A BENCHMARK FOR CAMERA CONTROLLABILITY IN INTERACTIVE WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive world models increasingly require precise camera control, yet existing benchmarks rarely treat commanded motion amplitude and delivery rate as jointly controlled experimental variables, and few calibrate their pose estimator against ground truth to establish which conditions are reliable enough to score. We propose a benchmark protocol for camera controllability in world models that takes explicit continuous pose commands as input and evaluates models against a structured set of commanded trajectories: axis-pure motions across the rotational camera degrees of freedom and a null condition for measuring hallucinated ego-motion, paired with a calibrated measurement protocol that establishes when the underlying pose estimator can be trusted. We apply this protocol to two models, evaluating WorldCam across all three rotational axes and both command directions together with a null-drift test, and evaluating SANA-WM under positive yaw motion, finding that amplitude error varies with commanded angle but is non-monotonic in rotation rate, that controllability failures are scene-dependent, and that the two models exhibit qualitatively different leakage patterns despite operating under different conditioning regimes.

1 INTRODUCTION

Camera control is a core capability for interactive world models deployed in simulation, gaming, and embodied AI. Applications requiring precise spatial navigation, robotic task simulation, synthetic data generation, and game engine emulation demand not only that the generated video looks plausible, but that the camera moves exactly as commanded. Despite this, existing benchmarks have not isolated motion amplitude as an evaluation axis: they test whether camera motion has the right geometry and direction, but not whether the model faithfully reproduces the commanded magnitude of rotation or translation.

Existing benchmarks evaluate whether a model moves in the correct direction and follows a geometrically plausible trajectory, but none combines a controlled sweep over commanded amplitude and delivery rate with documented pose-estimator reliability checks. Moreover, pose-estimator reliability is not audited before the estimator is used to score models: trajectory-level error is reported without characterizing estimator bias, coverage, or failure modes under controlled conditions. No unified evaluation protocol exists for pose-conditioned world models that directly measures motion amplitude under calibrated measurement conditions.

We address this gap with the following contributions:

1. a benchmark protocol for pose-conditioned world models that takes explicit continuous camera pose commands as input;
2. a principled commanded trajectory taxonomy—axis-pure rotational motions and a null condition for hallucinated ego-motion—modulated by commanded amplitude and delivery rate as first-class experimental axes;
3. a unified evaluation metric derived from relative pose comparison, producing per-axis recovery, leakage error and APE_{rot};
4. a pilot case study applying this protocol to two pose-conditioned world models: WorldCam, evaluated across all three rotational axes and both command directions with a null-drift test,

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

and SANA-WM, evaluated under positive yaw only, with the remaining axes and directions left to future work, revealing that rotation error depends jointly on commanded angle and rate, and that the two models exhibit qualitatively different leakage patterns.

2 Related Work

Camera-controllable video. Video generation has progressed from unconditional models (Vondrick et al., 2016; Skorokhodov et al., 2022) and text-conditioned diffusion models (Ho et al., 2022; Blattmann et al., 2023) toward increasingly precise control signals, including direct camera-pose conditioning (Wang et al., 2024; He et al., 2025). Evaluation typically follows an open-loop protocol over a fixed dataset of reference trajectories such as RealEstate10K (Zhou et al., 2018), measuring rotation and translation error against pre-recorded paths alongside visual-quality metrics such as FVD, PSNR, and SSIM. This framing treats video generation as a trajectory-following problem on a closed set of scenes, without requiring the model to maintain persistent scene knowledge or respond to novel user commands.

World models for games and simulation. World models extend video generation by coupling frame synthesis with memory of the current scene and the ability to react to arbitrary action inputs. Early neural game engines realised this idea within single games, demonstrating closed-loop playability but limited to narrow, fixed environments. As models generalised to open domains, two control paradigms emerged: action-conditioned systems such as Hunyuan-GameCraft (Li et al., 2025), Matrix-Game 2.0 (He et al., 2026), and Matrix-Game 3.0 (Wang et al., 2026) accept keyboard-and-mouse inputs for real-time interaction, while pose-conditioned systems such as LingBot-World Team et al. (2026) and WorldCam (Nam et al., 2026) accept explicit camera-pose inputs for precise viewpoint control. Our benchmark focuses on the pose-conditioned paradigm, where the commanded pose provides a natural ground truth against which executed motion amplitude can be measured.

