clips that belong to the evaluation set of Sec. 5.1 are excluded, leaving the 20,774 training clips of RealCOD-Rigid.

C Evaluation Details

C.1 Baselines

All four baselines run from their released checkpoints with their native sampling settings, resolution and clip length; MotionCtrl and Perception-as-Control generate 16 frames, which are mapped to the evaluation frames through each converter's frame table. Each converter reads the shared camera, depth and track estimates of Appendix A.1 and produces the baseline's native control at the capacity its implementation accepts. *MotionCtrl* receives the first frame through SparseCtrl, the shared camera through its camera module and up to eight track flows through its object module. *Perception-as-Control* receives up to 16 tracked 3D points rendered as spheres plus its camera tensor, and does not consume the text prompt. *SymphoMotion* receives the camera, per-object point sets built from the first-frame mask points and moved along the object trajectory, the 2D boxes it projects from them, a reference point-cloud render and the prompt. Since point-based controls natively take per-point trajectories, which are cumbersome to author one by one, moving the points rigidly along the object trajectory gives SymphoMotion the same motion that 4Director receives. *VerseCrafter* receives background RGB and depth renders, a per-object 3D Gaussian fitted at the first frame and moved along the object trajectory, and a merged mask.

C.2 METRICS

Visual quality. FID (Heusel et al., 2017) uses InceptionV3 pool3 features on 16 uniformly spaced frames per video, with the source clips of the same cases as the real set. FVD (Unterthiner et al., 2018) is computed with the StyleGAN-V (I3D) feature extractor over every native frame of each method. CLIP-SIM is the mean frame–text cosine similarity of CLIP ViT-B/32 (Radford et al., 2021) over 16 frames. VBench-I2V (Huang et al., 2024) is run on eight dimensions (subject consistency, background consistency, motion smoothness, dynamic degree, aesthetic quality, imaging quality, I2V subject, I2V background); its camera-motion dimension requires one of seven fixed camera instructions and does not apply to continuous trajectories.

Camera control. The camera trajectory of each generated video is estimated with the same MegaSaM pipeline that produced the target trajectories and aligned to the target by a $\mathrm{Sim}(3)$ transform, following the evaluation protocol of VGGSfM (Wang et al., 2024a). A case is successful if a trajectory was estimated and the alignment is valid (positive scale, proper rotation). Rotation error is the mean geodesic distance between aligned and target orientations in degrees; relative translation error is the mean camera-center distance divided by the target trajectory's maximum displacement from the first frame. Both errors are averaged over the cases in which a trajectory could be estimated.

Identity-Gated IoU. A frame of the source clip counts as visible if the ground-truth mask covers at least 10^{-4} of it, and the $N = 16$ test frames of Eq. 6 are taken at a uniform spacing over the visible frames of the first object of the case, mapped through each converter's frame table so that all methods are scored on the same moments; when the object is visible in fewer than 16 frames, N is the number of visible frames. Generated masks come from SAM3 prompted on the first generated frame with a point and a box derived from the ground-truth first-frame mask and propagated through the video. For each test frame a Qwen3-VL-2B-Instruct judge compares two RGB crops: the ground-truth frame and the generated frame, cut with the same window, which is the bounding box of the union of the two masks padded by 20% of its longer side and at least 96 pixels wide, and resized to a 448-pixel maximum edge. The judge never sees a mask overlay, so the gate reads appearance rather than segmentation. A frame is valid only for a `PASS` verdict with confidence at least 0.5; every other verdict, a low-confidence pass, and a frame in which the segmenter finds no object are invalid, and a frame without a generated mask is scored without calling the judge. The instruction is:

```
You are comparing the same target object at the same time in
two videos.  Image 1 is the ground-truth video crop.  Image
2 is the generated-video crop.  Both images use exactly the
same spatial crop window.  Judge only the main target object
shown in Image 1.  Ignore background changes, small texture
differences, minor blur, and location error.

Choose exactly one verdict:
- PASS: Image 2 preserves the target object's identity and
has no undeniable major structural defect.
- IDENTITY_FAIL: Image 2 shows a different semantic object
or loses the target object's defining identity.
- STRUCTURE_FAIL: It is the same object, but it has an
undeniable missing or duplicated major part, melted/fused
geometry, broken/floating major component, or impossible
attachment.
- BOTH_FAIL: Both identity and major structure fail.
- UNCERTAIN: visibility or image quality is insufficient to
decide.

Do not penalize a part hidden by the same natural occlusion
or frame boundary in Image 1.  Return one JSON object only,
with exactly three keys:  verdict (one label above),
confidence (a number from 0 to 1), and reason (one short
sentence).  The verdict must agree with the reason.
```

