

VerseCrafter: Dynamic Realistic Video World Model with 4D Geometric Control

Sixiao Zheng^{1,2}, Minghao Yin³, Wenbo Hu^{4†}, Xiaoyu Li⁴, Ying Shan⁴, Yanwei Fu^{1,2†}

¹Fudan University ²Shanghai Innovation Institute ³HKU ⁴ARC Lab, Tencent PCG

Project Page: https://sixiaozheng.github.io/VerseCrafter_page/



Figure 1. **VerseCrafter** enables precise control of camera motion and multi-object motion via a 4D Geometric Control representation built from a static background point cloud and per-object 3D Gaussian trajectories, producing videos that better follow the desired motion than Yume [61] and Uni3C [11] and closely match the ground-truth video.

Abstract

Video world models aim to simulate dynamic, real-world environments, yet existing methods struggle to provide unified and precise control over camera and multi-object motion, as videos inherently operate dynamics in the projected 2D image plane. To bridge this gap, we introduce **VerseCrafter**, a 4D-aware video world model that enables explicit and coherent control over both camera and object dynamics within a unified 4D geometric world state. Our approach is centered on a novel **4D Geometric Control** representation, which encodes the world state through a static background point cloud and per-object 3D Gaussian

trajectories. This representation captures not only an object’s path but also its probabilistic 3D occupancy over time, offering a flexible, category-agnostic alternative to rigid bounding boxes or parametric models. These 4D controls are rendered into conditioning signals for a pre-trained video diffusion model, enabling the generation of high-fidelity, view-consistent videos that precisely adhere to the specified dynamics. Unfortunately, another major challenge lies in the scarcity of large-scale training data with explicit 4D annotations. We address this by developing an automatic data engine that extracts the required 4D controls from in-the-wild videos, allowing us to train our model on a massive and diverse dataset.

[†]Corresponding authors.

1. Introduction

Video world models learn to simulate environmental dynamics by generating future frame sequences conditioned on past observations and control signals, such as actions or camera trajectory [13, 29, 40, 46, 61]. They provide a unified interface for visual prediction [31], navigation [7], and manipulation [23]. However, the reliance on video introduces a fundamental challenge: while an ideal world model should simulate the full 4D spatiotemporal space to reflect our physical reality, videos inherently operate dynamics in the projected 2D image plane.

To bridge this gap, recent works incorporate camera control into video generation using explicit 3D geometry [11, 115, 125], implicit pose embeddings [50], or learned movement embeddings [9, 12, 68]. However, these methods are often limited to static scenes or leave the motion of multiple objects uncontrolled. To control object motion, existing approaches typically rely on 2D cues such as point trajectories [96], optical flow [56], masks [123], or bounding boxes [89], which lack 3D awareness and often fail under large viewpoint changes. More advanced 3D-aware methods use depth maps [122], sparse 3D trajectories [15], 3D bounding boxes [92], or parametric human models like SMPL-X [11] to align camera and object motion in 3D space. Nevertheless, these control spaces are inadequate for representing multi-object dynamics in a compact, flexible, and editable 4D state that is also naturally aligned with camera control. For instance, sparse trajectories are often noisy and incomplete, 3D bounding boxes impose rigid constraints ill-suited to natural objects, and SMPL-X representations are category-limited. Besides, several existing works focus on synthetic game environments [40, 109, 113], where precise annotations are available for training, yet, the modeling of complex, realistic 4D worlds with multi-object dynamics remains underexplored.

Thus we propose *VerseCrafter*, a realistic, dynamic video world model that allows explicit and precise control over camera and multi-object motion in a unified 4D geometric control, as shown in Fig. 1. At its core is our novel *4D Geometric Control* representation, which encodes the world state using a static background point cloud for scene geometry and per-object 3D Gaussian trajectories to capture object dynamics. Each 3D Gaussian trajectory represents an object’s probabilistic 3D occupancy over time: its mean defines the motion path, while its covariance captures the object’s spatial extent and orientation. This probabilistic formulation provides a soft, flexible, and category-agnostic approach to modeling diverse object shapes and motions, overcoming the limitations of rigid 3D bounding boxes or category-specific parametric models. Crucially, the background point cloud and per-object 3D Gaussian trajectories share a common world coordinate system, enabling coherent and unified control over both camera and object motion.

By rendering our 4D Geometric Control into target views, we condition a frozen Wan2.1-14B video diffusion backbone [86] via a lightweight GeoAdapter, an adapter-style branch inspired by ControlNet [117]. This enables the generation of high-fidelity videos that accurately reflect the underlying 4D world state with specified camera and object dynamics. Unlike 2D control signals, our 4D Geometric Control is inherently 3D-aware, *i.e.* view-consistent and robust to occlusions, making it a more effective and reliable interface for video world modeling. Training VerseCrafter requires large-scale paired data of real-world videos and their corresponding 4D geometric controls. To this end, we constructed *VerseControl4D*, a large-scale real-world dataset with automatically annotated camera and object trajectories needed to construct our 4D geometric controls. This dataset allows us to train VerseCrafter on a massive and diverse set of real-world videos, significantly enhancing its generalization and performance.

Our contributions are threefold:

- We introduce a novel ***4D Geometric Control*** representation that unifies camera and multi-object motion in a shared 4D space. Its use of 3D Gaussian trajectories offers a flexible and category-agnostic way to control object dynamics, overcoming the limitations of rigid, category-specific models.
- We present ***VerseCrafter***, a geometry-driven video world model that leverages our 4D Geometric Control to offer explicit and precise control over both camera and object motion. This enables the creation of high-fidelity, view-consistent videos that accurately follow complex 4D instructions.
- We constructed an ***VerseControl4D***, a large-scale real-world dataset with automatically annotated camera and object trajectories. This breakthrough solves a key data bottleneck, enabling us to train our model on a massive and diverse real-world dataset for superior generalization.

2. Related Works

Video World Models. World models learn environment dynamics from observations by predicting future states for downstream simulation, planning, and control [29, 30, 46]. Early visual world models adopt recurrent and latent-variable architectures [16, 24, 28, 62, 66, 85], while recent approaches use large-scale transformer and diffusion backbones to roll out high-fidelity videos conditioned on actions, text, or camera trajectories [1, 2, 6, 9, 12, 20, 35, 39, 44, 48, 68, 86, 101, 108, 113], and further extend temporal horizons with explicit memories or long-sequence models [50, 71, 99]. Geometry-aware works such as DeepVerse [13], Voyager [40], and Yume [61] incorporate 3D structure to support 4D video generation and exploration, but are mainly controlled via text, actions, or camera tokens and do not expose a compact, editable 4D geometric state for

real multi-object dynamics. In contrast, VerseCrafter learns a geometry-driven mapping from 4D Geometric Control to dynamically realistic videos, enabling disentangled control over camera and multi-object motion.

3D World Generation. Recent work leverages powerful 2D generative priors to synthesize explorable 3D environments from text, images, or videos [47, 119]. Early methods mainly target object-level or single-scene generation [19, 37, 72, 116, 118, 120], distilling image diffusion models [77] into NeRFs [63], implicit fields, meshes, or 3D Gaussian splats [43], or optimizing per-scene geometry from multi-view or panoramic observations [18, 78, 111, 112, 114]. More recent approaches scale up to navigable 3D worlds [7], combining depth estimation [106], camera-guided video diffusion, iterative inpainting, and panoramic inputs to construct room- or city-scale Gaussian scenes for exploration [14, 52, 58, 59, 79, 84, 90, 109, 127]. However, these pipelines largely model static, synthetic-like geometry and offer limited explicit control over real multi-object dynamics, whereas VerseCrafter operates on real-world videos and uses a background point cloud plus per-object 3D Gaussian trajectories as an explicit 4D control state for world-consistent dynamic video generation.

Controllable Video Generation. Controllable video generation aims to steer camera and object motion via explicit conditioning signals. Camera-controlled models [3, 4, 33, 45, 51, 82, 104, 124] such as MotionCtrl [96] and CameraCtrl [32] inject camera extrinsics, Plücker-style encodings, or other 3D priors [11, 21, 27, 38, 73, 76, 97, 102, 115, 122, 125] into video diffusion models to achieve precise viewpoint control, but mostly assume static or weakly dynamic scenes. Object motion [10, 25, 34, 49, 53, 55, 60, 64, 65, 67, 74, 80, 81, 83, 87, 88, 94, 95, 98, 105, 107, 110, 121, 126] is typically controlled using 2D cues (bounding boxes, masks, trajectories, strokes, optical flow) as in Boximator [89], DragAnything [100], and MotionCanvas [103], or with more 3D-aware signals such as depth maps, sparse 3D trajectories, 3D boxes, or SMPL-X bodies in I2V3D [122], Uni3C [11], CineMaster [92], and Perception-as-Control [15]. While these methods substantially improve controllability, 2D controls remain view-dependent and fragile under large camera changes, and many 3D controls are category-specific, rigid, or tied to reconstruction-heavy pipelines. Recent approaches [15, 22, 26, 56, 92, 96, 103, 107, 125] begin to jointly control camera and object motion, but their control spaces are still fragmented rather than a unified, compact world state. VerseCrafter instead introduces *4D Geometric Control*: a lightweight, category-agnostic world state where a background point cloud and per-object 3D Gaussian trajectories in a shared frame jointly drive camera and multi-object motion.

3. Method

We propose **VerseCrafter**, a geometry-driven video world model that maps an explicit 4D geometric world state into dynamic, realistic videos with disentangled control over camera and multi-object motion. Our design has two key components: (i) a unified *4D Geometric Control* (Sec. 3.1) representation defined in a shared world coordinate frame, and (ii) a lightweight *GeoAdapter* (Sec. 3.2) that injects rendered geometric signals into a frozen Wan2.1-14B backbone, so that edits to the 4D state directly reshape the generated video while preserving Wan2.1’s strong visual prior. Given an input frame, a text prompt, and 4D Geometric Control, we model the world state as a static background point cloud and per-object 3D Gaussian trajectories, render them into multi-channel control maps, and feed these maps into GeoAdapter attached to Wan2.1.

3.1. 4D Geometric Control

We represent the state of the video world model as a 4D geometric world state, which we term *4D Geometric Control*. It is an explicit, editable state consisting of a static background point cloud P^{bg} and per-object 3D Gaussian trajectories $\{\mathcal{G}_o^t\}$, all defined in a shared world coordinate frame.

Background point cloud. As in Fig. 2, we start from the input image, estimate monocular depth with MoGe-2 [93], and obtain instance masks $\{M_o\}$ with Grounded SAM2 [75], where the user selects one or more objects to be controlled via text prompts or clicks. With camera intrinsics \mathbf{K} and extrinsics $(\mathbf{R}_1, \mathbf{t}_1)$, each pixel $\mathbf{u} = (u, v, 1)^T$ with depth $D_1(\mathbf{u})$ is back-projected as

$$\mathbf{p}(\mathbf{u}) = \mathbf{R}_1^T (D_1(\mathbf{u}) \mathbf{K}^{-1} \mathbf{u} - \mathbf{t}_1). \quad (1)$$

We use the instance masks to partition the reconstructed point cloud into per-object point clouds

$$P_o = \{\mathbf{x}_{o,k} \mid \mathbf{x}_{o,k} = \mathbf{p}(\mathbf{u}_k), \mathbf{u}_k \in M_o\}, \quad (2)$$

and a static background cloud

$$P^{\text{bg}} = \{\mathbf{p}(\mathbf{u}) \mid \mathbf{u} \notin \bigcup_o M_o\} = \{\mathbf{p}_i\}_{i=1}^{N_{\text{bg}}}. \quad (3)$$

During generation, the background at frame t is obtained by rendering P^{bg} with the camera pose, so that viewpoint changes are realized as rigid camera motion in a fixed 3D world rather than by hallucinating a new background at every frame.