Benchmarks for controllable generation. Existing benchmarks fall into several groups. Non-interactive suites such as VBench (Huang et al., 2023), EvalCrafter (Liu et al., 2024), and VideoPhy (Bansal et al., 2024) assess visual quality, text alignment, or physical commonsense without taking action inputs or evaluating closed-loop control. Interactive world-model benchmarks address this by conditioning on action inputs and evaluating whether the model responds correctly: WBench (Ying et al., 2026) evaluates multi-turn interaction quality, WorldMark (Xu et al., 2026b) measures navigation consistency using scale-free metrics, WorldRoamBench (Xu et al., 2026a) assesses long-horizon trajectory stability, and iWorld-Bench (Fang et al., 2026) probes directional controllability via categorical action primitives. Closest to our setting, concurrent work on pose-conditioned world models such as SANA-WM (Zhu et al., 2026) reports rotation and translation error against commanded pose on fixed evaluation trajectories, using an off-the-shelf pose estimator as ground truth. These benchmarks do not treat commanded amplitude or delivery rate as an experimental variable, and pose-estimator reliability is not audited before the estimator is used to score models: trajectory-level error is reported without characterizing estimator bias, coverage, or failure modes under controlled conditions. We address this gap by treating motion amplitude and rate as first-class experimental axes, alongside a calibrated measurement protocol that establishes when the pose estimator itself can be trusted.

3 Benchmark Protocol

3.1 Problem Formulation

Let $\mathcal{M}$ be an interactive world model that generates a video sequence $\mathbf{v} = (v_1, \ldots, v_T)$ conditioned on a starting frame $f_0 \in \mathbb{R}^{H \times W \times 3}$ and an explicit camera pose control signal

$$C = (c_1, \ldots, c_T), \quad c_t \in \mathrm{SE}(3),$$

where $\mathrm{SE}(3)$ denotes the group of rigid-body transforms in $\mathbb{R}^3$. The benchmark measures how accurately $\mathcal{M}$ implements C in the generated video's estimated camera trajectory. All evaluation is relative to f_0, which avoids requiring absolute scene coordinates and enables comparison across models and scenes.

3.2 COMMANDED TRAJECTORY TAXONOMY

All benchmark stimuli are drawn from the following trajectory types. A global motion scale factor controls the amplitude of commanded motion; each type is instantiated at multiple commanded amplitudes within the reliable operating range of the pose estimator (see Appendix A), allowing the benchmark to characterize how recovery error and leakage scale with motion magnitude.

Type 1: Axis-Pure Motions. Each stimulus commands motion along exactly one of the three camera degrees of freedom, holding the remaining two at zero:

- Pure yaw ($\Delta\phi$)
- Pure pitch ($\Delta\theta$)
- Pure roll ($\Delta\psi$)

Axis-pure stimuli enable a rotational leakage matrix: diagonal entries measure per-axis recovery accuracy. Off-diagonal entries measure unintended motion leaking onto the remaining axes. In the current pilot, diagonal recovery and off-diagonal leakage MAE are reported for WorldCam and SANA-WM under positive yaw commands only (Table 2); WorldCam pitch and roll commands additionally report APErot and off-diagonal leakage, but not a separate diagonal recovery MAE (Appendix B). All commanded trajectories must remain within the physical boundaries of the scene and be consistent with the environment's camera model.

Type 2: Null Condition. Zero commanded motion across all six DOFs. This measures hallucinated ego-motion decomposed per axis. We hypothesize that the pose estimator's noise floor under null motion is low relative to model-induced drift; this is characterised in Appendix A and should be verified before attributing observed drift solely to the world model.

Type 3: Naturalistic Roaming. Compound, multi-axis trajectories sampled from realistic camera motion to guarantee physical plausibility and boundary compliance. Evaluation shifts to full-trajectory fidelity measured by $\mathrm{APE}_{\mathrm{rot}}$. This trajectory type is defined here for completeness; evaluation is left to future work.

3.3 TRAJECTORY EXTRACTION VIA VIPE

Given the generated video, ViPE is applied to estimate per-frame camera poses, producing a sequence of $\mathrm{SE}(3)$ transforms. ViPE's calibration, bias, and motion-dependent uncertainty are characterized in Appendix A, which establishes the reliable operating range used to gate which motion-scale conditions are included in benchmark results.

