Captain Safari: A World Engine

Yu-Cheng Chou[1] Xingrui Wang[1] Yitong Li[2] Jiahao Wang[1] Hanting Liu[1]
Cihang Xie[3] Alan Yuille[1] Junfei Xiao[1]✉
[1]Johns Hopkins University [2]Tsinghua University [3]UC Santa Cruz

https://johnson111788.github.io/open-safari/

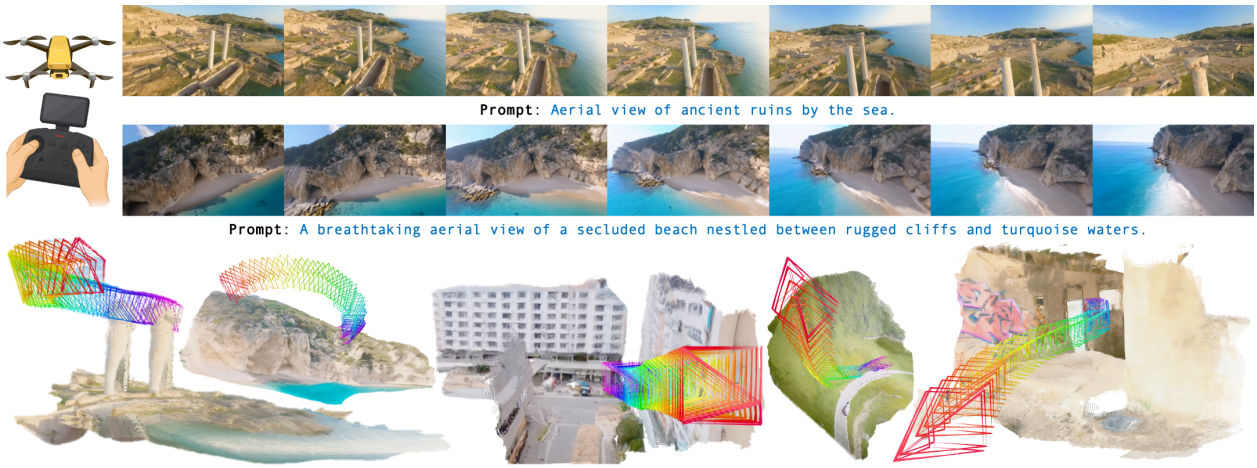

Figure 1. **Captain Safari** is a pose-aware world engine that generates long-horizon, 3D-consistent FPV videos from any user-specified camera trajectory. By retrieving pose-aligned world memory, it keeps geometry stable across large viewpoint changes and reconstructs crisp, well-formed structures while faithfully tracking aggressive 6-DoF motion.

Abstract

World engines aim to synthesize long, 3D-consistent videos that support interactive exploration of a scene under user-controlled camera motion. However, existing systems struggle under aggressive 6-DoF trajectories and complex outdoor layouts: they lose long-range geometric coherence, deviate from the target path, or collapse into overly conservative motion. To this end, we introduce Captain Safari*, a pose-conditioned world engine that generates videos by retrieving from a persistent world memory. Given a camera path, our method maintains a dynamic local memory and uses a retriever to fetch pose-aligned world tokens, which then condition video generation along the trajectory. This design enables the model to maintain stable 3D structure while accurately executing challenging camera maneuvers.*

To evaluate this setting, we curate OpenSafari*, a new in-the-wild FPV dataset containing high-dynamic drone videos with verified camera trajectories, constructed through a multi-stage geometric and kinematic validation pipeline. Across video quality, 3D consistency, and trajectory following,* Captain Safari *substantially outperforms state-of-the-art camera-controlled generators. It reduces MEt3R from 0.3703 to 0.3690, improves AUC@30 from 0.181 to 0.200, and yields substantially lower FVD than all camera-controlled baselines. More importantly, in a 50-participant, 5-way human study where annotators select the best result among five anonymized models,* ***67.6%*** *of preferences favor our method across all axes. Our results demonstrate that pose-conditioned world memory is a powerful mechanism for long-horizon, controllable video generation and provide* OpenSafari *as a challenging new benchmark for future world-engine research.*

1. Introduction

Simulating coherent 3D worlds through controllable video generation has long been a foundational challenge for augmented reality, embodied AI, and virtual agents [8–10, 13–16, 19, 20, 23, 26, 29, 40, 43–45, 50, 51, 57]. Classical game engines and physics simulators offer explicit geometry and precise control, but require heavy manual authoring and expensive computation [7, 28, 32]. More-

✉ Corresponding author: Junfei Xiao (xiaojf97@gmail.com)
Preprint, work in progress.

over, while they may achieve visual realism in specialized domains, they still fall short in capturing the richness and diversity characteristic of real world, such as natural scenes [27, 35, 58]. In contrast, modern video diffusion models synthesize high-fidelity, diverse videos from text or images, yet typically operate as feed-forward clip generators without persistent world state: *they struggle with long-range 3D consistency, complex trajectory following, and faithful reconstruction of diverse scenes* [18, 24]. In this work, we move toward bridging this gap with *Captain Safari*, a world engine that enables pose-conditioned modeling of 3D-consistent and diverse environments, surpassing the limitations of traditional game engines in terms of generality, diversity, and interactivity.

Contemporary video world models face three intertwined challenges. First, *long-horizon consistency* is limited by the temporal window of context frames; models often "forget" distant scenery or violate spatial coherence, leading to abrupt appearance changes that break the realism and continuity of the generated environment [8, 14, 45]. Second, achieving *complex camera maneuvers under strict 3D consistency* remains difficult: existing pose- or trajectory-conditioned methods typically work well only for slow, near-forward motions [16, 34, 49]. When the path involves fast 6-DoF movement, strong parallax, or sharp turns, models exhibit a trade-off—either dampening motion and restricting viewpoint changes to preserve geometry, or committing to the requested path at the cost of distortions, flicker, and structural drift. Third, current approaches underrepresent *complex outdoor layouts*. Much of the works focuses on structured, constrained settings (e.g., indoor tours, driving scenes, or real-estate videos), and models are seldom stress-tested in in-the-wild FPV scenarios where the camera weaves around buildings, vegetation, and varied terrain with substantial parallax [6, 22, 59, 60]. As a result, methods that look competitive in simplified environments often fail to preserve geometry and appearance when confronted with truly diverse, complex outdoor scenes.