3D Gaussian trajectories. A single 3D Gaussian $\mathcal{G}_o(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_o, \boldsymbol{\Sigma}_o)$ in the world frame compactly encodes an object’s position (through $\boldsymbol{\mu}_o$), approximate shape and size (through the eigenvalues of $\boldsymbol{\Sigma}_o$), and orientation (through its eigenvectors). A *3D Gaussian trajectory* for object o is then defined as a sequence of Gaussians

$$\{\mathcal{G}_o^t\}_{t=1}^T, \quad \mathcal{G}_o^t(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_o^t, \boldsymbol{\Sigma}_o^t), \quad (4)$$

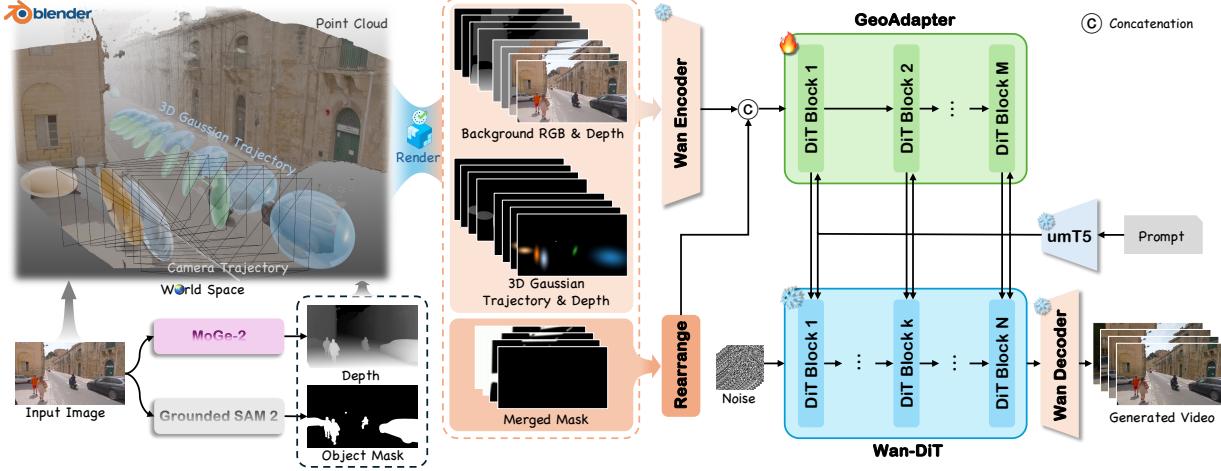


Figure 2. Framework of VerseCrafter. Given an input image and a text prompt, we first estimate depth and obtain user-specified object masks to construct a 4D Geometric Control state consisting of a static background point cloud and per-object 3D Gaussian trajectories in a shared world frame. This state is rendered into background RGB/depth, 3D Gaussian trajectory RGB/depth, and a soft control mask for each frame, forming multi-channel 4D control maps. The control maps are encoded and fed into the proposed GeoAdapter, which conditions a frozen Wan2.1-14B video diffusion backbone together with text embeddings from umT5, enabling geometry-consistent video generation with precise control over camera and multi-object motion.

whose means $\{\mu_o^t\}$ trace the motion path in 3D, while the covariances $\{\Sigma_o^t\}$ capture how the object’s spatial extent and orientation evolve over time. This probabilistic formulation describes the object’s 3D occupancy in a soft, continuous manner and yields a compact control space that is more flexible than rigid 3D bounding boxes and more category-agnostic than parametric body models.

To initialize the trajectory for each controllable object o , we fit a full-covariance Gaussian to its point cloud P_o obtained in the previous step:

$$\mu_o = \frac{1}{N_o} \sum_k \mathbf{x}_{o,k}, \Sigma_o = \frac{1}{N_o} \sum_k (\mathbf{x}_{o,k} - \mu_o)(\mathbf{x}_{o,k} - \mu_o)^\top, \quad (5)$$

which gives an initial Gaussian $\mathcal{G}_o(\mathbf{x})$

The low-dimensional parameters $\{\mu_o^t, \Sigma_o^t\}$ naturally support flexible, user-driven editing. In practice, we convert each \mathcal{G}_o^t into an ellipsoid mesh for visualization in a 3D editor such as Blender, and let the user specify or refine the trajectory by dragging and keyframing this ellipsoid in world space. The edited poses and shapes are mapped back to the $\{\mu_o^t, \Sigma_o^t\}$ as control signals. The ellipsoids are only a user interface; all conditioning maps used by our model are rendered directly from the underlying 3D Gaussians.

Rendering 4D control maps. Given 4D Geometric Control, we render per-frame conditioning maps in the target camera views. For each frame t , we generate three types of maps: (i) background RGB/depth, RGB_t^{bg} and $\text{Depth}_t^{\text{bg}}$, by projecting the static cloud P^{bg} with the camera pose $(\mathbf{R}_t, \mathbf{t}_t)$; (ii) 3D Gaussian trajectory RGB/depth, $\text{RGB}_t^{\text{traj}}$ and $\text{Depth}_t^{\text{traj}}$, by projecting the per-object Gaussians $\{\mathcal{G}_o^t\}$ into soft elliptical footprints and taking depth from the cor-

responding ellipsoid surfaces; (iii) a soft control mask M_t that indicates regions where the diffusion model should synthesize or overwrite content, obtained by inverting the valid background visibility and merging it with the projected 3D Gaussian footprints, followed by Gaussian smoothing. For the first frame $t = 1$, we replace RGB_1^{bg} with input image and set $M_1 = 0$, so that the first frame is preserved and only future frames are modified. Background and 3D Gaussian maps share the same world state but are rendered through *decoupled* channels, so camera edits only affect background branch and object edits only affect 3D Gaussian trajectory branch, enabling geometry-consistent control.

3.2. VerseCrafter Architecture

Backbone. We adopt Wan2.1-14B [86] as a frozen latent video diffusion / flow-matching backbone with a 3D VAE and a DiT-based denoiser. VerseCrafter treats Wan2.1 as a generic video prior: we do not change its architecture or weights, and instead attach a lightweight geometric adapter that conditions the backbone on our 4D control maps.

GeoAdapter. For each frame t , we take the rendered background and 3D Gaussian trajectory maps, RGB_t^{bg} , $\text{Depth}_t^{\text{bg}}$, $\text{RGB}_t^{\text{traj}}$, $\text{Depth}_t^{\text{traj}}$, together with the soft control mask M_t . The four RGB/depth maps are encoded by the same 3D VAE as the video latent, while M_t is reshaped and interpolated to the latent resolution, following the practice in [41, 86]. Stacking along the temporal dimension yields a spatio-temporal geometry tensor, which is concatenated channel-wise and aligned with the latent video tokens. GeoAdapter is a lightweight DiT-style branch that operates on this geometry tensor. It shares the same token dimen-

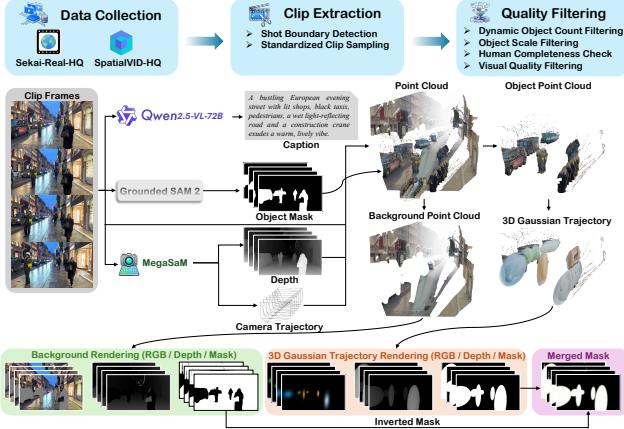


Figure 3. Starting from Sekai-Real-HQ and SpatialVID-HQ, we obtain 81-frame clips extraction, followed by quality filtering. For each retained clip, Qwen2.5-VL-72B, Grounded-SAM2, and MegaSAM provide captions, object masks, depth, and camera poses, which are lifted into background/object point clouds, fitted with 3D Gaussian trajectories, and rendered as background/trajec-tory maps plus a merged mask that constitute our 4D Geometric Control.

sionality as the Wan-DiT blocks, but uses far fewer layers. We interleave GeoAdapter blocks with the frozen Wan-DiT: every k -th DiT block in Wan2.1 is paired with a GeoAdapter block whose output is linearly projected back to the backbone width and added as a residual modulation to the corresponding DiT block. Text prompts are encoded by umT5 [17] into text embeddings, which are injected into both Wan’s DiT blocks and GeoAdapter through the same text-conditioning interfaces. This adapter-based conditioning injects 4D geometric information into Wan2.1 with only a small number of extra parameters, while keeping all backbone weights fixed.

Inference. At inference time, VerseCrafter supports both independent control of camera or object motion and joint control of both within a single unified framework. For *camera-only* control, we provide a camera trajectory and background control maps while setting all trajectory-related channels (RGB/depth/mask) to zero. For *object-only* control, we keep the camera pose fixed, render a static background branch (RGB/depth and its mask) from P^{bg} . For *joint* control, both branches are active and rendered from the same 4D world state, allowing VerseCrafter to adjust camera trajectory and multi-object motion in a coordinated, geometry-consistent manner.

4. VerseControl4D Dataset

To train and evaluate VerseCrafter on real, complex scenes with explicit 4D control, we construct **VerseControl4D**, a real-world video dataset with automatically derived 4D Ge-

ometric Control annotations. As shown in Fig. 3, VerseControl4D is built through four stages: data collection, clip extraction, quality filtering, and data annotation.

Data collection. VerseControl4D is built from two recent world-exploration datasets, Sekai-Real-HQ [54] and SpatialVID-HQ [91], which provide long in-the-wild videos with diverse outdoor and urban scenes, camera poses, and captions, but no object-motion labels. We take their high-resolution videos as the raw pool for constructing our 4D Geometric Control annotations.

Clip extraction. We apply PySceneDetect to detect shots in the videos. For each shot longer than 81 frames, we uniformly sample an 81-frame sub-clip and discard shorter shots, matching the default temporal length used by the Wan2.1 backbone.

Quality filtering. We apply an object-centric filtering pipeline to retain clips with clean geometry and controllable foreground. Using Grounded-SAM2 with prompts such as “*person . human . car . animal*”, we first obtain instance masks on the first frame and keep only clips whose controllable object count lies in [1, 6]. We then discard clips where any instance mask covers more than 20% of the image area. For human instances, we further remove clips whose masks touch image borders or whose aspect ratios fall outside [2, 4], as these typically correspond to severely truncated pedestrians. Finally, we apply visual-quality filtering (aesthetic and luminance scores) to exclude blurry or over-/under-exposed clips, yielding a set of visually clean, structurally reliable videos.

Data annotation. We then annotate each filtered clip with 4D Geometric Control. We first generate a descriptive caption using Qwen2.5-VL-72B [5], which serves as the text prompt during training. For geometry, we adopt MegaSAM as the base pipeline and replace its monocular and metric depth modules with MoGe-2 [93] and UniDepth V2 [70], respectively, to obtain more accurate and temporally consistent depth. Given the video frames, the depth, and the camera trajectory, we reconstruct a 3D point cloud for every frame. Applying Grounded-SAM2 instance masks on each frame to these point clouds yields per-object point clouds and a static background point cloud P^{bg} , as described in Sec. 3.1. For each object, we then fit per-frame 3D Gaussians and connect them into a 3D Gaussian trajectory $\{\mathcal{G}_o^t\}$. Finally, we render the 4D Geometric Control into model-ready signals. The background point cloud is rendered with the camera trajectory to obtain background RGB, depth and mask. The 3D Gaussian trajectories are rendered into trajectory RGB, depth and mask. We invert the background mask and merge it with the trajectory mask to produce a final merged mask that marks regions where the video diffusion model should synthesize content.