3.4 METRICS

Let commanded trajectory $C = \{c_1, \ldots, c_N\}$ and estimated trajectory $E = \{e_1, \ldots, e_N\}$ be sequences of $\mathrm{SE}(3)$ transforms, expressed as *camera-to-world* transforms in degrees. We anchor both to frame f_0:

$$\tilde{E}_t = E_0^{-1} \cdot E_t, \qquad \tilde{C}_t = C_0^{-1} \cdot C_t.$$

Each anchored transform has the form

$$\tilde{X}_t = \begin{bmatrix} \tilde{R}_t & \tilde{t}_t \\ \mathbf{0} & 1 \end{bmatrix}, \qquad \tilde{R}_t \in \mathrm{SO}(3).$$

Commanded trajectory schedule. For WorldCam, each stimulus conditions the model on a short clip (`COND_FRAMES` ground-truth frames), then commands rotation relative to the final conditioning frame's pose. For target amplitude Φ and rate r, the camera advances by exactly r degrees per frame for $n_{\mathrm{rot}} = \lfloor \Phi / r \rfloor$ frames, then holds the final rotated pose for the remainder; total generated frames are rounded up to the nearest multiple of 4. We retain only (Φ, r) pairs where r divides Φ exactly, so every condition reaches its target with no rounding error. Conditioning frames are not scored; metrics are computed over generated frames only.

For SANA-WM, each stimulus conditions on a single image (identity pose), then ramps linearly from 0° to Φ at rate r over $n_{\text{motion}} = \lfloor \Phi / r \rfloor$ frames before holding; total sequence length is rounded up to the nearest multiple of 8. As with WorldCam, only exactly-divisible (Φ, r) pairs are retained, and the conditioning frame is excluded from scoring.

Absolute rotation error (APE_{rot}). The primary measure of total rotational error at frame t is the $\mathrm{SO}(3)$ geodesic angle between the anchored estimated and commanded rotations:

$$\text{APE}_{\text{rot}}(t) \;=\; \arccos\left(\frac{\mathrm{tr}\left(\tilde{R}_t^{E\top}\,\tilde{R}_t^{C}\right) - 1}{2}\right), \tag{1}$$

averaged over all frames as

$$\text{APE}_{\text{rot}} = \frac{1}{N}\sum_{t=1}^{N} \text{APE}_{\text{rot}}(t).$$

Recovery and leakage error. For per-axis diagnostics, we decompose $\tilde{R}_t^E$ into Euler angles under the YXZ convention to recover the estimated yaw $\tilde{\phi}^E(t)$, pitch $\tilde{\theta}^E(t)$, and roll $\tilde{\psi}^E(t)$ at each frame, and similarly decompose $\tilde{R}_t^C$ into commanded $\phi^C(t)$, $\theta^C(t)$, $\psi^C(t)$.

For an axis-pure stimulus commanding axis $a^* \in \{\phi, \theta, \psi\}$, we distinguish two quantities. *Recovery error* is the signed difference between the estimated and commanded value on the commanded axis itself:

$$\varepsilon_{a^*}^{\text{rec}}(t) = \tilde{a}^{*E}(t) - a^{*C}(t).$$

Leakage error is the estimated value on each of the two *non*-commanded axes $a \in \{\phi, \theta, \psi\} \setminus \{a^*\}$, which is itself the signed leakage since the commanded value on those axes is zero:

$$\varepsilon_a(t) = \tilde{a}^E(t), \qquad a \neq a^*.$$

Under the pure-yaw stimulus reported in this paper ($a^* = \phi$), recovery error is $\varepsilon_\phi^{\text{rec}}(t) = \tilde{\phi}^E(t) - \phi^C(t)$, and leakage error is $\varepsilon_\theta(t) = \tilde{\theta}^E(t)$ and $\varepsilon_\psi(t) = \tilde{\psi}^E(t)$, as reported in Table 2. We report the mean absolute value (MAE) for recovery and leakage alike,

$$\text{MAE}_a = \frac{1}{N}\sum_{t=1}^{N} \big|\varepsilon_a(t)\big|,$$

where a ranges over a^* (recovery) or the two non-commanded axes (leakage) as appropriate. In the current pilot, we report leakage MAE for positive yaw commands only (Table 2); the full recovery/leakage matrix across all axes and directions is left to future work. APE_{rot} (Eq. 1) is the axis-agnostic total rotational geodesic error and is reported separately from recovery MAE.

Null drift. A special case of leakage where $\tilde{p}_t^C[a] = 0$ for all a and t, so $\varepsilon_a(t) = \tilde{p}_t^E[a]$ directly. APE_{rot} measures total hallucinated ego-motion; MAE_a decomposes it per axis as a diagnostic. In the current pilot, null-drift evaluation is conducted for WorldCam only; SANA-WM null-drift results are left to future work.