D LoRA Variant

We also train a much smaller adapter to test how much of the gain depends on adapter capacity. 4Director-LoRA keeps the released context blocks fixed and trains rank-16 LoRA updates of their attention and feed-forward matrices (15.3 M parameters) on the same data and schedule as the full adapter (Table 3). Table 4 reports its scores next to the full adapter on the metrics of Tables 1a and 2; it was not included in the user study. 4Director-LoRA exceeds all four baselines on FID, FVD and IG-IoU, while the full adapter leads it on every metric, including rotation error (4.02° to 3.65°) and translation error (0.131 to 0.122).

Table 4: **LoRA variant against the full adapter.** The metrics of Table 1a (left) and the VBench-I2V dimensions of Table 2 (right).

Method	FID ↓	FVD ↓	CLIP ↑	Rot. (°) ↓	Trans. ↓	Rec. ↑	IG-IoU ↑	Subj. cons.	Bkg. cons.	Motion smooth.	Dyn. degree	Aesth. quality	Imag. quality	I2V subj.	I2V bkg.
4Director-LoRA	44.4	393.2	31.4	4.02	0.131	94.1	57.3	90.0	91.5	98.7	92.0	56.8	63.2	96.2	96.9
4Director	**44.1**	**370.4**	**31.8**	**3.65**	**0.122**	**94.9**	**60.4**	**90.3**	**92.0**	**98.8**	**96.0**	**57.0**	**64.6**	**96.3**	**97.0**

E Object Permanence

Object permanence, where an object that leaves the field of view returns as the same object where its motion implies, is pursued with memory over past frames or trackers that advance out-of-view entities (Yu et al., 2025a; Xiao et al., 2025; Chen et al., 2026; Duan et al., 2026). A rigid 3D scene provides its geometric part by construction: the rigid trajectory of an object is defined at every frame whether or not the object projects inside the image, and the canonical mesh never changes, so where and in what shape the object re-enters is rendered rather than remembered; the generator only has to restore its appearance.

Fig. 7 shows a clip in which the object walks out of the frame and returns. Only 4Director follows the control out of view: the frame in which the object has left is empty, the background it occupied is filled in, and the same camel returns at the position the trajectory specifies. The baselines either keep the camel in view throughout, although the controls of SymphoMotion and VerseCrafter leave the frame empty while it is away, or, like Perception-as-Control, lose it and recover a deformed subject.

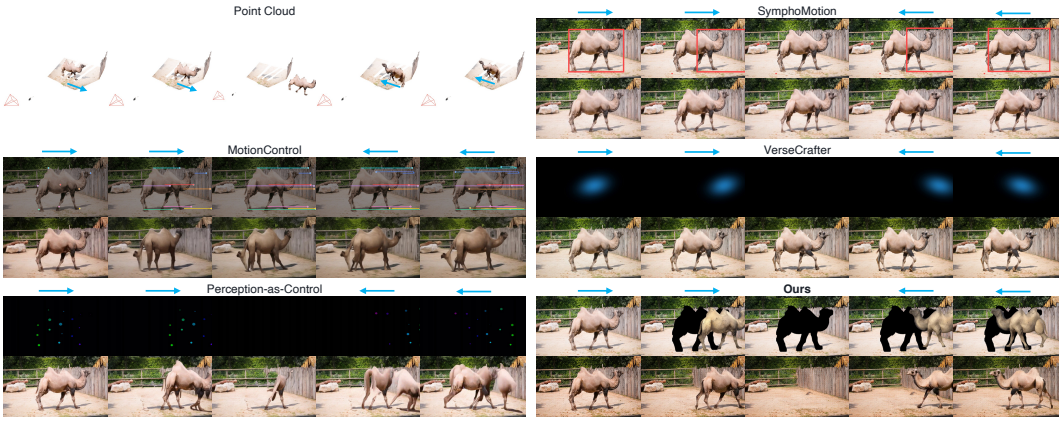


Figure 7: **Object permanence.** *Top left:* the rigid 3D scene at five times; the object leaves the scene and comes back (blue arrow) with the camera fixed. For each method the upper row is the control it receives and the lower row the video generated from it. Only 4Director empties the frame and then recovers the same object.