To address these issues, we introduce *Captain Safari*, a pose-aware world engine that explicitly maintains a persistent notion of world state to uphold *long-horizon* 3D consistency across strong parallax. Because storing and propagating a full long-term state is computationally prohibitive, we develop a retrieval mechanism that *selects and aggregates* only the most informative scene cues, thereby providing strong geometric guidance without incurring prohibitive cost. Crucially, this retrieval is *pose-aware*: given the target camera pose, it assembles a pose-aligned world prior that steers the generation process, enabling accurate tracking of *aggressive camera maneuvers* while preserving 3D-consistent structure in complex environments.

Furthermore, to close the gap in *complex outdoor layouts* and *aggressive camera motion*, we curate *OpenSafari*, a large-scale dataset of high-dynamic FPV drone videos with verified camera poses. Much of the literature targets structured, constrained settings (e.g., indoor tours, driving or real-estate videos), and even outdoor datasets typically feature slow, near-forward motion. In contrast, *OpenSafari* comprises in-the-wild FPV flights that weave around buildings and vegetation across uneven terrain, exhibiting large parallax, rapid 6-DoF maneuvers, and sharp viewpoint changes. Paired with verified camera trajectories, these videos present diverse, cluttered outdoor scenes and long-range motion, challenging models to maintain 3D consistency while faithfully tracking complex maneuvers.

We evaluate *Captain Safari* along three axes: *video quality*, *3D consistency*, and *trajectory following*. Across these criteria, our method consistently outperforms contemporary camera-controlled video generators on *OpenSafari*: Table 1 reports clear gains in 3D consistency and accurate tracking under complex maneuvers, while maintaining strong perceptual quality. Importantly, a large-scale human study (Table 2) shows that *Captain Safari* receives **67%** of votes in five-way comparisons, indicating that the improvements are perceptually salient. Qualitative comparisons (Fig. 4 and Fig. 5) further demonstrate stable geometry under long-range path and faithful adherence to sharp 6-DoF camera turns in cluttered outdoor scenes.

In summary, our contributions are:

1. We present *Captain Safari*, the first camera-controlled video generation method to enforce long-horizon 3D consistency while tracking aggressive FPV maneuvers.
2. We propose a *pose-guided, long-horizon retrieval* that efficiently reconciles strict 3D consistency with accurate tracking of complex maneuvers.
3. We curate *OpenSafari*, a large-scale in-the-wild FPV dataset with verified camera poses, featuring diverse, cluttered outdoor scenes and rapid 6-DoF motion that stress-test geometry-consistent camera control.
4. In *OpenSafari*, our pose-aware retrieval notably improves video quality, 3D consistency, and trajectory alignment, also achieving a **67%** human preference rate.

2. Related Work

2.1. 3D-Consistent World Models

Early image-to-3D approaches reconstruct geometry indirectly via multi-view consistency or implicit fields, but often fail to maintain coherent structure across large view changes [13, 21, 43, 50]. Recent efforts integrate 3D reasoning into the generative process. DiffusionGS [5] injects Gaussian Splatting into the diffusion denoiser, enforcing view consistency and enabling single-stage, scalable 3D generation. GenEx [23] and EvoWorld [40] extends this idea from static reconstruction to dynamic world creation, generating explorable 360° panoramic environments

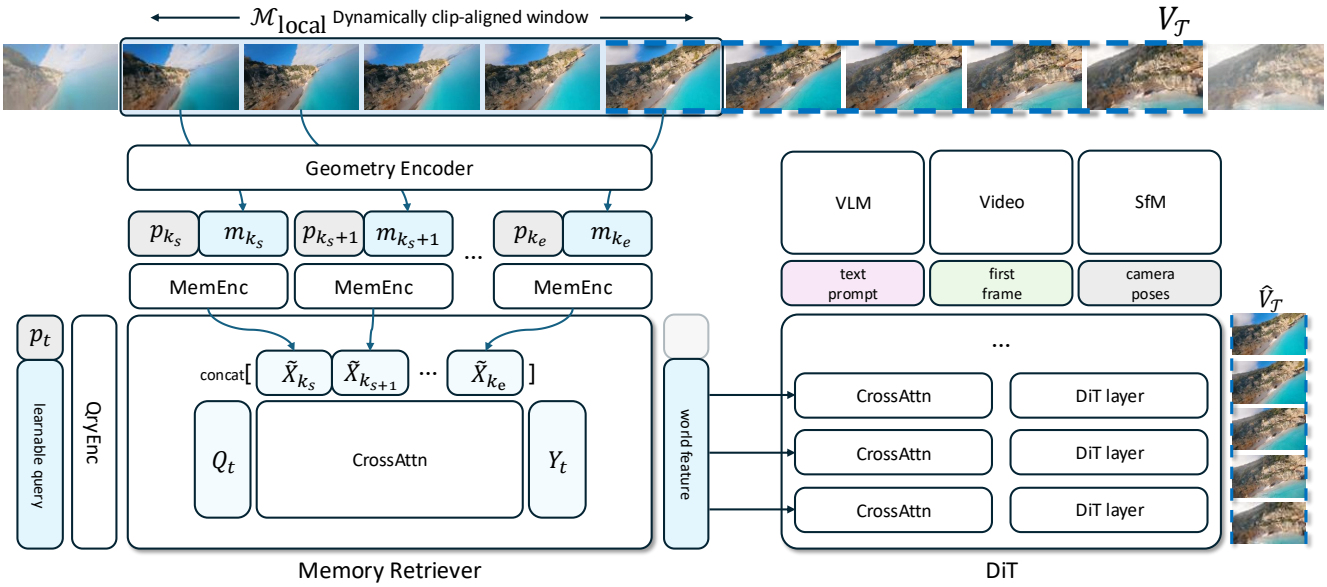

Figure 2. **Method overview.** *Captain Safari* builds a local world memory and, given a query camera pose, retrieves pose-aligned tokens that summarize the scene. These tokens then condition video generation along the user-specified trajectory, preserving a stable 3D layout.

