

Orbia: Do Generated Videos Form a Consistent 3D World?

Dongyue Lu^{1*}, Guian Fang^{1*}, Hur Lim¹, Yonggang Wu¹, Yanrong Wang¹, Hongshan Chen¹
 Shengju Qian², Yue Liao¹, Wei Chow¹, Lingdong Kong¹, Tianxin Huang³, Yu Yang⁴
 Hongji Yang⁵, Junchao Huang⁶, Alan Liang¹, Yihao Wang⁷, Zihan Wang⁷, Rong Li⁸
 Hanlin Chen¹, Xin Wang², Wei Tsang Ooi^{1†}, Mike Zheng Shou^{1†}, Shuicheng Yan^{1†}

¹NUS ²LIGHTSPEED ³HKU ⁴ZJU ⁵University of Macau ⁶CUHK(SZ) ⁷Aalto University ⁸HKUST(GZ)

Video generators serve as interactive worlds for exploration and simulation, yet realistic and visually appealing frames do not necessarily depict a consistent 3D world. Existing benchmarks assess selected geometric properties or use vision-language models to judge memory and consistency, leaving systematic 3D evaluation underexplored. We introduce **ORBIA**, the first benchmark dedicated to a systematic and comprehensive evaluation of **3D consistency** in interactive world models. Through controlled exploration and revisit, we organize the evaluation into three dimensions: input preservation, generated-content persistence, and 3D self-consistency. For input preservation, registered RGB-D references and annotated anchors support appearance, geometry, and spatial stability measurements. For generated-content persistence, we compare revisits to regions absent from the input image with their first generated observations to test whether models preserve what they create. For 3D self-consistency, cross-view depth and feature correspondence, regional scale agreement, and held-out Gaussian rendering test whether generated views describe the same 3D scene. Minute-long exploration and revisits further test long-term consistency. Our construction and annotation pipeline brings together public datasets, self-captured scenes, and simulation, yielding 720 instances across indoor and outdoor environments and six exploration-and-revisit protocols. Evaluation spans 27 text-, action-, and camera-conditioned models through native-interface adapters. Orbia provides a systematic basis for diagnosing 3D inconsistency and developing more persistent video world models. [Project Page](#) & [Benchmark](#) & [GitHub](#).

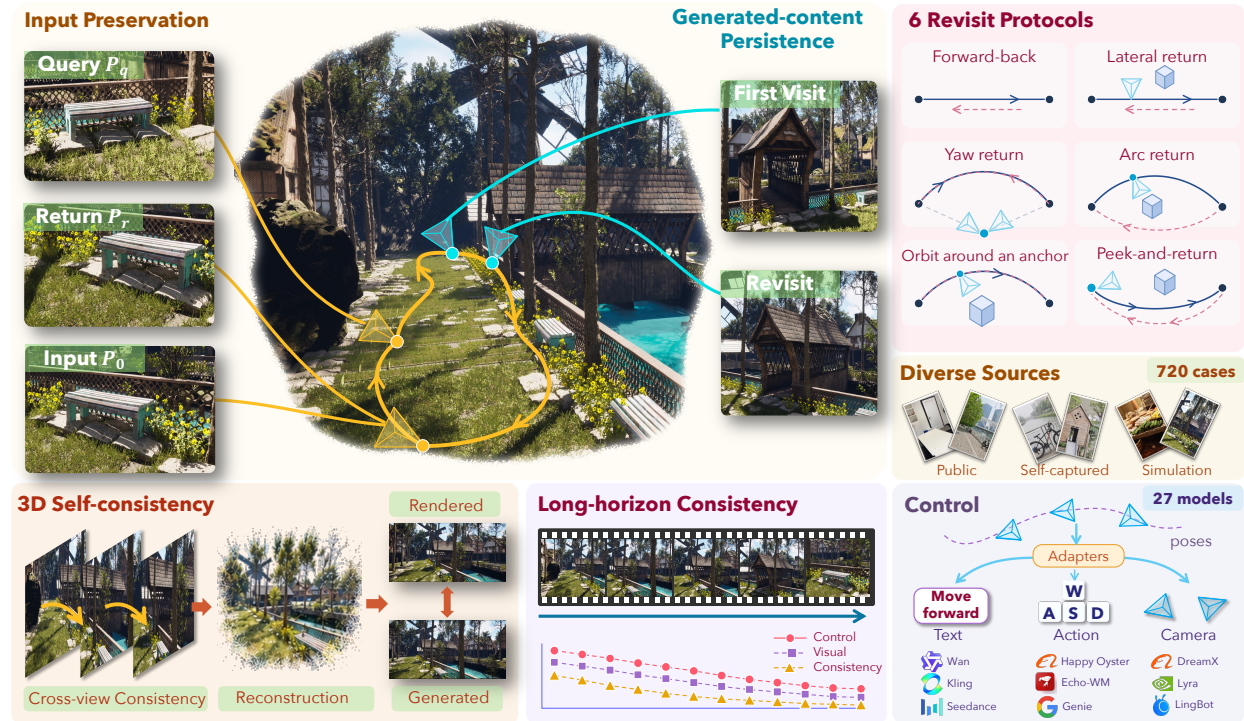


Figure 1. Overview of ORBIA. We introduce **ORBIA**, a benchmark for evaluating 3D consistency along three dimensions: ¹*input preservation*, ²*generated-content persistence*, and ³*3D self-consistency*. Six exploration-and-revisit templates support short and minute-long evaluation across *public datasets*, *self-captured scenes*, and *UE simulation*. We benchmark world models and video generators through *text*, *action*, and *camera-pose* interfaces.

1 Introduction

Video generation is expanding from visual synthesis toward world simulation (Brooks et al., 2024; Wan Team et al., 2025; Google DeepMind, 2025; Gao et al., 2025), supporting interactive games (Bruce et al., 2024; Valevski et al., 2024), open-domain exploration (Parker-Holder and Fruchter, 2025; Sun et al., 2025), generative rendering (Yu et al., 2024a; Shen et al., 2026), and embodied simulation (Yang et al., 2023; NVIDIA et al., 2025). These applications require observations across time and viewpoints to remain compatible with a persistent 3D world. A player should return to the same environment, a renderer should reveal mutually consistent surfaces, and an embodied agent should reuse previously observed spatial relationships. Visually convincing frames alone cannot establish this structural reliability across viewpoints and repeated visits. As generated videos become environments for exploration and interaction, a fundamental question arises: **Do generated videos form a consistent 3D world?**

Existing benchmarks address different aspects of this question. These protocols assess perceptual quality (Huang et al., 2024; Liu et al., 2024), local or scene-level geometric self-coherence (Duan et al., 2025; Ying et al., 2026; Xu et al., 2026b; Zhao et al., 2026; Liang et al., 2026), or memory and state retention upon revisit using same-view appearance comparisons, point-cloud comparisons, and vision-language judgments (Ye et al., 2026; Lu et al., 2026b; Xu et al., 2026a). Yet **systematic, quantitative evaluation of 3D consistency** remains underexplored. Geometric agreement does not guarantee content preservation, nor does a successful return establish consistency throughout exploration. A dedicated benchmark must jointly evaluate 3D structure and the preservation of input and generated content across viewpoints and over time during exploration and repeated revisits.

We introduce **ORBIA**, the first benchmark dedicated to a systematic and comprehensive evaluation of 3D consistency in interactive world models. We organize evaluation into **three dimensions** (Figure 1). **(1) Input preservation** assesses whether the model preserves the environment shown in the input as the viewpoint changes. Registered RGB-D references and annotated anchors measure the appearance, geometry, and spatial relationships of observed content. Exploration also reveals regions outside the input, whose appearance the model must generate. **(2) Generated-content persistence** compares newly revealed content across visits, using its first generated appearance as the reference. The model is free to create unseen content, but should preserve what it has created. Beyond content retention, generated views must agree in their underlying 3D structure. **(3) 3D self-consistency** tests whether generated views describe the same 3D scene through cross-view depth and feature correspondence, regional scale agreement, and held-out Gaussian rendering. These measurements examine both local surface compatibility and scene-level coherence. To assess resistance to temporal drift under longer exploration and more frequent revisits, we extend sequences to one minute and repeatedly return to previously visited regions. This tests whether models maintain scene geometry and content as the exploration history grows. We also examine how control accuracy and visual quality relate to the three consistency dimensions.

Joint evaluation requires cross-view references and controlled revisits. Our **construction pipeline** prepares registered RGB-D pairs, verifies anchor correspondence and surface support, and synthesizes trajectories connecting the reference poses. We select revisits to newly revealed regions and manually review anchors, reference views, and trajectories. To cover diverse scene distributions, we draw on ¹*real-world datasets*, ²*self-captured scenes*, and ³*simulated environments*, spanning indoor and outdoor settings in both real and virtual worlds. Processing over *110k* source samples and investing over *300* hours in annotation and curation yields **720 evaluation instances**, comprising 480 short cases and 240 paired minute-long extensions. Six exploration-and-revisit protocols cover trajectories with distinct spatial demands. Model-specific adapters translate the trajectories into text instructions, action sequences, or camera conditions for evaluation across control interfaces for the same exploration and revisit tasks.

To summarize, our contributions are fourfold:

- We introduce **ORBIA**, the first benchmark dedicated to systematic, quantitative evaluation of **3D consistency** in interactive world models across viewpoints and exploration horizons.
- We decompose 3D consistency into **input preservation**, **generated-content persistence**, and **3D self-consistency** to systematically assess content and geometry.
- We build a **unified annotation pipeline** yielding 720 instances with registered evidence and controlled revisits across real and simulated environments.
- We evaluate **27** text-, action-, and camera-conditioned models, examining how 3D consistency relates to control execution, visual quality, and exploration duration.

Table 1. Comparison with representative benchmarks. We compare input formats, data sources, evaluation references, and five consistency tests: ¹local 3D self-consistency, ²global 3D self-consistency, ³input-object preservation, ⁴revisit consistency, and ⁵long-term drift. Inputs:  text,  action,  camera pose. Data:  public real imagery,  self-captured,  simulated,  generated. References:  color images,  depth maps,  camera poses,  reference actions.

Benchmark Setup				Consistency Evaluation				
Benchmark	Input	Data	Reference	Local Consis.	Global Consis.	Input Preserv.	Revisit Consis.	Long-term Drift
WorldScore (Duan et al., 2025)	 	  		✓	×	×	×	×
WorldOlympiad (Zhao et al., 2026)	  	 		×	✓	×	×	×
WBench (Ying et al., 2026)	  	  		✓	×	✓	✓	×
WorldMark (Xu et al., 2026b)	  	  		✓	×	×	×	✓
iWorld-Bench (Fang et al., 2026)	  	 	 	×	×	×	✓	×
MIND (Ye et al., 2026)			 	×	×	×	✓	×
WRBench (Lu et al., 2026b)	 	-		×	×	✓	✓	×
WorldRoamBench (Xu et al., 2026a)	 	 		✓	×	✓	✓	✓
LoopBench (Wu et al., 2026b)		-	-	×	×	×	✓	✓
R2M-Bench (Gu et al., 2026b)	 	-	-	✓	×	×	✓	×
CycleBench (Gu et al., 2026a)		 		×	×	×	✓	✓
MemoBench (Chen et al., 2026b)	 	 	  	✓	×	✓	✓	×
ORBIA (Ours)	  	  	  	✓	✓	✓	✓	✓

2 Related Work

Interactive Video World Models. Large-scale video generators (Brooks et al., 2024; Polyak et al., 2024; Google DeepMind, 2025; Kong et al., 2024; Wan Team et al., 2025) provide the foundation for generative simulation. Interactive models introduce action-conditioned dynamics (Bruce et al., 2024; Valevski et al., 2024; Li et al., 2025b; He et al., 2025; Jiang et al., 2026), language-directed interaction (Mao et al., 2025; Tang et al., 2025), and explicit camera control (Wang et al., 2023; He et al., 2024). Recent systems extend these interfaces to streaming, open-domain exploration (Sun et al., 2025; Robbyant Team et al., 2026; Gao et al., 2026; DreamX Team et al., 2026; Zhu et al., 2025, 2026; Duan et al., 2026; Huang et al., 2026; Yang et al., 2026a). Despite richer interaction, evaluation largely prioritizes perceptual fidelity and control adherence, leaving the preservation of 3D structure and content through exploration insufficiently characterized.

3D Interactive World Models. Explicit scene representations enable navigable world generation and editing (Yu et al., 2024a; Liang et al., 2024; Szymanowicz et al., 2025; World Labs Team, 2025, 2026). Point-cloud and Gaussian renderings provide geometric guidance for view synthesis (Yu et al., 2024b; Ren et al., 2025; Huang et al., 2025b; Lu et al., 2026a; Yang et al., 2026b; InSpatio Team et al., 2026; Chen et al., 2026a; Li et al., 2025c), coupling generative priors with reconstructed geometry. Spatial memory further anchors exploration to prior observations through geometry-guided correspondence (Shen et al., 2026) and persistent geometric caches (AlayaWorld Team et al., 2026; Yin et al., 2026; Qian et al., 2026). These methods establish structured spatial representations, yet evaluation relies on 2D rendering quality, with geometric measurements covering selected aspects of self-consistency. A systematic account of input fidelity, persistence of generated regions, and geometric agreement across repeated exploration remains missing.

World Model Evaluation. Benchmarks assess perceptual quality (Huang et al., 2024; Liu et al., 2024), physical faithfulness (Zheng et al., 2025; Bansal et al., 2024; Meng et al., 2024; Li et al., 2025a), embodied behavior (Qin et al., 2024; Hu et al., 2025; Shang et al., 2026), and driving (Kong et al., 2025; Liang et al., 2026). Interactive tests cover control, geometry (Duan et al., 2025; Xu et al., 2026b; Fang et al., 2026; Zhao et al., 2026), and outcomes (Ying et al., 2026; Tang et al., 2025; Wu et al., 2026a). Diagnostics cover deformation (Gu et al., 2025), reconstructability (Fan and Favaro, 2026), and manipulation (Wang et al., 2026b). Memory tests cover returns and states (Ye et al., 2026; Lu et al., 2026b; Chen et al., 2026b,d), revisit geometry (Xu et al., 2026a), and calibrated similarity (Gu et al., 2026b). Further tests address drift (Wu et al., 2026b; Gu et al., 2026a) and agent-based transition and persistence judgments (Chen et al., 2026c). ORBIA jointly evaluates input preservation, generated-content persistence, and 3D self-consistency.

3 The Orbia Benchmark

ORBIA systematically evaluates the 3D consistency of interactive video world models. In this section, we introduce the evaluation principles and protocols (Section 3.1), the data construction and annotation pipeline (Section 3.2), and the dataset scale and coverage across scene sources and exploration horizons (Section 3.3).

3.1 Evaluation Principles and Protocols

Interactive World Generation. An interactive video world model generates observations from an initial image I_0 , a scene description s , and user inputs $\mathcal{U}$, yielding $\hat{V} = \mathcal{M}(I_0, s, \mathcal{U})$. Depending on the model’s interface, $\mathcal{U}$ comprises requested text instructions T , discrete actions A , or camera poses P .

Consistency Decomposition. Existing works examine aspects of consistency, including local geometric agreement or content recovery upon return (Duan et al., 2025; Ying et al., 2026). A systematic evaluation should account for what the inputs constrain, what the model is free to generate, and whether the generated content forms a coherent 3D scene. We evaluate three dimensions covering input content, newly generated regions, and 3D structure.

Input preservation

We test whether the model **preserves the scene shown in the initial image**. Distinctive static objects and structures are annotated as **anchors**, providing landmarks that should remain in place across viewpoint changes and reappear when the camera returns. Evaluating anchors individually prevents a missing landmark from being overlooked in a scene-wide average dominated by other visible objects and scene content.

Generated-content persistence

We evaluate **regions absent from the input image**. Their appearance is not prescribed, but the controls specify when the model should revisit them. We compare these revisits with the first generated observations to test whether the same content remains at corresponding locations when the camera revisits the region.

3D self-consistency

We examine the generated video itself, using recovered cameras to test **cross-view agreement** and whether a **shared 3D reconstruction** explains observations beyond those used to build it. These tests assess internal coherence independently of the requested camera trajectory used to produce the video.

Exploration-and-Revisit Protocols. Exploration-and-revisit trajectories test input preservation, generated-content persistence, and 3D self-consistency within one rollout. Six templates challenge preservation under changes in parallax and viewing direction: ¹*forward and backward*, ²*lateral return*, ³*yaw return*, ⁴*arc return*, ⁵*orbit around an anchor*, and ⁶*peek and return*. In peek-and-return, the camera moves past an anchor into an unseen region, returns, and revisits that region on a second excursion. Each trajectory starts at P_0 and returns to the same camera pose at P_r . For **input preservation**, we compare the return view against the initial RGB image and reference depth. A cross-view query at P_q uses registered RGB-D and annotated anchors to evaluate input-visible surfaces from another viewpoint. For **generated-content persistence**, we pair a first visit at P_{first} to a region outside the input view with a revisit at P_{revisit} after exploring elsewhere. The first generated observation provides the reference for comparing appearance and geometry across visits. For **3D self-consistency**, views sampled along the trajectory support comparisons of corresponding surfaces across views and reconstruction tests using held-out generated views.

3.2 Reference-Guided Data Construction

We combine ¹*public datasets*, ²*self-captured scenes*, and ³*UE simulation* to cover diverse indoor and outdoor environments. Each case pairs **registered RGB-D observations** from different viewpoints with a synthesized revisit trajectory that follows one of six templates and passes through both reference poses. This enables controlled exploration from existing observations without frame-by-frame RGB-D references. Figure 2 summarizes the four-stage construction and annotation pipeline, from selecting reference views to reviewing the completed cases.

Geometry-Guided View Selection. We retain existing geometry and recover missing annotations with ViPE (Huang

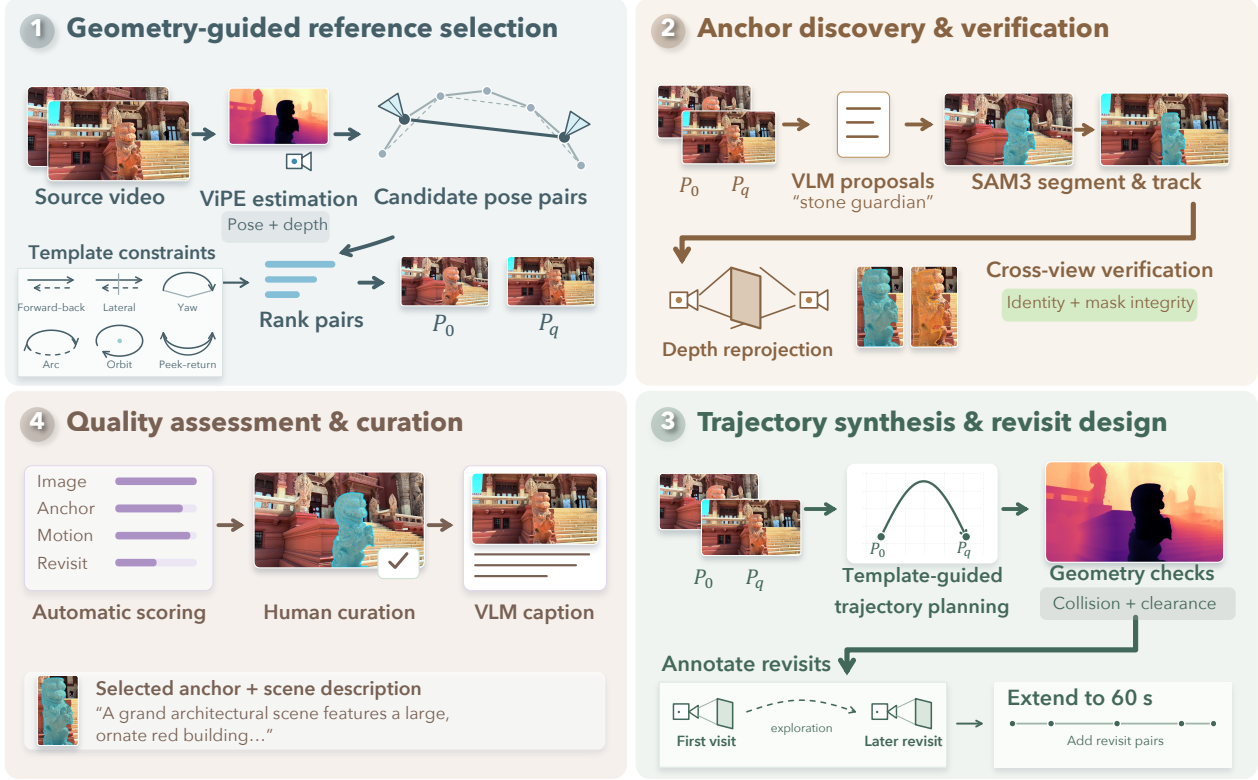


Figure 2. Data construction and annotation pipeline. ¹**Reference selection** uses geometry and template constraints. ²**Anchor verification** combines VLM proposals, segmentation, tracking, and cross-view surface checks. ³**Trajectory synthesis** defines geometrically valid exploration and revisits. ⁴**Quality review** includes ranking, human curation, and captioning.

et al., 2025a). We select input–query pairs (P_0, P_q) from registered poses using each template’s rotation and translation requirements. Ranking favors sufficient viewpoint change while balancing coverage across templates and scene content.

Anchor Discovery and Verification. A VLM (Qwen Team, 2026) proposes bounded, distinctive static landmarks visible at both endpoints. SAM3 (Carion et al., 2025) segments them independently at both endpoints. We filter masks by area, connectivity, and completeness, then track shortlisted objects to verify **cross-view identity**. Back-projecting input anchors with depth tests their visibility and depth agreement in the query view, verifying **cross-view surface support** for landmarks visible in both the input and query observations.

Trajectory Synthesis and Revisit Design. We fit template parameters to synthesize trajectories through P_0 and P_q that return to the initial pose. Paths are checked against scene geometry and revised for clearance. We select **temporally separated visits** to newly revealed regions based on pose proximity, novelty relative to the input, and shared visible surfaces. Minute-long extensions increase travel distance while retaining the short case’s **input, anchors, and reference poses**. Additional revisit pairs test whether appearance and geometry persist in these regions after further exploration.

Quality Assessment and Human Curation. We rank cases by image sharpness, anchor visibility, mask integrity, camera motion, and shared surface support across revisits. Reviewers inspect references, anchor overlays, projected support, and trajectories to correct or reject cases. A VLM produces factual scene captions from source frames to provide scene descriptions alongside the initial images as model inputs.