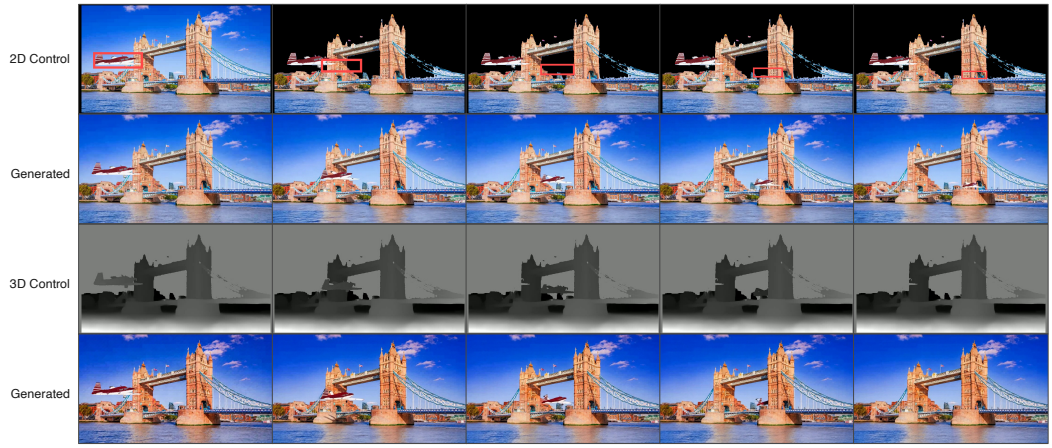


Figure 8: **Occlusion.** A plane is directed to fly under the bridge and behind its right tower. *Rows 1–2:* the control of SymphoMotion, a 2D box projected from the 3D trajectory onto the rendered point cloud (red), and the video it generates. *Rows 3–4:* the depth control rendered from the rigid 3D scene and the video 4Director generates from it. Only with the depth control does the tower hide the plane.

F Occlusion

A 2D control fixes where an object appears in the image, but not whether it passes in front of or behind the rest of the scene. In the depth video of a rigid 3D scene, each pixel takes the depth of the nearest surface (Sec. 3.1), so this order is fixed before generation. Fig. 8 directs a plane to fly under the bridge and on behind its right tower. SymphoMotion marks the plane by a 2D box on the rendered point cloud and conditions on its 3D trajectory, but its control gives no per-pixel visibility, and the plane it generates passes in front of the tower; with the rigid 3D geometry, whose depth video states which surface is nearest at every pixel, 4Director flies it under the deck and hides it behind the tower.

G Failure Case

One rigid transformation per object cannot express motion that is mostly articulated. Fig. 9 moves a breakdancer in a handstand to the right as one rigid body while the camera moves. The depth control fixes the silhouette of the whole body in every frame, and the generated dancer follows it: it keeps the handstand pose of the first frame throughout instead of dancing. Prescribing such motion requires articulated parts in the scene (Sec. 6).



Figure 9: **Failure case.** A breakdancer in a handstand is moved to the right as one rigid body while the camera moves. *Top:* the rigid 3D scene rendered to RGB, for visualization only, and to the depth video that conditions the adapter; pixels without geometry are black in the RGB render and uniform gray in the depth video. *Bottom:* the generated dancer keeps the pose of the first frame throughout, following the rigid silhouette, instead of dancing.