grounded in physical priors. Complementary to these generative reconstructions, Geometry Forcing [44] and Memory Forcing [14] explicitly couple training signals with geometric supervision and spatio-temporal memory, ensuring consistency during long rollouts. Meanwhile, open-world models such as Wonderland [20], WonderWorld [51], WonderTurbo [26], and EvoWorld [40] further integrate geometry-indexed or adaptive memories to maintain persistent world states across interactions. However, these approaches still use implicit, clip-bound memories, whereas we introduce an explicit pose-indexed world memory retrieved on demand for camera-controlled generation.

2.2. Camera Controlled Video Generation

Early T2V/I2V models learned camera motion implicitly and struggle to reliably repeat explicit trajectories [11, 42, 61]. Recent work such as CameraCtrl [10] treats camera parameters as explicit conditions, encoding camera extrinsics and trajectories or enforcing path constraints—to improve controllability and accuracy [2, 25, 41, 47, 55]. MotionPrompting [9] implement compositional control by point-track conditioning and MotionPro [56] use path-alignment losses that lower rotational and translational error; training-free control is also achieved by fitting a lightweight point-cloud and using a noise-layout prior to steer denoising [11]. Scene-preserving geometric priors further strengthen clip-level consistency. Cami2V[57] treats camera pose as a physical prior and exploits epipolar and multiview constraints; RealCam-I2V [19] recovers metric depth with DepthAnything v2 [48] to reconstruct a scale-stable scene; PoseTraj[16] employs pose-aware pretraining to obtain rotation-aligned motion. Compared with parameter-only conditioning, these priors reduce within-clip layout drift and better preserve local geometry under view changes. Further, recent work links camera control with world modeling. CVD [17], Cavia [46], and WoVoGen [22] jointly synthesize multi-view and multi-trajectory videos from a shared scene representation, enforcing cross-path consistency. Meanwhile, methods that conditioning from explicit renderable 3D representations (e.g., 3D Gaussians) can anchor geometry, improve cross-view 3D consistency and path adherence [15, 29, 33, 52, 53]. However, these approaches typically build one-off 3D scenes, whereas we unify long-horizon camera control with a persistent pose-indexed world memory shared across trajectories.

3. Captain Safari

We introduce *Captain Safari*, a memory-guided video generation framework. Sec. 3.1 presents an implicit world memory for stable scene representation, while Sec. 3.2 describes a pose-conditioned retrieval system that maps camera views to world tokens, guiding a DiT-based generator for coherent outputs along arbitrary trajectories.

3.1. Implicit Memory of World Geometry

Problem setup. We represent a video as $V = \{I_t\}_{t=0}^{T}$, where I_t is the frame at time step t. On the same time axis we define camera poses $\mathcal{C} = \{(R_t, T_t)\}_{t=0}^{T}$ and obtain a 3D-aware memory feature m_t at each time step t using a pretrained geometry encoder. All memory features form a global memory bank $\mathcal{M} = \{m_t\}_{t=0}^{T}$.

Given a text prompt p, the camera poses $\mathcal{C}$, and a target

clip time step $\mathcal{T} = [t_0, t_1]$, together with its associated local world memory $\mathcal{M}_{\text{local}} \subset \mathcal{M}$, our goal is to synthesize a video segment $\hat{V}_{\mathcal{T}}$ that (i) aligns with p, (ii) respects the prescribed poses $\{(R_t, T_t)\}_{t\in\mathcal{T}}$, and (iii) maintains a coherent 3D world across viewpoints.

Local world memory. Directly conditioning on the full memory bank $\mathcal{M}$ for every clip would be computationally expensive and dominated by temporally distant observations. Instead, for each target clip time step $\mathcal{T} = [t_0, t_1]$ we define a *local* memory $\mathcal{M}_{\text{local}} = \{m_\tau \mid \tau \in [k_s, k_e]\}$ whose endpoints are sampled under

$$\begin{aligned} t_0 - L \;&\le k_s \;\le\; t_0, \\ \max(k_s,\, t_0){+}1 \;&\le k_e \;\le\; \min(k_s{+}L,\, t_1), \end{aligned} \tag{1}$$

where L is a fixed bound and all time steps are integers. These constraints enforce that: (i) the memory window starts at most L seconds before the clip entrance t_0, tying it to nearby observations; (ii) its duration is at most L, which keeps the conditioning set compact; and (iii) its end time k_e always touches or overlaps t_0 while remaining within $[t_0, t_1]$, ensuring that each clip is supported by a temporally compatible world prior. All $\mathcal{M}_{\text{local}}$ are constructed as such dynamic clip-aligned window of the shared bank $\mathcal{M}$, so neighboring clips naturally share overlapping memory entries, constraining computation while coupling their generations to a 3D-consistent underlying world.