3.3 Dataset Scale and Coverage

We curate 240 cases from approximately 110,000 source samples across four public datasets, ScanNet++ (Yeshwanth et al., 2023), SpatialVID (Wang et al., 2026a), Sekai (Li et al., 2025d), and MiraData (Ju et al., 2024). For simulation, we collect 48 UE environments and render 365 candidate videos, retaining 180 cases after curation. We also record 85 videos ourselves and retain 60. Together, data annotation and **manual curation** required over **300 hours**.

ORBIA contains **720 cases** and **856 anchor evaluation targets**, with **480 short cases** and **240 paired minute-long extensions**, standardized at 720p and 16 FPS. The short cases cover 120 indoor and 360 outdoor settings, including 300 cases from real-world observations and 180 from simulation. Short trajectories last approximately 22 seconds with one revisit; their minute-long extensions include three revisits to test preservation through longer exploration and repeated returns. Source-specific construction procedures and dataset composition are detailed in Appendix B.

4 Evaluating 3D Consistency

Given a generated video, we estimate depth and poses with DA3 (Lin et al., 2026) and evaluate the designated observations against registered RGB-D references or earlier generated views. Cross-view agreement and held-out rendering assess internal 3D consistency. We describe the core consistency metrics and long-horizon evaluation below. Detailed metric definitions and computation procedures are provided in Appendix C.

■ **Input Preservation.** To assess input preservation at two checkpoints, we compare the return at P_r with the input at P_0 and the query at P_q with **registered RGB-D references on input-visible surfaces**. ¹*Scene Appearance Fidelity* quantifies appearance fidelity at return and query views, while ²*Anchor Appearance Fidelity* assesses the preservation of distinctive landmarks across viewpoints and on return. Both use DINO similarity (Oquab et al., 2024), PSNR, SSIM, and LPIPS on input-visible scene or anchor regions, with **equal weight per anchor**. ³*Scene Geometry Fidelity* quantifies geometric fidelity to the observed scene, while ⁴*Anchor Geometry Fidelity* assesses anchor geometry relative to the surrounding scene. Both measure depth AbsRel against the registered reference using a shared scene-level scale. ⁵*Anchor Identity Consistency* tests anchor identity retention during exploration. SAM3.1 (Carion et al., 2025) tracks input masks toward P_q . We average masked DINO cosine similarity to the input over four times and equally across anchors.

■ **Generated-Content Persistence.** For each annotated pair $(P_{\text{first}}, P_{\text{revisit}})$, **the first generated observation serves as the reference** for a revisit at the same pose or an overlapping viewpoint. ¹*Revisit Appearance Consistency* measures appearance retention across revisits, while ²*Revisit Depth Consistency* measures geometric consistency across visits. Both use **surfaces visible in both observations**, matched by projecting first-visit depth into the revisit view with estimated cameras; we compare DINO features for appearance and compute AbsRel for depth. At identical requested poses, we also compare images directly with DINO, PSNR, SSIM, and LPIPS, and depth maps at the shared estimator scale. ³*Revisit Surface Agreement* measures whether both visits reconstruct the same 3D surfaces. We compute a bidirectional nearest-point F-score between aligned point clouds, with matching distance relative to scene depth.

■ **3D Self-Consistency.** We assess whether generated views form **a coherent 3D scene** using cross-view correspondences from recovered cameras. ¹*Cross-View Depth Consistency* tests whether views agree on the depth of shared surfaces, while ²*Cross-View Feature Consistency* tests whether matched surfaces retain consistent visual features across views. Both use valid, unoccluded correspondences established by bidirectional depth reprojection for frame pairs at short and medium temporal intervals, measuring depth AbsRel and DINO cosine similarity, respectively. ³*Regional Scale Agreement* tests relative scale consistency across scene regions. We measure dispersion in observed-to-reprojected depth ratios across a 4×4 grid. ⁴*Reconstruction* tests **reconstruction accuracy on unseen views**. DA3 predicts a Gaussian scene (Kerbl et al., 2023) from **64 context frames**; we render **16 held-out views** and compare them with the corresponding generated frames using PSNR, SSIM, and LPIPS for pixel-level and perceptual agreement.

■ **Consistency over Longer Exploration.** We track ¹*visual quality*, ²*control*, and ³*local geometry* in **five-second windows** for both short and minute-long videos. Temporal stability combines the **average performance** with its **retention relative to the initial window** to assess whether a model sustains its capabilities during continued exploration. The final score weights short- and long-video means by **30%** and **70%**, respectively. We additionally evaluate **control accuracy** and **visual quality**. Appendix C details these metrics and evaluation procedures.

5 Experiments

We compare **27 models**: **19 camera-** or **action-conditioned world models** and **eight T12V models**. Adapters map requested trajectories to native camera conditions, discrete commands, or text instructions, starting from the same initial image. Table 2 reports all six tiers and the overall score. Appendices D and E provide adapters and component results.

Table 2. Comparison across six evaluation tiers. Gen. Persist. denotes generated-content persistence. **Temp. Stab.** denotes temporal stability. **3D Repr.** indicates scene geometry used by the generator (✓: present, ×: absent, -: unconfirmed). Models are sorted by **Overall** within each group. **1st**, **2nd**, and **3rd** mark the top three scores per column and group.

Model	Size (B)	Steps	3D Repr.	Control Align.	Visual Quality	Input Preserv.	Gen. Persist.	3D Self-Consist.	Temp. Stab.	Overall
World Models										
SolarWM-H3 (2026)	33	4	×	79.93	74.76	69.26	70.28	68.84	84.05	76.32
Lyra2 (2026)	14	4	✓	81.64	71.37	68.79	70.25	61.51	82.31	74.77
EVOKE (2026)	14	3	✓	66.47	76.48	68.81	66.43	68.02	83.18	72.70
Echo-WM (2026)	22	4	×	70.74	75.02	65.91	61.42	65.26	82.69	72.24
Gen3C (2025)	7	35	✓	75.51	67.55	75.23	68.58	60.16	78.96	71.95
Matrix game 3.5 (2026)	5	3	✓	64.18	74.03	66.93	60.82	54.61	76.41	68.07
AlayaWorld (2026)	15	4	✓	61.68	71.04	70.17	56.62	58.92	76.31	67.01
Genie3 (2025)	–	–	–	55.26	74.43	61.01	54.41	61.25	83.15	66.72
LingBot-World v2 (2026)	14	4	×	58.48	75.34	61.54	52.00	57.11	77.28	65.98
HY-WorldPlay (2025)	8	4	×	45.26	72.06	66.84	69.84	70.34	77.32	65.50
SANA-WM (2026)	2.6	4 + 3	×	59.39	73.22	63.19	51.95	54.10	75.03	65.08
HappyOyster	–	–	–	60.49	68.78	58.23	46.31	59.17	78.25	64.34
Cosmos3-Nano (2026)	16	30	×	35.70	74.25	66.78	66.29	70.10	74.71	62.75
LingBot-World v1 (2026)	14	4	×	49.12	73.84	59.56	50.71	58.93	75.37	62.73
Voyager (2025b)	13	50	✓	60.92	68.05	58.65	43.72	46.32	78.04	62.72
DreamX-World (2026)	5	4	×	48.99	74.96	60.96	48.70	48.73	74.82	61.79
FantasyWorld (2025)	14	50	✓	39.16	70.70	63.08	60.17	63.53	72.51	60.64
ABot-World (2026)	5	4	×	47.06	66.97	57.47	45.22	48.32	76.32	58.87
Astra (2025)	1.3	50	×	28.45	61.30	59.38	44.74	59.91	67.31	52.30
TI2V Models										
Seedance2.5 (2026)	–	–	–	43.23	76.97	72.84	67.70	72.54	80.39	67.44
Kling3.0 (2026)	–	–	–	38.58	77.38	69.14	63.32	71.30	79.30	65.23
Cosmos3-Super (2026)	64	4	–	27.92	79.04	66.04	69.75	74.55	79.73	63.72
HunyuanVideo1.5 (2025)	8.3	12	×	36.09	77.07	63.12	57.44	71.31	78.96	63.27
MiniMax-H3 (2026)	33	50	–	32.48	78.09	63.36	55.79	69.42	77.51	62.00
LTX2.5 (2026)	22	8 + 3	–	26.51	77.96	64.81	60.25	71.64	77.30	61.25
Wan2.7 (2025)	–	–	–	31.31	76.46	65.96	54.35	64.19	75.79	60.55
LongCat (2025)	13.6	16	×	23.76	77.30	58.89	47.77	64.11	74.22	57.18

5.1 Overall Performance and Capability Gaps

SolarWM-H3 leads overall. SolarWM-H3 ranks first overall at 76.32, with the strongest generated-content persistence and temporal stability among world models. Lyra2 and EVOKE rank second and third, leading control and visual quality, respectively. Gen3C leads input preservation, while HY-WorldPlay leads 3D self-consistency. Seedance2.5 leads the TI2V group overall, whereas Cosmos3-Super leads its visual quality, generated-content persistence, and 3D self-consistency. No model leads all six tiers.

Several leading world models use explicit 3D memory. The 3D Repr. column in Table 2 shows that three of the top five world models use explicit geometry: Lyra2, EVOKE, and Gen3C. Lyra2 uses per-frame geometry to retrieve relevant past frames and establish correspondences with target views (Shen et al., 2026). EVOKE stores observed geometry in an external world state bank and renders it from the current camera pose (Yin et al., 2026). Gen3C maintains a 3D cache whose renderings guide generation along the requested camera trajectory (Ren et al., 2025). By linking past observations to 3D locations, these mechanisms let models reuse scene content at a requested viewpoint even after it leaves the recent video context. This may help explain their strong control and content-preservation results.

Camera control is the largest gap between the two model groups. World models average 57.29 in control, compared with 32.48 for TI2V models, a gap of 24.80 points. Even the best TI2V model, Seedance2.5 (43.23), falls well below Lyra2 (81.64). Control also has the largest standard deviation across world models, indicating substantial differences even among models with native camera or action interfaces. These results identify accurate execution of requested camera motion as a major obstacle to using current video generators as interactive worlds.

Leading world models approach TI2V visual quality. TI2V models have higher mean visual quality (77.53 versus



Figure 3. Input preservation and generated persistence expose different failures. **Top two rows:** reference input P_0 and query P_q versus model returns and queries. Cyan contours mark *anchors*. **Bottom two rows:** first visits and revisits on another case. Headers report the evaluated models and their corresponding **case-level tier scores** for each example.

71.80), but EVOKE reaches 76.48, only 1.05 points below the TI2V mean. The best models in the two groups differ by 2.56 points. Within the TI2V group, the standard deviation is just 0.75, the smallest of its six tiers. Under the current metrics, quality distinguishes TI2V models less clearly than control and consistency.

TI2V models generate strong 3D self-consistency despite weak control. TI2V models average 69.88 in 3D self-consistency, exceeding world models by 10.14 points. Cosmos3-Super (74.55) also surpasses the best world model, HY-WorldPlay (70.34). Its strong depth, cross-view feature, and regional-scale scores show that this advantage extends beyond reconstruction fidelity (Table 11). Thus, the TI2V control gap does not stem simply from an inability to generate internally consistent scenes. The challenge is to make those scenes unfold along the requested camera trajectory.

Models differ substantially in generated-content persistence. The group means are close for both input preservation (64.30 for world models and 65.52 for TI2V) and generated-content persistence (57.29 and 59.55). Gen3C leads input preservation at 75.23, while SolarWM-H3, Lyra2, and HY-WorldPlay match or exceed the best TI2V persistence score. Within each group, however, generated-content persistence varies more than input preservation: their standard deviations are 9.22 versus 4.79 for world models and 6.80 versus 3.91 for TI2V. **Models differ more in their ability to maintain newly generated content than to preserve the input scene.**

Leading world models match or exceed TI2V temporal stability. SolarWM-H3, EVOKE, and Echo-WM score 84.05, 83.18, and 82.69 in temporal stability, respectively, all above the best TI2V score of 80.39 from Seedance2.5. These world models generate autoregressively in causal chunks (Huang et al., 2026; Yin et al., 2026; Zhang et al., 2026), whereas TI2V generators such as LTX2.5 use bidirectional temporal attention within each clip (Lightricks, 2026). These causal models already match or exceed the temporal stability of bidirectional TI2V generators.

Table 3. Tier 3: Detailed input preservation results. Scores use a 0–100 scale (↑). Models are ordered by **Tier Score** within each group. 1st, 2nd, and 3rd mark the top three scores per column and group.

Model	3D Repr.	Return	Query	Identity	Tier Score
World Models					
Gen3C	✓	81.24	63.93	84.28	75.23
AlayaWorld	✓	66.56	60.47	86.71	70.17
SolarWM-H3	×	54.73	65.30	89.06	69.26
EVOKE	✓	57.72	61.79	89.25	68.81
Lyra2	✓	60.33	62.82	85.19	68.79
Matrix game 3.5	✓	52.40	62.86	86.89	66.93
HY-WorldPlay	×	57.26	55.88	91.04	66.84
Cosmos3-Nano	×	54.91	56.17	92.79	66.78
Echo-WM	×	50.07	62.32	86.55	65.91
SANA-WM	×	47.17	60.44	82.89	63.19
FantasyWorld	✓	45.98	56.64	88.77	63.08
LingBot-World v2	×	41.69	56.85	87.65	61.54
Genie3	–	45.53	53.91	85.98	61.01
DreamX-World-5B	×	42.54	54.22	88.34	60.96
LingBot-World v1	×	42.92	52.75	85.27	59.56
Astra	×	41.31	52.92	86.06	59.38
Voyager	✓	40.07	57.25	79.09	58.65
HappyOyster	–	42.64	47.58	88.02	58.23
ABot-World	×	40.11	50.88	83.61	57.47
TI2V Models					
Seedance2.5	–	74.16	57.38	92.14	72.84
Kling3.0	–	63.64	54.86	93.69	69.14
Cosmos3-Super	–	54.13	54.43	93.44	66.04
Wan2.7	–	54.58	54.09	93.16	65.96
LTX2.5	–	52.47	53.50	92.24	64.81
MiniMax-H3	–	48.04	54.04	91.11	63.36
HunyuanVideo1.5	×	45.69	54.08	92.60	63.12
LongCat	×	40.55	51.09	87.64	58.89

5.2 Complementarity of 3D Consistency Metrics

Input preservation measures fidelity to the input scene. Gen3C leads input preservation and Return. Figure 3 shows its advantage in both return and query views. In the first row, Gen3C’s return view most closely matches the input reference, while SolarWM-H3 and LingBot-World v2 show substantial forgetting. In the second row, Gen3C preserves both the appearance and spatial placement of key anchors, including the cabinet and lamp, closely matching the query reference, while other models place these anchors in incorrect relative positions.

Input preservation requires complementary checks. SolarWM-H3 leads Query, while Cosmos3-Nano leads Identity (Table 3). Gen3C instead leads Return, showing that the three components favor different models. Query checks appearance and geometry at a new viewpoint, Identity tracks anchors during the input-to-query interval, and Return tests recovery of the initial scene after the full exploration trajectory. The three checks jointly assess input preservation during exploration and after return, covering failures that any one component alone would miss.

Semantic similarity alone does not ensure visual fidelity. High semantic similarity between a first visit and revisit can coexist with poor perceptual fidelity (Figure 3, bottom two rows). MiniMax-H3, for example, scores 77.27 on different-view DINO but only 17.32 on normalized same-view LPIPS (Table 4). A revisit can match the scene’s general content while changing textures and object appearance. Our tests therefore combine semantic, perceptual, and pixel-level comparisons to check both whether the same content reappears and whether its visual details are preserved.

Appearance consistency does not establish spatial agreement. Lyra2 leads several revisit-appearance measures, whereas HY-WorldPlay leads reprojected depth consistency and SolarWM-H3 leads surface agreement (Table 4). The appearance tests compare how the revisited content looks; the depth and surface tests check whether it occupies the

Table 4. Tier 4: Detailed generated-content persistence results. Scores use a 0–100 scale (↑). Models are ordered by **Tier Score** within each group. **1st**, **2nd**, and **3rd** mark the top three scores per column and group. **Same** and **Diff.** denote direct comparisons at identical and different requested poses. **Reproj.** denotes geometry-aligned comparison.

3D		Revisit Appearance					Revisit Depth		Revisit Surface	Tier
Model	Repr.	Same PSNR	Same LPIPS	Same DINO	Diff. DINO	Reproj. DINO	Same AbsRel	Reproj. AbsRel	F-score	Score
World Models										
SolarWM-H3	×	43.68	47.45	62.88	87.33	62.59	82.40	80.76	59.54	70.28
Lyra2	✓	53.88	54.30	67.72	82.78	64.44	80.23	78.73	58.12	70.25
HY-WorldPlay	×	44.18	44.82	59.34	85.48	62.06	82.13	84.24	59.06	69.84
Gen3C	✓	55.64	51.52	64.19	81.86	61.85	84.91	76.98	53.12	68.58
EVOKE	✓	47.53	44.61	63.19	84.12	59.83	81.98	74.36	49.14	66.43
Cosmos3-Nano	×	53.56	42.64	56.92	81.25	58.82	81.11	77.31	52.85	66.29
Echo-WM	×	37.11	34.83	53.10	82.07	55.65	76.78	70.25	45.20	61.42
Matrix game 3.5	✓	41.68	42.79	59.46	78.42	58.07	74.27	64.94	42.67	60.82
FantasyWorld	✓	34.42	32.52	52.29	78.65	54.90	77.16	71.66	42.84	60.17
AlayaWorld	✓	34.91	30.74	53.41	73.74	53.54	70.07	63.97	40.42	56.62
Genie3	–	36.73	22.95	47.16	71.63	47.68	63.50	75.41	35.00	54.41
LingBot-World v2	×	22.34	16.70	41.38	73.30	47.35	66.59	66.51	34.38	52.00
SANA-WM	×	35.27	24.40	44.74	76.55	48.62	71.64	54.43	29.63	51.95
LingBot-World v1	×	20.83	16.71	41.39	68.48	46.90	61.95	68.50	33.99	50.71
DreamX-World-5B	×	31.04	22.14	44.26	76.47	46.63	62.31	50.91	23.70	48.70
HappyOyster	–	42.68	17.64	39.55	64.13	42.51	46.99	64.42	23.23	46.31
ABot-World	×	46.29	20.84	42.32	57.58	44.37	63.19	53.52	21.12	45.22
Astra	×	36.99	14.19	37.64	54.08	39.27	74.58	62.49	19.26	44.74
Voyager	✓	13.52	14.44	43.39	61.92	41.94	61.93	49.34	25.61	43.72
TI2V Models										
Cosmos3-Super	–	36.85	41.66	60.33	88.60	61.41	85.15	82.96	57.57	69.75
Seedance2.5	–	40.41	37.59	57.08	86.79	57.92	79.65	83.76	55.56	67.70
Kling3.0	–	34.59	32.08	52.33	82.83	53.45	77.90	82.64	46.89	63.32
LTX2.5	–	37.47	29.93	51.78	78.80	50.16	78.43	76.91	42.05	60.25
HunyuanVideo1.5	×	19.81	20.71	48.34	80.75	51.06	69.26	73.71	40.99	57.44
MiniMax-H3	–	19.68	17.32	43.91	77.27	46.31	70.33	79.04	37.73	55.79
Wan2.7	–	20.18	21.02	45.81	74.97	48.09	73.87	70.31	34.15	54.35
LongCat	×	18.27	14.95	38.19	67.91	42.38	65.31	66.08	25.23	47.77

same 3D locations. Preserving visual details and preserving the 3D layout therefore require separate checks.

Local cross-view agreement does not ensure coherent scene reconstruction. FantasyWorld ranks second among world models in cross-view feature consistency at both temporal intervals, but second to last in reconstruction PSNR (Appendix E.1.5). HY-WorldPlay likewise leads depth consistency and regional scale agreement without achieving the best reconstruction score. These ranking differences show that strong agreement between pairs of views does not necessarily mean that the generated views can be reconstructed into a coherent 3D scene. Reconstruction consequently provides a separate test of whether the video depicts one coherent scene throughout exploration.

5.3 Relationships among Quality, Control, and Consistency

We examine Spearman correlations between tier scores. Figure 5 reports model-level correlations separately for short and long videos within each model group.

Visual quality does not ensure 3D consistency. LingBot-World v2 ranks second among world models in visual quality, but twelfth in both input preservation and generated-content persistence, and fourteenth in 3D self-consistency. Gen3C shows the reverse: seventeenth in visual quality but first in input preservation. Across all models, visual quality has a rank correlation of only 0.33 with the mean of the three consistency tiers among world models and 0.21 among Ti2V models. Figure 4 illustrates this discrepancy: LingBot-World v2 generates a detailed view but yields a less faithful

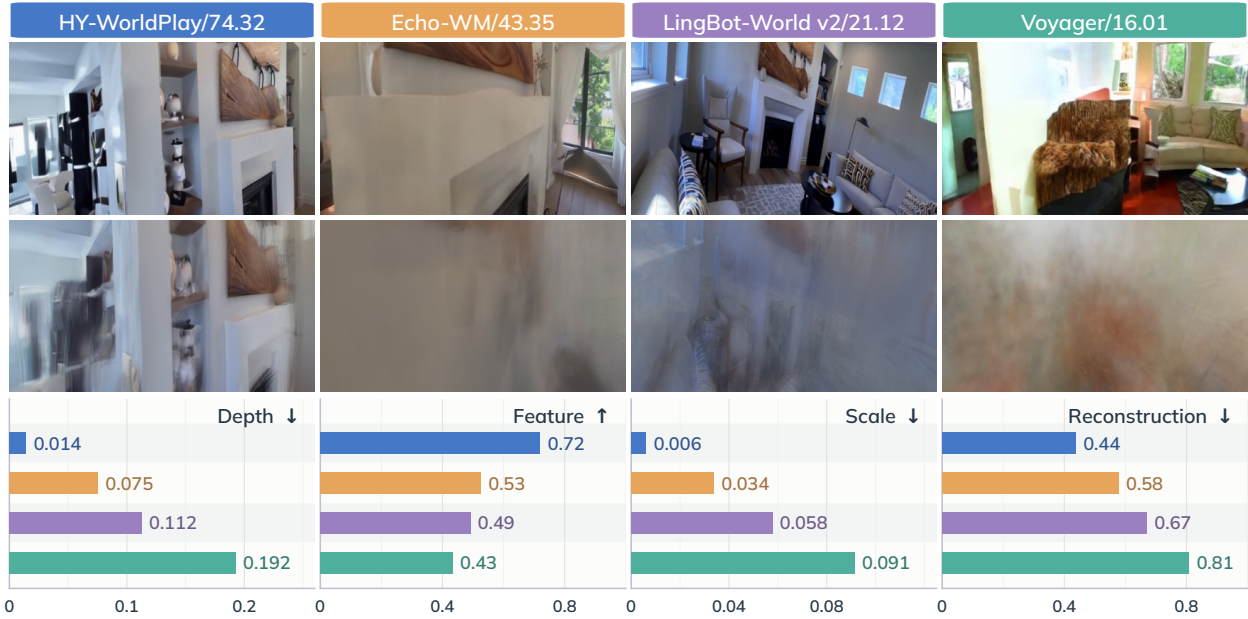


Figure 4. 3D self-consistency reveals reconstruction failures. **Top two rows:** held-out generated views and their Gaussian reconstructions. **Bottom:** depth, feature, scale, and reconstruction metrics. Header and bar colors identify models. Header values summarize the **case-level tier scores** for the corresponding evaluated videos.

reconstruction and weaker depth agreement than HY-WorldPlay. Visual quality alone therefore gives little assurance that a model preserves scene content or maintains a consistent 3D scene.