H Interfaces

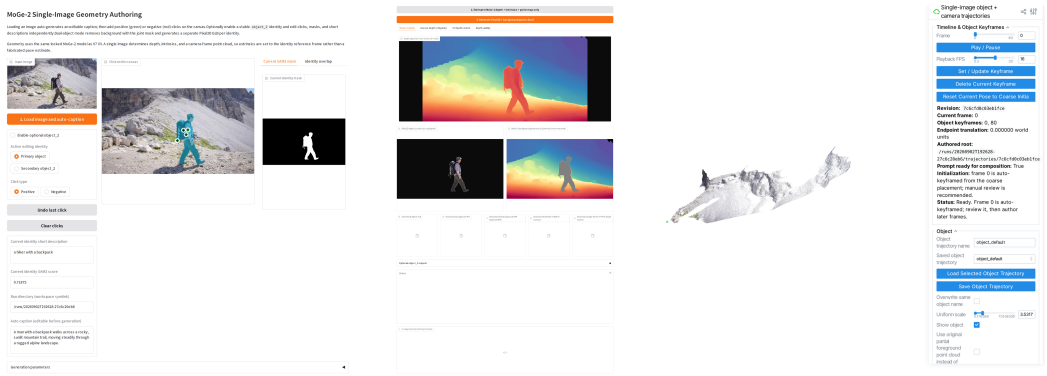
H.1 Authoring Interface

One image is enough to build a rigid 3D scene, whose camera and object trajectories are then authored in a 3D viewer; this is how the cases of Appendix I were produced. The interface has three stages.

(1) Identity. The user loads an image, which is captioned automatically, and marks each object with positive and negative clicks; the segmentation and the resulting identity mask are shown beside the canvas (Fig. 10a).

(2) Geometry. Depth and intrinsics are estimated for the frame, the canonical mesh of each identity is reconstructed, and the background is lifted into a point cloud with the object masks removed. The panel previews the result and exports the scene (Fig. 10b).

(3) Trajectories. The scene opens in a viewer where the camera trajectory and each object's rigid trajectory are set by keyframes on a timeline (Fig. 10c). The authored scene is rendered by the same renderer as a recovered one (Sec. 3.1) and the resulting depth video conditions the adapter.



(a) Identity (b) Geometry (c) Trajectories

Figure 10: **Authoring interface.** The three stages of building and directing a rigid 3D scene from a single image; in (c) the trajectories are set by keyframes on a timeline.

H.2 User Study Interface

The study of Sec. 5.3 is a form of nine cases, each on its own page. A case states the target control in words, shows the input image beside a 3D visualization of the target motion, and shows the five results as an anonymized grid labelled A–E, in which our model occupies a different letter in each case. Three matrix questions then ask for a rating from 5 (excellent) to 1 (poor) for every result: *visual quality* (realism, sharpness, artifacts, overall perceptual quality), *control accuracy* (whether the intended object and camera motion is realized, judged on motion alone), and *temporal and input consistency* (flicker and unintended changes of subject identity, appearance, geometry or background). The nine cases cover object rotation (five), camera yaw (one), object translation with a following camera (two), and non-rigid dynamics (one). Fig. 11 shows one page.

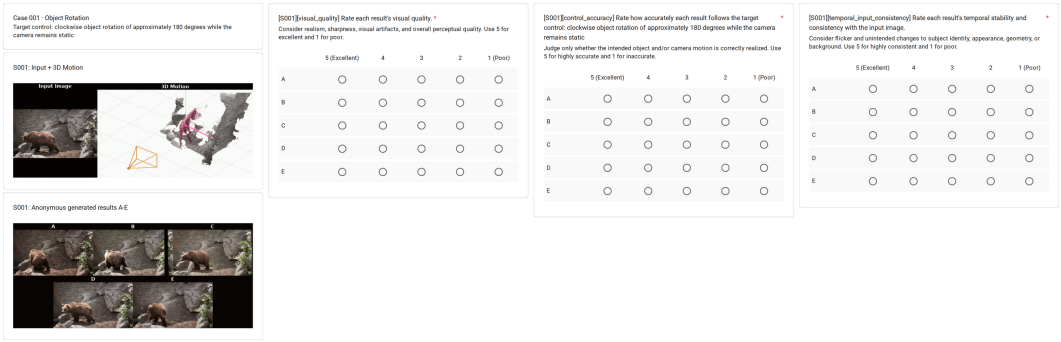


Figure 11: **One page of the user study.** *Left:* the case and the five anonymized results. *Right:* the three rating questions.