In total, VerseControl4D contains 35,000 training samples and 1,000 validation samples. In the training set, about

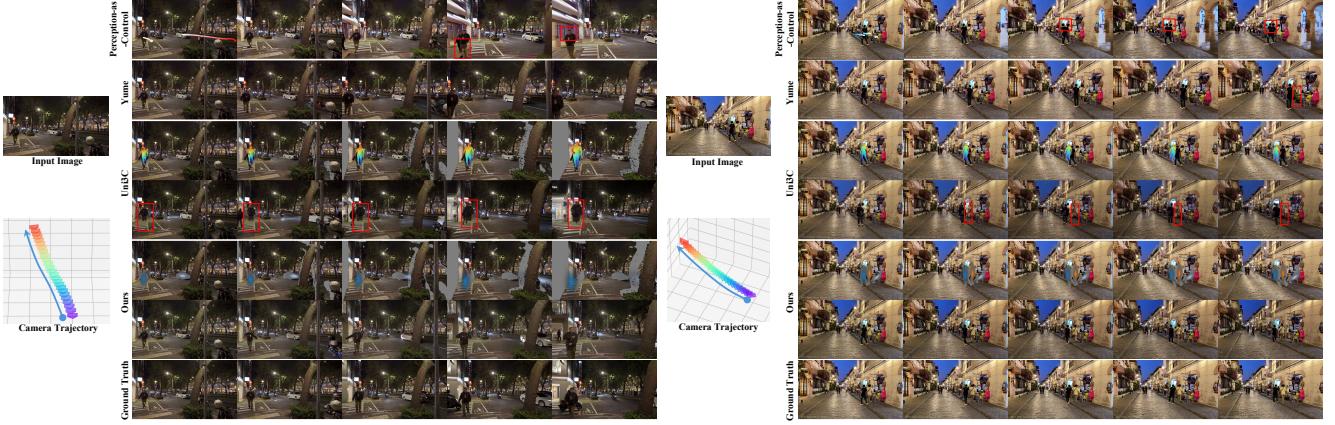


Figure 4. **Qualitative comparison of joint camera and object motion control.** Perception-as-Control often yields low-fidelity frames with inaccurate camera motion, Yume roughly follows the text-described motion but lacks precise control, and Uni3C is limited to human motion. VerseCrafter more faithfully follows both the camera trajectory and multi-object motion while maintaining sharp appearance and geometrically consistent backgrounds.

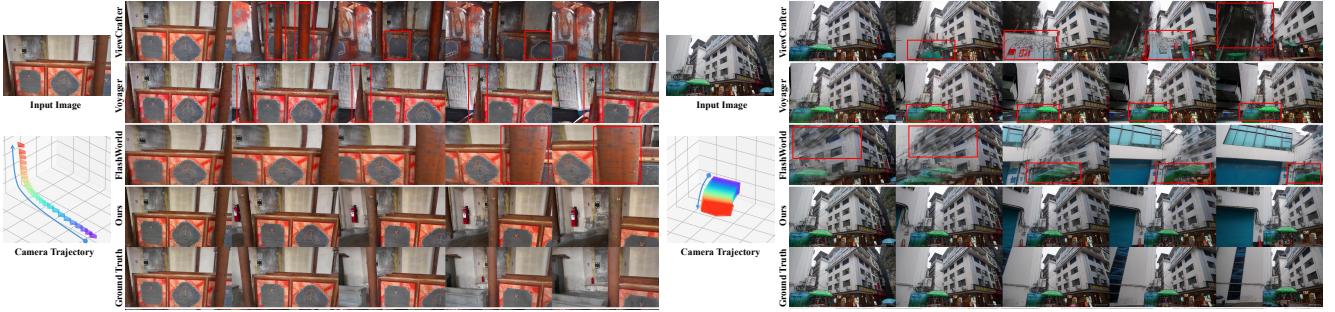


Figure 5. **Qualitative comparison of camera-only motion control on static scenes.** ViewCrafter, Voyager, and FlashWorld often exhibit distorted facades, drifting structures, or inconsistent parallax along the camera path. VerseCrafter better follows the target trajectory while preserving sharp details and globally consistent 3D geometry.

26% of samples are sourced from Sekai-Real-HQ and 74% from SpatialVID-HQ, and 20% of the samples depict static scenes, encouraging VerseCrafter to learn both camera-only world exploration and coupled camera–object dynamics. The validation set additionally includes 250 static-scene samples to specifically assess camera-only control.

5. Experiments

Implementation Details. We build VerseCrafter on top of the Wan2.1 T2V-14B model. The Wan backbone is kept frozen and only the GeoAdapter is updated. Each GeoAdapter block is initialized from the weights of its paired DiT block in Wan2.1 to stabilize training, and we set $k = 5$ so that every 5-th DiT block in Wan2.1 is paired with a GeoAdapter block. We use the Adam optimizer with a learning rate of $2e - 5$, 100 warmup steps, and a constant-with-warmup learning-rate schedule. All experiments are conducted on 16 96GB GPUs with a global batch size of 16. Training is performed in two stages: we first train for 2,500 iterations on 480P clips, and then fine-tune the same model

for another 2,500 iterations on 720P clips. The total wall-clock training time is about 380 hours. We adopt classifier-free guidance during training by randomly dropping the text condition with probability 0.1. At inference time, we use 50 denoising steps and a classifier-free guidance scale of 5.0. Generating an 81-frame 720P video clip on 8 96GB GPU takes about 1152 seconds, with a peak memory usage of about 90 GB.

Evaluation Metrics. We evaluate overall video quality using VBench-I2V. For camera control, we follow prior camera-control work [32] and report rotation error (RotErr) and translation error (TransErr). For object-motion control, we adopt ObjMC proposed in MotionCtrl [96]. Given a generated video, we run the same geometry annotation pipeline as in VerseControl4D to estimate its camera trajectory and 3D Gaussian trajectories, and compare them with the corresponding ground-truth trajectories from our dataset. ObjMC is computed as the average Euclidean distance between the estimated and ground-truth 3D Gaussian means over all controlled objects and frames.

Table 1. **Joint camera and object motion control on VerseControl4D.** We report VBench-I2V scores and 3D control metrics (RotErr, TransErr, ObjMC;). VerseCrafter achieves the best overall video quality and the most accurate joint control of camera and object motion.

	Overall Score ↑	Imaging Quality ↑	Aesthetic Quality ↑	Dynamic Degree ↑	Motion Smoothness ↑	Background Consistency ↑	Subject Consistency ↑	I2V Background ↑	I2V Subject ↑	RotErr ↓	TransErr ↓	ObjMC ↓
Perception-as-Control [15]	83.66	66.81	53.34	73.91	96.89	93.19	94.02	96.35	94.78	5.006	8.767	6.556
Yume [61]	85.47	71.16	52.39	72.24	98.96	95.66	96.43	98.51	98.39	7.560	8.735	7.959
Uni3C [11]	83.55	68.06	53.16	66.09	98.94	93.74	94.19	97.19	97.05	1.361	7.731	5.883
Ours	88.10	72.70	57.49	86.26	98.79	95.69	96.48	98.76	98.65	0.890	3.103	2.507

Table 2. **Camera-only motion control on static scenes.** On the static subset of VerseControl4D, we report VBench-I2V scores and camera control metrics RotErr / TransErr. VerseCrafter achieves the best overall visual quality while substantially reducing camera pose errors.

	Overall Score ↑	Imaging Quality ↑	Aesthetic Quality ↑	Dynamic Degree ↑	Motion Smoothness ↑	Background Consistency ↑	Subject Consistency ↑	I2V Background ↑	I2V Subject ↑	RotErr ↓	TransErr ↓
ViewCrafter [115]	84.04	69.56	55.52	68.02	97.86	92.09	94.25	97.70	97.29	2.101	9.868
Voyager [40]	78.12	55.48	49.80	65.34	99.39	92.31	91.55	86.02	85.03	3.557	3.880
FlashWorld [52]	81.80	68.94	53.72	58.26	98.81	91.88	94.44	94.40	93.93	2.748	10.010
Ours	86.80	74.57	54.78	80.34	97.62	94.88	95.55	97.86	98.79	0.650	2.587

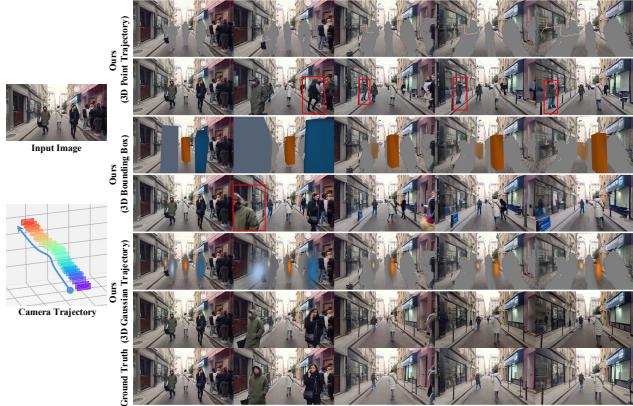


Figure 6. **Ablation on object-motion representations.** We compare controlling objects with 3D point trajectory (top), 3D bounding box (middle), and 3D Gaussian trajectory (fourth). 3D point trajectory and 3D bounding box often cause scale drift and misaligned motion (red boxes), whereas 3D Gaussian trajectory track the intended camera trajectory and preserve plausible shapes and background interactions.

5.1. Joint Camera and Object Motion Control

We first evaluate joint control of camera and object motion on VerseControl4D. As shown in Table 1, VerseCrafter achieves the best VBench-I2V scores among Perception-as-Control, Yume, Uni3C, and our model, with clear gains in Overall Score, Imaging Quality, Aesthetic Quality, and both Subject/Background consistency. On 3D control metrics, VerseCrafter substantially reduces rotation, translation, and object-motion errors compared with the strongest baseline, reflecting much tighter alignment with the target 4D trajectories. Qualitative comparisons in Fig. 4 further highlight these differences: Perception-as-Control often produces low-quality frames with inaccurate camera motion; Yume, driven only by text descriptions of motion, roughly follows the desired direction but lacks precise trajectory control; and Uni3C, relying on SMPL-X, can control human



Figure 7. **Ablation on depth-aware control.** We compare VerseCrafter without depth inputs (*Ours (w/o depth)*, top) and with RGB+depth control (middle) under the same camera trajectory. Without depth, the model often misorders foreground and background, e.g., lampposts are pulled in front of distant buildings—and occlusion boundaries drift over time (red boxes). Adding depth restores consistent parallax and occlusion, producing geometry much closer to the ground truth.

motion but fails to handle other categories such as vehicles. In contrast, VerseCrafter keeps multiple objects attached to their 3D Gaussian trajectories while accurately following the specified camera path, yielding sharp and temporally coherent videos.

5.2. Camera-Only Motion Control

We evaluate camera-only control on the static-scene subset of VerseControl4D, where objects remain stationary and only the camera moves. As shown in Table 2, VerseCrafter achieves the best VBench-I2V performance among ViewCrafter, Voyager, FlashWorld, and our model, with consistent gains in Overall Score, Imaging Quality, and both background and subject consistency, while maintaining motion smoothness comparable to prior methods. On 3D camera metrics, VerseCrafter substantially reduces rotation and translation errors relative to the strongest baseline, indicating that it follows the target camera trajectory much more faithfully in static scenes. Qualitative comparisons in Fig. 5 further confirm these trends: baselines often exhibit bending walls, misaligned windows, or unstable paral-

Table 3. **Ablation study on 3D representation, depth, and decoupled controls.** We compare different variants of VerseCrafter using VBench-I2V and 3D control metrics (RotErr, TransErr, ObjMC). Our full model with 3D Gaussian trajectories, depth-aware rendering, and decoupled background/foreground controls achieves the best visual quality and the most accurate camera and object motion control.