Translational scale. Monocular pose estimation is scale-ambiguous: the absolute metric scale of $\tilde{t}_t^E$ is not recoverable from image data alone. In principle, if ViPE applied a consistent scale factor relative to the commanded translation, the bias could be characterised and corrected. However, calibration against ground-truth trajectories (Appendix A) reveals that the scale factor is not consistent across conditions: estimated translation magnitude varies substantially. We therefore treat translation magnitude as an unreliable metric in the current benchmark and restrict quantitative amplitude evaluation to rotation, where ViPE is well-calibrated.

4 Case Study: Multi-Model Evaluation

4.1 Setup

Models. The pilot evaluation covers two pose-conditioned world models, WorldCam and SANA-WM. WorldCam is a Diffusion Transformer (DiT) model fine-tuned on CS:GO gameplay and is

initialized from a short video clip. SANA-WM is an open-source model trained on roughly 213K public video clips. Both models accept explicit 6-DoF camera-pose inputs, so commanded amplitude can be evaluated without an action-adapter layer. We run SANA-WM as released, without CS:GO fine-tuning.

Scenes. WorldCam is evaluated on two CS:GO scenes sampled from held-out gameplay sequences, and is conditioned on the initial video clip of each scene. SANA-WM is evaluated on two example images from its public repository, each depicting a game-style scene, and is conditioned on a single image. The two models are therefore evaluated on different scenes, and we treat the comparison as descriptive. All trajectories are anchored to the starting frame (Section 3.1). Each scene is evaluated under two seeds, and all reported results are averaged across scenes and seeds.

Conditions. For both models, the pilot covers yaw at commanded amplitudes $\Phi \in \{10^\circ, 30^\circ, 60^\circ\}$ and rates $r \in \{1, 2, 5, 10\}$ deg/frame, positive direction only; ViPE calibration has been established for this regime (Appendix A). For WorldCam, pitch, roll, and negative-direction yaw are additionally evaluated across the same amplitudes and rates. These results are reported in Appendix B and Appendix C as exploratory, since estimator reliability for pitch, roll, and negative yaw has not been verified against ground truth. SANA-WM is evaluated on positive yaw only. Negative-direction yaw evaluation for SANA-WM is left to future work. Results are reported as $\mathrm{APE}_{\mathrm{rot}}$ averaged across scenes and seeds in Table 1.

Table 2 reports cross-axis leakage onto pitch (θ) and roll (ψ) under positive yaw commands. Leakage is small at low amplitudes and rates but rises at Φ=60°.

Null condition. For WorldCam, the null arm holds the final conditioning frame's pose constant throughout the generated sequence, commanding no rotation. Any rotation ViPE reports under this condition is estimated null-motion drift, which may include residual estimator noise in addition to genuine model-induced ego-motion. Results are in Table 3. Drift varies substantially by scene, with Scene 1 reaching $\mathrm{APE}_{\mathrm{rot}} = 1.07^\circ$ and $\mathrm{MAE}_\theta = 0.64^\circ$, comparable to the cross-axis leakage in Table 2. No null condition was run for SANA-WM.

Table 1: $\mathrm{APE}_{\mathrm{rot}}$ (averaged across scenes and seeds) as a function of commanded amplitude Φ and rotation rate r, for WorldCam and SANA-WM under positive-direction yaw commands. Negative-direction WorldCam yaw results are provided in Appendix C as exploratory. Lower is better.

		Rate r (deg/frame)			
Model	Φ	1	2	5	10
WorldCam	10°	1.91	1.03	3.40	2.84
	30°	2.83	2.70	8.77	9.66
	60°	15.70	8.04	12.08	19.69
SANA-WM	10°	5.27	5.73	7.89	8.57
	30°	1.31	6.58	16.49	22.83
	60°	2.66	12.60	23.91	31.27

4.2 Failure Mode Analysis

Beyond the quantitative profile in Table 1, we document qualitative failure modes per model:

- **Distributional failures:** at larger commanded amplitudes, the generated frames drift from the conditioning scene toward the end of the sequence, with the model producing content inconsistent with the initial environment as accumulated rotation moves the viewpoint beyond the training distribution.
- **Scene sensitivity:** WorldCam's behavior is strongly scene-dependent in ways that do not generalize across conditions. At low rates ($r = 1$ deg/frame), Scene 1 overshoots the commanded amplitude, producing rotations substantially exceeding the target (up to 90° for a 60° command), while Scene 0 tracks accurately under the same conditions. Conversely,

Table 2: Recovery MAE onto yaw (ϕ, diagonal) and cross-axis leakage MAE onto pitch (θ) and roll (ψ, off-diagonal) under positive-direction yaw commands, as a function of commanded amplitude Φ and rotation rate r, for WorldCam and SANA-WM. WorldCam results under commanded pure pitch and pure roll motions are provided in Appendix B as exploratory.