I Authored Trajectories and Insertions

The 61 authored cases are public clips under re-authored camera and rigid trajectories, single-image object insertions and user-supplied cases (Fig. 1). They have no source video under the authored trajectories to compare against, and the training data contain no insertion pairs, so we report them qualitatively; Figs. 12 and 13 show seven.

J Additional Qualitative Results

Figs. 14 and 15 compare 4Director with the four baselines on six further clips, grouped by the control they receive. In every panel the input image and the target trajectories are on the left; the rows are MotionCtrl (labelled MotionControl), Perception-as-Control, SymphoMotion, VerseCrafter and 4Director (Ours).

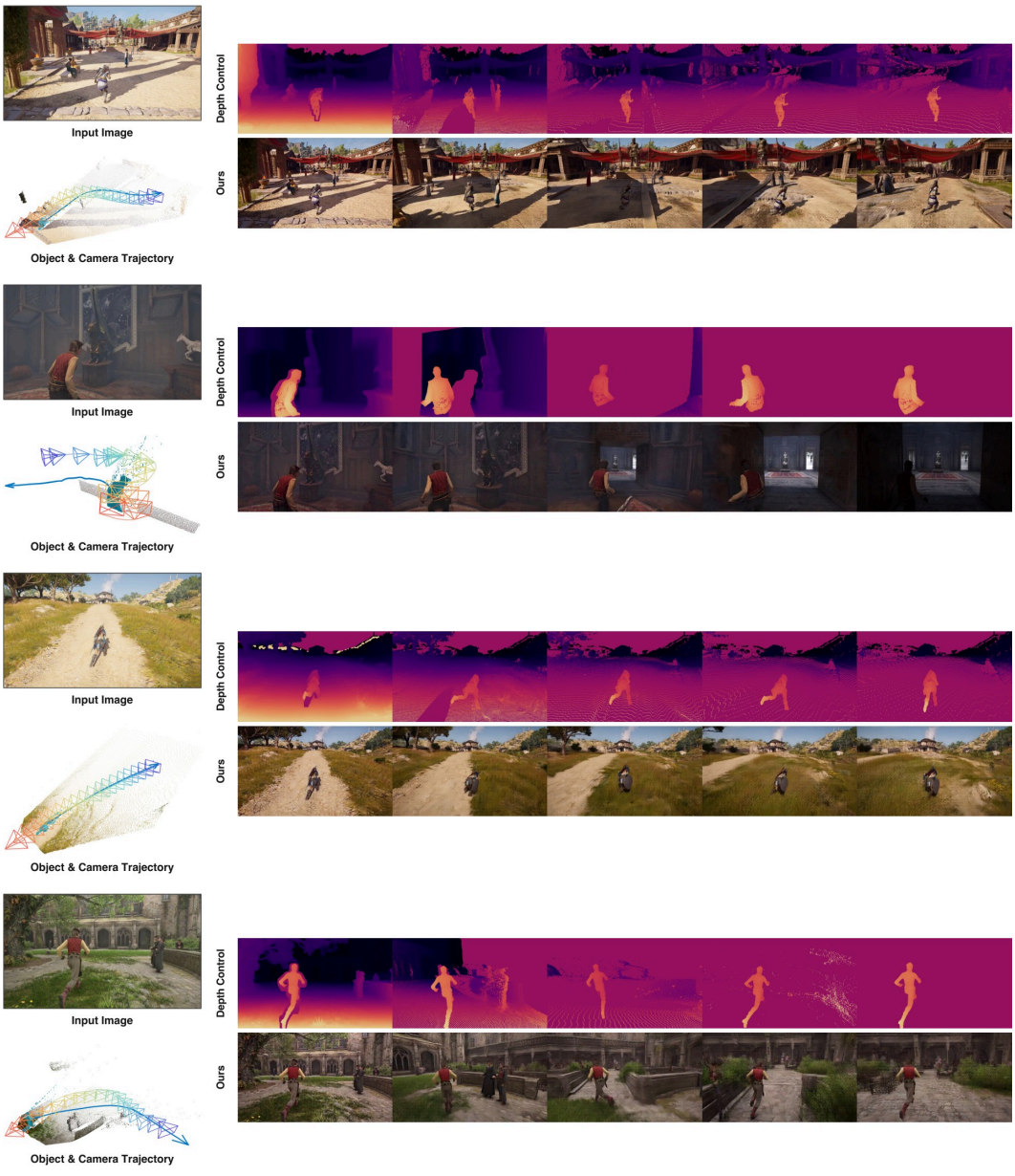


Figure 12: **Authored camera and object trajectories (1/2).** Each panel shows the input image and the authored trajectories on the left, the depth control rendered from the rigid 3D scene in the upper row, and the generated video below it. None of these controls comes from a source video.