Strong visual quality does not guarantee faithful execution of the requested scene and camera motion. Quality and control have weak rank agreement among world models (0.08 for short videos and 0.30 for long videos), and negative correlations among TI2V models (-0.43 and -0.26). Short-video TI2V quality also correlates negatively with input preservation (-0.62). In Table 2, Cosmos3-Super leads quality but ranks sixth in control, while Seedance2.5 leads input preservation but ranks seventh in quality. Higher visual quality does not imply more accurate camera motion, and in short TI2V videos it can coincide with poorer input preservation.

Control and 3D self-consistency measure different capabilities. Lyra2 ranks first in control but seventh in 3D self-consistency, while HY-WorldPlay ranks sixteenth and first, respectively. Across world models, the rank correlation between the two tiers is only 0.09 in short videos and -0.04 in long videos (Figure 5). A model can generate views that agree with one another even when the camera follows the wrong path. Internal geometric agreement and accurate execution of camera commands are thus separate requirements for interactive world generation.

Preserving the input does not ensure preserving newly generated content. World-model input preservation and generated-content persistence have strongly aligned model rankings, but much weaker agreement across cases within each model. For short videos, the model-level correlation is 0.87, compared with a mean within-model correlation of 0.35; the same pattern holds for long videos. Strong models tend to perform well on both tasks, yet a successful input-preservation case can still lose newly generated content at revisit. A model’s ability to recover the input scene therefore does not reliably predict whether it will preserve the regions it creates during exploration.

Following the camera path is not enough to preserve content during long exploration. Among world models, the correlation between control and input preservation falls from 0.63 in short videos to 0.32 in long videos; for generated-content persistence, it falls from 0.53 to 0.22. Longer exploration requires models to preserve scene content across additional movements and revisits, even when they follow the requested camera path.

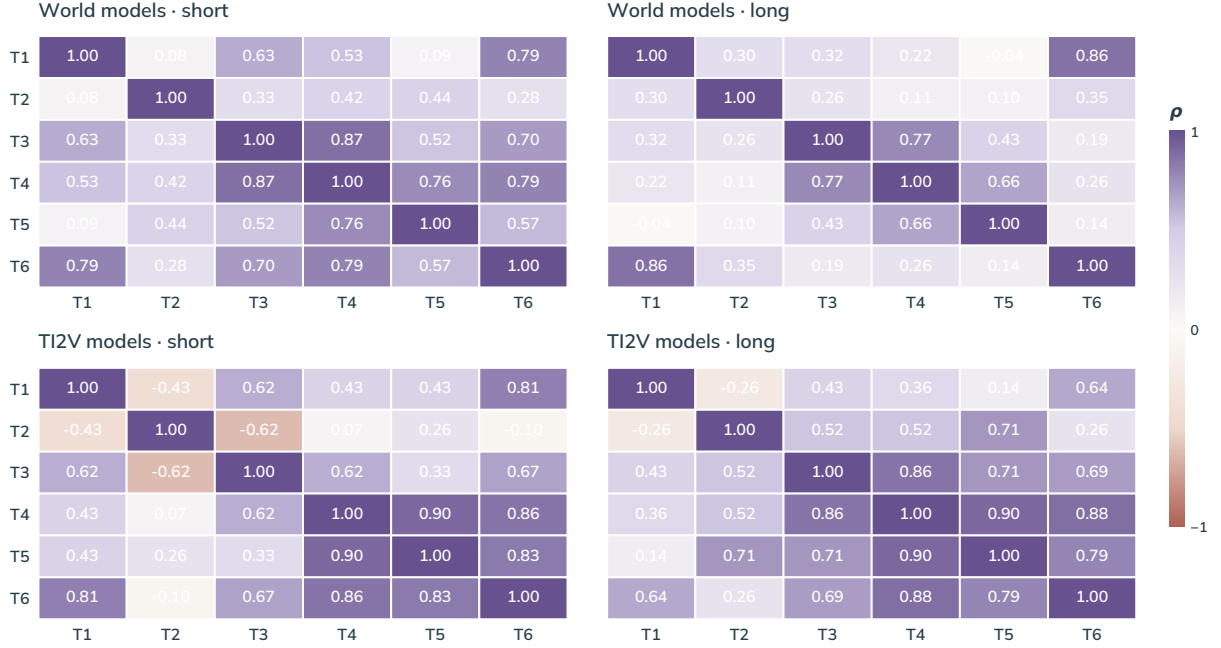


Figure 5. Cross-tier ranking correlations. Top: world models. Bottom: TI2V models. Left: short videos. Right: long videos. Cells show Spearman correlations between model-level tier scores.

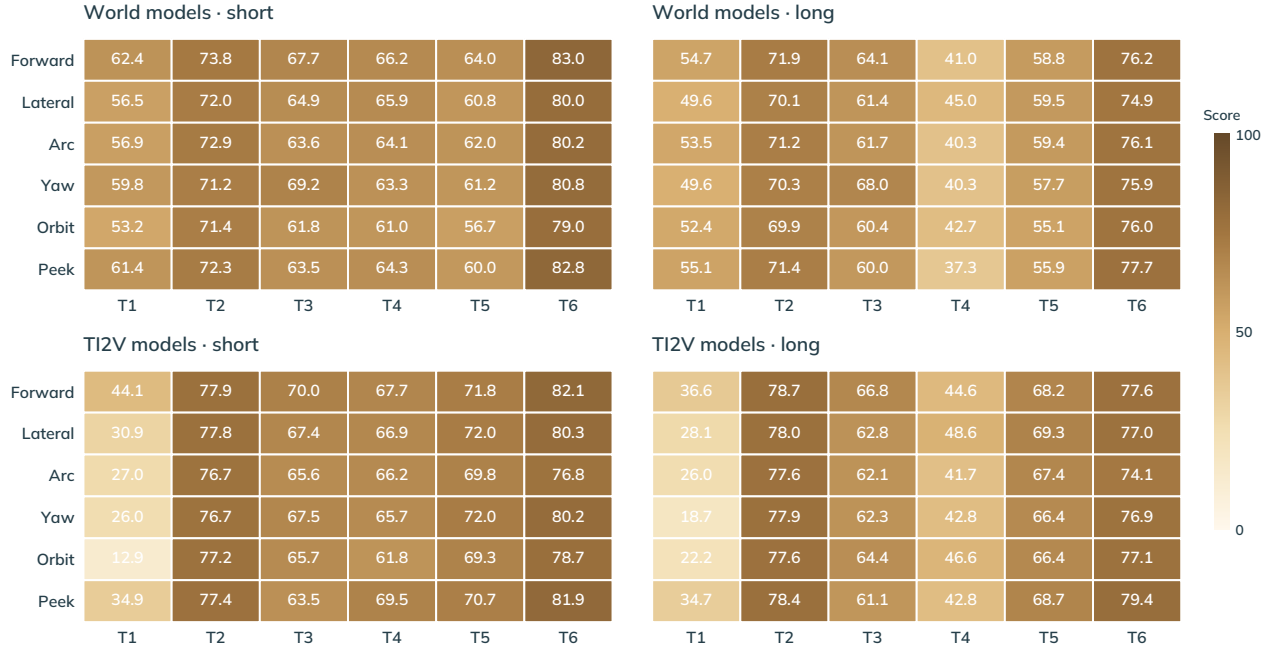


Figure 6. Performance across trajectory templates. Top: world models. Bottom: TI2V models. Left: short videos. Right: long videos. Cells show mean tier scores with equal model weights within each evaluated model group.

5.4 Control and Consistency across Trajectory Templates

Figure 6 compares all six trajectory templates using equal model weights within each group. The templates range from reversing a translation to circling an anchor or revealing an occluded region. These comparisons show which camera motions expose failures that are less apparent on a simple forward-backward path.

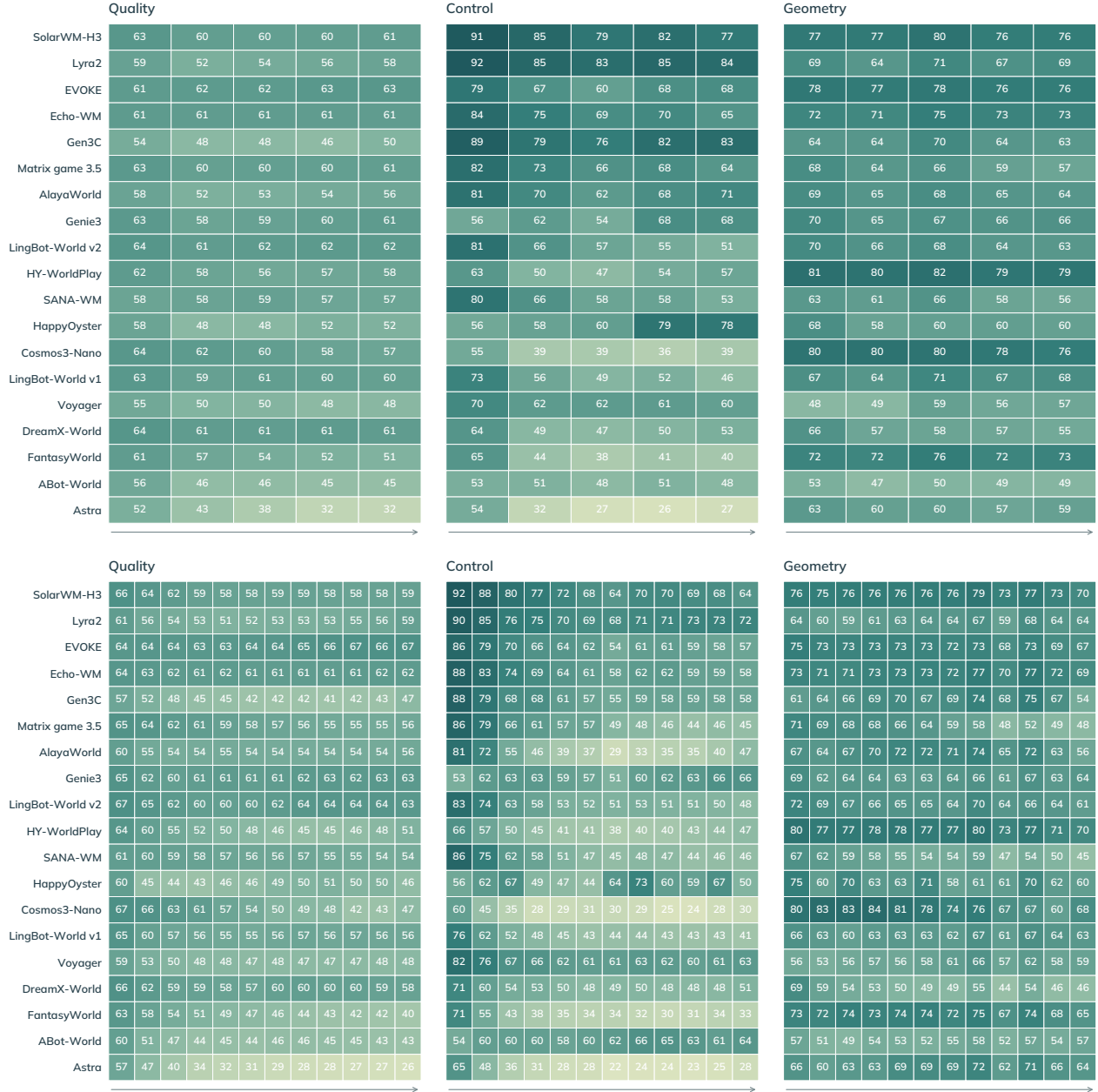


Figure 7. Temporal performance for world models. Top: short videos. Bottom: long videos. Rows follow the model order in Table 2. Time increases from left to right. Cells show mean window scores for **quality**, **control**, and **local geometry** on a shared 0–100 scale, rounded to integers. Darker colors indicate higher scores within each evaluated time window.

TI2V models fall behind in camera control on every trajectory template. TI2V models score below world models in control on all six templates, for both short and long videos. On short forward–backward trajectories, their mean control is 44.05, compared with 62.41 for world models. On orbit trajectories, the gap widens to 12.95 versus 53.15. Meanwhile, TI2V visual quality remains within 76.70–77.94 across the short-video templates. These models can produce good-looking frames across camera motions, but translating a text instruction into the specified trajectory remains difficult, especially when position and viewing direction must change together.

Orbit motion challenges both camera control and scene preservation. For short world-model videos, orbit has the lowest mean control, input-preservation, and generated-content-persistence scores among the six templates. Control falls

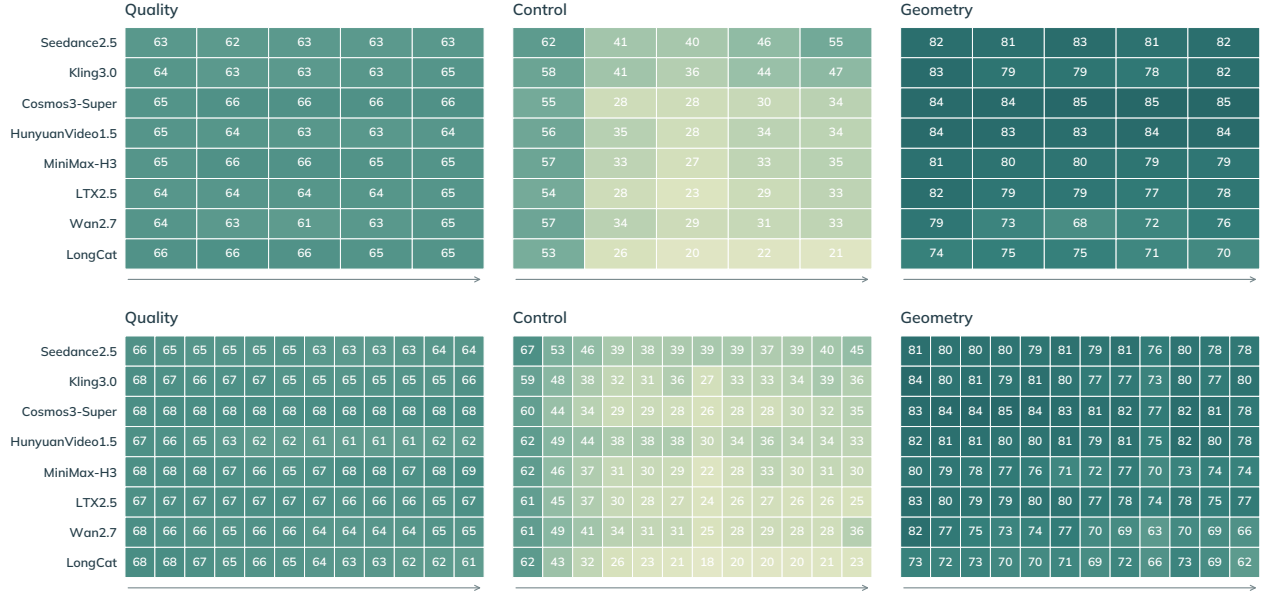


Figure 8. Temporal performance for TI2V models. Top: short videos. Bottom: long videos. Rows follow the model order in Table 2. Time increases from left to right. Cells show mean window scores for **quality**, **control**, and **local geometry** on a shared 0–100 scale, rounded to integers. Darker colors indicate higher scores within each evaluated time window.

from 62.41 on forward–backward motion to 53.15 on orbit, while input preservation and generated-content persistence reach 61.83 and 60.97. Circling an anchor repeatedly changes visible surfaces and occlusions. Simple reversal tests therefore underestimate the challenge of preserving a scene across different viewpoints.

Complex trajectories can still end in forgetting despite stronger control. For long world-model videos, peek-and-return has the highest control score (55.09) but the lowest generated-content persistence (37.32). Models can therefore follow this camera motion yet fail to preserve the content revealed during the peek. Preserving newly revealed content through changes in visibility remains difficult even when camera control is comparatively strong.

5.5 Performance Degradation during Long Exploration

Figures 7 and 8 show temporal performance in visual quality, control, and local geometry for all models in successive five-second windows. Local geometry includes depth, feature, and regional-scale consistency.

Control declines most over time in leading models from both groups. Across SolarWM-H3, Lyra2, EVOKE, Echo-WM, and Gen3C, control drops by an average of 26.98 points from the first to the final five-second window of minute-long videos. For Seedance2.5, Kling3.0, and Cosmos3-Super, the corresponding decline is 23.25 points. Control declines more than visual quality or local geometry in every one of these eight models. SolarWM-H3 falls from 91.91 to 64.50, and Cosmos3-Super from 60.26 to 34.94. Sustaining accurate camera motion remains difficult for both groups.

TI2V visual quality remains stable while local geometry deteriorates. During minute-long generation, local geometry declines more than visual quality in seven of the eight TI2V models. The mean declines are 7.13 points for local geometry and only 2.08 for visual quality. Wan2.7, for example, loses 16.26 points in local geometry but only 3.11 in quality. Cosmos3-Super’s quality improves slightly while its geometry falls by 5.34 points. Visually stable frames therefore do not ensure that geometric agreement is maintained during long exploration.

Long trajectories make newly generated content much harder to preserve. Figure 6 shows lower generated-content persistence (Tier 4) on long videos for every trajectory template in both model groups. On forward–backward trajectories, scores fall from 66.21 to 40.96 for world models and from 67.74 to 44.58 for TI2V models. On peek-and-return, they fall from 64.29 to 37.32 and from 69.52 to 42.83, respectively. The drop spans roughly 15–27 points across the six templates. Long exploration exposes forgetting even on simple forward–backward paths.

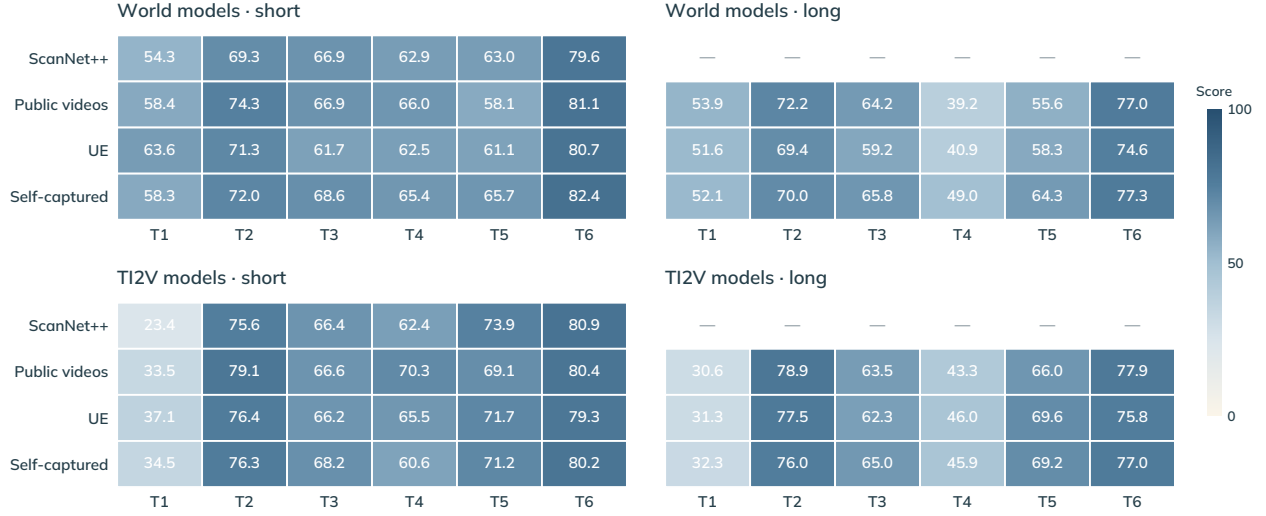


Figure 9. Performance across data sources. Top: world models. Bottom: TI2V models. Left: short videos. Right: long videos. Cells show **mean tier scores** with equal model weights. Dashes indicate *unavailable groups*.

5.6 Performance Variation across Scene Sources

Figure 9 compares ScanNet++ (with sensor depth and reconstructed meshes), a public video group (SpatialVID, Sekai, and MiraData), UE simulation, and self-captured scenes.

Regular UE trajectories are easier to follow in short videos. UE scenes have the highest short-video control scores in both groups: 63.6 for world models and 37.1 for TI2V models. The trajectories are authored directly from the six templates, then previewed and refined in the simulator. Regular motion may make these paths easier to follow.

The public video group yields the highest visual quality. Both model groups achieve their highest visual-quality scores on the public video group in both short and long videos. The scores are 74.3 and 72.2 for world models on short and long videos, and 79.1 and 78.9 for TI2V models. This advantage may reflect a closer match between these real-video scenes and the visual distributions learned by current video generators.

5.7 Alignment with Human Judgments

We assess whether automatic scores reproduce human preferences in the overall ordering of models across the six evaluation tiers. We select 90 cases spanning six models and distribute the annotation tasks among 20 raters. Raters rank the available videos separately for all six tiers. Ties are allowed. The interface provides the input and query references, requested trajectory, generated videos, and revisit clips appropriate to each judgment.

We convert rankings into relative preference scores, assigning 1 for a win, 0.5 for a tie, and 0 for a loss, and average across raters and cases. Human and automatic scores are compared on the same available cases. We compute the Spearman correlation between human and automatic model-level scores for each tier. Correlations range from 0.859 to 0.921 across the six tiers, with **temporal stability** showing the highest agreement. These results support the benchmark’s ability to distinguish models in agreement with human judgments.