			Rate r (deg/frame)			
Model	Φ	Axis	1	2	5	10
WorldCam	10°	ϕ	1.79	0.69	3.25	2.78
		θ	0.25	0.45	0.38	0.21
		ψ	0.34	0.30	0.36	0.27
	30°	ϕ	2.59	2.36	8.60	9.41
		θ	0.64	0.43	0.91	0.74
		ψ	0.52	1.08	0.80	0.67
	60°	ϕ	15.53	7.64	11.70	19.11
		θ	0.34	1.29	1.30	1.21
		ψ	1.43	1.09	0.87	1.08
SANA-WM	10°	ϕ	5.10	5.73	7.89	8.56
		θ	0.96	0.02	0.02	0.02
		ψ	0.36	0.16	0.16	0.16
	30°	ϕ	0.54	6.03	16.49	22.82
		θ	0.31	1.99	0.06	0.06
		ψ	1.01	1.14	0.39	0.40
	60°	ϕ	1.88	12.03	23.83	31.27
		θ	0.93	0.62	0.32	0.18
		ψ	1.18	1.44	1.15	0.58

Table 3: WorldCam null-drift: $\mathrm{APE}_{\mathrm{rot}}$ and per-axis MAE under zero commanded motion, across two scenes and two seeds. Higher values indicate more hallucinated ego-motion.

Run	$\mathrm{APE}_{\mathrm{rot}}$	MAE_{ϕ}	MAE_{θ}	MAE_{ψ}
Scene 0, Seed 0	0.336	0.326	0.047	0.020
Scene 0, Seed 1	0.055	0.046	0.017	0.010
Scene 1, Seed 0	1.070	0.843	0.636	0.152
Scene 1, Seed 1	0.837	0.813	0.112	0.159
Mean	0.575	0.507	0.203	0.085

at high rates ($r = 10$ deg/frame), Scene 0 undershoots severely and produces blurry outputs, while Scene 1 remains coherent. This cross-scene reversal suggests that amplitude error reflects an interaction between the commanded motion parameters and scene-specific properties of the conditioning frame, rather than a consistent model-level bias.

ViPE calibration and biases are documented in Appendix A.

5 Discussion

What the benchmark reveals. Under pure yaw commands, $\mathrm{APE}_{\mathrm{rot}}$ depends jointly on commanded amplitude and rate (Table 1). For both models, error is generally higher at higher rates, and error does not grow consistently with commanded angle. Both models track the largest command ($60°$) poorly at the fastest rate ($r = 10$ deg/frame), while WorldCam is markedly more accurate at low rates for the same amplitude. An evaluation that fixes a single amplitude and rate would miss these interactions, which motivates treating both as experimental variables.

Cross-axis leakage. Leakage onto pitch and roll under pure yaw commands is small for both models (Table 2), but the two differ in how it behaves. For SANA-WM, roll leakage tends to increase with commanded yaw amplitude, though not monotonically at every rate. Pitch leakage does not follow a consistent trend across amplitudes or rates, with the largest values appearing at isolated conditions (r=1 deg/frame, Φ=10° and r=2 deg/frame, Φ=30°) rather than at the largest commanded angle. For WorldCam, leakage on both axes stays low and does not follow a consistent trend with commanded amplitude. We speculate that such differences may partly reflect training data, since WorldCam is fine-tuned on CS:GO while SANA-WM is trained on general public video.

6 Conclusion and Future Work

We propose a benchmark for precise camera controllability in interactive world models that treats commanded amplitude and delivery rate as first-class experimental axes. It compares each generated camera trajectory against the commanded one, reporting rotational absolute pose error (APErot) and per-axis recovery and leakage error. Because these scores depend on the pose estimator, we first calibrate it on engine-rendered ground truth — using still-frame null videos to establish its noise floor — and score only conditions where it is reliable. A pilot study of WorldCam and SANA-WM under pure positive yaw commands reports recovery and leakage error alongside APErot, showing that rotation error depends jointly on commanded angle and rate and is non-monotonic in both. Cross-axis leakage onto pitch and roll is small for both models; SANA-WM's roll leakage tends to increase with commanded yaw amplitude, while pitch leakage and WorldCam's leakage on both axes show no consistent trend. WorldCam is additionally evaluated on negative-direction yaw, pitch, and roll commands, extending single-axis coverage beyond the positive-direction case study.