Pose-retrieved memory. Within a given clip time step $\mathcal{T}$, we treat the local memory $\mathcal{M}_{\text{local}}$ as a static hypothesis of the surrounding world built from key frames. Each time step τ provides a pose token p_τ (derived from (R_τ, T_τ)) and a set of 3D-aware memory tokens $m_{\tau,1}, \ldots, m_{\tau,M}$. The collection $\{(p_\tau, m_{\tau,1:M})\}_\tau$ forms an implicit world table: pose token indicates *where* the camera has observed the scene, while memory tokens encode *what* the world looks like from those configurations. For any target time step $t \in \mathcal{T}$, we derive its camera pose to a query pose token p_t, embed it as $q_t = \phi_p(p_t)$, and use a dedicated retrieval module to read from this static table in a pose-dependent manner. Concretely, q_t is concatenated with a bank of learnable query tokens and processed into retrieval queries, which perform cross-attention over the encoded memory $\tilde{X}^{\text{mem}}$ (defined in Sec. 3.2), yielding a set of world tokens

$$w_t = \text{Agg}\Big(\text{CrossAttn}(Q_t,\; \tilde{X}^{\text{mem}})\Big). \tag{2}$$

corresponding to the updated learnable queries. These pose-aligned world tokens w_t are directly used as the reconstructed memory at pose t. Thus, all frames in $\mathcal{T}$ access local memory through pose-conditioned queries instead of raw time indices, encouraging multi-view observations to remain tied to a consistent static 3D world.

3.2. Memory Retrieval and Conditioning

Memory retriever design. As shown in Figure 2, given the local memory, we represent each time step τ by a pose token p_τ and its associated memory tokens $m_{\tau,1:M}$. Our retriever is designed to (i) jointly encode pose–memory pairs into a coherent world representation, and (ii) extract, for any query pose, a compact set of pose-aligned tokens that summarize the most relevant parts of this local world.

We first embed pose and memory features into a shared space and form a joint sequence per time step:

$$\hat{X}_\tau = \big[\phi_p(p_\tau),\; \phi_m(m_{\tau,1}), \ldots, \phi_m(m_{\tau,M})\big], \tag{3}$$

where ϕ_p and ϕ_m denote learnable embeddings for pose and memory tokens, respectively. A stack of transformer blocks (MemEnc) with 3D-aware positional encoding refines these sequences,

$$\tilde{X}_\tau = \text{MemEnc}\big(\hat{X}_\tau\big), \tag{4}$$

and we obtain the encoded local world memory by concatenation

$$\tilde{X}^{\text{mem}} = \big[\tilde{X}_{k_s}, \ldots, \tilde{X}_{k_e}\big], \tag{5}$$

optionally masked to exclude padded or non-key entries.

For a target time step t, we derive the query pose token p_t, embed it as $q_t = \phi_p(p_t)$, and concatenate it with M learnable query tokens $r_1, \ldots, r_M$,

$$\hat{Q}_t = [q_t, r_1, \ldots, r_M]. \tag{6}$$

This sequence is refined by transformer blocks sharing the same architecture as MemEnc, denoted as QryEnc, yielding pose-aware retrieval queries

$$Q_t = \text{QryEnc}\big(\hat{Q}_t\big). \tag{7}$$

We then perform cross-attention from Q_t to the encoded memory $\tilde{X}^{\text{mem}}$,

$$Y_t = Q_t \;+\; \text{CrossAttn}(Q_t,\; \tilde{X}^{\text{mem}}), \tag{8}$$

and take the subset of tokens in Y_t corresponding to the learnable queries as the retrieved world tokens

$$w_t = [w_{t,1}, \ldots, w_{t,M}], \tag{9}$$

which form a pose-aligned world feature for time t. During training, a linear head maps w_t back to the original memory space to reconstruct the target memory tokens at the query pose. Stacking multiple retrieval blocks iteratively refines both the queries and the retrieved tokens, enabling the model to softly route each query pose to the most relevant subset of past observations, instead of relying on a rigid temporal neighborhood or a single nearest frame.

Memory-conditioned DiT. For a given target clip time step $\mathcal{T}$, the retriever consumes $\mathcal{M}_{\text{local}}$ and the query pose p_t and outputs a pose-aligned set of world tokens $w_t \in \mathbb{R}^{M\times d_m}$, which summarize the static local world relevant to this segment. These tokens are mapped into the DiT hidden space by the memory embedding MLP

$$W_{\mathcal{T}} = \phi_w(w_t) \in \mathbb{R}^{M\times D}. \tag{10}$$

Figure 3. ***OpenSafari***. A new in-the-wild FPV dataset with rigorously verified camera trajectories, designed to stress-test geometry-consistent, camera-controllable video generation. We curate clips through a compact, multi-stage pipeline that filters, reconstructs, and verifies trajectories, yielding clean, motion-rich videos with reliable camera paths.

The latent clip is encoded as a single spatio-temporal token sequence $Z \in \mathbb{R}^{L_z \times D}$, obtained by patchifying all frames in $V_{\mathcal{T}}$. At each DiT layer l, we first apply self-attention over the full sequence and then inject the world tokens through a dedicated memory cross-attention:

$$Z^{(l+1)} = Z^{l} + \mathrm{CrossAttn}(Z^{l}, W_{\mathcal{T}}, W_{\mathcal{T}}). \quad (11)$$

The clip-level world tokens $W_{\mathcal{T}}$ are reused as keys and values across all layers, providing a stable, 3D-consistent prior that shapes the denoising of every spatio-temporal token.

4. OpenSafari

4.1. Video Data Curation

Existing camera-conditioned datasets do not match our target regime. RealEstate10K [59] focuses on slow, mostly indoor real-estate walkthroughs with gentle motion and clean, quasi-static scenes, while Minecraft [4] is a synthetic voxel world with simplified geometry and engine-constrained dynamics. Neither captures aggressive, in-the-wild 6-DoF drone flight with strong parallax, large elevation changes, and complex outdoor layouts that truly stress long-horizon 3D consistency. We therefore propose *OpenSafari*, a new dataset of real-world FPV-style drone videos with verified camera trajectories tailored to this challenging setting.