6 Conclusion

ORBIA provides 720 exploration-and-revisit cases to evaluate input preservation, generated-content persistence, and 3D self-consistency. Across 27 models, strong visual quality and control do not ensure 3D consistency, while local agreement can coexist with poor scene reconstruction. These distinct failure patterns demonstrate the value of complementary tests of content and geometry. We hope **ORBIA** supports more reliable evaluation and guides the development of world models that preserve coherent scenes throughout exploration.

Appendix

A	Dataset Overview and Annotated Examples	19
A.1	Dataset Statistics	19
A.2	Exploration-and-Revisit Templates	19
A.3	Annotated Dataset Examples	20
B	Dataset Construction and Annotation	25
B.1	From Source Observations to Evaluation Cases	25
B.2	Construction from Different Data Sources	25
B.2.1	Public Videos: View Selection and Anchor Verification	25
B.2.2	ScanNet++: Mesh-Guided Reference Construction	27
B.2.3	UE Simulation: Controlled Capture and Rendered References	28
B.2.4	Self-Capture: Landmark-Guided Acquisition	30
B.3	Extending Exploration to One Minute	32
B.4	Quality Ranking and Human Review	32
B.5	Released Case Package	34
C	Evaluation Protocol and Metrics	34
C.1	Tier 1: Control Alignment	34
C.2	Tier 2: Visual Quality	35
C.3	Tier 3: Input Preservation	35
C.4	Tier 4: Generated-Content Persistence	37
C.5	Tier 5: 3D Self-Consistency	37
C.6	Tier 6: Temporal Stability	38
C.7	Evaluation Pipeline and Score Aggregation	39
D	Model Adapters and Generation Settings	40
D.1	Camera-Conditioned Models	40
D.2	Action-Conditioned Models	41
D.3	TI2V Models	44
E	Extended Results and Analysis	48
E.1	Detailed Results by Evaluation Tier	49
E.1.1	Tier 1: Control Alignment	49
E.1.2	Tier 2: Visual Quality	50
E.1.3	Tier 3: Input Preservation	51
E.1.4	Tier 4: Generated-Content Persistence	51
E.1.5	Tier 5: 3D Self-Consistency	53
E.1.6	Tier 6: Temporal Stability	53
E.2	Additional Annotated Examples and Review Interface	53

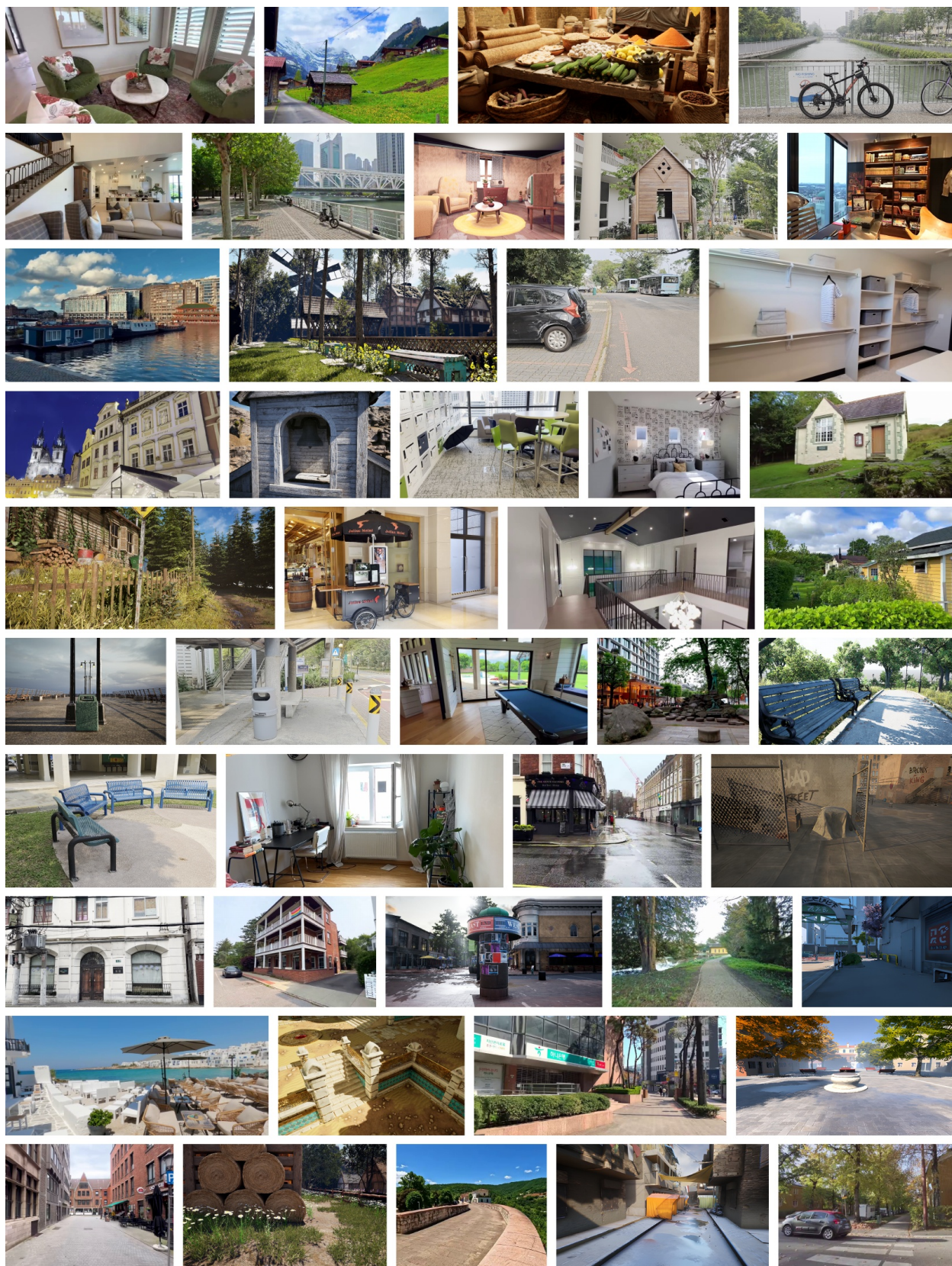


Figure 10. Dataset gallery. Selected input images from the ORBIA benchmark illustrate diverse indoor and outdoor environments.

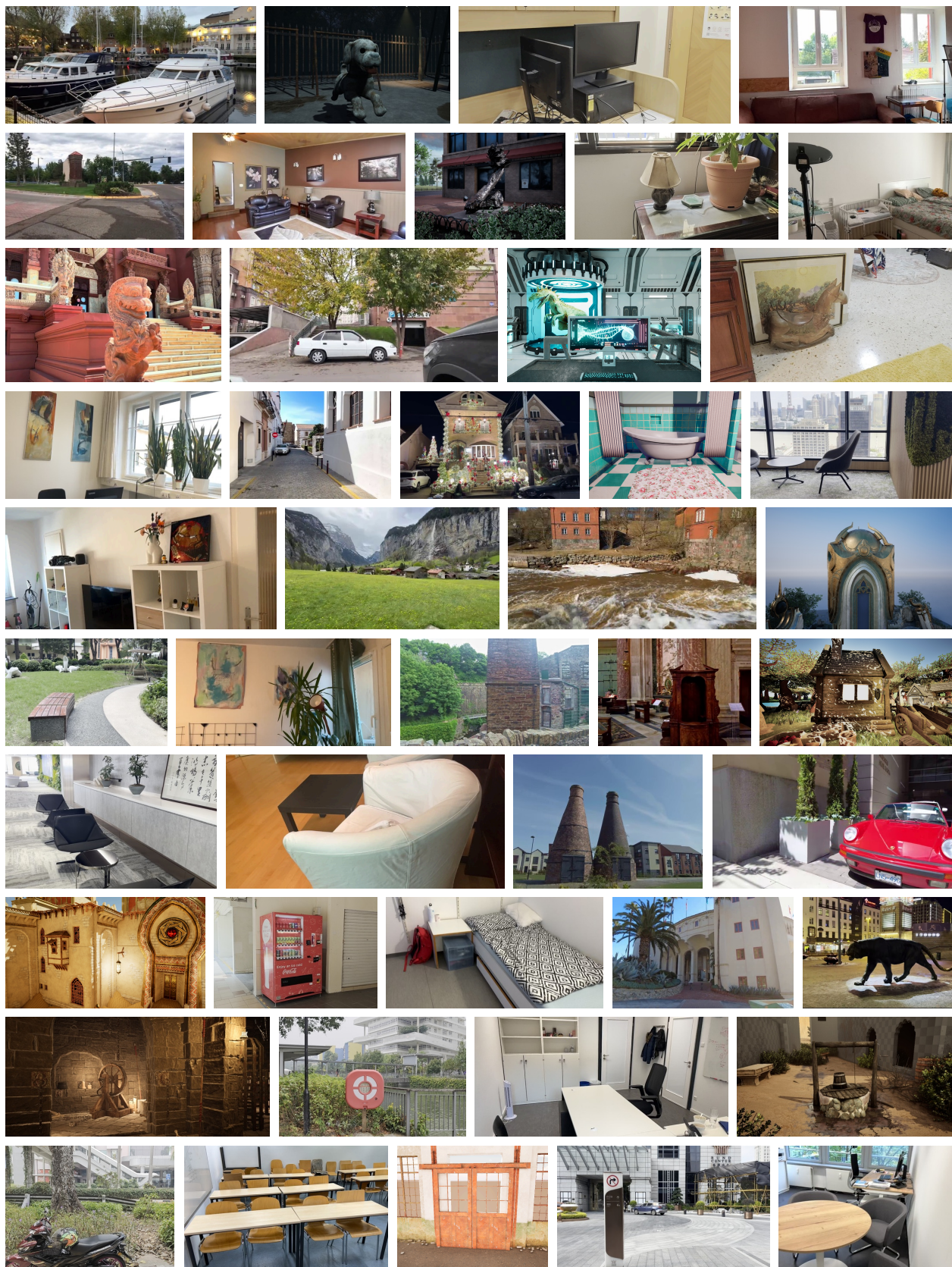


Table 5. Dataset composition.

Source	Data Type	Raw Data	Cases		Anchors		Geometry
			Short	Long	Short	Long	
ScanNet++ (Yeshwanth et al., 2023)	Real	906	60	0	106	0	Measured
SpatialVID (Wang et al., 2026a)	Real	100,000	134	74	177	90	Estimated
Sekai (Li et al., 2025d)	Real	8,000	25	25	34	34	Estimated
MiraData (Ju et al., 2024)	Real	1,500	21	21	28	28	Estimated
Our recordings	Self-captured	85	60	30	60	30	Estimated
UE simulation	Simulated	365	180	90	180	90	Rendered
Total		110,856	480	240	585	272	

Table 6. Short and long protocols. Median interval measures time between first visits and their paired revisits.

Specification	Short	Long
RGB resolution	1280 × 720	
Frame rate	16 FPS	
Frame count	357	961
Duration	22.25 s	60 s
Revisit pairs	1	3
Median interval	8.5 s	29.5 s

Table 7. Median input–query motion. Translation is camera-center distance in the original reference-pose units.

Template	Rotation	Translation
Forward–backward	1.1°	1.36
Lateral return	5.4°	1.47
Yaw return	31.3°	0.02
Arc return	11.6°	1.57
Orbit around an anchor	47.5°	1.26
Peek-and-return	11.2°	0.75

A Dataset Overview and Annotated Examples

This appendix describes the composition and reference annotations of **ORBIA**. We report dataset statistics in Section A.1, describe the exploration templates in Section A.2, and present complete annotated examples in Section A.3.

A.1 Dataset Statistics

ORBIA contains **720 evaluation cases** and **856 anchor evaluation targets**. The dataset combines varied scenes with registered reference views and repeatable camera trajectories, enabling comparisons of how models preserve observed content and maintain newly generated regions during exploration.

Scene coverage. Public datasets cover indoor rooms, streets, and natural environments. Self-captured videos add outdoor landmarks, while UE provides virtual scenes with rendered depth and known camera poses. This combination supports evaluation across real and simulated environments. We scanned 110,406 public videos and scenes, reduced them to 10,944 machine-selected candidates and an 881-case diversity pool, and retained 240 public-data cases. Table 5 lists the raw sample counts and final source composition.

Annotated references. Each case provides an input image and a registered query view, with depth and anchor masks. **Anchors** occupy a median of 5.2% of the input image and 5.4% of the query image, so changes to these objects can be overlooked in whole-image scores. Their masks and depth enable object-level appearance and geometry comparisons. At the query view, input preservation is evaluated on surfaces visible in the input; newly revealed regions are evaluated at later registered revisits to the corresponding scene surfaces.

Exploration protocols. Each template contributes 80 short and 40 long cases. Forward–backward and lateral trajectories primarily change camera position, while yaw and orbit introduce larger orientation changes (Table 7). **Long extensions** preserve the short case’s initial scene and references while adding more revisits and longer intervals between them (Table 6), allowing us to assess content preservation as exploration continues.

A.2 Exploration-and-Revisit Templates

The six trajectory templates vary how the camera explores a scene and revisits previously observed regions (Figure 12). Each case starts at P_0 and includes a registered query pose P_q and a return pose P_r . The paired poses ($P_{\text{first}}, P_{\text{revisit}}$)

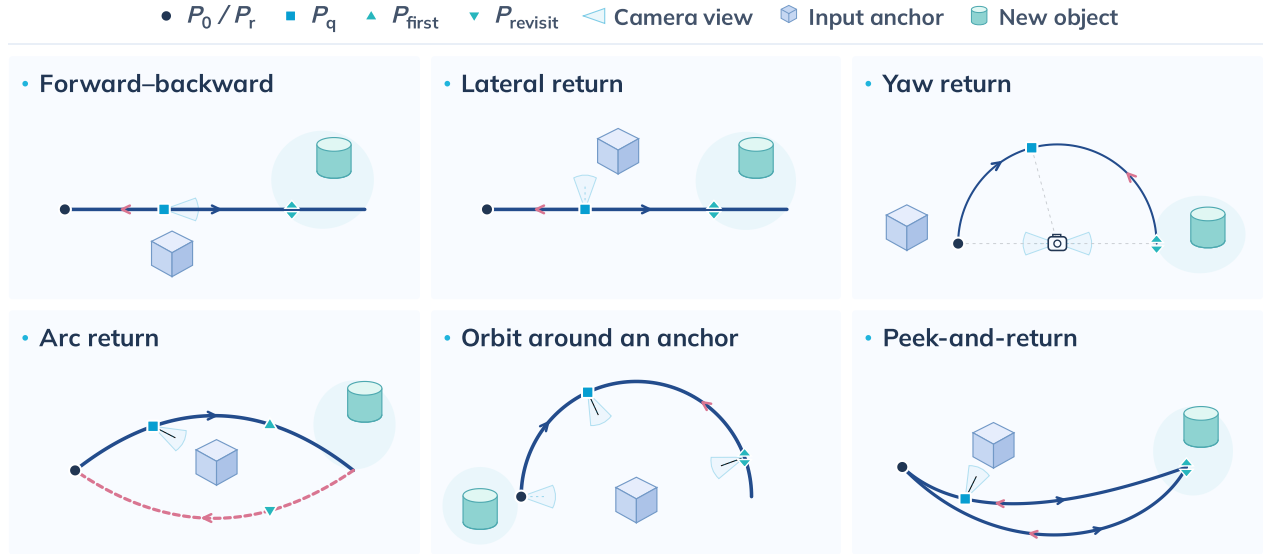


Figure 12. Exploration-and-revisit templates. Blue arrows show outward motion and rose arrows show return motion. In yaw return, the camera stays at a fixed center and the symbols along the arc indicate viewing directions.

provide two observations of content revealed during exploration. Section A.3 shows the trajectories and reference views for annotated examples from each of the four scene source groups.

- **Forward-backward.** The camera advances into the scene and retraces its path while maintaining its viewing direction. Changes in object size and occlusion test whether scene depth and object placement remain consistent as the camera approaches and retreats.
- **Lateral return.** The camera moves sideways and returns while keeping a similar viewing direction. Nearby and distant surfaces shift by different amounts, testing whether the model preserves their relative depth and recovers surfaces that become temporarily occluded.
- **Yaw return.** The camera rotates to reveal another part of the scene and then turns back, with little translation. This tests whether content remains recognizable after leaving the field of view and appearing again.
- **Arc return.** The camera follows a curved trajectory through the scene before returning. The changing position and viewing direction reveal different sides of objects, testing whether their appearance and geometry remain consistent across changing viewpoints along the curved trajectory.
- **Orbit around an anchor.** The camera moves around an anchor while looking toward it. Observing the same object from multiple directions tests whether its shape and position relative to surrounding objects remain consistent when viewed from different sides and at different points along the trajectory.
- **Peek-and-return.** The camera passes an anchor near the starting view, explores a previously unseen region, and returns. A second excursion revisits that region, providing the pair $(P_{\text{first}}, P_{\text{revisit}})$ to test whether newly generated content is retained between visits.

A.3 Annotated Dataset Examples

The following cards show model inputs, requested trajectories, and evaluator-only references. Query RGB, depth, and anchor masks are withheld from generation. We retain one figure for each source group: ScanNet++, public videos, UE, and self-captured videos. Additional trajectory examples appear in Appendix E.2. Trajectory panels show the requested camera positions and trajectories in a top-down view of the scene.

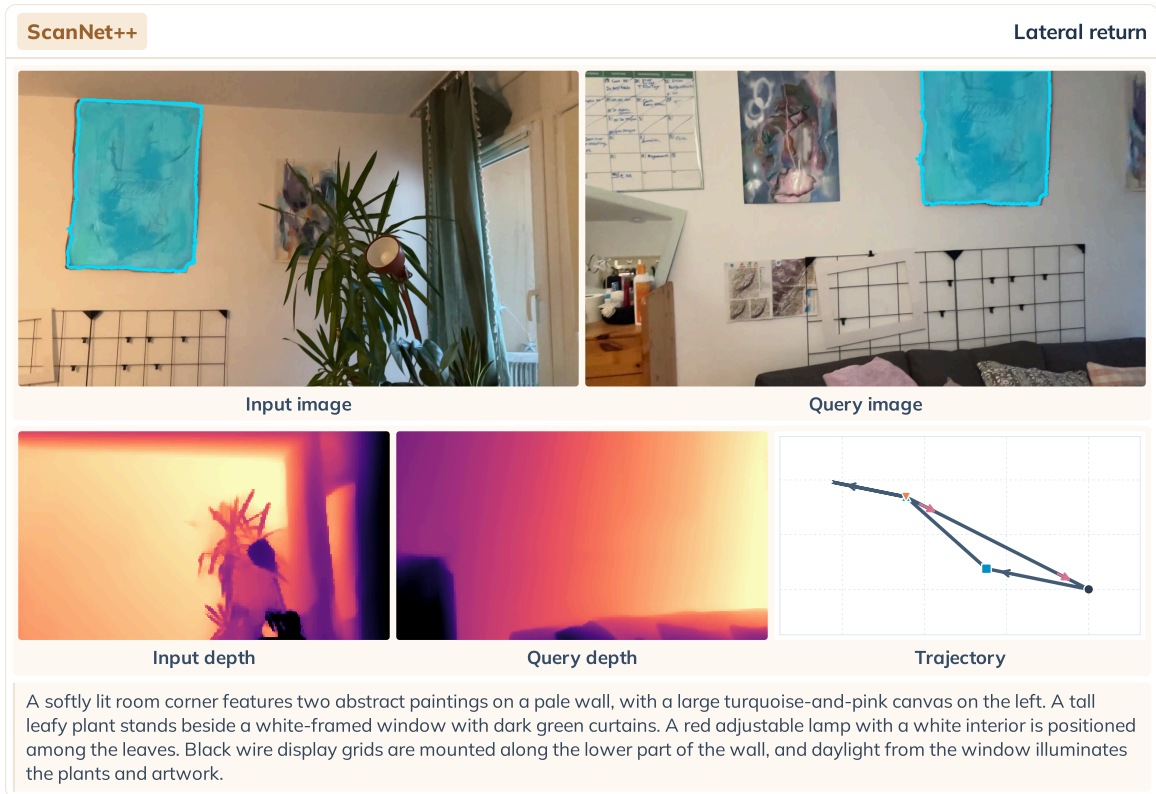
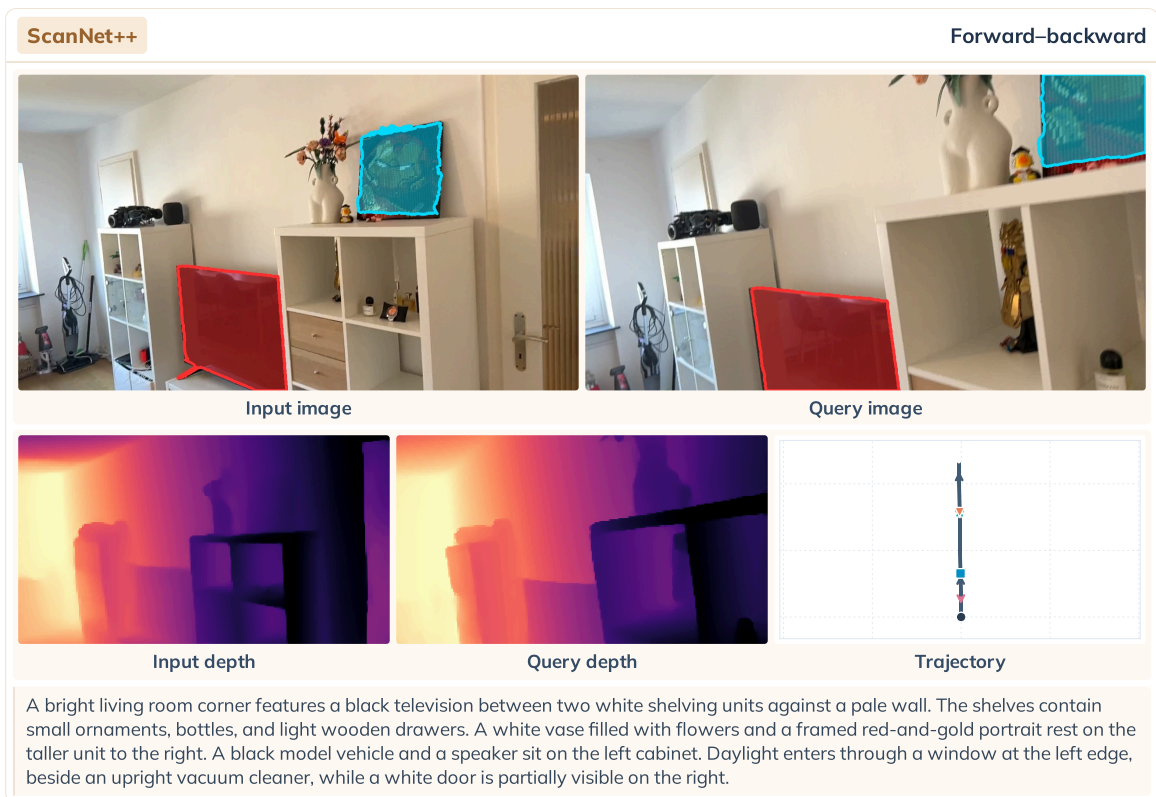


Figure 13. Examples of the forward-backward (top) and lateral return (bottom) templates from ScanNet++.