	Overall Score ↑	Imaging Quality ↑	Aesthetic Quality ↑	Dynamic Degree ↑	Motion Smoothness ↑	Background Consistency ↑	Subject Consistency ↑	I2V Background ↑	I2V Subject ↑	RotErr ↓	TransErr ↓	ObjMC ↓
Ours (3D Bounding Box)	85.45	69.23	55.70	78.57	98.70	92.92	93.27	97.74	97.48	1.350	3.805	4.520
Ours (3D Point Trajectory)	85.57	70.29	55.27	78.23	98.63	94.00	92.75	97.85	97.55	1.298	3.281	6.896
Ours (w/o depth)	85.64	70.19	55.00	80.60	98.66	92.07	92.83	98.07	97.69	1.177	3.900	4.929
Ours (BG & FG Merged)	85.72	69.19	54.86	83.72	98.65	91.15	92.86	97.93	97.41	1.080	3.803	3.726
Ours	88.10	72.70	57.49	86.26	98.79	95.69	96.48	98.76	98.65	0.890	3.103	2.507



Figure 8. **Ablation on decoupled background / foreground controls.** We compare merging background and foreground controls into a single map (*Ours (BG & FG Controls Merged)*, top) with our default decoupled design (middle). When controls are merged, object motion control performance significantly degrades (red box), while the separation design preserves the static background and produces sharper, more stable object motion.

lax along the path, whereas VerseCrafter preserves straight structures, stable depth relationships, and an appearance closer to the ground-truth video, evidencing precise camera control in a static 3D world.

5.3. Ablation Study

We conduct ablations to analyze three key design choices in VerseCrafter: (i) the object 3D representation in the control space, (ii) the use of depth in control maps, and (iii) the decoupling of background and foreground controls. All variants share the same training data, backbone, and optimization settings; only the control representation is changed.

3D representation of object motion. To isolate the effect of our motion representation, we derive two ablations from each per-frame 3D Gaussian: (1) an **oriented 3D bounding box** whose axes follow the Gaussian’s principal directions and whose side lengths scale with its principal spreads; and (2) a **3D point trajectory** that retains only the Gaussian centroid. The rest of the pipeline is unchanged—we simply rasterize cuboids (for boxes) or tiny disks/spheres (for points) instead of Gaussian ellipses. As reported in Table 3, replacing Gaussians with boxes slightly hurts both visual quality and control accuracy (Overall Score ↓ from 88.10 to 85.45; ObjMC ↑ from 2.51 to 4.52), while point trajectories give the weakest object-motion consistency (ObjMC = 6.90). Qualitatively (Fig. 6), points and boxes often yield scale artifacts and misaligned motion, whereas 3D Gaussian trajectories better track the intended paths and preserve plausible object shapes.

Effect of depth. To evaluate the effect of depth, we removed the depth channel from the background and trajectory controls (“Ours (w/o depth)” in Table 3). This variation resulted in a lower overall score and significantly worse 3D control (higher RotErr and ObjMC values). As shown in Figure 7, without depth, the model frequently misorders foreground and background: vertical structures like streetlights appear next to shelves in the foreground, while buildings that should be behind the character are positioned elsewhere, and occlusion boundaries drift over time. With RGB+depth control, With RGB+depth control, VerseCrafter recovers more consistent parallax and occlusion, producing geometry much closer to the ground truth.

Decoupled vs. merged controls. We further compare our decoupled design with a variant that merges background and 3D Gaussian trajectory maps into a single control stream (*Ours (BG & FG Merged)* in Table 3). Although this variant still benefits from the explicit 4D state, it consistently underperforms the full model on VBench, with a particularly noticeable drop in object-motion accuracy (ObjMC increases from 2.51 to 3.73). As shown in Fig. 8, the merged control leads to a clear degradation in motion control for moving people. In contrast, keeping decoupled design preserves static geometry while producing more precise and stable object motion, which is crucial for accurate and geometry-consistent control.

6. Conclusion

We presented **VerseCrafter**, a geometry-driven video world model that exposes an explicit 4D Geometric Control state, built from a static background point cloud and per-object 3D Gaussian trajectories in a shared world frame. Coupled with the GeoAdapter that conditions a frozen Wan2.1 backbone, this design enables high-fidelity video generation with precise, disentangled control over camera and multi-object motion. To support training and evaluation, we constructed **VerseControl4D**, a large-scale real-world dataset with automatically annotated camera and object trajectories. Experiments and ablations show that VerseCrafter delivers superior visual quality and more accurate 3D control than existing controllable video generators and world models, highlighting 4D Geometric Control as a promising interface for future work on dynamic world simulation and editing.

References

- [1] 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. 2
- [2] 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. 2
- [3] 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. 3
- [4] 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 Computer Vision and Pattern Recognition Conference*, pages 22875–22889, 2025. 3
- [5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5
- [6] Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Gharamani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models. 2025. 2
- [7] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15791–15801, 2025. 2, 3
- [8] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023. 1
- [9] Jake Bruce, Michael D 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 *Forty-first International Conference on Machine Learning*, 2024. 2
- [10] Ryan Burger, 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*, pages 13–23, 2025. 3
- [11] 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. *arXiv preprint arXiv:2504.14899*, 2025. 1, 2, 3, 7
- [12] Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. *arXiv preprint arXiv:2411.00769*, 2024. 2
- [13] 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*, 2025. 2
- [14] Luxi Chen, Zihan Zhou, Min Zhao, Yikai Wang, Ge Zhang, Wenhao Huang, Hao Sun, Ji-Rong Wen, and Chongxuan Li. Flexworld: Progressively expanding 3d scenes for flexible-view synthesis. *arXiv preprint arXiv:2503.13265*, 2025. 3
- [15] Yingjie Chen, Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. Perception-as-control: Fine-grained controllable image animation with 3d-aware motion representation. *arXiv preprint arXiv:2501.05020*, 2025. 2, 3, 7
- [16] Silvia Chiappa, Sébastien Racaniere, Daan Wierstra, and Shakir Mohamed. Recurrent environment simulators. *arXiv preprint arXiv:1704.02254*, 2017. 2
- [17] Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharun Narang, and Orhan Firat. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *arXiv preprint arXiv:2304.09151*, 2023. 5
- [18] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023. 3
- [19] Dana Cohen-Bar, Elad Richardson, Gal Metzer, Raja Giryes, and Daniel Cohen-Or. Set-the-scene: Global-local training for generating controllable nerf scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2920–2929, 2023. 3
- [20] Etched Decart, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. URL: <https://oasis-model.github.io>, 2024. 2
- [21] 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*, 2024. 3
- [22] 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*, 2024. 3
- [23] Stefano Ferraro, Pietro Mazzaglia, Tim Verbelen, and Bart Dhoedt. Focus: object-centric world models for robotic manipulation. *Frontiers in Neurorobotics*, 19:1585386, 2025. 2
- [24] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *Advances in neural information processing systems*, 29, 2016. 2
- [25] 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. 3
- [26] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, et al. Motion prompting: Controlling video generation with motion trajectories. *arXiv preprint arXiv:2412.02700*, 2024. 3
- [27] 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*, pages 1–12, 2025. 3
- [28] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018. 2
- [29] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018. 2
- [30] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pages 2555–2565. PMLR, 2019. 2
- [31] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 2
- [32] 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. 3, 6, 4
- [33] 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*, 2025. 3
- [34] Xuehai He, Shuhang Wang, Jianwei Yang, Xiaoxia Wu, Yiping Wang, Kuan Wang, Zheng Zhan, Olatunji Ruwase, Yelong Shen, and Xin Eric Wang. Mojito: Motion trajectory and intensity control for video generation. *arXiv preprint arXiv:2412.08948*, 2024. 3
- [35] 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*, 2025. 2
- [36] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1
- [37] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 3
- [38] Chen Hou, Guoqiang Wei, Yan Zeng, and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024. 3
- [39] 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. 2
- [40] 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. 2, 7
- [41] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 4, 1
- [42] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 1
- [43] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 3
- [44] 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. 2
- [45] 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. 3
- [46] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022. 2
- [47] Haoran Li, Haolin Shi, Wenli Zhang, Wenjun Wu, Yong Liao, Lin Wang, Lik-hang Lee, and Peng Yuan Zhou. Dreamscene: 3d gaussian-based text-to-3d scene generation via formation pattern sampling. In *European Conference on Computer Vision (ECCV)*, 2024. 3