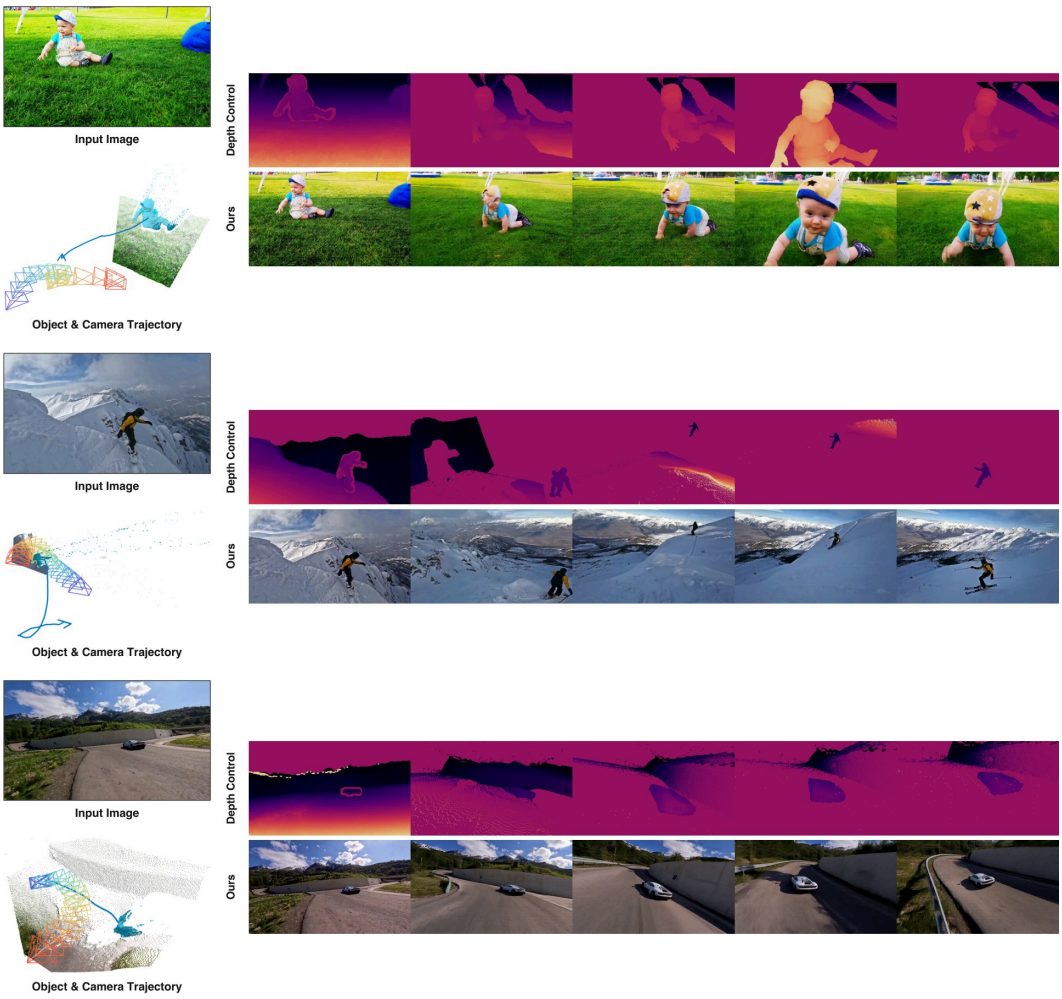


Figure 13: **Authored camera and object trajectories (2/2).** As in Fig. 12. The generator supplies the non-rigid dynamics, illumination and background that the rigid control leaves unspecified.