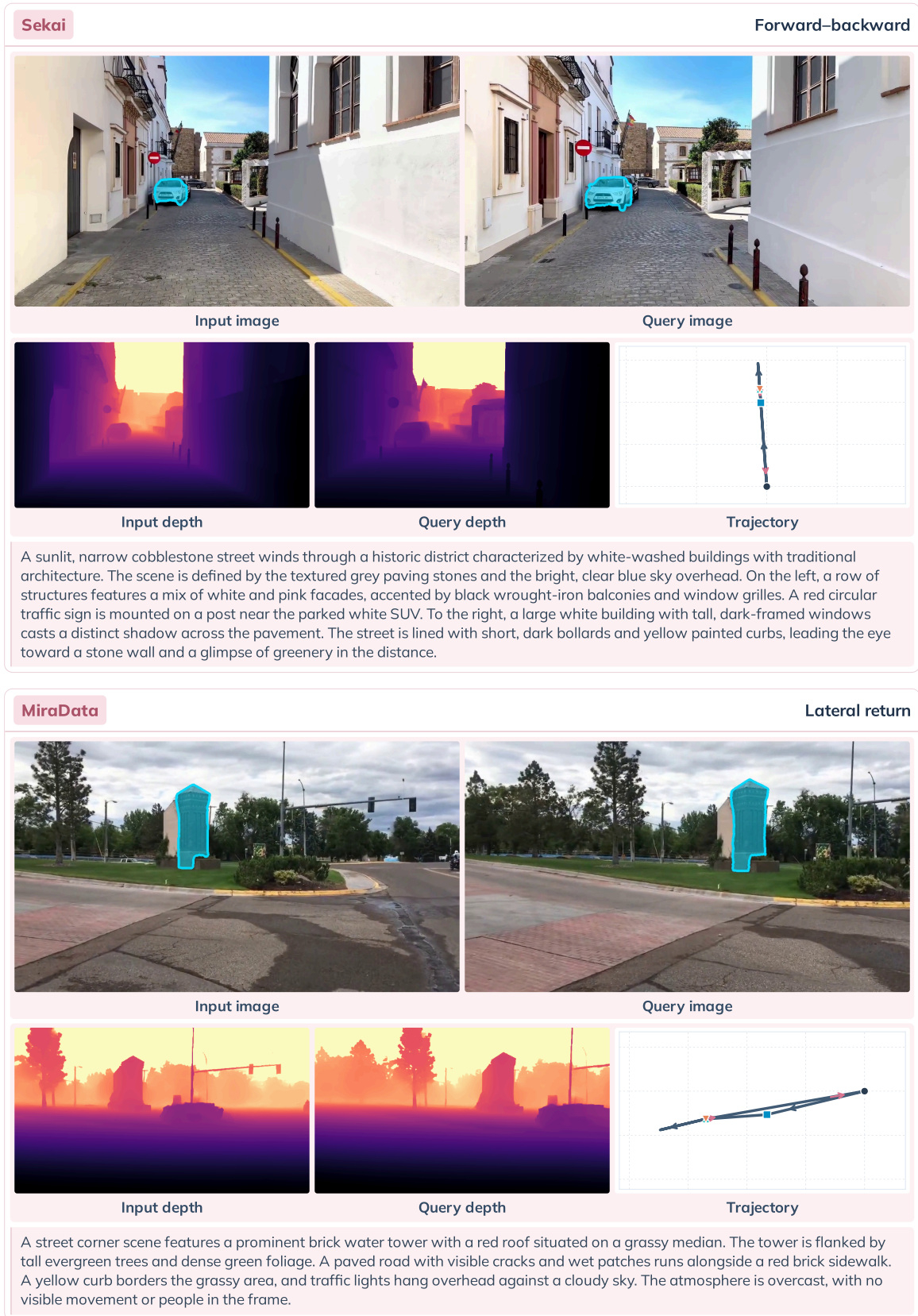


Figure 14. Examples of the forward-backward template from Sekai (top) and the lateral return template from MiraData (bottom).

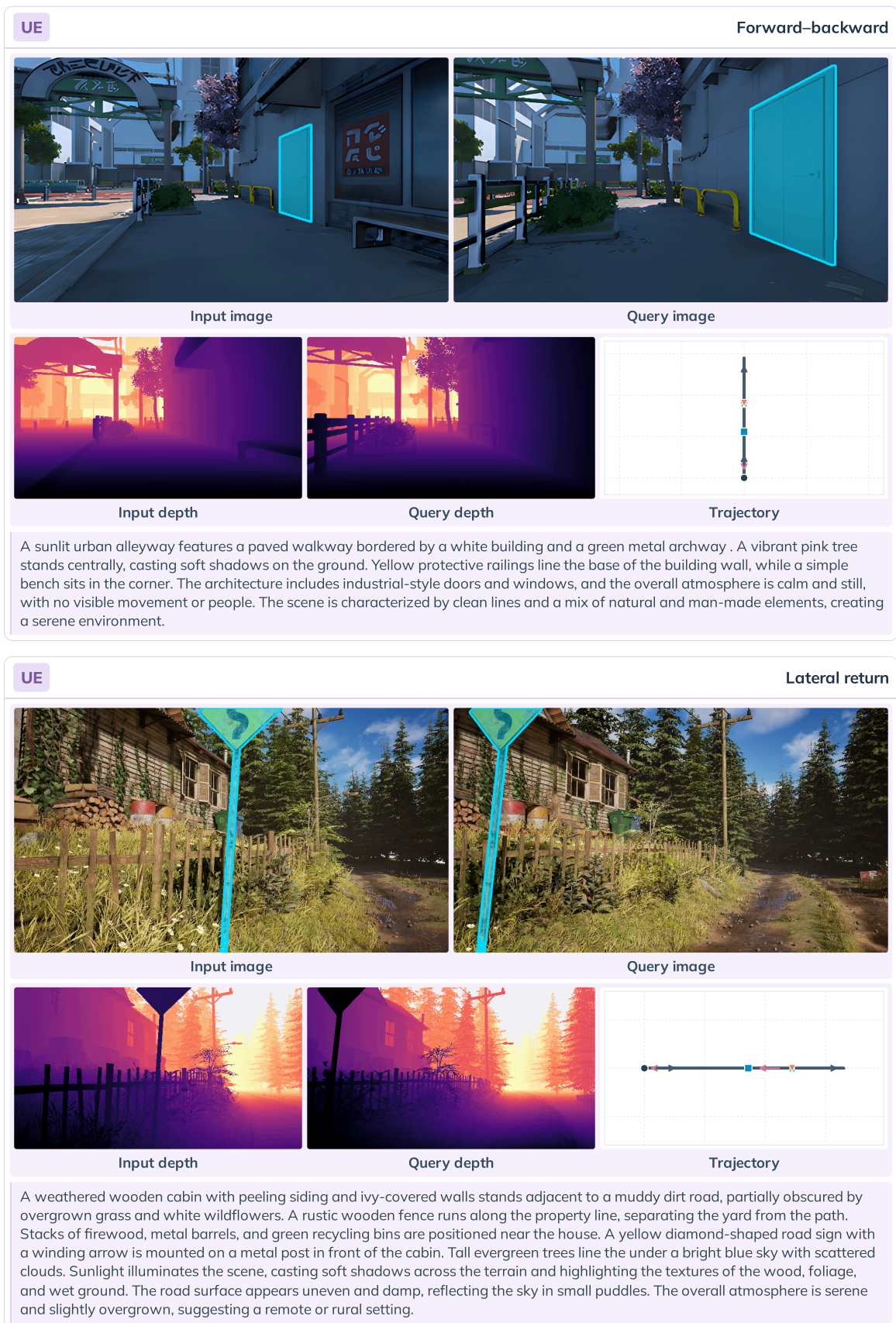


Figure 15. Examples of the forward-backward (top) and lateral return (bottom) templates from UE.

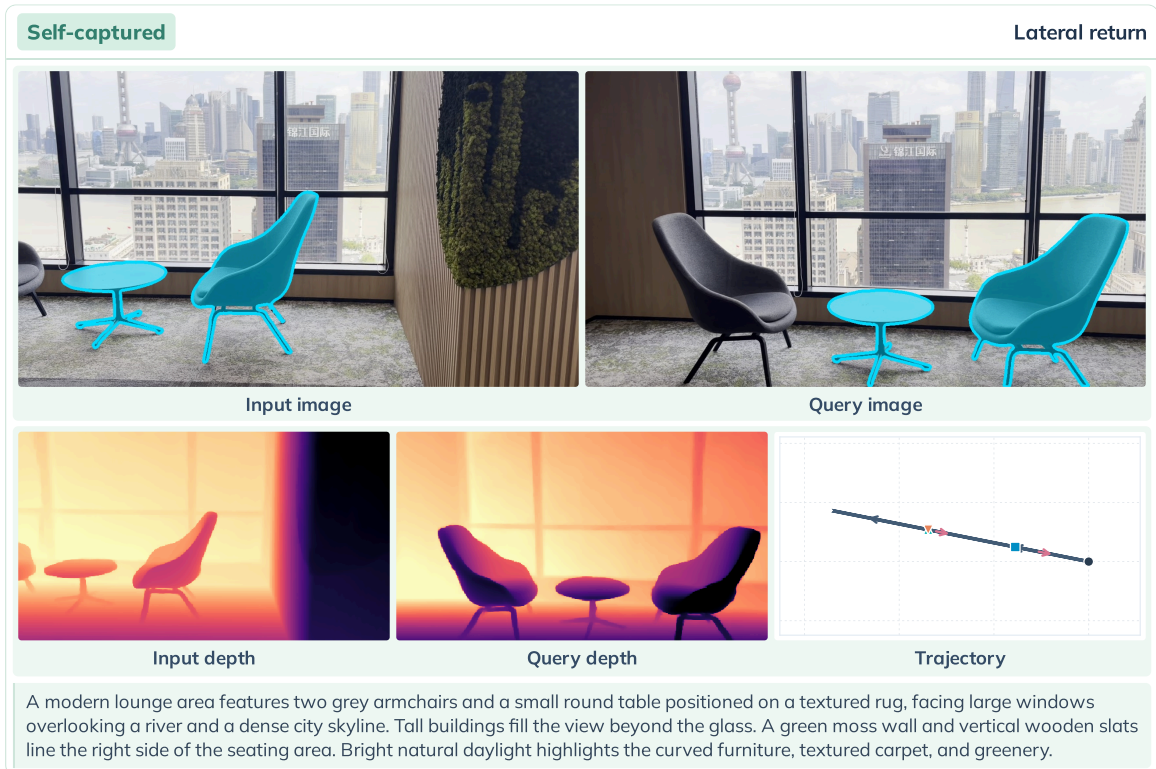


Figure 16. Examples of the forward-backward (top) and lateral return (bottom) templates from self-captured videos.

B Dataset Construction and Annotation

Constructing an **ORBIA** case requires a pair of registered reference views, anchors that can be verified across those views, and a camera trajectory that supports exploration and revisits. We describe how these components are assembled from different sources, extended to one minute, and reviewed before release. Table 5 reports dataset statistics, and Appendix A.3 presents annotated examples.

B.1 From Source Observations to Evaluation Cases

Following Section 3.2, our pipeline connects **observed scene references** with **designed camera motion** in four stages. Two registered RGB-D views provide the input and query references at P_0 and P_q . A synthesized trajectory passes through these poses and returns to P_0 at P_r , enabling controlled exploration without recording every intermediate view.

- **Select reference views.** Prepare depth and camera annotations, then identify input–query pairs compatible with the translation and rotation of a trajectory template. Candidate pairs should provide a meaningful viewpoint change while keeping a recognizable landmark visible in both views.
- **Verify anchors across views.** Obtain an object mask at each endpoint and check that both masks identify the same static object or structure. Project the input anchor into the query view using depth and cameras to identify the surfaces that can be compared across views.
- **Construct exploration and revisits.** Fit a smooth trajectory through the reference poses and check its motion against the available scene geometry. Select a first visit and a later revisit to a region outside the input view, recording the intervening exploration and the surfaces visible at both visits.
- **Rank, review, and annotate.** Rank candidates by image, anchor, trajectory, and revisit quality. Human reviewers inspect the proposed annotations and trajectories, correct unsuitable choices, and select the final cases. A VLM produces a factual scene caption for model conditioning.

B.2 Construction from Different Data Sources

The sources differ in which components are available initially. Public videos require view selection from recorded camera motion. ScanNet++ additionally provides a semantic mesh for anchor and visibility annotation. For UE and self-capture, we choose landmarks and trajectory templates during acquisition. The procedures below adapt the common pipeline to these differences.

B.2.1 Public Videos: View Selection and Anchor Verification

SpatialVID (Wang et al., 2026a), Sekai (Li et al., 2025d), and MiraData (Ju et al., 2024) broaden the range of real-world environments. Their recorded camera paths usually lack the revisits required by our protocols, so we extract suitable reference views and synthesize a new trajectory through them.

■ **Prepare geometry and rank view pairs.** We retain SpatialVID’s released depth and cameras and recover missing geometry for Sekai and MiraData with ViPe (Huang et al., 2025a). SpatialVID’s sparse geometry indices are mapped to the corresponding video frames, while the Sekai/MiraData adapter uses per-frame geometry. We filter registered pose pairs by each template’s translation and rotation requirements. Ranking favors motion compatible with the template and sufficient viewpoint change, while balancing template coverage and distributing candidates across different scenes and times within each video. Ranked alternatives allow another pair to be examined when the first fails annotation.

■ **Propose and segment landmarks.** A VLM (Qwen Team, 2026) examines five frames spanning the candidate pair, including both endpoints and three approximately evenly spaced intermediate frames. Figure 17 presents the anchor-selection instructions. It proposes static, bounded landmarks that are visually distinctive and recognizable at both endpoints. SAM3 (Carion et al., 2025) segments each proposal independently in the two reference images. We check mask area, connectivity, and completeness, remove duplicate proposals, and track shortlisted objects between views to verify that the endpoint masks identify the same physical landmark.

■ **Filter fragmented masks and identity mismatches.** The SpatialVID strict anchor profile accepts endpoint masks covering 0.3%–50% of the image and requires tracking agreement of at least 0.8 IoU at both endpoints. Mask integrity is assessed through connected components and internal gaps. The largest component must contain at least 85% of the

Input. Five ordered source frames spanning P_0 to P_q .

Anchor selection

Only for a usable scene, propose 3-8 visual anchors. A valid anchor must:

- refer to one static physical object or one coherent bounded structure;
- refer to the same physical instance at P_0 and P_q ;
- be clearly visible and independently identifiable at both endpoints;
- preferably be a substantial foreground or middle-distance target, rather than a small distant detail, while not occupying almost the entire image;
- look coherent and bounded enough to segment as one target;
- be distinctive enough not to be confused with nearby instances.

Reject a target that is dynamic, ambiguous, unbounded, severely occluded, too small or distant, easily interchangeable with repeated details, or not the same physical instance at both endpoints. Trees and other vegetation are not valid anchors even when one individual tree is prominent. Prefer larger, closer, more complete and more distinctive targets. Do not judge mask connectivity, holes, or tracking; downstream segmentation performs those checks.

Centered-anchor soft preference

After satisfying identity, staticness, two-endpoint visibility, completeness, and distinctiveness, rank otherwise valid anchors by visual usefulness. Prefer foreground or middle-distance targets with moderate-to-large visible area whose visible mask is near the central region of each image. P_0 is the model input, so give somewhat greater priority to a substantial, centered and complete P_0 view; P_q must still show the same instance clearly and completely, and should also be reasonably centered when possible. Order candidates from strongest to weakest under these preferences.

Centering is only a tie-breaking preference, never a reason to choose a person, moving object, ambiguous region, unbounded surface, severe occlusion, or a different physical instance. Do not reject an otherwise usable pair merely because no valid anchor is perfectly centered.

Segmentation prompts

For each anchor, provide 2-4 synonymous noun phrases for SAM3.1. Each phrase must contain 1-4 words, name the same physical target, and contain no spatial directions, coordinates, boxes, mask IDs, or multiple semantic objects. Prompt variants must remain at the same semantic target and scale: for example, a viewing platform cannot use ground or dirt path as a synonym.

Output contract

Return exactly one JSON object and no other text. It must contain:

- `scene_hazards`: an object containing exactly these boolean fields: `shot_boundary`, `endpoint_no_coview`, `low_light_or_blur`, `severe_weather_or_visibility_loss`, `salient_person_or_crowd`, `moving_traffic_dominant`, `persistent_text_overlay`, `open_or_generic_scene`, `no_distinctive_shared_anchor`, `severe_anchor_occlusion`;
- `pair_usable`: boolean, true exactly when every scene hazard is false;
- `pair_reason`: one short explanation of the scene-level decision;
- `candidates`: an empty list for a rejected pair, otherwise 3-8 objects.

Each candidate object must contain only `label`, `anchor_kind`, and `sam_prompts`. `anchor_kind` must be "instance" or "compact_structure". `sam_prompts` must contain 2-4 synonymous 1-4 word noun phrases for the same target. Do not output endpoint descriptions, boxes, coordinates, mask IDs, markdown, or additional keys.

Figure 17. VLM prompt for initial anchor selection. The preceding task-purpose and initial scene-gate instructions are omitted here. The full prompt also rejects unsuitable scenes before proposing anchors.

mask, with at most one significant component and an internal-gap ratio no greater than 20%. A component is significant when it contains at least the larger of 16 pixels and 0.2% of the mask area. These checks remove fragmented regions and accidental merges of neighboring objects. Tracking then checks that independently segmented endpoint masks follow the same object through the video.

■ **Verify shared anchor surfaces.** We back-project pixels inside the input mask into 3D and project them into the query view. Visibility and depth agreement determine which parts of the query anchor correspond to surfaces observed in the input. This selects anchors that support a meaningful cross-view comparison even when the query reveals substantial new scene content. The resulting masks and projection overlays are retained for human inspection. For each projected

Algorithm 1 Shared-surface and revisit annotation

Input: Registered depth, intrinsics K_i , camera-to-world poses P_i , endpoint anchor masks A_0, A_q , scene geometry G , and a requested camera trajectory through the input, query, and revisit poses.

Output: Query anchor support M_q , a first-visit/revisit pair, shared unseen surface masks, and geometric diagnostics.

A. Input–query shared anchor surfaces

1. Back-project valid input-anchor pixels into world coordinates. For camera-axis depth, $X = P_0[D_0(u)K_0^{-1}\tilde{u}; 1]$; calibrated source adapters handle other depth representations.
2. Transform X to the query camera, $x_q = P_q^{-1}X$, and project to $v = \pi(K_q x_q)$. Discard samples behind the camera or outside the image. Resolve competing projected samples by nearest depth.
3. Retain samples inside A_q with valid query evidence and depth agreement within the source-specific tolerance. Their query pixels form M_q . Record support size and depth residuals for anchor verification.

B. Shared surfaces outside the input view

4. For a designated outward visit i , choose a later return-branch pose j with small translation and rotation error, respecting the template’s event order and temporal-separation requirements.
5. Raycast G at P_0, P_i, P_j . Quantize valid first-hit surface points on a common world lattice to obtain visible surface sets V_0, V_i, V_j for the input, outward-visit, and revisit camera poses, respectively.
6. Remove cells within exclusion radius r of input-visible cells and intersect the remaining sets:

$$U_i = V_i \setminus \mathcal{N}_r(V_0), \quad U_j = V_j \setminus \mathcal{N}_r(V_0), \quad S = U_i \cap U_j.$$

Here $\mathcal{N}_r$ uses distances between voxel cells, rather than only their centers.

7. Map S back to first-hit pixels in both event views. Record shared-pixel fractions among valid geometry hits, valid-support counts, and pose errors. Apply the source-specific validity checks and coverage grading before retaining the pair for review.

Geometry: G is the scene mesh or a depth-derived proxy. Visibility and unseen-surface labels refer to this available geometry at the input, outward-visit, and revisit camera poses used in the annotation.

sample, we check positive camera depth, image bounds, and agreement with the query depth. When several samples project to the same pixel, the nearest surface determines visibility. The retained pixels identify the part of an anchor that the input actually reveals. This distinction matters when a viewpoint change exposes another side of the same object: its identity can remain clear while only some of its surfaces have an input reference.

■ **Synthesize the trajectory and select revisits.** We fit the chosen template through P_0 and P_q , preserving both reference poses and the final return. Available depth supports checks of camera motion and proximity to observed surfaces. Revisit selection considers temporal separation, camera-pose proximity, and shared surfaces outside the input view. Curved trajectories permit different viewpoints of the same region, while peek trajectories use repeated visits near the outward waypoint. Reviewers inspect the paired views together with their geometric overlap before accepting them.

Algorithm 1 summarizes the shared-surface checks and event-pair annotation. Tolerances and acceptance criteria follow the geometric checks configured separately for each data source.

B.2.2 ScanNet++: Mesh-Guided Reference Construction

ScanNet++ (Yeshwanth et al., 2023) provides indoor RGB-D observations, calibrated cameras, and semantic instance meshes. These annotations allow anchors and cross-view visibility to be derived directly from the reconstructed scene. The visibility operations in Algorithm 1 use this mesh, with sensor-depth validity checks for registered reference pixels.