We construct Safari-FPV from FPV-style drone videos collected on AirVuz[1] and YouTube[2], and retain only clips that pass a strict multi-stage preprocessing pipeline. As shown in Figure 3, we: (i) download the highest available resolution for each URL and discard sources below the target resolution; (ii) normalize all videos to 720p, 24 fps, and a fixed 16:9 center crop, removing letterboxing and black borders so that subsequent camera estimation operates on a clean field of view; (iii) run scene detection to obtain single-shot segments; (iv) split segments into fixed-length T videos via uniform temporal slicing.

We then filter videos with a single diagnostic based on motion. Specifically, we run RAFT [36] to estimate optical-flow magnitude; videos with too little motion are removed, while videos with stable, coherent motion are kept to emphasize informative, parallax-rich trajectories rather than static views. Only videos satisfying the motion constraint enter the final dataset. This yields a large-scale, in-the-wild drone corpus explicitly tailored to stress-test geometry-aware, trajectory-following video generation.

4.2. Camera Trajectory Reconstruction

For each curated video, we estimate camera intrinsics and extrinsics at 4 fps using Hierarchical Localization [30, 31]. We extract local features, build exhaustive image pairs within each video, run feature matching, and reconstruct a COLMAP-style SfM model; from this model we export

[1] https://www.airvuz.com/
[2] https://www.youtube.com/

Table 1. **Benchmark camera-controlled video generation.** *Captain Safari* ranks first in 3D consistency and trajectory following with competitive video quality. Compared to the ablated variant without memory, *Captain Safari* substantially improves 3D consistency and trajectory following, with only a slight trade-off in video quality. (Recon. = reconstruction rate. CosSim = cosine similarity.)

Model	Video Quality		3D consistency		Trajectory Following		
	FVD ↓	LPIPS ↓	MEt3R ↓	Recon. ↑	AUC@30 ↑	AUC@15 ↑	CosSim ↑
Geometry Forcing [44]	2662.75	0.667	0.4834	0.877	0.168	0.056	0.429
Real-CamI2V [19]	1585.61	0.513	<u>0.3703</u>	<u>0.923</u>	0.174	0.051	0.296
Wan2.2-5B-Control-Camera [38]	1387.75	0.545	0.3932	0.767	0.181	0.054	0.420
Captain Safari w/o *Mem.*	**998.47**	**0.504**	0.3720	0.912	<u>0.193</u>	**0.068**	<u>0.508</u>
Captain Safari	<u>1023.46</u>	<u>0.512</u>	**0.3690**	**0.968**	**0.200**	**0.068**	**0.563**

Table 2. **Human preference.** Users overwhelmingly prefer *Captain Safari* across all criteria, capturing **67%** of total votes. The memory-removed variant ranks a distant second, while baselines competitors receive single-digit preference.

Model	Video Quality	3D consistency	Trajectory Following	Average
Geometry Forcing [44]	0.20%	0.00%	0.20%	0.13%
Real-CamI2V [19]	4.20%	6.40%	4.40%	5.00%
Wan2.2-5B-Control-Camera [38]	3.20%	3.80%	6.40%	4.47%
Captain Safari w/o *Mem.*	25.00%	24.20%	20.00%	23.07%
Captain Safari	**67.40%**	**65.60%**	**69.00%**	**67.33%**

per-frame camera parameters as initial trajectories.

To obtain deployment-ready data, we apply a three-stage verification-and-fix pipeline to every reconstructed trajectory. First, *database check* consumes SfM statistics (inlier counts and ratios) to flag potentially unreliable transitions. Next, s*geometric check* revisits suspicious pairs using stored keypoints and matches, recomputes essential matrices, and thresholds symmetric epipolar errors. Last, *kinematics check* analyzes the pose sequence for translation spikes, rotation jumps, forward-direction flips, and higher-order smoothness violations, using robust MAD-based scores to detect implausible motion.

The per-transition decisions are fused into a binary bad-index, which drives a strict policy. If bad transitions are sparse and localized, we invoke a targeted fix: we linearly interpolate camera centers and apply SLERP to rotations with a capped interpolation angle, optionally extrapolating at video boundaries. The fixed segments are then re-validated by the same database/geometric/kinematics criteria. If post-fix validation succeeds, the trajectory is exported into the final dataset. If the bad-index is too dense, violations are too severe, or fixed trajectories still fail verification, the entire video is discarded.

The resulting *OpenSafari* couples high-dynamic, in-the-wild FPV drone video with rigorously verified camera trajectories. It departs from existing benchmarks by emphasizing aggressive 6-DoF motion, strong parallax, and complex outdoor layouts, while enforcing strict geometric and kinematic validation. This makes *OpenSafari* a challenging testbed for camera-controllable video generation.

5. Experiments

5.1. Implementation Details

Training recipe. We adopt a two-stage recipe. We first warm up the pose-conditioned memory retriever using pose-aligned memory tokens m_t. We then jointly train the retriever and DiT end-to-end, updating the DiT via LoRA [12]. Memory cross-attention is initialized from the corresponding context cross-attention weights, and other new layers use standard initialization.

Dataset. We extract overlapping clips with 1 s stride, yielding 51,997 training candidates. A diversity-based trajectory filter removes clips with near-static motion, resulting in 11,481 final training clips. We additionally construct a non-overlapping test set of 787 clips for evaluation. For each clip, we generate a single descriptive caption using Qwen2.5-VL-7B [3] and use it as the text condition.

Configuration and notation. We generate $\mathcal{T} = 5\,\text{s}$ clips at 24 fps from $T = 15\,\text{s}$ videos. Camera poses and memory features are sampled at 4 fps. For a target 5 s clip with interval $[t_0, t_1]$, we use the terminal pose p_{t_1} as the query. The memory window is limited to $L = 5\,\text{s}$. We use Wan2.2-Fun-5B-Control-Camera [38] as our base DiT with a hidden dimension $D = 3072$. Retriever and DiT are trained with 1 and 5 epochs, respectively. For each video, we extract 3D-aware memory feature from a pretrained StreamVGGT [62]. We select four layers $\{4, 11, 17, 23\}$; at each layer, the feature contains 782 tokens. Concatenating across the four layers yields $M = 4 \times 782$ and $d_m = 1024$ memory tokens per frame.