- ence on Computer Vision, pages 214–230. Springer, 2024.
- [48] 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*, 2025.
- [49] Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. *arXiv preprint arXiv:2503.16421*, 2025.
- [50] Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. *arXiv preprint arXiv:2506.18903*, 2025.
- [51] 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*, 2025.
- [52] Xinyang Li, Tengfei Wang, Zixiao Gu, Shengchuan Zhang, Chunchao Guo, and Liujuan Cao. Flashworld: High-quality 3d scene generation within seconds. *arXiv preprint arXiv:2510.13678*, 2025.
- [53] Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Yuexian Zou, and Ying Shan. Image conductor: Precision control for interactive video synthesis. *arXiv preprint arXiv:2406.15339*, 2024.
- [54] Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. *arXiv preprint arXiv:2506.15675*, 2025.
- [55] 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.
- [56] Xinyao Liao, Xianfang Zeng, Liao Wang, Gang Yu, Guosheng Lin, and Chi Zhang. Motionagent: Fine-grained controllable video generation via motion field agent. *arXiv preprint arXiv:2502.03207*, 2025.
- [57] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- [58] Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang, Yixuan Yuan, Yawei Luo, and Chunchao Guo. Worldmirror: Universal 3d world reconstruction with any-prior prompting. *arXiv preprint arXiv:2510.10726*, 2025.
- [59] Taiming Lu, Tianmin Shu, Junfei Xiao, Luoxin Ye, Jiahao Wang, Cheng Peng, Chen Wei, Daniel Khashabi, Rama Chellappa, Alan Yuille, et al. Genex: Generating an explorable world. *arXiv preprint arXiv:2412.09624*, 2024.
- [60] 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*, pages 1–11, 2024.
- [61] Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wen-shuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025.
- [62] Willi Menapace, Stephane Lathuiliere, Sergey Tulyakov, Aliaksandr Siarohin, and Elisa Ricci. Playable video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10061–10070, 2021.
- [63] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- [64] Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. Revideo: Remake a video with motion and content control. *arXiv preprint arXiv:2405.13865*, 2024.
- [65] 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*, pages 111–128. Springer, 2025.
- [66] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. *Advances in neural information processing systems*, 28, 2015.
- [67] Karran Pandey, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, Niloy J Mitra, and Paul Guerrero. Motion modes: What could happen next? *arXiv preprint arXiv:2412.00148*, 2024.
- [68] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplani, 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. 2024.
- [69] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.
- [70] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025.
- [71] 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.
- [72] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.
- [73] 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. 3
- [74] 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. 3
- [75] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 3
- [76] Xuanchi Ren, Tianshang 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*, pages 6121–6132, 2025. 3
- [77] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 1
- [78] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Hermann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9420–9429, 2024. 3
- [79] Manuel-Andreas Schneider, Lukas Hölein, and Matthias Nießner. Worldexplorer: Towards generating fully navigable 3d scenes. *arXiv preprint arXiv:2506.01799*, 2025. 3
- [80] 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*, pages 1–11, 2024. 3
- [81] Xincheng Shuai, Henghui Ding, Zhenyuan Qin, Hao Luo, Xingjun Ma, and Dacheng Tao. Free-form motion control: A synthetic video generation dataset with controllable camera and object motions. *arXiv preprint arXiv:2501.01425*, 2025. 3
- [82] 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. 3
- [83] Maham Tanveer, Yang Zhou, Simon Niklaus, Ali Mahdavi Amiri, Hao Zhang, Krishna Kumar Singh, and Nanxuan Zhao. Motionbridge: Dynamic video inbetweening with flexible controls. *arXiv preprint arXiv:2412.13190*, 2024. 3
- [84] 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. 3
- [85] Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V Le, and Honglak Lee. High fidelity video prediction with large stochastic recurrent neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 2
- [86] 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. 2, 4, 1
- [87] Zhang Wan, Sheng Tang, Jiawei Wei, Ruize Zhang, and Juan Cao. Dragentity: Trajectory guided video generation using entity and positional relationships. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 108–116, 2024. 3
- [88] 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. *arXiv preprint arXiv:2412.15214*, 2024. 3
- [89] 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*, 2024. 2, 3
- [90] Jiahao Wang, Luoxin Ye, TaiMing Lu, Junfei Xiao, Jiahao Zhang, Yuxiang Guo, Xijun Liu, Rama Chellappa, Cheng Peng, Alan Yuille, et al. Evoworld: Evolving panoramic world generation with explicit 3d memory. *arXiv preprint arXiv:2510.01183*, 2025. 3
- [91] Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. Spatialvid: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025. 5
- [92] 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 *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–10, 2025. 2, 3
- [93] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025. 3, 5
- [94] 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. 3
- [95] Zhouxia Wang, Yushi Lan, Shangchen Zhou, and Chen Change Loy. Objctrl-2.5 d: Training-free object control with camera poses. *arXiv preprint arXiv:2412.07721*, 2024. 3
- [96] 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*, pages 1–11, 2024. 2, 3, 6, 4
- [97] 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*, 2025. 3
- [98] Jianzong Wu, Xiangtai Li, Yanhong Zeng, Jiangning Zhang, Qianyu Zhou, Yining Li, Yunhai Tong, and Kai Chen. Motionbooth: Motion-aware customized text-to-video generation. *Advances in Neural Information Processing Systems*, 37:34322–34348, 2024. 3
- [99] 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. 2
- [100] 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*, pages 331–348. Springer, 2025. 3
- [101] Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024. 2
- [102] 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. 3
- [103] Jinbo Xing, Long Mai, Cusuh Ham, Jiahui Huang, Anirudha 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*, pages 1–11, 2025. 3
- [104] 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. 3
- [105] Tianshuo Xu, Zhifei Chen, Leyi Wu, Hao Lu, Yuying Chen, Lihui Jiang, Bingbing Liu, and Yingcong Chen. Motion dreamer: Realizing physically coherent video generation through scene-aware motion reasoning. *arXiv preprint arXiv:2412.00547*, 2024. 3
- [106] 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, 2024. 3
- [107] 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*, pages 1–12, 2024. 3
- [108] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenqi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2
- [109] Zhongqi Yang, Wenhong 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. 2, 3
- [110] 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. 3
- [111] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6658–6667, 2024. 3
- [112] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5916–5926, 2025. 3
- [113] 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*, 2025. 2
- [114] Jason J Yu, Fereshteh Forghani, Konstantinos G Derpanis, and Marcus A Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7094–7104, 2023. 3
- [115] 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. 2, 3, 7
- [116] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024. 3
- [117] 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*, pages 3836–3847, 2023. 2
- [118] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 3
- [119] Qihang Zhang, Chaoyang Wang, Aliaksandr Siarohin, Peiye Zhuang, Yinghao Xu, Ceyuan Yang, Dahua Lin, Bolei Zhou, Sergey Tulyakov, and Hsin-Ying Lee. Towards text-guided 3d scene composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6829–6838, 2024. 3
- [120] Shengjun Zhang, Jinzhao Li, Xin Fei, Hao Liu, and Yueqi Duan. Scene splatter: Momentum 3d scene generation from

- single image with video diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6089–6098, 2025. 3
- [121] 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. *arXiv preprint arXiv:2407.21705*, 2024. 3
- [122] Zhiyuan Zhang, Dongdong Chen, and Jing Liao. I2v3d: Controllable image-to-video generation with 3d guidance. *arXiv preprint arXiv:2503.09733*, 2025. 2, 3
- [123] Zhongwei Zhang, Fuchen Long, Zhaofan Qiu, Yingwei Pan, Wu Liu, Ting Yao, and Tao Mei. Motionpro: A precise motion controller for image-to-video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27957–27967, 2025. 2
- [124] 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. 3
- [125] 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. 2, 3
- [126] 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. *arXiv preprint arXiv:2408.11475*, 2024. 3
- [127] 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*, pages 8535–8546, 2025. 3

VerseCrafter: Dynamic Realistic Video World Model with 4D Geometric Control

Supplementary Material

A. Preliminary: Video Diffusion Models

Modern video diffusion models operate in a compact latent space learned by a spatio-temporal VAE. Given a video $x \in \mathbb{R}^{T \times H \times W \times 3}$, the encoder E maps it to latents $z_0 = E(x) \in \mathbb{R}^{T' \times C \times H' \times W'}$, on which the generative process is defined [8, 77]. A standard forward diffusion process gradually perturbs z_0 into noisy variables z_t via

$$q(z_t | z_0) = \sqrt{\alpha_t} z_0 + \sqrt{1 - \alpha_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad (6)$$

and a DiT-based denoiser ϵ_θ is trained to predict the noise under time step t and conditioning signal c (e.g., text prompts, reference frames) as

$$\mathcal{L}_{\text{diff}}(\theta) = \mathbb{E}_{z_0, t, \epsilon} [\|\epsilon_\theta(z_t, t, c) - \epsilon\|_2^2], \quad (7)$$

following the DDPM formulation [36]. Recent models further adopt continuous-time flow matching: given clean latents z_0 and Gaussian samples z_1 , one constructs interpolants $z_\tau = (1 - \tau)z_0 + \tau z_1$ with $\tau \in [0, 1]$ and learns a velocity field v_θ by

$$\mathcal{L}_{\text{flow}}(\theta) = \mathbb{E}_{z_0, \tau, \epsilon} [\|v_\theta(z_\tau, \tau, c) - (z_1 - z_0)\|_2^2], \quad (8)$$

as in recent flow-matching and ODE-based generative formulations [42, 57]. These objectives are naturally implemented with Diffusion Transformers (DiT), which operate on spatio-temporal latent tokens and inject (t, c) through attention [69], forming the backbone of current foundation video generators.

Wan2.1 instantiates the above latent video diffusion / flow-matching paradigm with a 3D VAE and a DiT-based denoiser, together with rich multi-modal conditioning trained on large-scale, diverse video-text data [86]. Throughout this work, we adopt Wan2.1-14B as a *frozen* latent video backbone and treat it as a generic video prior: we keep its encoder, decoder, and DiT-based denoiser unchanged, and only attach lightweight geometry-aware control interfaces on top of its DiT blocks. The detailed architecture of these control modules is provided in the Sec. B.

B. Model Architecture Details

VerseCrafter is built on top of Wan2.1 [86], a latent video diffusion / flow-matching model with a 3D VAE (Wan Encoder and Wan Decoder) and a DiT-based denoiser (Wan-DiT). We keep the Wan2.1 backbone frozen, and introduce a geometry-aware conditioning pathway together with a lightweight GeoAdapter that injects 4D geometric control signals into selected Wan-DiT blocks. We instantiate

VerseCrafter on top of the Wan2.1 T2V-14B backbone, resulting in a 14B-parameter controllable video world model. Fig. 9 illustrates the geometry-aware conditioning pathway and the integration of GeoAdapter into the Wan-DiT backbone, while Table 4 summarizes the input resolution, number of Wan-DiT layers, hidden dimension, GeoAdapter injection pattern, and fine-tuning configuration of VerseCrafter.

Geometry encoding and tokenization. For each frame t , we render background RGB/depth RGB_t^{bg} , $\text{Depth}_t^{\text{bg}}$, trajectory RGB/depth $\text{RGB}_t^{\text{traj}}$, $\text{Depth}_t^{\text{traj}}$, and a soft control mask M_t that marks regions where the diffusion model should synthesize or overwrite content (for $t=1$ we replace RGB_1^{bg} with the input image and set $M_1=0$). The four RGB/depth maps are passed through the frozen 3D VAE encoder to obtain latent features at the VAE resolution, while the mask $M \in \mathbb{R}^{1 \times T \times H \times W}$ is rearranged to align with the 3D VAE latent grid (the “Rearrange” module in Figure 9). Let s_t , s_h , and s_w denote the temporal and spatial strides of Wan’s 3D VAE (we use $s_t=4$ and $s_h=s_w=8$). Following the practice in [41, 86], we drop the singleton channel dimension, split the spatial dimensions into $s_h \times s_w$ sub-cells, and fold these sub-cells into the channel dimension via a reshape-permute operation, yielding a tensor of shape $C_M \times T \times H' \times W'$ with $C_M=s_h s_w$, $H'=H/s_h$, and $W'=W/s_w$. We then downsample the temporal dimension using nearest-neighbor interpolation to match the latent depth $T' = (T + s_t - 1)/s_t$, producing $\hat{M} \in \mathbb{R}^{C_M \times T' \times H' \times W'}$. Finally, \hat{M} is concatenated channel-wise with the encoded background and 3D Gaussian trajectory latents to obtain a spatio-temporal geometry feature $\mathcal{G} \in \mathbb{R}^{T' \times H' \times W' \times C_G}$. We follow Wan-DiT for tokenization: the latent grid \mathcal{G} is divided into non-overlapping 3D patches, and each patch is linearly projected into a token embedding, yielding a sequence of geometry tokens $\mathbf{g} \in \mathbb{R}^{L \times D}$, where $L = T' H' W'$ and D matches the hidden width of Wan-DiT. Because we use identical strides, positional encodings, and patch sizes, the geometry tokens are spatially and temporally aligned with the latent video tokens processed by Wan-DiT.

GeoAdapter integration. GeoAdapter is a lightweight DiT-style branch operating on the geometry tokens \mathbf{g} . It shares the same token dimensionality and positional encodings as Wan-DiT, but contains far fewer layers. Let $\{\mathcal{B}_0, \dots, \mathcal{B}_{N-1}\}$ denote the N Wan-DiT blocks of Wan2.1, and let $\{\mathcal{G}_1, \dots, \mathcal{G}_M\}$ denote the M GeoAdapter blocks. We attach GeoAdapter as a residual modulation branch to a subset of Wan-DiT blocks. Concretely, we choose a stride k and inject GeoAdapter after every k -th Wan-DiT block;