Future work includes expanding scene and seed coverage; extending negative-direction and pitch/roll evaluation to SANA-WM with verified sign conventions and calibrated estimator coverage; evaluating both models on a shared scene set; broadening the case study to additional pose-conditioned world models; adding naturalistic roaming sequences; and developing principled scale inference for translation, extending the calibrated evaluation regime beyond rotation.

AI Use Statement

In this work, we used generative AI tools to aid and polish writing and for retrieval and discovery of related work. We have not used generative AI tools for research ideation, execution, drafting sections, or generating synthetic datasets. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including all text, claims, and results.

References

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation, 2024. URL https://arxiv.org/abs/2406.03520.

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, 2023. URL https://arxiv.org/abs/2304.08818.

Jianjie Fang, Yingshan Lei, Qin Wan, Ziyou Wang, Yuchao Huang, Yongyan Xu, Baining Zhao, Weichen Zhang, Chen Gao, Xinlei Chen, and Yong Li. iworld-bench: A benchmark for interactive world models with a unified action generation framework, 2026. URL https://arxiv.org/abs/2605.03941.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation, 2025. URL https://arxiv.org/abs/2404.02101.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Size Wu, Wei Li, Xuchen Song, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 2.0: An open-source

real-time and streaming interactive world model, 2026. URL https://arxiv.org/abs/2508.13009.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. URL https://arxiv.org/abs/2204.03458.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models, 2023. URL https://arxiv.org/abs/2311.17982.

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, 2025. URL https://arxiv.org/abs/2506.17201.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models, 2024. URL https://arxiv.org/abs/2310.11440.

Jisu Nam, Yicong Hong, Chun-Hao Paul Huang, Feng Liu, JoungBin Lee, Jiyoung Kim, Siyoon Jin, Yunsung Lee, Jaeyoon Jung, Suhwan Choi, Seungryong Kim, and Yang Zhou. Worldcam: Interactive autoregressive 3d gaming worlds with camera pose as a unifying geometric representation, 2026. URL https://arxiv.org/abs/2603.16871.

Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2, 2022. URL https://arxiv.org/abs/2112.14683.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models, 2026. URL https://arxiv.org/abs/2601.20540.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics, 2016. URL https://arxiv.org/abs/1609.02612.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation, 2024. URL https://arxiv.org/abs/2312.03641.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, Yidan Xietian, Jiangbo Pei, Liang Hu, Boyi Jiang, Hua Xue, Zidong Wang, Haofeng Sun, Wei Li, Wanli Ouyang, Xianglong He, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory, 2026. URL https://arxiv.org/abs/2604.08995.

Ting-Bing Xu, Jiacheng Sui, Zhe Gao, Kewei Shi, Wenjin Yang, Zhicheng Liu, Zhaoxu Sun, Mingchao Sun, Hongyu Pan, Fan Jiang, Mu Xu, Qi Fan, Yang Gao, Yong Li, and Baoquan Chen. Worldroambench: An open-world benchmark for long-horizon stability of interactive world models, 2026a. URL https://arxiv.org/abs/2606.31672.

Xiaojie Xu, Zhengyuan Lin, Kang He, Yukang Feng, Xiaofeng Mao, Yuanyang Yin, Yongtao Ge, and Kaipeng Zhang. Worldmark: A unified benchmark suite for interactive video world models, 2026b. URL https://arxiv.org/abs/2604.21686.

Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Henghui Ding. Wbench: A comprehensive multi-turn benchmark for interactive video world model evaluation, 2026. URL https://arxiv.org/abs/2605.25874.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images, 2018. URL https://arxiv.org/abs/1805.09817.

Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. Sana-wm: Efficient minute-scale world modeling with hybrid linear diffusion transformer, 2026. URL https://arxiv.org/abs/2605.15178.

A ViPE Calibration and Uncertainty Estimation

We calibrate ViPE against engine-rendered ground truth to characterise its bias, variance, and failure modes before using it as a measurement instrument. Each calibration sequence translates the camera by a fixed displacement along a lateral direction, returns to the origin, then executes a commanded yaw rotation of $\Phi \in \{30°, 60°\}$. Conditions are crossed with three initial camera orientations ($0°$, $30°$, $60°$ yaw offset) and four lateral translation directions ($\pm X$, $\pm Z$). Summary results are reported in Table 4.

Table 4: ViPE $\mathrm{APE}_{\mathrm{rot}}$ as a function of commanded amplitude Φ and initial camera yaw offset. Lower is better.

	Initial yaw offset		
Φ	$0°$	$30°$	$60°$
$30°$	0.196	0.192	0.197
$60°$	0.378	0.379	0.383

Rotational accuracy. Under most conditions ViPE recovers the commanded rotation with sub-degree $\mathrm{APE}_{\mathrm{rot}}$, establishing it as a reliable rotation estimator across the tested range.