■ **Select registered RGB-D endpoints.** We use the iPhone RGB-D stream and calibrated poses to find view pairs compatible with each template. Selection checks image sharpness, valid depth, and camera displacement. The input view must contain a clearly identifiable static instance that remains visible from the query pose. RGB frames are center-cropped and resized together with their camera calibration. We evaluate candidate pairs using the selected instance’s visible surface area, allowing the rest of the query image to reveal new parts of the room. When several anchors are present, each retains its own masks and geometric checks so that a well-observed object can be selected independently of less suitable instances.

■ **Render instance masks and visible surfaces.** We render anchor masks from the semantic mesh, assigning each

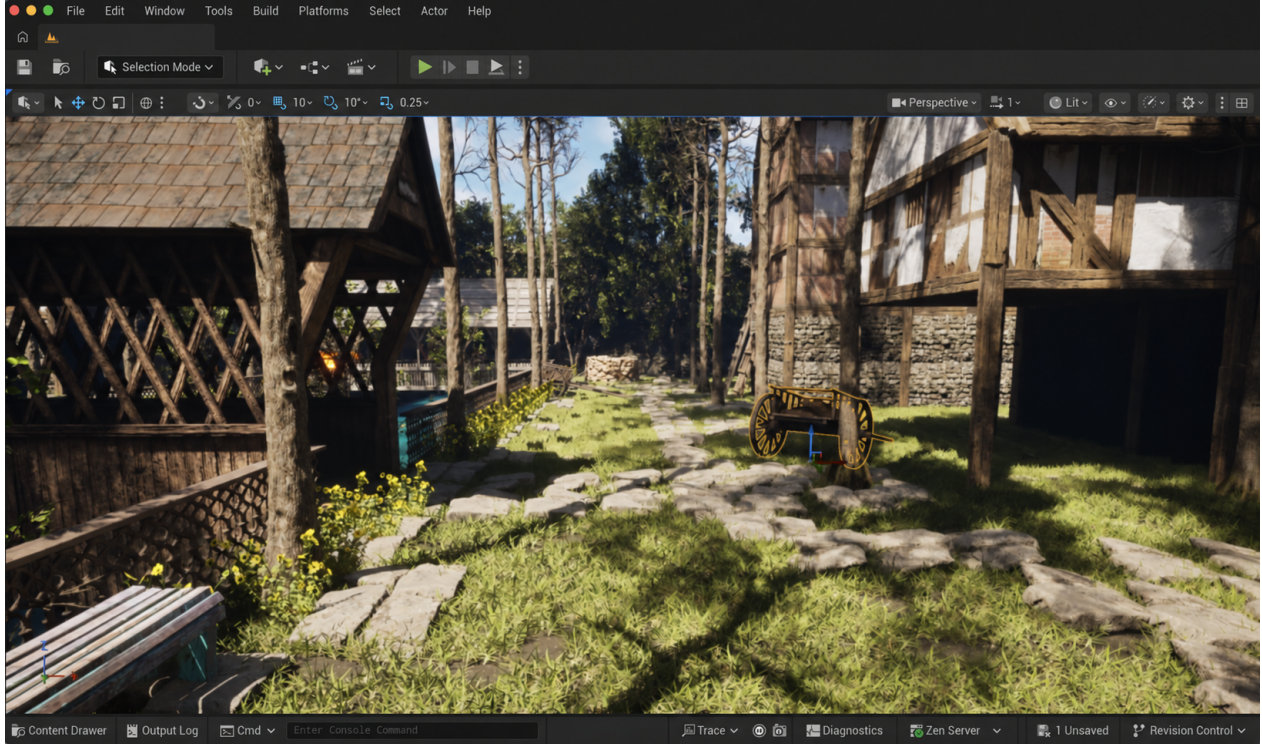


Figure 18. Anchor selection in the UE editor. A cart is selected as a bounded static landmark, with its scene Actor highlighted in yellow. Its location and bounds guide the initial camera framing and trajectory design.

pixel to the first visible surface along its camera ray. Anchors must have a recognizable silhouette, reliable depth, and enough common surface area for cross-view comparison. We exclude unbounded walls, floors, and ceilings, as well as transparent objects and masks that ambiguously combine several instances. Projecting the input-visible anchor surface into P_q identifies the part of the query mask used for input-preservation evaluation.

- **Plan and check the camera trajectory.** Translation and rotation are interpolated through the registered endpoints. The mesh supports checks of camera clearance, near-plane intersections, floor and ceiling distance, and abrupt motion. Validation examines the designated evaluation poses, 32 additional uniformly spaced poses, and the segments between sampled positions. Failed trajectories are replanned with different intermediate waypoints. Mesh visibility also identifies regions outside the initial view, from which we select temporally separated visits sharing visible surfaces.

- **Associate revisits with scene surfaces.** We record the mesh surfaces visible from P_0 and compare this set with the surfaces visible at candidate revisit poses. Surfaces shared by two visits but absent from the input provide locations for generated-content evaluation. Ray visibility handles occlusion as the camera moves through the room. Pixels with invalid depth or unreliable mesh boundaries are excluded from the comparison. This produces reference masks tied to particular scene surfaces and lets reviewers inspect where the camera leaves the input-visible region and later returns.

B.2.3 UE Simulation: Controlled Capture and Rendered References

UE provides synchronized images and geometric annotations along prescribed trajectories. We collect 48 virtual environments and render 365 candidate videos, retaining 180 cases after curation. Acquisition starts with a static scene actor and a trajectory template selected for the scene layout and its landmark.

Anchor selection and controlled rendering. We design each capture around a selected landmark and an exploration template. Scene selection, camera placement, and motion are reviewed together so that the resulting video provides both a recognizable anchor and meaningful changes in the surrounding view.

- **Choose the scene and anchor.** Annotators inspect a virtual environment and select a static object or structure with clear boundaries and a distinctive appearance. Figure 18 illustrates this selection in the UE editor. We favor

#10 Rejected by reviewer 59.3

Scene VLM PASS · The scene is a continuous outdoor urban street view with no editing cuts, no prominent moving people, and no dynamic content interfering with observation. The environment is well-lit, clear, and features distinct built structures.
No scene hazards detected

Image 4.3/5 · Anchor 34.7/35 · Trajectory 19.8/20 · Completion 0.5/10 · Total score 59.3/70

Q0/QE sharpness 923/996 · Completion weak

Manual review

Keep scene and all anchors Keep scene; choose anchors Reject scene Undecided Reject

Review individual anchors

A1 brick building Reject

Area / placement 13.0/13 · Integrity 11.9/12 · depth 9.8/10

Keep anchor Reject anchor Undecided

A2 metal sculpture Keep

Area / placement 6.5/13 · Integrity 0.0/12 · depth 3.0/10

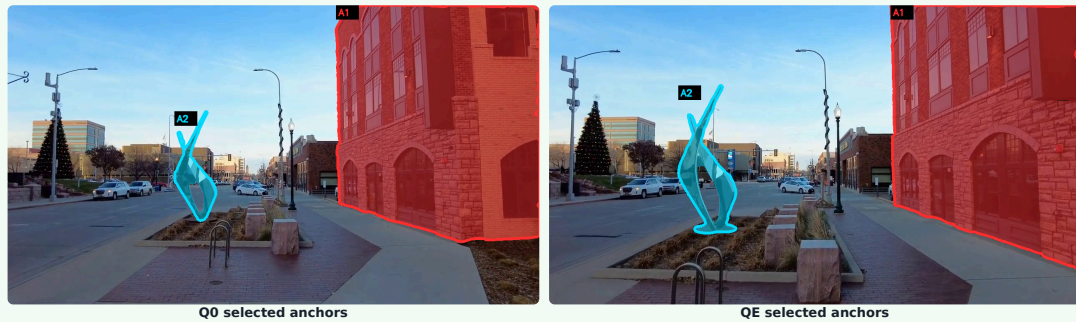
Keep anchor Reject anchor Undecided

Notes

e.g., fragmented mask, cropped object, keep A2, or other observations...

Candidate anchors

A1 brick building A2 metal sculpture



Show score breakdown and supporting evidence

Anchor scores

- A1 brick building:** Area / placement 13.0/13 · Integrity 11.9/12 · depth 9.8/10
- A2 metal sculpture:** Area / placement 6.5/13 · Integrity 0.0/12 · depth 3.0/10

Trajectory quality

Motion extent 14.9/15 · Upright / vertical behavior 4.9/5

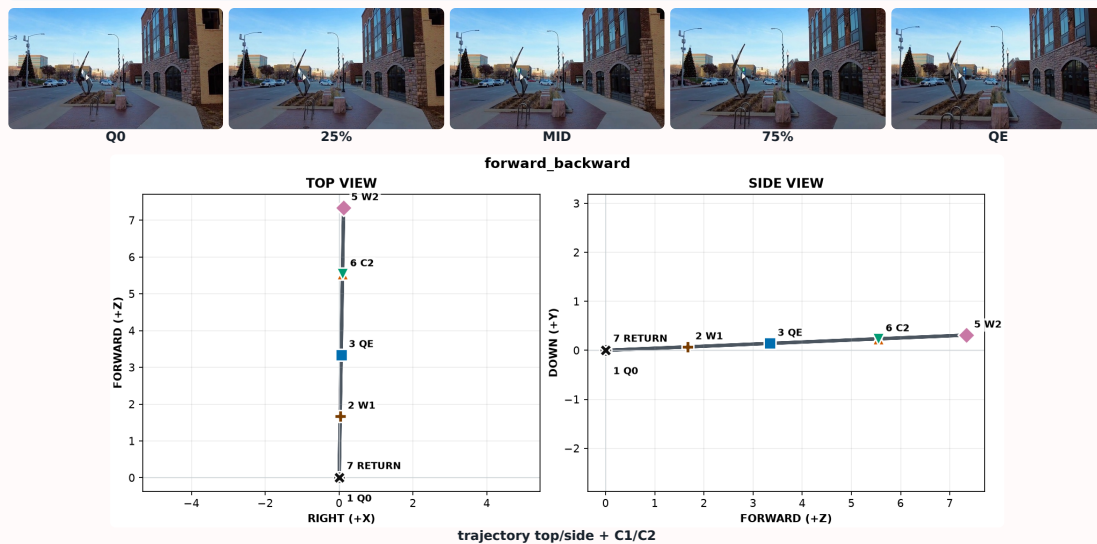


Figure 19. Manual review of a MiraData forward-backward case. The interface combines scene review, annotation scores, per-anchor decisions, reference overlays, sampled frames, and camera trajectories.

bounded landmarks that remain recognizable under viewpoint changes, while avoiding thin repetitive geometry, foliage, transparent surfaces, and large unbounded regions. The anchor is associated with its scene Actor, or a group of Actors forming one semantic object, so that its identity remains fixed throughout rendering. Its location and bounds guide camera placement. The initial view should show the anchor clearly together with enough surrounding scene content to support exploration around the selected landmark and into the surrounding scene.

- **Assign a trajectory and set its scale.** We choose a template suited to the scene layout while balancing coverage across the six motion types. The anchor’s size, camera field of view, and an authored starting view determine the initial framing and a suitable travel distance or rotation range. Forward and lateral paths require room to translate, yaw requires a useful angular sweep, and arc and orbit paths require clearance around the landmark. Peek paths place the anchor near the initial view and extend into a region revealed during exploration.

- **Preview and refine the motion.** Candidate paths are checked in UE for camera collisions and inspected through sparse renders of the input, intermediate exploration, query region, and return. Reviewers assess anchor scale, foreground occlusion, background content, and whether the motion visibly follows the intended template. Collision checks alone are insufficient because some visible geometry has no collision enabled. We therefore inspect the rendered views and adjust the starting position, travel direction, radius, or viewing direction when a trajectory is obstructed or poorly framed.

- **Render the accepted sequence.** The camera path is placed in a Level Sequence and rendered with Movie Render Queue. RGB, depth, and object IDs share the same camera sequence, and evaluated camera poses and intrinsics are recorded for geometric annotation. Object IDs provide the selected anchor’s visible mask at each frame; depth is converted to metric camera-axis Z . We allow assets and lighting to settle before capture, control exposure, and disable motion blur and depth of field to preserve clear reference images. The rendered video is then reviewed for smooth motion, loading artifacts, and anchor visibility before reference selection.

Post-processing captured videos. We process the captured RGB videos together with their rendered depth, object masks, and camera poses to construct benchmark cases. This stage selects the input and query references, identifies visits to newly revealed regions, and maps the accepted observations to the benchmark timeline.

- **Check reference-view quality.** For the query, candidates must contain a visible anchor covering at least 0.5% of the image and valid anchor depth. Projected input surfaces must explain at least 40% of the query anchor. Reviewers choose among valid candidates using the RGB images, masks, and projected-surface overlays. The default query search also requires the anchor’s image-area fraction to retain at least half its value in the input. This avoids choosing a distant or heavily occluded observation in which the landmark is too small to evaluate.

- **Select visits to newly revealed regions.** Rendered depth and cameras identify surfaces absent from the input and visible at both candidate visits. The general selector requires more than 8 seconds between visits, a camera-center distance within 1 meter, and an orientation difference within 15° . Arc trajectories allow a larger positional difference when both views observe the same region. Peek trajectories use repeated visits near the outward waypoint. Candidates are ranked by newly observed surface content, shared visible surfaces, and temporal separation.

- **Align the selected observations.** We map RGB, depth, masks, and camera poses to the 357-frame, 16-FPS short protocol using the same source-frame indices. The mapping preserves the input, return, and selected evaluation events.

B.2.4 Self-Capture: Landmark-Guided Acquisition

We record 85 outdoor videos around selected landmarks to broaden scene coverage beyond the public collections. The intended anchor and template are specified before recording, and 60 cases are retained after review.

- **Record the scene and recover geometry.** Acquisition follows the assigned exploration template around a distinctive static landmark. We estimate depth and cameras with DA3 (Lin et al., 2026), preserving the correspondence between source timestamps, RGB frames, and geometric annotations. The input is selected near the beginning of the recording, with adjustments for blur or exposure.

- **Annotate the designated landmark.** Query candidates are selected according to the known template. SAM3 segments the anchor independently at the input and query views, and reviewers verify identity and choose the reference pair. Checks consider mask area, completeness, valid depth, and the query surfaces recovered by projecting the input anchor. The nominal projected-surface threshold is 40% of the query mask. Here, the acquisition record already supplies the intended landmark. Review therefore concentrates on choosing sharp reference frames and confirming that segmentation

Input. Five ordered raw source frames, without annotation overlays or quality scores.

Review the five ordered video frames: P_0 , 25%, middle, 75%, and P_q .

First classify the scene as indoor or outdoor. Then decide whether the scene should be accepted or rejected.

Reject the scene if any of the following is true:

- The frames cross a shot or editing boundary, or do not show one continuous scene.
- One or more people are central, prominent, being followed, presenting to the camera, or visibly moving through the main part of the scene across multiple frames. A single such person is enough to reject the scene, even when the surrounding room or building occupies more pixels.
- Moving people, vehicles, or other dynamic content strongly interfere with observation of the static scene.
- The image is materially dark, blurred, foggy, obscured, or difficult to inspect.
- Heavy rain, snow, lens contamination, glare, or other weather significantly reduces visibility.
- Any subtitle, channel name, logo, or watermark is overlaid on the video. Carefully inspect all four corners in every frame. Fixed screen-position text is an overlay even when it is small, faint, or semi-transparent. Physical text on signs, products, or buildings is not an overlay.
- The scene is mainly open or natural scenery such as road, ground, snow, water, mountain, forest, or generic vegetation, without clear distinctive built objects.

Do not reject only because:

- one or two people remain small and peripheral in every frame;
- a vehicle is clearly parked;
- physical text appears on an object or building;
- vegetation appears around an otherwise distinctive built scene.

If any rejection condition is present, set `scene_status` to "reject". Otherwise set it to "pass".

Return exactly one JSON object and no markdown:

```
{
  "environment": "indoor",
  "environment_confidence": "high",
  "scene_hazards": {
    "shot_boundary_or_no_coview": false,
    "central_or_prominent_people": false,
    "dynamic_content_dominant": false,
    "low_light_or_visibility_loss": false,
    "severe_weather_or_lens_contamination": false,
    "subtitle_or_watermark": false,
    "open_or_nature_dominated": false
  },
  "scene_status": "pass",
  "reason": "brief evidence-based reason",
}
```

Figure 20. VLM prompt for dedicated scene filtering. The illustrative description value in the JSON schema is shortened for display. The decision instructions are reproduced in full.

follows that landmark in both views. The reviewer checks both endpoint images against the source recording to confirm that the masks follow the same physical landmark in both reference views.

■ **Regularize motion and select revisits.** We synthesize a smooth trajectory through the selected reference poses, preserving the final return while removing irregularities in handheld motion. Estimated scene geometry supports selection of visits to newly revealed regions. Reviewers inspect the trajectory and paired views, retaining the selected anchors, reference frames, and geometric diagnostics.

Input. Five ordered source frames: P_0 , 25%, middle, 75%, and P_q .

Write one dense English scene caption from the five ordered frames.

Describe only persistent visible architecture, terrain, vegetation, water, furniture, materials, lighting, weather, atmosphere, and stable spatial relations supported across the frames. Do not mention people, animals, human actions, moving vehicles, transient subjects, or any other temporary event. A stationary parked vehicle may be mentioned only when it is clearly persistent. Never mention the camera, viewpoint, framing, video, trajectory, temporal change, future frames, generation, rendering, games, animation, media, or these instructions. Do not use image-plane directions such as left, right, centre, foreground, or background. Do not speculate beyond visible evidence. Begin directly with scene language such as "A ... scene where ..."; do not write "This image/video...". Use 80-150 English words.

Return exactly one JSON object and no markdown: `{"dense_caption":"..."}`

Figure 21. VLM prompt for final scene captions. This separate captioning pass describes the persistent architecture, objects, and surroundings used to condition model generation on the initial scene.

B.3 Extending Exploration to One Minute

Long cases retain the short case’s input image, anchors, and registered reference views, providing a paired setting for examining degradation during extended exploration. Each extension contains 961 camera poses over 60 seconds at 16 FPS and three pairs of generated-region revisits.

- **Preserve the registered references.** The standard extension copies the short trajectory from P_0 through P_q , including its pose sequence. This keeps the initial exploration and cross-view checkpoint fixed. Some peek cases place P_q at the intermediate retraction between two excursions. Their extension lengthens the first excursion and adjusts the query time while preserving the registered P_0 and P_q poses.
- **Extend motion within the same template.** We increase travel distance or angular coverage while preserving the template’s characteristic motion. Timing follows the short trajectory’s median nonzero translation or angular speed. Forward and lateral paths explore farther before returning; yaw expands its sweep to 180° – 300° ; and arc paths retain their curved shape. Orbit preserves its anchor-centered motion, while peek preserves the outward movement into an unseen region and the subsequent return to the starting view.
- **Register three separated revisits.** We select three locations along the extended trajectory and revisit their camera poses after intervening exploration, with at least 5 seconds between paired observations and a final return to P_0 . Checks verify pose agreement within each pair, and geometric overlays show the shared surfaces and their visibility from the input and query views.
- **Validate and revise the extension.** We check reference-pose preservation, the final return, revisit alignment, and motion continuity. Endpoint depth supports surface-clearance checks along the trajectory, using a radius of 1.5% of the median endpoint depth and a sampling step no greater than half this radius. Failed trajectories are revised by adjusting their direction, heading, or template parameters. Reviewers inspect the resulting path and revisit regions before accepting the extended case into the final evaluation set.

B.4 Quality Ranking and Human Review

Automatic scoring prioritizes candidates for inspection. Human review then resolves choices that require scene understanding, such as whether a mask isolates a meaningful landmark or two visits show the intended region.

- **Rank annotation quality.** The large candidate pool makes exhaustive manual inspection costly: each case requires checking reference images, anchor masks, depth, and camera motion. We therefore use automatic quality scores to prioritize promising candidates, followed by human inspection and final selection. The score considers four aspects: ¹*image sharpness*; ²*anchor quality*, including area, placement, mask integrity, and depth consistency; ³*trajectory quality*, including motion extent and uprightness or vertical stability; and ⁴*revisit quality*, including shared unseen surfaces and pose agreement. These criteria favor clear reference views, well-defined landmarks, and informative exploration with verifiable revisits. Table 8 gives the component weights in the 70-point scheme.

Table 8. Annotation-quality ranking for public-video candidates. The SpatialVID scoring scheme assigns up to 70 points.

Category	Component	Points	Factors considered
Image	Endpoint sharpness	5	Sharpness of input and query RGB.
Anchor	Area and placement	13	Visible mask area at both endpoints, balance between their areas, and image-space placement of the two endpoint masks.
	Integrity and shape	12	Largest connected component, fragmentation, internal gaps, substantial border contact, aspect ratio, and solidity.
Trajectory	Depth consistency	10	Fraction of projected samples with consistent depth, and fraction falling inside the opposite-view mask.
	Motion extent	15	Maximum rotation from P_0 for yaw, or maximum horizontal excursion for the other translation-based trajectory templates.
	Upright and vertical behavior	5	Camera tilt, vertical-to-horizontal excursion ratio, and the vertical slope of the camera path between consecutive poses.
Revisit	Shared unseen surfaces	10	Shared surface fraction outside the input view at both visits, support validity, and pose mismatch.
Total		70	

For cases with multiple reliable anchors, the highest-scoring anchor determines the anchor term. Scores guide review order, while reviewers make the final acceptance and anchor-selection decisions.

- **Review scene suitability.** We apply a dedicated VLM review after anchor, segmentation, and geometric checks have reduced the candidate pool. Early anchor search processes thousands of candidate frames, so postponing detailed scene review avoids spending additional inference on candidates that fail those checks. The anchor-selection prompt already includes an initial scene gate; this later pass rechecks scene suitability independently of anchor proposals and numerical quality scores (Figure 20). Separating this pass allows detailed review to concentrate on candidates that survive annotation and geometry checks. The scene reviewer receives five ordered raw frames and checks for shot changes, prominent moving subjects, poor visibility, severe weather or lens contamination, and intrusive subtitles or watermarks. It also records indoor or outdoor scene type. Anchor overlays, trajectory plots, and numerical scores are excluded from this review to keep the judgment focused on scene content. The pass/reject decision is recorded separately from the numerical ranking. Passing cases are ordered by their automatic scores within the template and indoor/outdoor groups for manual inspection.
- **Inspect and correct annotations.** Figure 19 shows the manual review interface for a MiraData case; the Sekai orbit example is shown in Figure 43 in Appendix E.2. The review interface presents input and query images, anchor overlays, projected input surfaces, alternative query frames, trajectory plots, and revisit pairs. Reviewers check *anchor identity and mask completeness*, *template fit and camera motion*, and *the region shared across revisits*. They can replace an anchor, change the query selection, or reject a case. Accepted manual choices are preserved during export. Top and side trajectory views help identify paths that rise through a ceiling, approach an observed wall, or make an abrupt turn. Reference overlays show missing object parts, neighboring objects included in a mask, and surfaces that become occluded at the query. The review record stores the chosen frame, retained anchor IDs, acceptance decisions, and comments or failure tags. These choices determine the final reference frames, anchor masks, and annotations exported for evaluation.
- **Curate diversity and add captions.** We remove rejected candidates, group duplicates by source provenance, and balance trajectory templates while avoiding repeated use of the same scene content. A VLM annotates factual scene captions from sampled source frames for model conditioning, using the dedicated instructions in Figure 21. The final record stores the selected references, anchors, review decisions, and source bindings. Annotation and manual curation required over 300 hours across the collection.