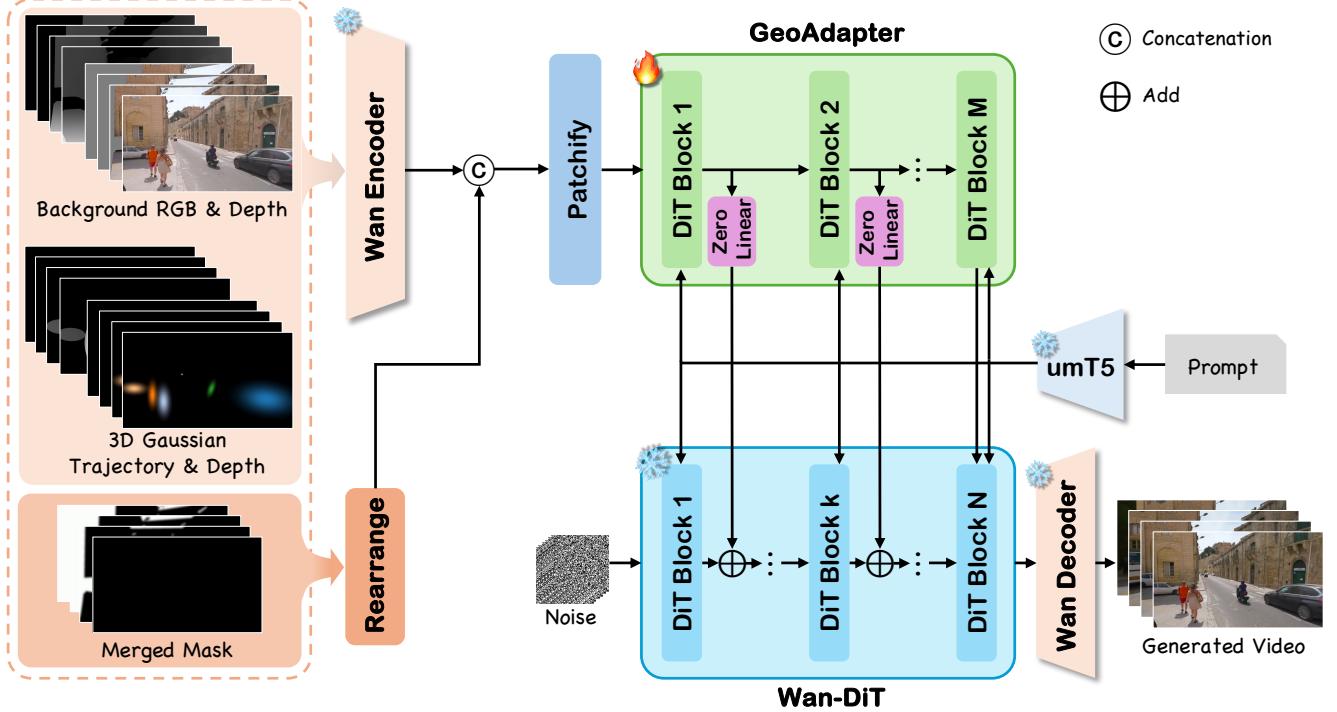


Figure 9. **Detailed architecture of VerseCrafter.** Background RGB & depth and 3D Gaussian trajectory RGB & depth are first encoded by the frozen 3D VAE. The soft control mask is rearranged into latent-aligned channels, and all geometry latents are then concatenated along the channel dimension to form a unified spatio-temporal geometry feature. This feature is patchified into tokens and processed by the GeoAdapter branch. At selected Wan-DiT blocks, GeoAdapter outputs are passed through zero-initialized linear layers and added as residual modulations to the backbone tokens, enabling 4D geometry-consistent camera and object control.

see Table 4 for the exact injection pattern and configuration. For each Wan-DiT block \mathcal{B}_n whose index n belongs to the injection set, with input tokens $\mathbf{x}_n \in \mathbb{R}^{L \times D}$ and geometry tokens \mathbf{g} , we add a geometry-conditioned residual of the form

$$\mathbf{x}_{n+1} = \mathcal{B}_n(\mathbf{x}_n) + \mathcal{G}_m(\mathbf{g}) \mathbf{W}_0^{(m)}, \quad (9)$$

where \mathcal{G}_m is the corresponding GeoAdapter block and $\mathbf{W}_0^{(m)} \in \mathbb{R}^{D \times D}$ is a zero-initialized linear projection. All entries of $\mathbf{W}_0^{(m)}$ are initialized to zero, so at the beginning of training VerseCrafter behaves identically to the original Wan2.1 backbone. During fine-tuning, $\mathbf{W}_0^{(m)}$ gradually learns to inject geometry information as a residual modulation, in the spirit of zero-initialized adapter designs in ControlNet-style architectures [117].

C. VerseControl4D Dataset Details

We construct **VerseControl4D**, a large-scale in-the-wild dataset for real-world 4D geometric control, where each clip is annotated with camera trajectories and multi-object 3D Gaussian trajectories. Control signals are automatically derived by the pipeline in main paper, producing background/3D Gaussian trajectories RGB and depth maps to-

gether with a merged mask.

VerseControl4D contains 35,000 training clips and 1,000 validation/test clips. Table 5 summarizes the data distribution by source and scene type. Overall, 26% of the clips come from Sekai-Real-HQ and 74% from SpatialVID-HQ, reflecting their complementary scene coverage. To support both camera-only world exploration and coordinated camera–object control, VerseControl4D includes *dynamic scenes* (clips with salient foreground object motion together with camera motion) and *static scenes* (clips with negligible object motion and only camera movement). About 20% of the training clips are *static scenes*, and the validation set additionally includes 250 static-scene clips for dedicated camera-only evaluation. Representative samples and their rendered 4D control signals are shown in Fig. 10.

D. Evaluation Metrics

D.1. VBench-I2V

We evaluate image-conditioned video quality using the VBench Image-to-Video (I2V) evaluation suite, denoted as VBench-I2V. For each generated clip, we follow the official VBench-I2V protocol: the conditioning image and its corresponding generated video are fed into the evaluation

Table 4. **Model configuration of VerseCrafter.** Settings include input resolution, number of Wan-DiT layers, GeoAdapter injection blocks, pre-trained backbone, and fine-tuning configuration.

	VerseCrafter
Resolution	720P
Num layers of Wan-DiT	40
GeoAdapter injection blocks	[0, 5, 10, 15, 20, 25, 30, 35]
Pre-trained backbone	Wan2.1 T2V-14B
Hidden dimension	5120
Batch size	16
Training iterations	5,000

Table 5. **VerseControl4D data split and scene-type statistics.** We report the number of clips from each source dataset and split. *Dynamic scenes* contain coupled camera and foreground object motion, while *static scenes* have negligible object motion and are used for camera-only evaluation.

Split	Sekai-Real-HQ		SpatialVID-HQ	
	Dynamic Scenes	Static Scenes	Dynamic Scenes	Dynamic Scenes
Train	9,071	7,000	18,929	
Val/Test	468	250	282	

pipeline, which computes a set of learned, human-aligned metrics that jointly capture video-image consistency and perceptual video quality. In our experiments, we report the following eight VBench-I2V dimensions, and define the *Overall Score* as the simple arithmetic mean of these eight normalized scores, where higher values indicate better performance:

- **Imaging Quality.** This metric measures low-level image fidelity, including sharpness and the absence of artifacts such as blur, noise, or overexposure. VBench uses an image quality predictor (e.g., MUSIQ), averaging scores across frames to obtain a video-level imaging quality score.
- **Aesthetic Quality.** This dimension assesses the artistic and aesthetic appeal of individual frames, including composition, color harmony, and realism. VBench applies an aesthetic quality predictor (e.g., the LAION aesthetic model) to each frame and averages the predictions over the clip.
- **Dynamic Degree.** This metric quantifies how dynamic the generated video is. Optical flow magnitudes (e.g., estimated by RAFT) are used to measure the amount of motion; the score reflects whether the model produces sufficiently active (non-static) content.
- **Motion Smoothness.** This metric evaluates whether subject and camera motion evolves smoothly and respects reasonable physical dynamics. VBench leverages a pre-trained video frame interpolation prior to assess how well intermediate motion can be interpolated, with smoother and more physically plausible motion achieving higher scores.
- **Background Consistency.** This dimension measures

temporal stability of the background layout and textures. Frame-level features (e.g., CLIP) are compared across time; large feature variations indicate flickering or unstable backgrounds and lead to lower scores.

- **Subject Consistency.** This dimension evaluates temporal consistency of the foreground subject *within* the video, regardless of the input image. VBench computes subject-region features across frames and measures their similarity over time to penalize identity drift or sudden appearance changes.
- **I2V Background (Video–Image Background Consistency).** This metric evaluates how well the global background in the video matches the background in the input image, especially for scene-centric inputs. VBench uses background-sensitive features (e.g., DreamSim) and aggregates image-frame and inter-frame similarities into a single background consistency score.
- **I2V Subject (Video–Image Subject Consistency).** This metric measures how well the main subject in the generated video matches the subject in the input image. VBench extracts high-level visual features (e.g., DINO) from the conditioning image and from each video frame, and combines image-frame similarities with inter-frame similarities into a weighted average subject consistency score.

Formally, given these eight per-dimension scores $\{s_k\}_{k=1}^8$ returned by VBench-I2V for a video, we define

$$\text{Overall Score} = \frac{1}{8} \sum_{k=1}^8 s_k, \quad (10)$$

which is the value reported as “Overall Score” in the main paper.

D.2. Rotation Error (RotErr)

To measure how well the generated camera motion follows the ground-truth camera trajectory, we adopt the camera-alignment metric from CameraCtrl [32]. For each generated video, we estimate its camera trajectory using the same geometry-annotation pipeline as for VerseControl4D, yielding rotation matrices $\{\mathbf{R}_{\text{gen}}^j\}_{j=1}^n$ and translation vectors $\{\mathbf{T}_{\text{gen}}^j\}_{j=1}^n$, where n is the number of frames. Let $\{\mathbf{R}_{\text{gt}}^j\}_{j=1}^n$ denote the corresponding ground-truth rotation matrices. The rotation error is computed by comparing the ground-truth and generated rotation matrices at each frame:

$$\text{RotErr} = \sum_{j=1}^n \arccos \left(\frac{\text{tr}(\mathbf{R}_{\text{gen}}^j \mathbf{R}_{\text{gt}}^j) - 1}{2} \right), \quad (11)$$

where $\text{tr}(\cdot)$ denotes the matrix trace. A lower RotErr indicates better alignment between the generated and ground-truth camera orientations.

D.3. Translation Error (TransErr)

We also evaluate the accuracy of the generated camera positions. Let $\{\mathbf{T}_{\text{gt}}^j\}_{j=1}^n$ and $\{\mathbf{T}_{\text{gen}}^j\}_{j=1}^n$ be the ground-truth and generated camera translation vectors for a video with n frames. Following CameraCtrl [32], the translation error is defined as the sum of per-frame Euclidean distances between the translation vectors:

$$\text{TransErr} = \sum_{j=1}^n \|\mathbf{T}_{\text{gt}}^j - \mathbf{T}_{\text{gen}}^j\|_2, \quad (12)$$

where \mathbf{T}_{gt}^j and $\mathbf{T}_{\text{gen}}^j$ denote the ground-truth and generated camera translation vectors at frame j , respectively. Smaller TransErr indicates that the generated camera positions more closely match the ground-truth camera positions.

D.4. Object Motion Control (ObjMC)

For object-motion control, we follow the ObjMC metric proposed in MotionCtrl [96] and extend it to the multi-object setting in our 3D Gaussian trajectory space. Given a generated video, we run the same geometry-annotation pipeline as in VerseControl4D to estimate per-object 3D Gaussian trajectories, and compare them with the corresponding ground-truth trajectories from our dataset.

Let N_{gt} and N_{pred} denote the numbers of ground-truth and predicted controlled objects in a sample, and let T be the number of frames. For each ground-truth object $o \in \{1, \dots, N_{\text{gt}}\}$ and frame $t \in \{1, \dots, T\}$, we denote the ground-truth 3D Gaussian mean by $\boldsymbol{\mu}_o^{(t)} \in \mathbb{R}^3$ and the estimated mean from the generated video by $\hat{\boldsymbol{\mu}}_k^{(t)} \in \mathbb{R}^3$ for a predicted object k .

Multi-object matching. Since N_{gt} and N_{pred} may differ, we first define the trajectory distance between a ground-truth object o and a predicted object k as the average Euclidean distance of their 3D Gaussian means over time:

$$d(o, k) = \frac{1}{T} \sum_{t=1}^T \|\hat{\boldsymbol{\mu}}_k^{(t)} - \boldsymbol{\mu}_o^{(t)}\|_2. \quad (13)$$

We then build a cost matrix $\mathbf{C} \in \mathbb{R}^{N_{\text{gt}} \times N_{\text{pred}}}$ with entries $C_{ok} = d(o, k)$. To handle unmatched objects, we pad this matrix with dummy rows/columns and fill them with a constant penalty λ (set to 10.0 m in our experiments). Finally, we apply the Hungarian algorithm [?] to this padded matrix to obtain an optimal one-to-one matching between ground-truth and predicted trajectories. This step assigns each ground-truth object either to a predicted trajectory (matched) or to a dummy entry (missed), and symmetrically accounts for spurious predicted objects.