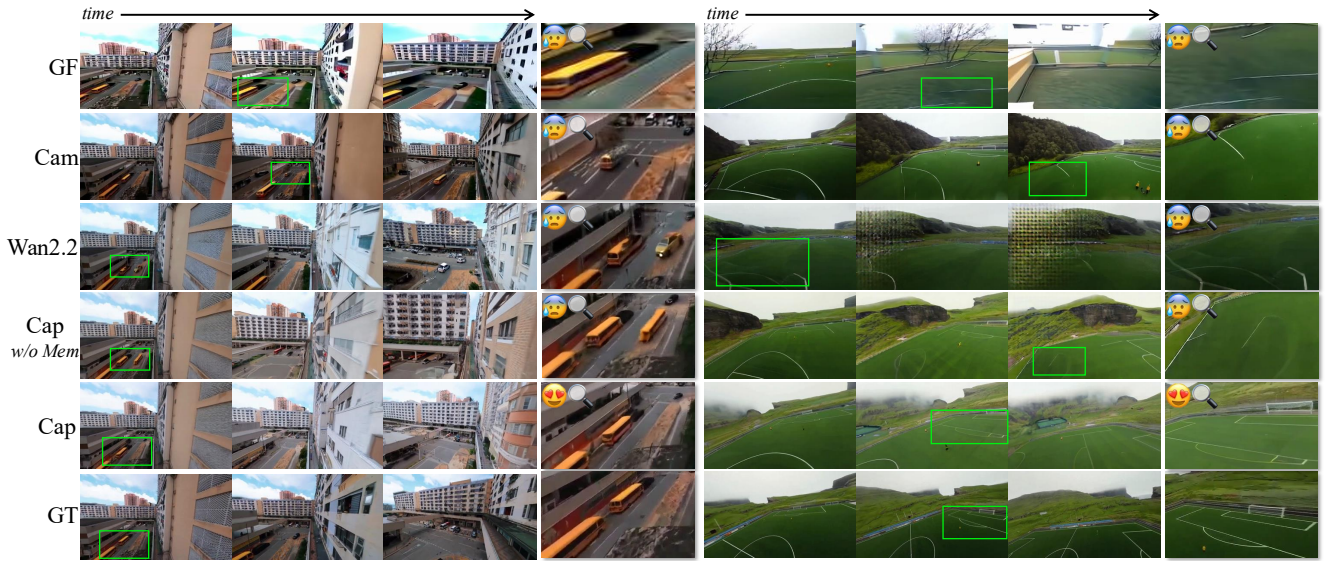

Figure 4. **Qualitative comparisons. Left**: Baselines—including the memory-removed variant—exhibit abrupt popping/vanishing of the school bus, and GF is low-quality. *Captain Safari* alone renders the bus smoothly exiting the frame. **Right**: Baselines distort or lose field marking, with Wan2.2 collapsing under large camera motion, affirming the challenge of 3D consistency under rapid trajectories. *Captain Safari* preserves crisp markings and coherent layout while following the fast 6-DoF path.

5.2. Benchmark

Metrics. We evaluate video generation along three complementary axes: video quality, 3D consistency, and trajectory following. For video quality, we report FVD [37] and LPIPS [54]. For 3D consistency, we use MEt3R [1], computed between GT and generated videos at matched time steps and a reconstruction rate that measures the percentage of frames successfully registered in the recovered 3D model [30, 31]. For trajectory following, we report camera relocation accuracy (AUC [39]) and the cosine similarity between the flattened camera pose, capturing how the model adheres to the desired camera parameters over time.

Baselines. We compare against representative camera-controllable video generation models, including Geometry Forcing [44], Real-CamI2V [19, 57], and Wan2.2-5B-Control-Camera [38], which cover geometry-constrained, reconstruction-driven, and large-scale diffusion-based approaches to trajectory-conditioned video synthesis.

Human Study. We conduct a human study with 50 participants. Each participant is presented with 10 cases, where each case contains the GT video and five anonymized model-generated videos (three baselines, our model, and its ablated variant). For every case, participants are asked to select the best video under three criteria: Video Quality, 3D Consistency, and Trajectory Following. In total, the study collects $50 \times 10 \times 3 = 1,500$ human preference votes.

5.3. Generation Quality

As shown in Table 1, our *Captain Safari* attains a substantially lower FVD (1023.46 vs. 1387.75) and a slightly improved LPIPS score (0.512 vs. 0.513) compared to the SOTA baseline, demonstrating more stable temporal dynamics and sharper spatial details. Moreover, the human study in Table 2 indicates that **67.40%** of participants prefer our videos over all competing methods, highlighting the perceptual realism and overall fidelity of our generations.

Qualitative comparisons in Figure 4 further reveal that *Captain Safari* produces visually compelling, realistic, and highly authentic scene dynamics. These findings are also consistent with the samples shown in Figure 1, where our method delivers vivid, coherent, and natural-looking drone videos that closely resemble real-world captures.

5.4. 3D Consistency

Captain Safari achieves state-of-the-art 3D consistency. As shown in Table 1, our method lowers MEt3R by 0.0013 (0.3690 vs. 0.3703) and raises the reconstruction rate by 0.045 (0.968 vs. 0.923) copared to the strongest baseline. Consistently, the human study in Table 2 shows that **65.60%** of participants prefer *Captain Safari* for 3D consistency, substantially surpassing all competing approaches.