B.5 Released Case Package

Each case separates model-facing inputs from evaluator-only references, with shared metadata linking both sets to the scene and specifying the camera poses and observation times used for evaluation.

Model inputs. The public package contains an initial RGB image, a scene caption, and a requested camera trajectory with intrinsics. Images are standardized to 1280×720 , with matching geometric annotations, and camera poses are expressed relative to the initial view. Short and long cases contain 357 and 961 poses, respectively, at 16 FPS.

Evaluation references. A separate reference package contains query RGB, registered depth, anchor masks and identities, and camera calibration. These annotations support comparisons at the query and revisit events and are withheld from generation. Query images are also excluded from generated-camera recovery and reconstruction context.

Event and source metadata. Each case records the frame indices and timestamps of the query, return, and first-visit/revisit pairs, together with their source-frame bindings. Long cases retain a link to the paired short case, preserving their common input and registered references.

C Evaluation Protocol and Metrics

This appendix details the metrics used in each of the six evaluation tiers, including their definitions, implementation, and tier-level aggregation. We also explain score normalization and aggregation into tier scores and the overall score.

C.1 Tier 1: Control Alignment

Control alignment measures the agreement between requested and recovered camera motion. Let $P_t = [R_t \mid \mathbf{p}_t]$ and $\hat{P}_t = [\hat{R}_t \mid \hat{\mathbf{p}}_t]$ be their camera-to-world poses at frame t . A hat denotes a quantity recovered from generation; E denotes an error and S a similarity or normalized score. We align the initial poses and fit a single nonnegative scale to all camera positions recovered from the entire generated video.

1. Translation and Rotation Error

We measure the spatial and angular accuracy of the recovered camera trajectory.

■ **Translation Error.** We measure the mean distance between corresponding camera positions, normalized by the total length of the requested camera trajectory over the evaluated video:

$$E_{\text{trans}} = \frac{1}{NL} \sum_{t=1}^N \|s^* \hat{\mathbf{p}}_t - \mathbf{p}_t\|_2. \quad (1)$$

Here N is the number of evaluated poses, L is the total requested path length, and s^* minimizes the sum of squared position errors over the sequence. This alignment resolves the recovered trajectory’s global scale, while the remaining error measures its spatial deviation. Pure-yaw trajectories are excluded from translation evaluation.

■ **Rotation Error.** We average the angular distance between camera orientations:

$$E_{\text{rot}} = \frac{1}{N} \sum_{t=1}^N d_R(\hat{R}_t, R_t). \quad (2)$$

The geodesic distance d_R is the angle of the relative rotation $R_t^\top \hat{R}_t$, reported in degrees. Both pose errors use observations at matching timestamps, so they measure the execution of the requested motion throughout the video.

2. Discrete Action Accuracy

We extract local motion directions from the requested and recovered cameras and compare their direction sets at request-active intervals. Both sequences use identical extraction settings.

■ **Move Accuracy.** We measure camera-local displacement over two frames. An interval is active when its speed exceeds 1% of the sequence’s mean speed. Directions are selected by unit-displacement components with magnitude above 0.5, using their signs to distinguish opposite directions. Multiple directions may be active together.

■ **Look Accuracy.** We obtain the local rotation vector over four frames. Pitch and yaw components with magnitude above 0.4° determine the active look directions, with their signs distinguishing opposite rotations. These intervals correspond to the benchmark’s 16 fps timeline.

For action type $a \in \{\text{Move}, \text{Look}\}$, let $\mathcal{U}_a$ be the request-active intervals and $B_t^a, \hat{B}_t^a$ the requested and recovered direction sets. Accuracy requires an exact match of the full set:

$$S_a = \frac{100}{|\mathcal{U}_a|} \sum_{t \in \mathcal{U}_a} \mathbf{1}[\hat{B}_t^a = B_t^a]. \quad (3)$$

We compute accuracy within each video and then average valid video scores. An action without requested activity is omitted for that case.

Tier score. We convert translation and rotation errors to percentage scores and combine the applicable pose and action scores, assigning equal weight to each component evaluated for that video.

C.2 Tier 2: Visual Quality

Visual quality measures frame appearance and temporal smoothness. We adopt five components from VBench and WBench (Huang et al., 2024; Ying et al., 2026), each expressed on a 0–100 scale:

1. Aesthetic Quality

The LAION aesthetic predictor (Schuhmann et al., 2022) scores sampled frames on a 0–10 scale. We average the frame scores and multiply by ten.

2. Imaging Quality

MUSIQ (Ke et al., 2021) measures perceptual image quality. We average its native 0–100 scores over the sampled frames in each video, giving every sampled frame the same weight.

3. Temporal Flickering

We measure the mean absolute RGB difference between adjacent sampled frames and invert it as $S_{\text{flicker}} = 100(1 - \text{MAE}/255)$. Frames are sampled every two original frames, corresponding to 8 fps for the 16 fps videos. Higher scores indicate smaller frame-to-frame appearance changes.

4. Motion Smoothness

AMT predicts an intermediate frame from its two neighbors. Agreement with the observed middle frame measures local temporal continuity; we multiply the normalized consistency score by 100. Triplets use consecutive frames $(t, t + 1, t + 2)$, with starting indices spaced eight frames apart.

5. Human Preference Score

HPSv3 scores 20 uniformly sampled frames for short videos and 54 for minute-long videos, using an empty text prompt. Following WBench, we map the mean raw reward $\bar{r}$ as

$$S_{\text{HPS}} = 100 \text{ clip}\left(\frac{\bar{r} + 5.21}{13.82}, 0, 1\right). \quad (4)$$

The fixed endpoints -5.21 and 8.61 are shared across models.

Tier score. The five components receive equal weight within each video, and videos are averaged equally regardless of frame count. Table 10 reports the component scores.

C.3 Tier 3: Input Preservation

We measure scene and anchor fidelity at Return and Query, and anchor identity during exploration.

1. Scene Appearance Fidelity

We compare the generated frame J with its reference I using PSNR, SSIM, LPIPS, and DINO. The reference is the input image at Return and the registered query image at Query. We score valid pixels in region Ω , restricting Query to surfaces visible in the input.

Regional DINO uses the comparison region’s bounding crop. At Query, DINO instead matches patch features within a one-patch neighborhood and averages the best cosine similarities.

2. Scene Geometry Fidelity

Let D and $\bar{D}$ denote reference and generated depth, and let $\Omega_D \subseteq \Omega$ contain pixels with valid depth in both. We align generated depth to the reference with a median scale $a^* = \text{median}_{u \in \Omega_D} D(u)/\bar{D}(u)$, then compute the following scale-aligned AbsRel over all scene pixels with valid reference and generated depth:

$$E_{\text{scene}} = \frac{1}{|\Omega_D|} \sum_{u \in \Omega_D} \frac{|a^* \bar{D}(u) - D(u)|}{D(u) + \epsilon}. \quad (5)$$

Here ϵ is a small positive constant. Scaling corrects global depth ambiguity but leaves local geometric errors intact, so errors on individual objects and other visible surfaces remain measurable.

3. Anchor Appearance Fidelity

Let A_k be the reference mask for anchor k . We evaluate it within $\Omega_k = \Omega \cap A_k$, using the same four appearance measurements as for the scene. For N_A anchors and each measurement m , the case score is the equal-weight mean

$$m_{\text{anchor}} = \frac{1}{N_A} \sum_{k=1}^{N_A} m(I, J; \Omega_k). \quad (6)$$

We average each measurement equally across anchors.

4. Anchor Geometry Fidelity

We measure depth fidelity within each anchor using the AbsRel definition in Eq. 5. For anchor k , the comparison region is $V_k = \Omega_k \cap \Omega_D$. We average errors over valid pixels within each anchor, then weight anchors equally. All anchors reuse the scene-level scale a^* , preserving relative scale differences between objects and their surroundings.

5. Anchor Identity Consistency

We track each input anchor with SAM3.1 (Carion et al., 2025) and compare its DINO descriptor with the input descriptor at one-quarter, one-half, three-quarters, and the end of the input-to-query interval. Descriptors are formed by mask-area-weighted pooling of DINO patches followed by unit normalization. With input descriptor $f_{k,0}$ and tracked descriptor $\hat{f}_{k,t}$, identity consistency is

$$S_{\text{id}} = \frac{1}{N_A} \sum_{k=1}^{N_A} \frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} f_{k,0}^\top \hat{f}_{k,t}. \quad (7)$$

Here $\mathcal{T}_k$ contains valid scoring times for anchor k . Tracking samples every fourth frame and includes all four scoring times; samples with insufficient mask or feature support are omitted.

Tier score. Each checkpoint averages appearance and depth equally, with equal scene and anchor weights within each modality. Appearance averages normalized PSNR, SSIM, LPIPS, and DINO. After converting identity similarity to a percentage, we combine the three components:

$$S_{\text{T3}} = 0.30S_{\text{return}} + 0.40S_{\text{query}} + 0.30S_{\text{identity}}. \quad (8)$$

C.4 Tier 4: Generated-Content Persistence

We compare generated frames at the registered first-visit and revisit times, using one pair for short cases and three for long cases. The selected pairs revisit surfaces outside the input view.

1. Revisit Appearance Consistency

We compare DINO features of geometrically corresponding patches using Eq. 12. The mean cosine similarity measures appearance retention on co-visible surfaces.

For identical requested poses, we also compare the images directly using PSNR, LPIPS, and DINO similarity between patches at corresponding image locations. For different requested poses, the direct comparison uses global DINO descriptors pooled over the complete first-visit and revisit images.

2. Revisit Depth Consistency

We project first-visit depth into the revisit camera and compute AbsRel on co-visible surfaces using Eq. 11. For identical requested poses, we additionally compare depth directly at corresponding pixels. Both comparisons use the same recovered global scale for the first-visit and revisit observations.

3. Revisit Surface Agreement

For same- and different-pose revisit pairs, we back-project both depth maps into point clouds in the recovered coordinates. Precision p_τ is the fraction of first-visit points with a revisit neighbor within distance τ ; recall r_τ measures the reverse. Surface agreement is their harmonic mean:

$$F_\tau = \frac{2p_\tau r_\tau}{p_\tau + r_\tau}. \quad (9)$$

We use $\tau = \max(0.02, 0.05d_{\text{med}})$, where d_{med} is the median depth across both clouds. Points are sampled at stride eight and voxel-downsampled at $\tau/2$, retaining the sequence alignment. Nonempty clouds with no matches receive zero.

Tier score. Registered pairs are averaged equally within each case. For same-pose pairs, the direct score weights RGB fidelity, patch DINO similarity, and direct depth consistency equally, with PSNR and LPIPS sharing the RGB contribution. Different-pose pairs use global DINO. The two branches are pooled by their valid case counts to obtain the direct-comparison score S_{direct} across the registered revisit pairs.

The geometric score averages normalized reprojected DINO, depth consistency, and surface F-score equally. The tier combines direct and geometric scores as

$$S_{\text{T4}} = \frac{1}{2}(S_{\text{direct}} + S_{\text{geometry}}). \quad (10)$$

C.5 Tier 5: 3D Self-Consistency

This tier tests whether generated views agree with a common 3D scene. Short- and medium-range frame pairs are fixed on the video timeline. Recovered depth and cameras establish correspondences between their shared visible surfaces.

1. Cross-View Depth Consistency

We measure relative depth disagreement on surfaces visible in both views:

$$E_D(s \rightarrow t) = \frac{1}{|\Omega_{s \rightarrow t}|} \sum_{u \in \Omega_{s \rightarrow t}} \frac{|\hat{D}_{s \rightarrow t}(u) - \hat{D}_t(v(u))|}{\hat{D}_t(v(u)) + \epsilon}. \quad (11)$$

Here $\hat{D}_t$ is the estimated depth at frame t , $\hat{D}_{s \rightarrow t}(u)$ is the source point’s depth in the target camera, and $v(u)$ is its projected pixel. The set $\Omega_{s \rightarrow t}$ contains valid correspondences after image-bound and occlusion checks; ϵ is a small positive constant. Each frame pair receives the mean of the two directional errors. Lower error indicates stronger agreement about the depth of shared surfaces.

We average the two directional errors for each frame pair, then average pairs within the short- and medium-range groups. These groups use offsets of 8 and 32 frames, with at most 48 and 16 pairs, respectively. Pixels are sampled at stride four,

with a relative target-occlusion tolerance of 0.10. Short-range pairs measure local depth stability; medium-range pairs cover larger viewpoint changes.

2. Cross-View Feature Consistency

We compare DINOv2 (Oquab et al., 2024) descriptors at geometrically matched locations:

$$S_\phi(s \rightarrow t) = \frac{1}{|\mathcal{C}_{s \rightarrow t}|} \sum_{(a,b) \in \mathcal{C}_{s \rightarrow t}} \phi(\hat{I}_s, a)^\top \phi(\hat{I}_t, b), \quad (12)$$

Here $\hat{I}_t$ is the generated RGB frame, $\phi(I, a)$ is the unit-normalized DINO descriptor for patch a in image I , and $\mathcal{C}_{s \rightarrow t}$ contains unique patch pairs induced by geometric correspondences.

We evaluate at most 16 pairs in each temporal group and average both projection directions. Each matched patch pair contributes once, preventing dense pixel correspondences from overweighting the same feature pair.

3. Regional Scale Agreement

This metric measures whether different image regions imply a consistent relative depth scale. We divide the target image into a 4×4 grid and compute the median ratio $\rho_j = \text{median}(\hat{D}_t / \hat{D}_{s \rightarrow t})$ in each sufficiently supported cell j . The score is

$$E_{\text{scale}} = \text{median}_{j \in \mathcal{J}} |\log \rho_j - \text{median}_{k \in \mathcal{J}} \log \rho_k|. \quad (13)$$

Here $\mathcal{J}$ contains sufficiently supported cells; lower values indicate more consistent regional scale.

Subtracting the median log-ratio removes a common multiplicative offset between the two depth maps. The remaining spread measures disagreement among regions. For example, if foreground furniture and a background wall require different relative scales, their cell ratios separate and the score increases. The median statistics reduce the influence of a few extreme depth samples; cells without enough correspondences are excluded.

4. Reconstruction

This metric evaluates whether a Gaussian scene reconstructed from generated views reproduces other views of the same video. DA3 predicts the scene from 64 context frames, which we render at 16 disjoint held-out cameras. We compare each rendered image with its corresponding generated frame using PSNR, SSIM, and LPIPS.

Context and held-out frames are uniformly sampled from the estimated sequence. DA3 (Lin et al., 2026; Kerbl et al., 2023) predicts Gaussians in a feedforward pass using context RGB, recovered poses, and intrinsics. Each fidelity measure is averaged over the 16 held-out views on pixels with finite positive rendered depth.

Tier score. Depth, feature, regional-scale, and reconstruction scores receive equal weight:

$$S_{T5} = \frac{1}{4}(S_{\text{depth}} + S_{\text{feature}} + S_{\text{scale}} + S_{\text{GS}}). \quad (14)$$

Depth and feature scores each average their normalized short- and medium-range components equally. The reconstruction score averages normalized PSNR, SSIM, and LPIPS equally.

C.6 Tier 6: Temporal Stability

We evaluate both short and minute-long videos in five-second windows, measuring both the average performance and its retention relative to the initial window. Each window uses the score mappings and full-sequence camera alignment defined for the preceding tiers. We evaluate three dimensions:

- **Quality.** HPS, MUSIQ, and aesthetic quality.
- **Control.** translation, rotation, Move Accuracy, and Look Accuracy over applicable components.
- **Geometry.** cross-view depth, feature, and regional-scale consistency. Depth and feature each combine their two temporal groups with equal weight for the short- and medium-range comparisons.

We retain the final complete five-second window. Actions are scored on request-active intervals within each window. Each indicator requires a valid initial window and at least one later window.

1. Performance Level and Retention

Let $s_{m,b}$ be the percentage score of indicator m in window b , with $b = 0$ denoting the initial window and $\mathcal{B}_m$ the valid later windows. The average level includes all valid windows:

$$A_m = \frac{s_{m,0} + \sum_{b \in \mathcal{B}_m} s_{m,b}}{1 + |\mathcal{B}_m|}. \quad (15)$$

Retention measures the mean positive decline from the initial score:

$$R_m = 100 - \frac{1}{|\mathcal{B}_m|} \sum_{b \in \mathcal{B}_m} \max(0, s_{m,0} - s_{m,b}). \quad (16)$$

Later improvements contribute zero decline. Combining level and retention captures both the average scores achieved during exploration and the amount of decline relative to the initial five-second window.

Tier score. Within each video, we average indicators equally within each dimension to obtain A_d and R_d . The dimension score is

$$S_d^{\text{long}} = 0.30A_d + 0.70R_d. \quad (17)$$

Tier 6 weights quality, control, and geometry equally within each video. We average videos separately by horizon, then combine the short- and long-video means with weights of 30% and 70%.

C.7 Evaluation Pipeline and Score Aggregation

The evaluator recovers geometry and constructs Gaussian and feature representations for metric comparison. Measurements are normalized and combined into tier scores and the overall score.

■ **Recover cameras and depth.** We decode each generated video on its requested timeline, sample frames at stride two, and include the registered query, return, and revisit frames. DA3 Giant 1.1 (Lin et al., 2026) estimates poses, intrinsics, depth, and confidence at processing resolution 504. Sequences with at most 240 sampled frames are processed jointly. Longer sequences use 120-frame streaming chunks with 30-frame overlap and SALAD loop closure. Chunk alignment places the estimates in a common coordinate system. Each recovered observation is stored with its video-frame index for subsequent event and cross-view comparisons.

■ **Predict Gaussians and render held-out views.** From the sampled sequence, we select 64 context frames and 16 disjoint held-out frames. The context RGB images and recovered cameras are passed to DA3’s feedforward Gaussian prediction. Images are resized to 504×504 with matching intrinsics, and held-out cameras are expressed in the reconstruction’s coordinates and scale. We render RGB and depth at the 16 held-out cameras with gsplat, using a black background. These renders and their corresponding generated frames supply the reconstruction measurements in Tier 5.

■ **Extract appearance and identity features.** We extract DINOv2 (Oquab et al., 2024) global and patch descriptors from the selected images. Camera and depth correspondences identify patch pairs for cross-view and revisit comparisons. Registered reference masks define scene and anchor regions for input preservation; SAM3.1 (Carion et al., 2025) tracks input anchors to the four identity-scoring times. PSNR, SSIM, and LPIPS compare the image pairs specified by each metric. Visual-quality predictors operate on their own frame and triplet schedules, as described in Tier 2.

■ **Compute event and temporal measurements.** Recovered cameras are aligned to the requested trajectory for pose and action scoring. Depth, DINO correspondences, and point clouds provide the geometric measurements; query and return observations use registered references, while revisit measurements compare the selected generated observations. The same video estimates support five-second window measurements for Tier 6. We first reduce frame, anchor, and frame-pair observations to case-level components, then apply the score mappings and aggregation below.

Normalization and aggregation. We convert metric components to a common 0–100 scale, with higher scores indicating better performance. Unit-range similarities are scaled directly; errors and PSNR use fixed linear mappings calibrated with equal weight per model. For Tiers 3–5, we normalize case-level measurements, average each component across valid cases, and apply the tier weights defined above. Tier 6 completes its temporal aggregation within each video before averaging across videos. The final six tier scores are combined as

$$S_{\text{overall}} = 0.25S_{T1} + 0.25S_{T2} + 0.10S_{T3} + 0.10S_{T4} + 0.10S_{T5} + 0.20S_{T6}. \quad (18)$$

D Model Adapters and Generation Settings

We group the evaluated models by the control interface used in our experiments: **camera-conditioned**, **action-conditioned**, and **text-conditioned**. For each model, we describe below its generation setting. The benchmark provides an initial image I_0 , a scene caption, and a camera trajectory $\{P_t\}$. Adapters convert the trajectory to the model’s coordinate convention and temporal sampling. Image resizing or cropping is accompanied by the corresponding intrinsic transformation when the model accepts calibration. For chunked generators, we extend the terminal camera pose to the required length and trim the extra output frames. Models with fixed generation windows continue from their own generated images or retained context. Evaluated videos use the benchmark’s 16-FPS timeline; when a model uses a different native frame rate, its controls or text timings follow that native clock before output conversion.

D.1 Camera-Conditioned Models

These models receive numerical camera conditions derived from $\{P_t\}$, expressed relative to the initial camera. We use the camera interface directly when an interactive model also offers an action-to-camera front end. This applies to Echo-WM and Matrix game 3.5; HY-WorldPlay also receives camera poses and derives its auxiliary action labels internally. Cosmos3-Nano’s continuous camera increments belong to this group, although the native implementation labels these continuous numerical camera conditions as actions.

 **SolarWM-H3** (Huang et al., 2026) generates videos autoregressively with camera conditioning. We use 4 denoising steps, 5 latent frames per chunk, and a 6-chunk attention window. The requested trajectory enters its camera-conditioning path with the native logd4 translation encoding and fixed camera intrinsics. The input image and caption initialize generation, and the history cache is carried across rollout chunks.

 **Lyra2** (Shen et al., 2026) combines video generation with geometric conditioning. We use the DMD-distilled checkpoint with 4 steps at 832×480 . Requested poses are rebased to P_0 , converted to world-to-camera matrices, and passed through the custom-trajectory interface with pose scale 1. Intrinsics follow the image resize, while the native geometry pipeline aligns estimated depth to MoGe scale. Short runs generate 401 frames and retain the first 357.