ObjMC score. Given the optimal matching, we define the per-object trajectory error for a ground-truth object o as

$$d_o = \begin{cases} d(o, k) & \text{if } o \text{ is matched to a predicted object } k, \\ \lambda & \text{if } o \text{ is unmatched,} \end{cases} \quad (14)$$

and compute the final ObjMC score as the average over all ground-truth controlled objects:

$$\text{ObjMC} = \frac{1}{N_{\text{gt}}} \sum_{o=1}^{N_{\text{gt}}} d_o. \quad (15)$$

Lower ObjMC values indicate more accurate multi-object 3D motion control, and the unmatched-penalty λ ensures that both missed objects and spurious trajectories are appropriately penalized.

E. Additional Qualitative Results

We provide additional qualitative comparisons on VerseControl4D, following the same evaluation settings and baselines as in the main paper. Figures 11 and 12 showcase *dynamic scenes* with joint camera–object control. Perception-as-Control often yields low-clarity frames and noticeable camera mis-tracking, while Yume may capture coarse motion intent but fails to precisely align object trajectories with the target camera path. Uni3C is restricted to human motion and struggles to generalize to multi-object dynamics. In contrast, VerseCrafter consistently adheres to both camera trajectories and multiple object motions, preserving object identity and shape over time and maintaining geometrically coherent backgrounds.

Figures 13 and 14 present *static scenes* for camera-only exploration. We observe that ViewCrafter, Voyager, and FlashWorld can introduce structural distortions,

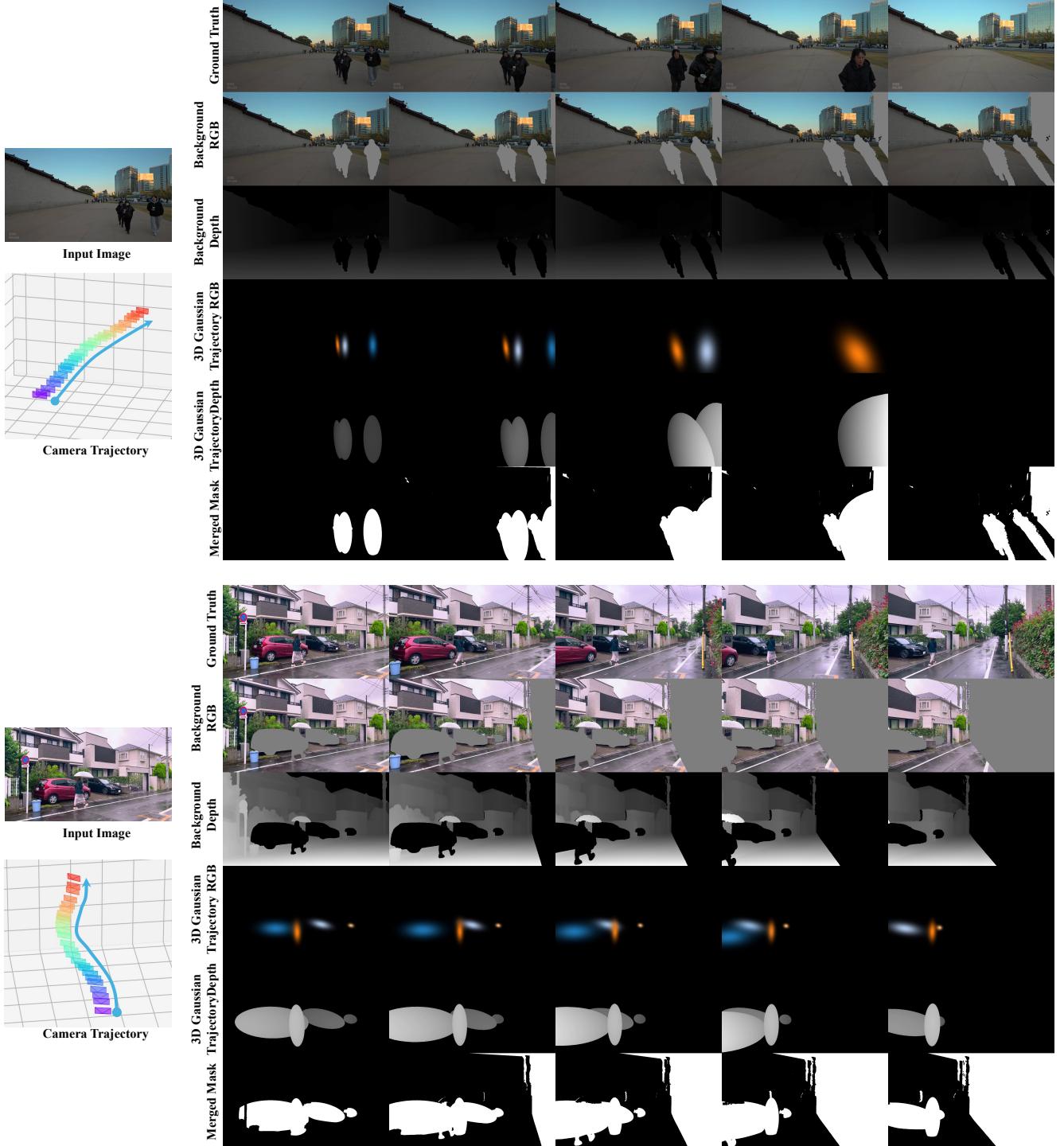


Figure 10. VerseControl4D dataset examples. For each clip, we visualize the input image and target camera trajectory (left), followed by several frames of ground-truth video and our rendered control signals (right): background RGB/depth, 3D Gaussian trajectory RGB/depth for controlled objects, and the final merged mask. These signals are automatically derived by our annotation pipeline in main paper.

depth/parallax instability, or temporal flicker when following long or curved camera paths. VerseCrafter produces smoother camera motion with faithful parallax, keeping

background layout stable and details sharp across frames. These additional cases further confirm VerseCrafter’s robustness in real-world 4D control for both dynamic and

static settings.

F. Limitations and Future Work

Despite the encouraging results, VerseCrafter has several limitations that suggest promising directions for future work. First, while VerseCrafter enforces 4D geometric consistency through explicit camera and 3D Gaussian trajectory controls, it does not impose *explicit physical constraints* during generation. As a result, the model may occasionally produce motion that is geometrically plausible yet physically imperfect, such as subtle sliding, interpenetration, or dynamics that deviate from real-world contact and inertia. In future work, integrating stronger physics priors—e.g., collision-aware losses, contact/ground constraints, or differentiable physics guidance—could improve physical realism and controllability in complex interactions.

Second, VerseCrafter is computationally expensive at high resolution and long horizons, since it conditions a large frozen video diffusion backbone and renders multi-channel 4D controls per frame. Our current 81-frame 720P generation requires substantial GPU memory and runtime, limiting interactive use. Future work may explore more efficient backbones, distilled or cached control encoding, and streaming/long-video synthesis to scale VerseCrafter to faster and longer world rollouts.

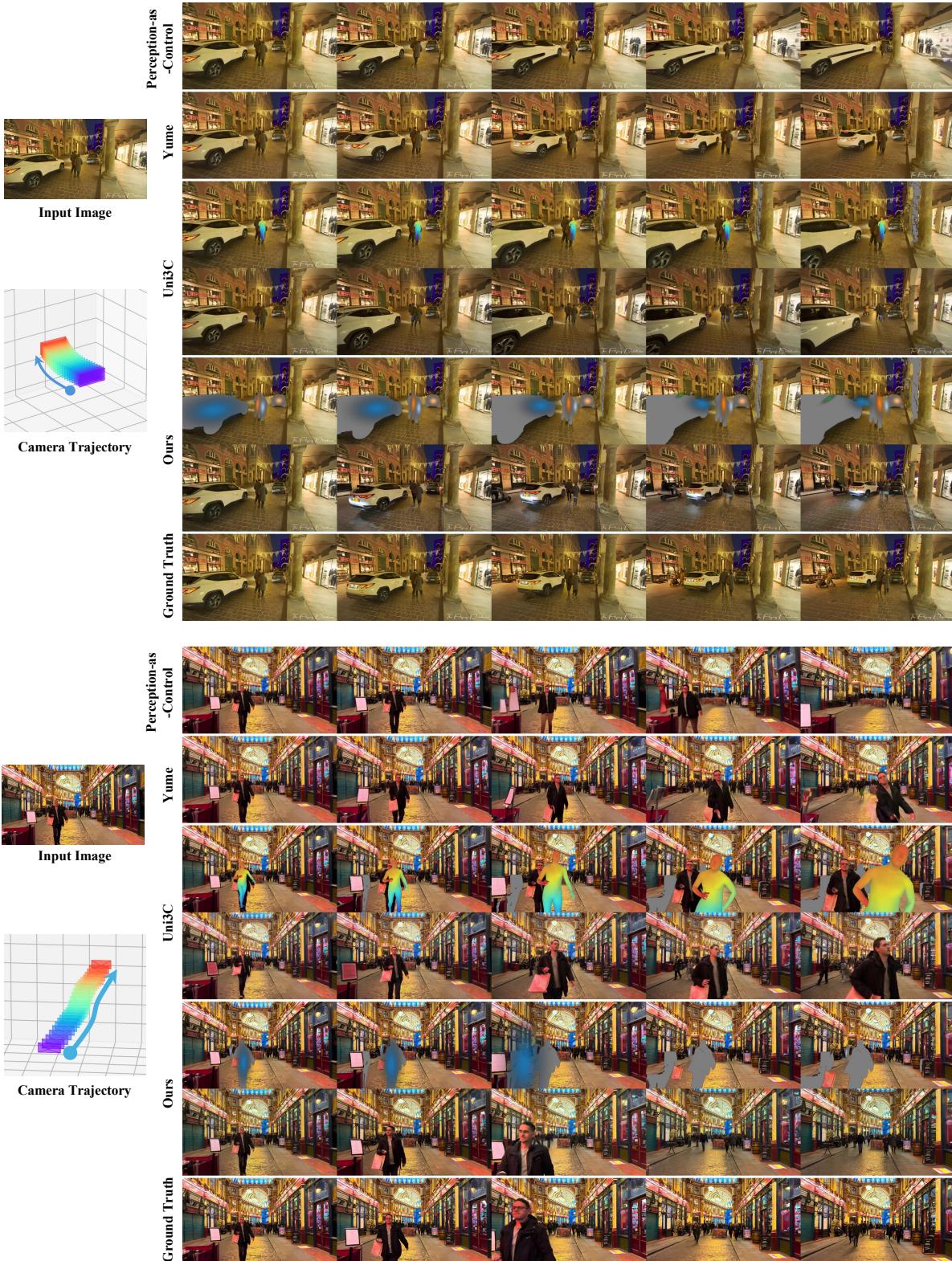


Figure 11. Additional qualitative comparison of joint camera and object motion control on dynamic scenes. Perception-as-Control often produces low-fidelity frames with inaccurate camera motion; Yume roughly follows text-described motion but lacks precise geometric control; Uni3C is mainly limited to human-centric motion. VerseCrafter more faithfully follows both the target camera trajectory and multi-object motions while maintaining sharp appearance and geometrically consistent backgrounds.