Null-motion control. To measure how much rotation ViPE reports in the absence of any camera motion, we construct a null-motion control by sampling three distinct frames from each of three CS:GO scene videos and, for each sampled frame, repeating it 60 times to produce a still video. We run ViPE on all nine resulting static videos, three per scene. Because the input contains no motion whatsoever, any rotation reported by ViPE is attributable solely to estimator noise. Results are in Table 5. ViPE recovers near-zero rotation, with an $\mathrm{APE}_{\mathrm{rot}}$ of $0.005°$ and per-axis mean absolute errors of at most $0.003°$. This is more than an order of magnitude below the calibration errors in Table 4 ($\geq 0.19°$), suggesting that rotations of the magnitudes considered here are unlikely to be explained by estimation error alone. This control covers static scenes only and does not characterise ViPE's behaviour under scene dynamics.

Table 5: ViPE null-motion control: three distinct static frames per scene, each repeated 60 times, across three CS:GO scenes. Any reported rotation is attributable solely to estimator noise.

	$\mathrm{APE}_{\mathrm{rot}}$	MAE_{ϕ}	MAE_{θ}	MAE_{ψ}
ViPE	0.005	0.002	0.003	0.002

Translation alignment failures. ViPE degrades severely when the translation direction is roughly aligned with the camera's forward axis. In a first-person shooting scenario where the camera faces negative-Z and translates along the same axis, `calib_yaw0_negZ` produces yaw, pitch, and roll errors of $65.7°$, $7.2°$, and $91.9°$ respectively. At $30°$ initial yaw, the (pos-X, neg-Z) and (neg-X, pos-Z) translations both move along $3.4°$ relative to the camera forward axis, triggering the same failure: `yaw30_posX_negZ` produces a yaw error of $74.1°$ with pitch and roll excursions exceeding $44°$ and $75°$, and `yaw30_negX_posZ` produces pitch and roll excursions exceeding $33°$. In all cases the translation vector projects substantially onto the optical axis, providing insufficient lateral baseline for monocular scale recovery.

Translational inconsistency across camera orientations. Even outside catastrophic failure cases, the recovered world-frame translation is inconsistent across initial yaw offsets, as shown in Table 6. For a nominally identical positive-X displacement, the estimated world-X magnitude varies from

1.153 at yaw=0° to 3.383 at yaw=30° and 2.757 at yaw=60°. Cross-axis leakage onto the world-Z component follows the same pattern: negligible at yaw=0° (-0.160) and yaw=60° (-0.009), but rising to -0.584 at yaw=30°. This inconsistency confirms that translation estimation is a function of the angle between the commanded direction and the camera's forward axis, not of the commanded displacement alone.

Table 6: Estimated world-frame translation for a commanded positive-X displacement of 10 units across three initial yaw offsets. The ground-truth displacement is 10 units along world-X and zero along world-Z in all cases.

Condition	Est. X	Est. Z
`yaw0_posX`	1.153	−0.160
`yaw30_posX`	3.383	−0.584
`yaw60_posX`	2.757	−0.009

Reliable operating range. Based on these results, we treat ViPE yaw estimates as reliable. Conditions where the translation vector aligns with the camera forward axis are flagged as unreliable and excluded from benchmark scoring rather than discarded silently. Translation magnitude is additionally treated as an unreliable metric due to the inconsistency documented in Table 6; quantitative amplitude evaluation is therefore restricted to rotation. A potential direction for resolving this is to infer a per-condition scale factor by comparing the estimated translation trajectory against the commanded camera pose, using the known metric displacement as a reference; we leave this to future work.

B WorldCam Pitch and Roll Results

Tables 7 and 8 report $\mathrm{APE}_{\mathrm{rot}}$ and cross-axis leakage for WorldCam under pure positive and negative pitch and roll commands. SANA-WM is not evaluated on these axes in the current pilot. Roll $\mathrm{APE}_{\mathrm{rot}}$ is substantially larger than pitch $\mathrm{APE}_{\mathrm{rot}}$ at every condition (e.g. $49.27°$ vs. $39.01°$ at $\Phi = 60°$, $r = 10$ deg/frame). Comparing leakage onto the axis common to both commands, yaw (ϕ), leakage is generally worse under roll commands than under pitch commands at large commanded rotation: at $\Phi = 60°$, $r = 10$ deg/frame , MAE_{ϕ} reaches $4.00°$ under roll commands, versus $1.78°$ under pitch commands, with all other variables held fixed. Because these axes have not been calibrated against ground truth (Appendix A covers yaw only), we report these results as exploratory and do not attribute this pattern to a specific cause such as training-data composition.