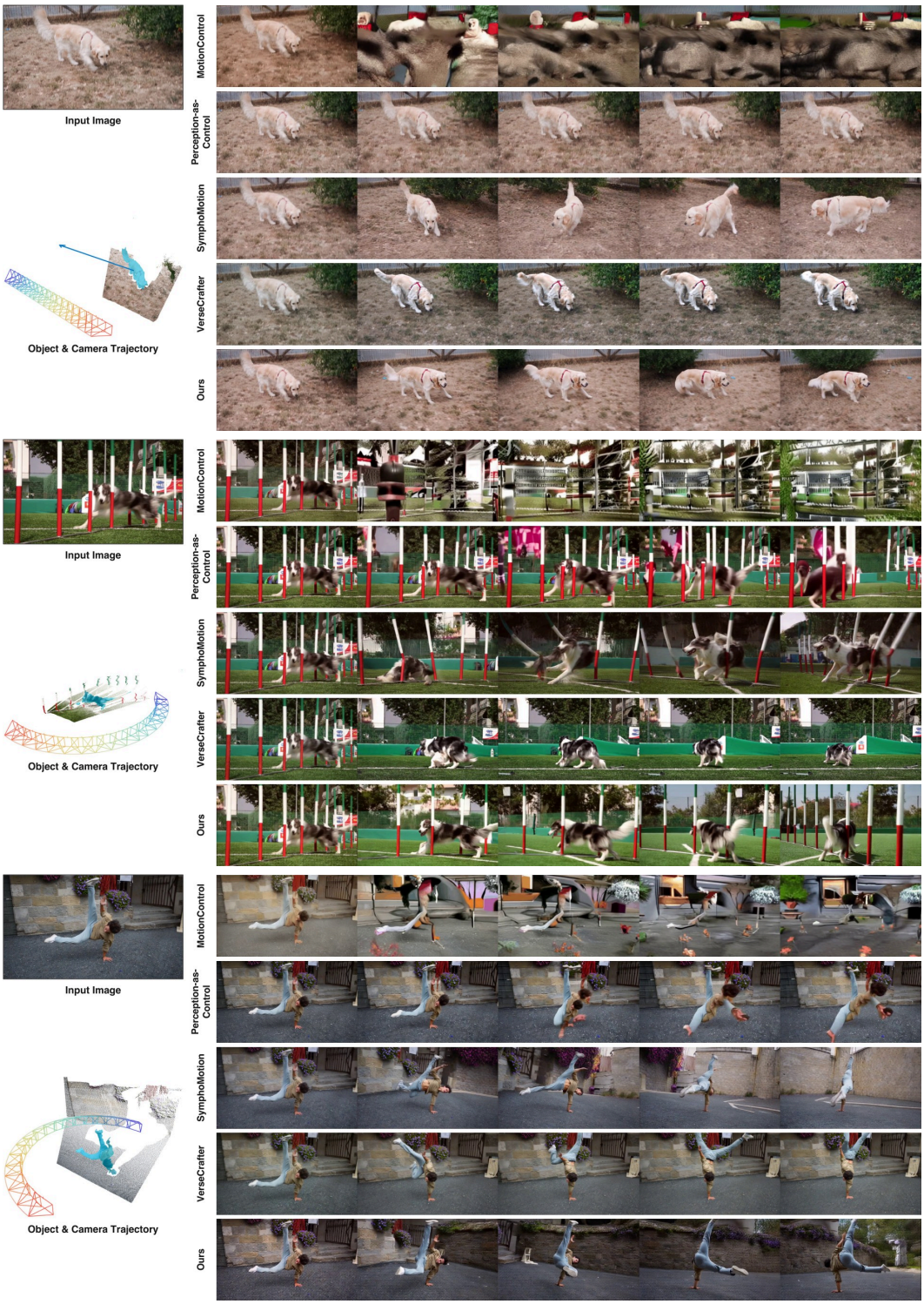


Figure 14: **Camera and object motion.** In the first two clips the camera follows a trajectory (the coloured frusta) while the object is moved by its own rigid trajectory. In the third the camera follows an arc around the subject and the object keeps its place, so the non-rigid dynamics come from the generator alone.