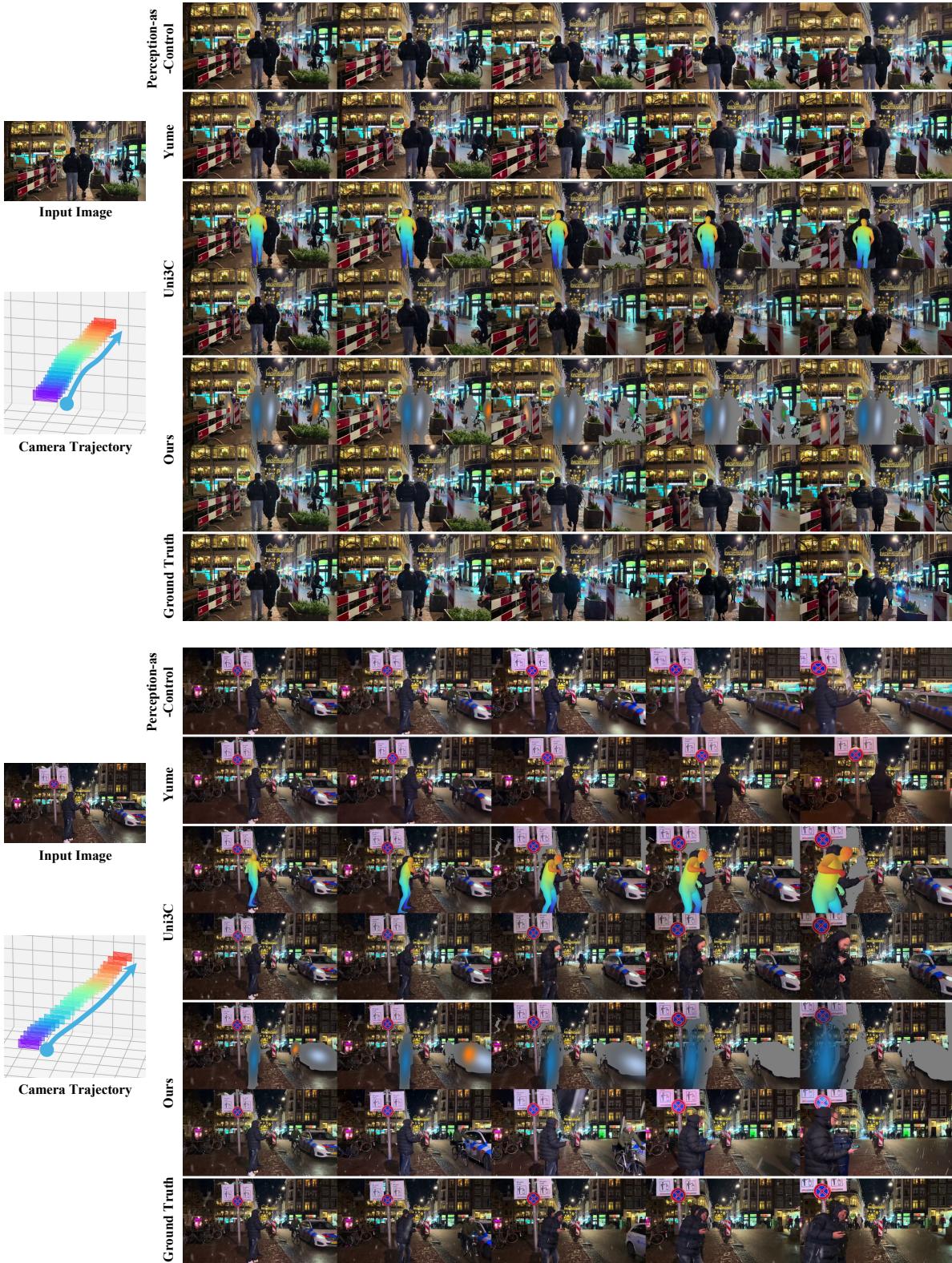


Figure 12. Additional qualitative comparison of joint camera and object motion control on dynamic scenes. Across diverse real-world cases, baselines frequently suffer from camera drift, motion misalignment, or object identity/shape inconsistency. VerseCrafter preserves scene geometry and object coherence over time, yielding accurate multi-object 3D motion along the specified camera path.

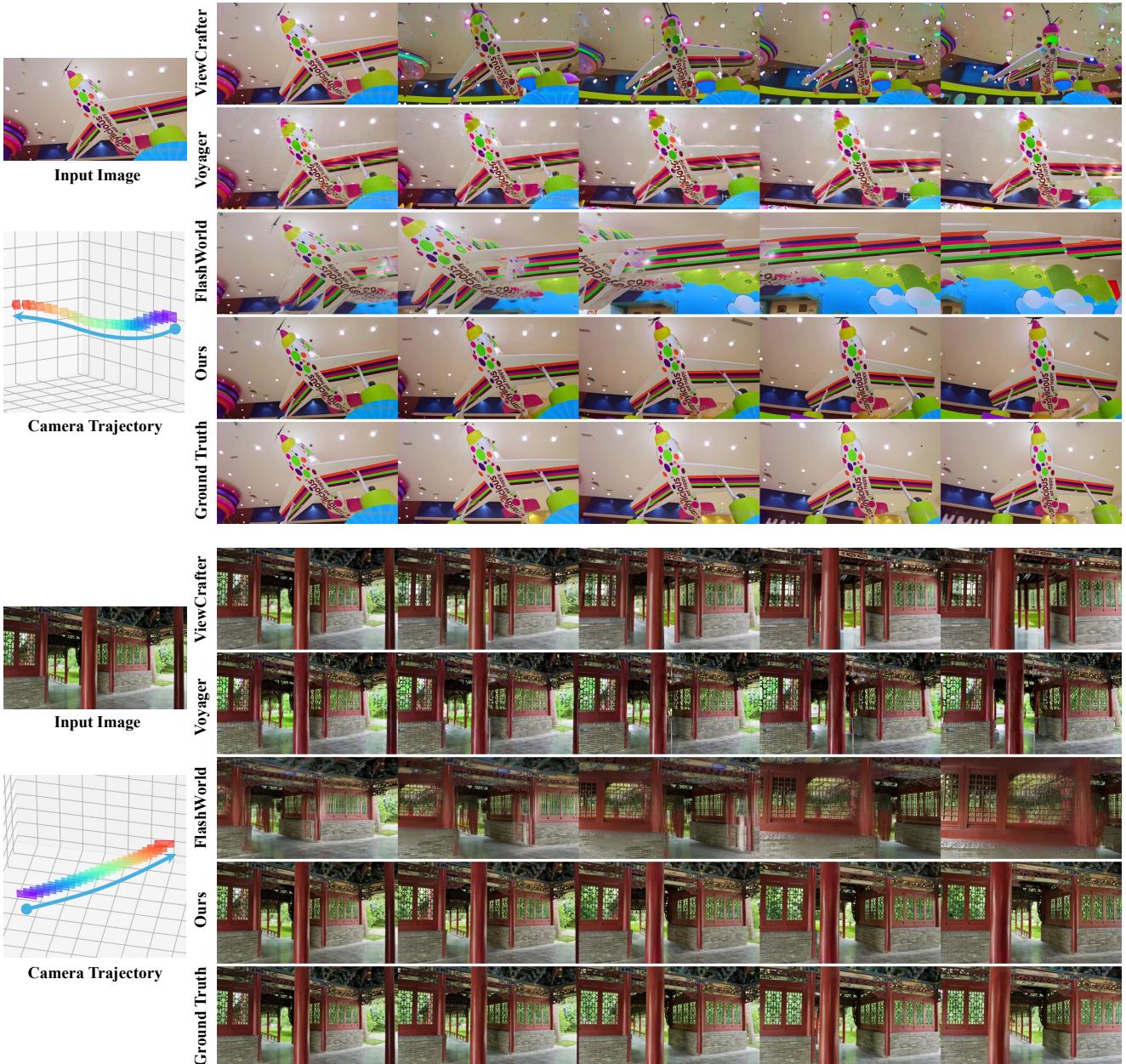


Figure 13. Additional qualitative comparison of camera-only motion control on static scenes. ViewCrafter, Voyager, and FlashWorld often exhibit distorted facades, drifting structures, or inconsistent parallax along the camera path. VerseCrafter better follows the target trajectory while preserving sharp details and globally consistent 3D geometry.

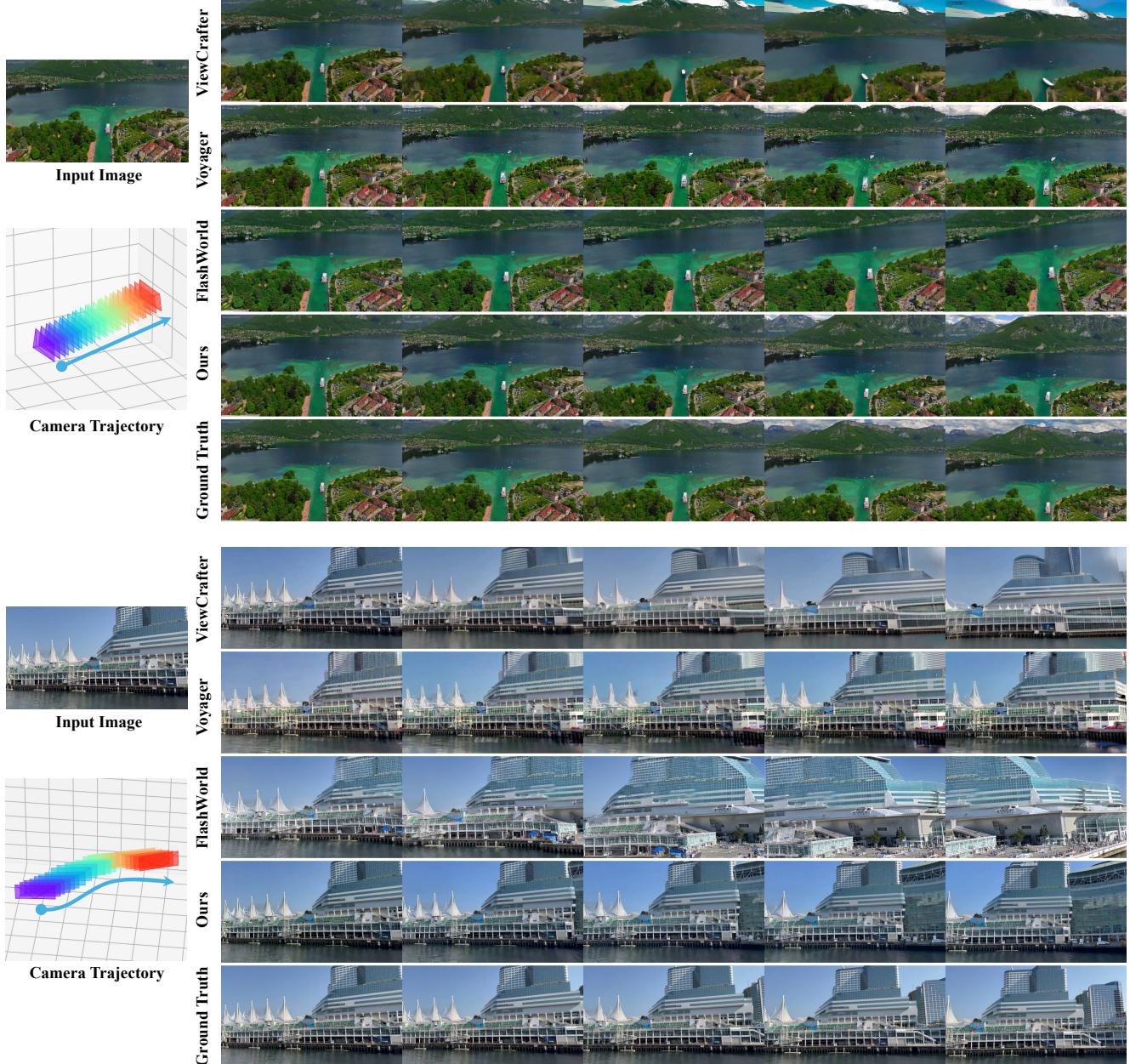


Figure 14. **Additional qualitative comparison of camera-only motion control on static scenes.** Baselines may introduce structural warping, background flicker, or unstable depth cues when exploring long camera paths. VerseCrafter maintains stable parallax and texture consistency, producing smooth camera motion with faithful 3D scene structure.