Table 7: WorldCam $\mathrm{APE}_{\mathrm{rot}}$ (averaged across scenes and seeds) under pure pitch and roll commands, both positive (+) and negative (−) directions. SANA-WM not evaluated. Lower is better.

			Rate r (deg/frame)			
Axis	Φ	Dir.	1	2	5	10
Pitch	10°	+	3.78	3.02	5.73	5.84
		−	5.49	5.81	5.60	4.83
	30°	+	6.73	12.44	15.85	14.47
		−	12.64	11.84	12.59	16.21
	60°	+	21.32	20.76	30.59	39.01
		−	22.13	20.73	23.37	36.43
Roll	10°	+	7.15	8.83	9.60	10.12
		−	7.11	8.54	9.88	10.55
	30°	+	17.96	19.29	24.37	25.02
		−	18.48	19.44	24.31	28.04
	60°	+	32.75	36.09	39.00	49.27
		−	34.11	36.89	41.89	47.18

Table 8: WorldCam cross-axis leakage (MAE, degrees) under pure pitch and roll commands, both positive (+) and negative (−) directions. Lower is better.

				Rate r (deg/frame)			
Cmd.	Φ	Dir.	Leaked axis	1	2	5	10
Pitch	10°	+	MAE_ϕ	0.68	0.29	0.87	0.62
		+	MAE_ψ	0.47	0.58	0.97	1.07
		−	MAE_ϕ	0.52	0.23	0.32	0.50
		−	MAE_ψ	0.35	0.53	0.54	0.53
	30°	+	MAE_ϕ	0.82	1.00	0.88	1.83
		+	MAE_ψ	1.06	1.74	2.62	1.61
		−	MAE_ϕ	0.53	0.45	0.90	0.76
		−	MAE_ψ	1.34	1.94	0.99	2.13
	60°	+	MAE_ϕ	1.23	1.10	0.64	1.78
		+	MAE_ψ	4.44	2.96	2.70	3.20
		−	MAE_ϕ	1.70	2.37	2.37	2.88
		−	MAE_ψ	3.42	4.98	3.54	1.34
Roll	10°	+	MAE_ϕ	0.42	0.38	0.47	0.69
		+	MAE_θ	1.42	2.78	2.65	1.98
		−	MAE_ϕ	0.47	0.58	0.87	0.89
		−	MAE_θ	0.97	2.25	3.39	3.45
	30°	+	MAE_ϕ	0.68	0.56	0.90	1.11
		+	MAE_θ	4.74	6.17	9.32	7.48
		−	MAE_ϕ	1.26	1.38	2.56	6.17
		−	MAE_θ	5.19	5.98	8.31	5.12
	60°	+	MAE_ϕ	1.71	2.40	2.51	4.00
		+	MAE_θ	10.27	12.12	12.21	20.59
		−	MAE_ϕ	3.07	6.74	2.31	12.38
		−	MAE_θ	12.32	14.02	14.06	11.36

C WorldCam Negative Yaw Results

Table 9 reports $\mathrm{APE}_{\mathrm{rot}}$ and Table 10 reports recovery and cross-axis leakage MAE for WorldCam under negative yaw commands. As with Appendix B, these results are exploratory: ViPE calibration (Appendix A) was established for positive yaw only, so we do not treat estimator error as characterized for the negative direction.

Table 9: WorldCam $\mathrm{APE}_{\mathrm{rot}}$ (averaged across scenes and seeds) under negative-direction yaw commands.

	Rate r (deg/frame)			
Φ	1	2	5	10
−10°	1.82	5.43	5.20	6.47
−30°	2.62	9.97	14.85	13.88
−60°	8.98	10.94	20.12	25.64

Table 10: WorldCam recovery MAE onto yaw (ϕ) and cross-axis leakage MAE onto pitch (θ) and roll (ψ) under negative-direction yaw commands.

		Rate r (deg/frame)			
Φ	Axis	1	2	5	10
$-10°$	ϕ	1.70	5.39	5.14	6.44
	θ	0.19	0.34	0.18	0.21
	ψ	0.34	0.12	0.33	0.17
$-30°$	ϕ	2.29	9.80	14.81	13.65
	θ	0.58	0.34	0.36	0.43
	ψ	0.75	0.51	0.14	0.55
$-60°$	ϕ	8.90	8.99	20.05	25.51
	θ	0.37	2.43	0.39	0.44
	ψ	0.53	1.06	0.29	0.43