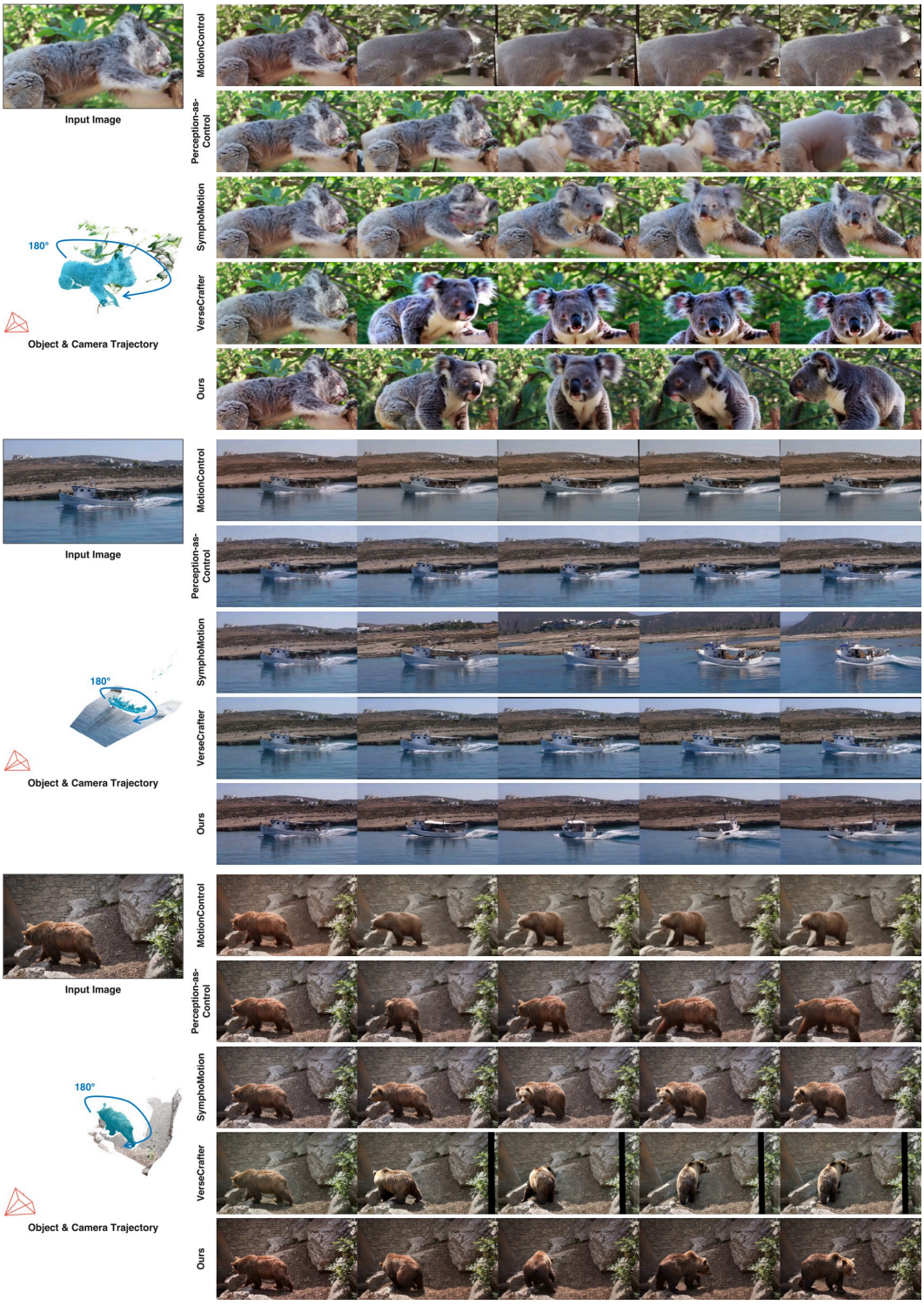


Figure 15: **Object rotation with a static camera.** The object is turned by 180° in place, with the camera held at the single frustum on the left. The baselines keep the original heading or lose the subject, whereas 4Director completes the turn and shows the side that turns into view.