Qualitative visualizations further confirm these quantitative gains. In Figure 1, structures such as the Greek-style columns remain geometrically stable across large viewpoint changes. In Figure 4, our model produces (*left*) a school bus that smoothly moves out of the frame, and (*right*) preserves crisp, globally consistent field markings on the soccer pitch, whereas baselines exhibit distortions and disappearance. Figure 5 and Figure 1 further show that our reconstructions yield sharper façades and well-formed windows

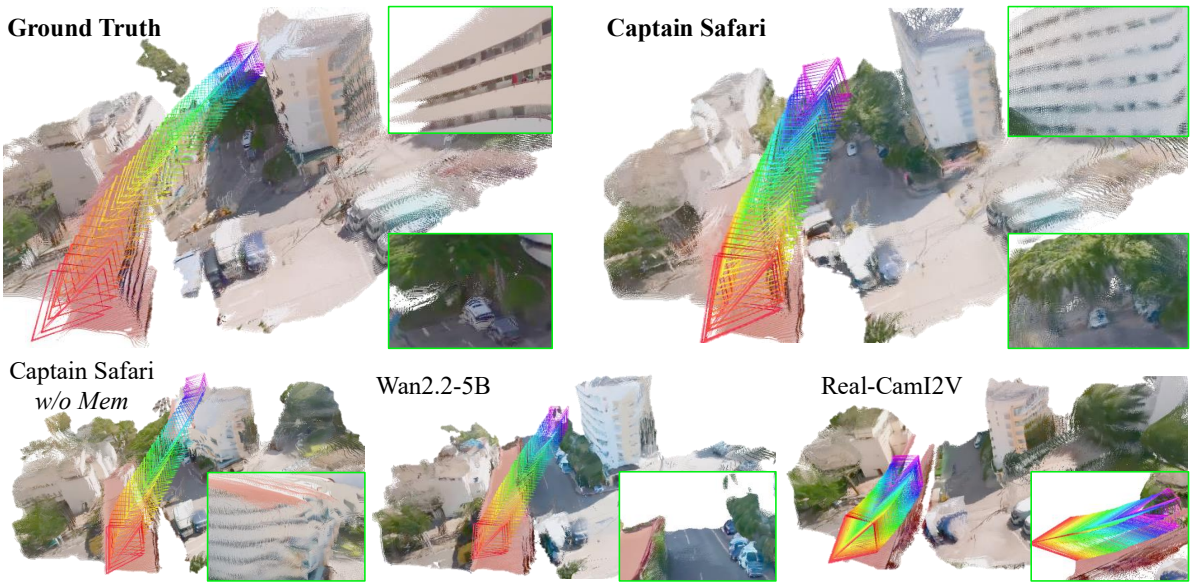

Figure 5. **Scene reconstruction and camera trajectory.** With pose-aligned memory, *Captain Safari* reconstructs a well-structured building façade (the memory-removed variant blurs/warps it), demonstrating the benefit of memory. It also preserves fine details—parked cars and the tree on their roofs—that Wan2.2-5B fails to retain. Meanwhile, Real-CamI2V follows only a short path, whereas *Captain Safari* covers the full trajectory with stable 3D structure, highlighting the challenge of maintaining 3D consistency under fast motion.

without collapsing geometry. Together, these results validate that the implicit world memory and pose-conditioned retrieval of *Captain Safari* effectively stabilize the underlying 3D world under aggressive camera motion.

5.5. Trajectory Following

Captain Safari delivers the most accurate trajectory following among all competing models. As shown in Table 1, our method achieves the highest AUC@30 (0.200) and AUC@15 (0.068), along with the best cosine similarity (0.563), outperforming the strongest baseline by clear margins. The human study in Table 2 further reinforces this observation, with **69.00%** of participants identifying our model as the most faithful to the target camera path.

Figure 5 provides a clear visualization of these improvements. *Captain Safari*'s predicted trajectory closely aligns with the ground-truth path, while the ablated variant deviates and flies over the rooftop, and RealCam-I2V fails to follow the intended forward motion, advancing only slightly rather than committing to the prescribed trajectory. Furthermore, our method demonstrates stable and coherent generation under challenging viewpoint changes with complex camera maneuvers in Figure 1. These results highlight the effectiveness of our memory-augmented, pose-conditioned design for precise trajectory adherence.

5.6. Ablation Study

Our results highlight the importance of the proposed pose-conditioned world memory. As shown in Table 1, adding memory yields substantial improvements in both 3D consistency and trajectory following. These gains confirm that retrieving pose-aligned world features at the target frame provides the model with an explicit understanding of what the scene *should* look like, enabling stable geometry and accurate motion alignment.

Qualitative comparisons in Figure 4 and Figure 5 further illustrate these effects. With memory, the generated scenes preserve global structure, maintain consistent geometry across viewpoints. In contrast, the ablated variant often drifts and exhibits geometric inconsistencies. Together, these results validate the effectiveness of our memory-augmented design in stabilizing the underlying 3D world and guiding precise camera motion.

6. Conclusion

We introduced *Captain Safari*, a pose-conditioned world engine built on a world memory that enables long-range, 3D-consistent video generation under complex FPV trajectories. Together with *OpenSafari*, our curated dataset of in-the-wild drone videos with verified camera poses, this establishes a rigorous benchmark for controllable video generation. *Captain Safari* markedly improves 3D consistency and trajectory accuracy over prior methods while maintaining strong visual fidelity. Although the system incurs non-trivial inference overhead, future work will explore real-time world engines with lightweight memory and faster generative backbones. We hope *Captain Safari* and *OpenSafari* encourage further research in persistent world models and long-horizon controllable video generation.

References

[1] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: Measuring multi-view consistency in generated images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6034–6044, 2025. 7

[2] Sherwin Bahmani, Xian Liu, Wang Yifan, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, et al. Tc4d: Trajectory-conditioned text-to-4d generation. In *European Conference on Computer Vision*, pages 53–72. Springer, 2024. 3

[3] 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. 6

[4] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022. 5

[5] Yuanhao Cai, He Zhang, Kai Zhang, Yixun Liang, Mengwei Ren, Fujun Luan, Qing Liu, Soo Ye Kim, Jianming Zhang, Zhifei Zhang, et al. Baking gaussian splatting into diffusion denoiser for fast and scalable single-stage image-to-3d generation and reconstruction. *arXiv preprint arXiv:2411.14384*, 2024. 2

[6] Yaru Cao, Zhijian He, Lujia Wang, Wenguan Wang, Yixuan Yuan, Dingwen Zhang, Jinglin Zhang, Pengfei Zhu, Luc Van Gool, Junwei Han, et al. Visdrone-det2021: The vision meets drone object detection challenge results. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 2847–2854, 2021. 2

[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1

[8] Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. Adaworld: Learning adaptable world models with latent actions. *arXiv preprint arXiv:2503.18938*, 2025. 1, 2

[9] Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, et al. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1–12, 2025. 3

[10] 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. 1, 3

[11] Chen Hou and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024. 3

[12] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6

[13] Ronghang Hu, Nikhila Ravi, Alexander C Berg, and Deepak Pathak. Worldsheet: Wrapping the world in a 3d sheet for view synthesis from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12528–12537, 2021. 1, 2

[14] Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft. *arXiv preprint arXiv:2510.03198*, 2025. 2, 3

[15] 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. 3

[16] Longbin Ji, Lei Zhong, Pengfei Wei, and Changjian Li. Posetraj: Pose-aware trajectory control in video diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22776–22785, 2025. 1, 2, 3

[17] 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

[18] 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. 2

[19] Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yepan Xiong, Min Chen, et al. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 28785–28796, 2025. 1, 3, 6, 7

[20] Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N Plataniotis, Sergey Tulyakov, and Jian Ren. Wonderland: Navigating 3d scenes from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 798–810, 2025. 1, 3

[21] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 2

[22] Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *European Conference on Computer Vision*, pages 329–345. Springer, 2024. 2, 3

[23] 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. 1, 2

[24] Andrew Melnik, Michal Ljubljanac, Cong Lu, Qi Yan, Weiming Ren, and Helge Ritter. Video diffusion models: A survey. *arXiv preprint arXiv:2405.03150*, 2024. 2

[25] Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. Revideo: Remake a video with motion and content control. *Advances in Neural Information Processing Systems*, 37:18481–18505, 2024. 3

[26] Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Weijie Wang, Haoyun Li, Guosheng Zhao, Jie Li, Wenkang Qin, Guan Huang, and Wenjun Mei. Wonderturbo: Generating interactive 3d world in 0.72 seconds. *arXiv preprint arXiv:2504.02261*, 2025. 1, 3

[27] Ava Pun, Gary Sun, Jingkang Wang, Yun Chen, Ze Yang, Sivabalan Manivasagam, Wei-Chiu Ma, and Raquel Urtasun. Neural lighting simulation for urban scenes. *Advances in Neural Information Processing Systems*, 36:19291–19326, 2023. 2

[28] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021. 1

[29] 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*, pages 6121–6132, 2025. 1, 3

[30] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 5, 7

[31] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 5, 7

[32] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019. 1

[33] Manuel-Andreas Schneider, Lukas Höllein, and Matthias Nießner. Worldexplorer: Towards generating fully navigable 3d scenes. *arXiv preprint arXiv:2506.01799*, 2025. 3

[34] 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*, pages 12449–12458, 2025. 2

[35] Shuhan Tan, Kelvin Wong, Shenlong Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Scenegen: Learning to generate realistic traffic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 892–901, 2021. 2

[36] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 5

[37] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 7

[38] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 6, 7

[39] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 7

[40] Jiahao Wang, Luoxin Ye, TaiMing Lu, Junfei Xiao, Jiahan 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. 1, 2, 3

[41] 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

[42] 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. 3

[43] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7467–7477, 2020. 1, 2

[44] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025. 3, 6, 7

[45] Jianzong Wu, Liang Hou, Haotian Yang, Xin Tao, Ye Tian, Pengfei Wan, Di Zhang, and Yunhai Tong. Vmoba: Mixture-of-block attention for video diffusion models. *arXiv preprint arXiv:2506.23858*, 2025. 1, 2

[46] Dejia Xu, Yifan Jiang, Chen Huang, Liangchen Song, Thorsten Gernoth, Liangliang Cao, Zhangyang Wang, and Hao Tang. Cavia: Camera-controllable multi-view video diffusion with view-integrated attention. *arXiv preprint arXiv:2410.10774*, 2024. 3

[47] 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

[48] 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

[49] Xiaoda Yang, Jiayang Xu, Kaixuan Luan, Xinyu Zhan, Hongshun Qiu, Shijun Shi, Hao Li, Shuai Yang, Li Zhang, Checheng Yu, et al. Omnicam: Unified multimodal video generation via camera control. *arXiv preprint arXiv:2504.02312*, 2025. 2

[50] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 1, 2

[51] 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. 1, 3

[52] Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv preprint arXiv:2503.05638*, 2025. 3

[53] 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. 3

[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[55] 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 Computer Vision and Pattern Recognition Conference*, pages 2063–2073, 2025. 3

[56] 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. 3

[57] 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. 1, 3, 7

[58] Mengqi Zhou, Yuxi Wang, Jun Hou, Shougao Zhang, Yiwei Li, Chuanchen Luo, Junran Peng, and Zhaoxiang Zhang. Scenex: Procedural controllable large-scale scene generation. *arXiv preprint arXiv:2403.15698*, 2024. 2

[59] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 2, 5

[60] Yunsong Zhou, Michael Simon, Zhenghao Peng, Sicheng Mo, Hongzi Zhu, Minyi Guo, and Bolei Zhou. Simgen: Simulator-conditioned driving scene generation. *Advances in Neural Information Processing Systems*, 37:48838–48874, 2024. 2

[61] Zhenghong Zhou, Jie An, and Jiebo Luo. Latent-reframe: Enabling camera control for video diffusion models without training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12779–12789, 2025. 3

[62] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025. 6