 **EVOKE** (Yin et al., 2026) uses camera and geometric conditions for autoregressive scene exploration. We use the Stage3 post-distillation checkpoint in its image-conditioned mode, with 3 steps, guidance scale 1, and 640×384 output. The initial image, caption, and requested cameras enter the native camera and geometry branches. Short sequences use 10 rollout chunks; longer sequences extend the same rollout with the corresponding camera windows.

 **Echo-WM** (Zhang et al., 2026) is a causal video generator with unified camera conditioning. We use the Flash checkpoint with 4 steps, 3 latent frames per chunk, and 1280×704 output. The adapter bypasses its Action DSL and supplies initial-camera-relative c2w poses directly to UCPE, applying the native translation factor $1/30$ and 8-frame camera sampling. Short and long runs share the same causal history mechanism, with a 19-frame latent cache and 7 sink frames carried from one rollout chunk to the next during generation.

 **Gen3C** (Ren et al., 2025) uses a persistent 3D cache to condition subsequent generated views. We use 35 denoising steps at 1280×704 . The input image initializes the scene and the requested camera sequence drives the native autoregressive generation path. Inference proceeds in 121-frame windows with 1 shared boundary frame, retaining the 3D cache across windows.

 **Matrix game 3.5** (Qian et al., 2026) combines autoregressive generation with geometric patch memory. We use the distilled first-person model with 3 denoising steps and 1280×704 output. The camera adapter supplies c2w extrinsics and image-aligned pixel intrinsics directly. The rollout consumes consecutive camera windows and retains its native memory of generated views.

 **AlayaWorld** (AlayaWorld Team et al., 2026) uses ViGeo to condition generation on scene geometry and camera motion. We use v1.1 Stage3, combining the Stage2b base with the Stage3 LoRA, at 960×544 with 4 steps and guidance scale 1. The adapter supplies poses relative to P_0 with scale 1 and intrinsics adjusted to the resized input. Short videos use 12 rollout rounds; the same interface processes the extended trajectories.

 **LingBot-World v2** (Gao et al., 2026) generates camera-conditioned video with causal attention and cached history. We use the causal-fast checkpoint with 4 denoising steps, 4 latent frames per chunk, and the 480p configuration. Requested poses and intrinsics enter its native camera preprocessing, including its relative-motion normalization. For short cases, terminal padding produces 365 frames, of which the first 357 are retained.

 **HY-WorldPlay** (Sun et al., 2025) supports autoregressive generation from camera and action conditions. We use its AR-distilled checkpoint with 4 steps, chunk size 4, and 848×480 output. The adapter supplies sampled c2w poses from the requested trajectory; the native pipeline converts them to its camera representation and derives the associated action labels. Thus, our input uses the numerical camera interface. Short generation produces 365 frames before trimming to 357; native 24-FPS output is played at 16 FPS.

 **SANA-WM** (Zhu et al., 2026) provides streaming video generation with camera conditioning. We use the streaming checkpoint, 4 generator steps followed by 3 refiner steps, guidance scale 1, and flow shift 8. The requested poses are rebased to P_0 and converted into the native ray and Plücker conditions. The 1280×720 input is cropped to 1280×704 to satisfy spatial divisibility, with intrinsics updated accordingly. Streaming generation retains the model’s causal decoding and refinement stages.

 **Voyager** (Huang et al., 2025b) generates new views conditioned on an image and camera motion. We use 50 steps, and 768×512 output. The requested trajectory is divided into 49-frame generation windows with a 1-frame overlap. Each window receives the corresponding camera segment and starts from the preceding generated last frame, extending the model to the benchmark’s short and long durations.

 **LingBot-World v1** (Robbyant Team et al., 2026) is evaluated with its Fast camera-control checkpoint and 4-step causal sampling. The adapter provides the initial image, scene caption, requested poses, and intrinsics to the native camera pipeline. Short runs retain the full available history; long runs use a 21-frame local attention window and 1 sink frame. The evaluated batches use the recorded 720p and 480p configurations, with the same pose-conditioning interface.

 **Cosmos3-Nano** (NVIDIA, 2026) accepts continuous camera-motion conditions. We use 30 denoising steps. Consecutive requested poses are converted to local SE(3) increments, encoded as 3 translation components scaled by 10 and a 6-dimensional rotation representation. These 9-dimensional vectors are supplied directly to the camera-action input. Generation adds 60 frames per chunk, conditioned on the preceding generated last frame, at an inference frame rate of 16 FPS.

 **DreamX-World** (DreamX Team et al., 2026) is evaluated with the 5B autoregressive checkpoint and 4 denoising steps at 1280×704 . We use 3 latent frames per block, and the native 4-step schedule. Requested poses and normalized intrinsics pass through the official chunk-relative camera helper. Each chunk’s cameras are expressed relative to the preceding chunk endpoint, while calibration follows the benchmark input image.

 **FantasyWorld** (FantasyWorld Team, 2025) generates camera-conditioned video from an initial image. We use 50 denoising steps, and 592×336 output. The requested trajectory is partitioned into 81-frame windows with a 1-frame overlap. Each window receives camera conditions for its trajectory segment and starts from the preceding window’s final generated frame, continuing generation from the model’s own output.

 **Astra** (Zhu et al., 2025) uses relative camera motion for blockwise video generation. We use 50 steps, and 832×480 output. The adapter converts the requested trajectory to consecutive world-to-camera increments at the native 4-frame sampling interval; this interface has no intrinsic conditioning. Camera windows advance with the history origin, and long runs extend the temporal indexing while retaining the native recent-history context. Short runs retain 357 frames from a 385-frame generation.

D.2 Action-Conditioned Models

ABot-World, Genie3, and HappyOyster receive discrete movement and look commands. We compile these commands from the requested camera trajectory and execute them through the model’s streaming or browser interface.

Pose-to-action conversion. We smooth reconstructed camera paths while retaining the initial, query, return, and revisit waypoints. The adapter regularizes positions and continuous-sign quaternions with a nominal 0.2-Hz smoothing setting. For camera-to-world poses P_t , the local increment $\Delta P_t = P_t^{-1} P_{t+1}$ supplies translation directions and yaw/pitch increments. After removing near-stationary motion, we normalize translation directions and use signed horizontal and forward components above 0.5 to identify movement keys. A direction is selected when its net vote exceeds half the action window. Movement and turning can be active together; yaw-return cases disable translation. We use model-specific turning responses to convert requested yaw and pitch into discrete action blocks or timed key presses.

Control evaluation. For ABot-World, Genie3, and HappyOyster, converting camera trajectories into discrete actions introduces differences in motion scale, timing, and turning response. Their results therefore serve primarily as diagnostic evaluations under the adapted action interface.

Segment 1 (15.0 seconds). initial image I_0 .

<Scene caption>

Use one continuous shot with a fixed focal length.

The times below are local to this video segment.

Move at a relaxed, steady walking pace, easing smoothly into stops and reversals.

From 0.0 to 7.5 seconds, move steadily to the right.

From 7.5 to 15.0 seconds, reverse direction and move steadily to the left.

By 15.0 seconds in this segment, smoothly finish the specified camera movement and come to a stop.

Preserve the scene layout and objects throughout, including when revisiting. Do not cut or zoom.

Figure 22. Kling3.0 prompt for short-horizon lateral return. The scene caption is represented by a placeholder.

Segment 1 (15.0 seconds). initial image I_0 .

<Scene caption>

Use one continuous shot with a fixed focal length.

The times below are local to this video segment.

Move at a relaxed, steady walking pace, easing smoothly into stops and reversals.

From 0.0 to 15.0 seconds, move steadily to the right.

Preserve the scene layout and objects throughout, including when revisiting. Do not cut or zoom.

Segment 2 (15.0 seconds). preceding generated last frame.

<Scene caption>

Use one continuous shot with a fixed focal length.

Continue from the supplied image.

The times below are local to this video segment.

Move at a relaxed, steady walking pace, easing smoothly into stops and reversals.

From 0.0 to 5.0 seconds, move steadily to the right.

From 5.0 to 15.0 seconds, reverse direction and move steadily to the left.

Preserve the scene layout and objects throughout, including when revisiting. Do not cut or zoom.

Segment 3 (10.0 seconds). preceding generated last frame.

<Scene caption>

Use one continuous shot with a fixed focal length.

Continue from the supplied image.

The times below are local to this video segment.

Move at a relaxed, steady walking pace, easing smoothly into stops and reversals.

From 0.0 to 10.0 seconds, move steadily to the left.

By 10.0 seconds in this segment, smoothly finish the specified camera movement and come to a stop.

Preserve the scene layout and objects throughout, including when revisiting. Do not cut or zoom.

Figure 23. Kling3.0 prompt for long-horizon lateral return. The same scene caption precedes each segment. Continuation requests use the preceding generated last frame.

These interfaces cannot accept camera poses or directly specify the requested displacement and rotation at each frame. Comparing DA3-recovered poses with the reference trajectory would therefore mix directional control with differences in action quantization and motion scale. For these 3 models, we use only **Move Accuracy** and **Look Accuracy** in the reported control score. DA3 recovers motion from the generated video, and we compare its movement and look direction sets with those derived from the original requested trajectory, following Appendix C.1. Accuracy is computed over request-active intervals within each video and then averaged across videos. Tier 1 averages the 2 action scores; the control dimension of Tier 6 uses the same components within each temporal window.

Table 9. Tier 1: Detailed control alignment results. Scores use a 0–100 scale (↑). Models are ordered by Tier Score within each group. Apricot, gold, and gray mark the three highest scores per column and group.

Model	3D Repr.	Translation Score	Rotation Score	Move Accuracy	Look Accuracy	Tier Score
World Models						
Lyra2	✓	79.89	75.61	83.46	87.59	81.64
SolarWM-H3	×	71.12	84.43	74.90	89.27	79.93
Gen3C	✓	72.56	74.52	70.11	84.86	75.51
Echo-WM	×	60.71	74.74	64.29	83.21	70.74
EVOKE	✓	63.92	61.26	65.96	74.75	66.47
Matrix game 3.5	✓	46.41	77.14	51.94	81.22	64.18
AlayaWorld	✓	61.74	56.93	59.59	68.45	61.68
Voyager	✓	49.68	61.42	61.35	71.22	60.92
HappyOyster	–	(42.64)	(38.00)	61.11	59.86	60.49
SANA-WM	×	51.39	57.35	56.16	72.65	59.39
LingBot-World v2	×	50.94	43.22	79.45	60.30	58.48
Genie3	–	(52.47)	(35.42)	68.97	41.55	55.26
LingBot-World v1	×	49.03	32.31	70.08	45.08	49.12
DreamX-World-5B	×	45.98	46.95	57.06	45.96	48.99
ABot-World	×	(43.76)	(46.14)	62.13	31.99	47.06
HY-WorldPlay	×	49.67	48.21	59.43	23.72	45.26
FantasyWorld	✓	30.05	46.90	36.51	43.17	39.16
Cosmos3-Nano	×	34.79	38.92	47.42	21.65	35.70
Astra	×	18.68	33.16	18.82	43.12	28.45
TI2V Models						
Seedance2.5	–	44.08	38.00	58.74	32.11	43.23
Kling3.0	–	37.62	35.92	53.22	27.57	38.58
HunyuanVideo1.5	×	32.56	39.11	59.40	13.29	36.09
MiniMax-H3	–	28.15	24.64	49.15	27.96	32.48
Wan2.7	–	26.89	30.00	37.01	31.36	31.31
Cosmos3-Super	–	23.37	41.87	28.81	17.61	27.92
LTX2.5	–	21.97	27.16	27.63	29.27	26.51
LongCat	×	22.42	18.95	27.76	25.91	23.76

ABot-World (Jiang et al., 2026) uses the 0-5B-LF streaming checkpoint with 4 steps and 1280×704 output. Its input consists of I_0 , the scene caption, and a sequence of 8-dimensional WASD/IJKL action vectors. Before the benchmark run, we generated 64 controlled clips covering static, translation, and yaw actions, followed by 8 pitch clips. DA3 recovered cameras from these generated clips to estimate turning responses. We adopted symmetric nominal increments of 20° for yaw and 17° for pitch per action block. A turn is issued only if the accumulated requested angle and the nominal emitted angle differ by at least half a block’s increment, and the resulting turn points in the same direction as the requested camera rotation for that action block.

The native decoder produces 9 frames in the initial block and 12 thereafter. We therefore use 30 blocks for a 357-frame short video and 81 blocks for a 969-frame long rollout, retaining the first 961 frames. Each case initializes a fresh stream from I_0 ; actions are updated before generating each block, while the model retains its history within the case. The native 12-FPS frame sequence is published at 16 FPS.

Genie3 (Parker-Holder and Fruchter, 2025) is run through the Project Genie webpage in first-person mode. We upload I_0 and the scene description, create the world, and begin the action schedule once the interactive view is ready. The trajectory is summarized in 12-frame analysis windows. Movement directions become WASD states, while yaw and pitch become timed look commands. A separate turning pilot checked response directions and approximate angular motion; the final scheduler uses nominal yaw and pitch rates of $70^\circ/\text{s}$ and $40^\circ/\text{s}$. These rates determine hold durations, with translation and look commands dispatched independently.

Browser-side scheduling uses an absolute clock to issue key-down and key-up events while the page remains in the foreground. Short and long cases use 20- and 54-second action schedules, followed by a 2-second key-release interval.

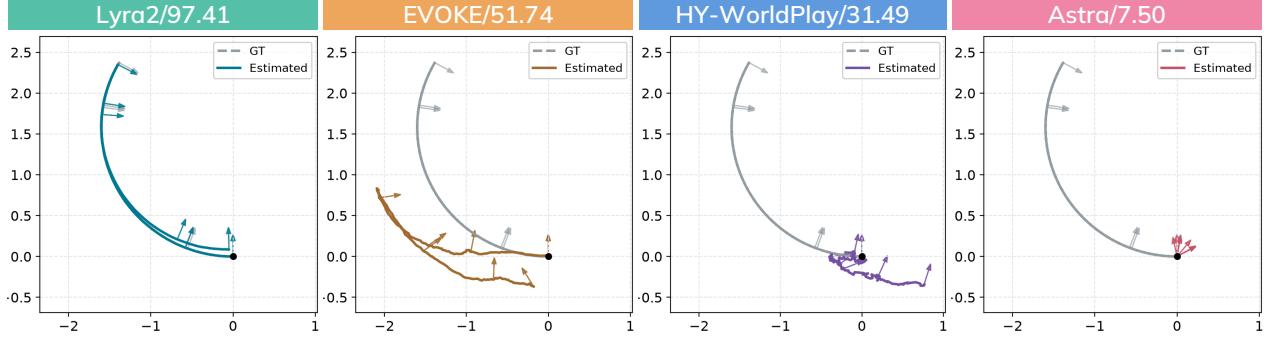


Figure 24. Control alignment on a UE orbit example. Requested and recovered camera trajectories are shown on shared axes, with arrows indicating viewing direction. Astra fails to recover the requested trajectory. Displayed scores are for this case.

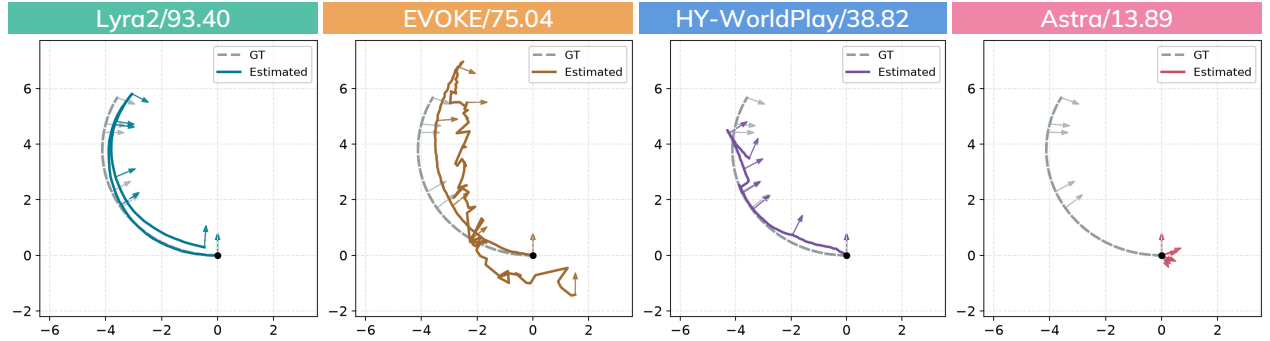


Figure 25. Control alignment on a UE orbit example. Requested and recovered camera trajectories are shown on shared axes, with arrows indicating viewing direction. Astra fails to recover the requested trajectory. Displayed scores are for this case.

The runner exits the world and saves the webpage’s downloadable video, together with the planned and dispatched actions. For evaluation, we normally remove the final 2 seconds and uniformly select 357 or 961 frames, then encode them at 16 FPS.

HappyOyster is run through its browser SDK in first-person Adventure mode at 720p. We initialize the world with I_0 and the scene description, wait for the video stream, and execute a local action plan. A turning pilot was analyzed with DA3 to estimate visible angular response. We use nominal yaw and pitch rates of approximately $18.23^\circ/\text{s}$ and $20.97^\circ/\text{s}$ to allocate turn durations. Local motion is grouped into 12-frame analysis windows and mapped to a 15-second short or 40-second long interaction schedule.

The SDK maintains separate held states for translation and rotation. Our scheduler updates a state only when it changes, so releasing a turn does not interrupt ongoing translation. We record the incoming video track with MediaRecorder, retain the action and session timestamps, and release all controls before closing the session. Evaluated outputs retain the first 357 or 961 captured frames and play them at 16 FPS.

D.3 T12V Models

T12V models receive I_0 and a prompt combining the scene caption with a motion script derived from the requested trajectory. We first construct the script, then divide it according to each model’s generation windows. This keeps the motion instructions consistent with the frames generated by each request. Prompt processing and enhancement follow the model-specific inference configurations.

From poses to motion descriptions. We analyze translation and rotation separately, smooth their trends over approximately 0.6 seconds, and group them into motion phases: travel, turning, pauses, reversal, and return. The script states the direction and duration of each phase, with turn angles derived from the corresponding poses. For curved trajectories, the script describes both the path and the viewing direction; orbit prompts name the public anchor when

Table 10. Tier 2: Detailed visual quality results. Scores use a 0–100 scale (↑). Models are ordered by Tier Score within each group. Apricot, gold, and gray mark the three highest scores per column and group.

Model	3D Repr.	Aesthetic Quality	Imaging Quality	Temporal Flickering	Motion Smoothness	HPS	Tier Score
World Models							
EVOKE	✓	53.23	72.29	94.48	98.69	63.69	76.48
LingBot-World v2	×	53.01	73.54	91.95	97.05	61.14	75.34
Echo-WM	×	53.36	74.72	93.73	97.14	56.17	75.02
DreamX-World-5B	×	52.48	71.27	93.09	98.03	59.94	74.96
SolarWM-H3	×	51.18	74.21	93.90	97.86	56.62	74.76
Genie3	–	49.71	71.58	92.67	97.27	60.95	74.43
Cosmos3-Nano	×	52.18	70.36	97.16	99.18	52.37	74.25
Matrix game 3.5	✓	51.80	74.29	92.29	98.25	53.50	74.03
LingBot-World v1	×	52.24	69.32	93.55	97.15	56.95	73.84
SANA-WM	×	52.21	69.37	94.37	98.92	51.24	73.22
HY-WorldPlay	×	49.17	70.69	95.08	98.90	46.46	72.06
Lyra2	✓	47.52	71.92	92.74	98.20	46.45	71.37
AlayaWorld	✓	49.27	70.82	92.72	98.11	44.30	71.04
FantasyWorld	✓	50.85	64.34	95.75	98.94	43.64	70.70
HappyOyster	–	46.80	64.92	94.31	98.47	39.38	68.78
Voyager	✓	50.65	62.12	92.36	97.56	37.55	68.05
Gen3C	✓	47.65	63.27	94.02	98.82	33.98	67.55
ABot-World	×	48.79	60.24	95.43	97.71	32.66	66.97
Astra	×	41.78	51.86	95.87	98.45	18.52	61.30
T12V Models							
Cosmos3-Super	–	55.69	74.55	96.26	99.16	69.51	79.04
MiniMax-H3	–	54.16	76.43	93.55	98.70	67.59	78.09
LTX2.5	–	53.94	74.33	95.22	99.35	66.95	77.96
Kling3.0	–	53.47	74.34	94.93	98.83	65.35	77.38
LongCat	×	54.38	76.91	92.10	98.36	64.73	77.30
HunyuanVideo1.5	×	52.26	74.05	95.33	99.40	64.32	77.07
Seedance2.5	–	53.52	73.13	95.65	99.27	63.26	76.97
Wan2.7	–	51.93	75.64	93.11	97.47	64.14	76.46

available. Brief fluctuations are merged into longer phases, while meaningful pauses and direction changes are retained. This simplification changes the text description, leaving the requested trajectory used for evaluation unchanged.

Window-specific prompts. Each generation window receives only the motion phases that overlap its frame range. We recompute its local times and turning angles at the model’s native frame rate, then append the resulting motion text to the scene caption. A continuation prompt starts from its supplied generated image or video context and describes the next part of the trajectory. Return instructions appear when the corresponding return phase falls within that window. Figures 22 and 23 show the short- and long-horizon scripts for the same lateral-return case in Kling3.0.

Segmented generation. The first request is conditioned on I_0 . Subsequent requests use the preceding generated last frame, except for LongCat, which retains 13 context frames. We remove repeated conditioning frames at segment boundaries and trim terminal padding to obtain 357 or 961 frames at 16 FPS. API models are run as dependent requests: the current segment must finish before its generated endpoint can initialize the next. We retain each segment’s prompt and output alongside the assembled video.

Seedance2.5 (ByteDance Seed, 2026) is evaluated through its first-frame image-to-video API at 720p with audio disabled. Short cases use a 15-second request; long cases use 2 consecutive 20-second requests at the native 24-FPS rate. The long prompts cover requested frames 0–480 and 480–960, respectively, each expressed as a local 0–20-second script. The second request uses the first segment’s generated last frame. Short output is trimmed to 357 frames; long segments are joined after removing the repeated boundary frame to obtain 961 frames.

Kling3.0 (Kling AI, 2026) uses the image-to-video API at 720p and native 24 FPS, with audio and multi-shot generation disabled. Short cases request 15 seconds. Long cases request 15, 15, and 10 seconds, with a separate prompt



Figure 26. Visual quality on a MiraData orbit example. Rows show three matched timestamps, with models arranged in columns. Displayed scores are for this case.

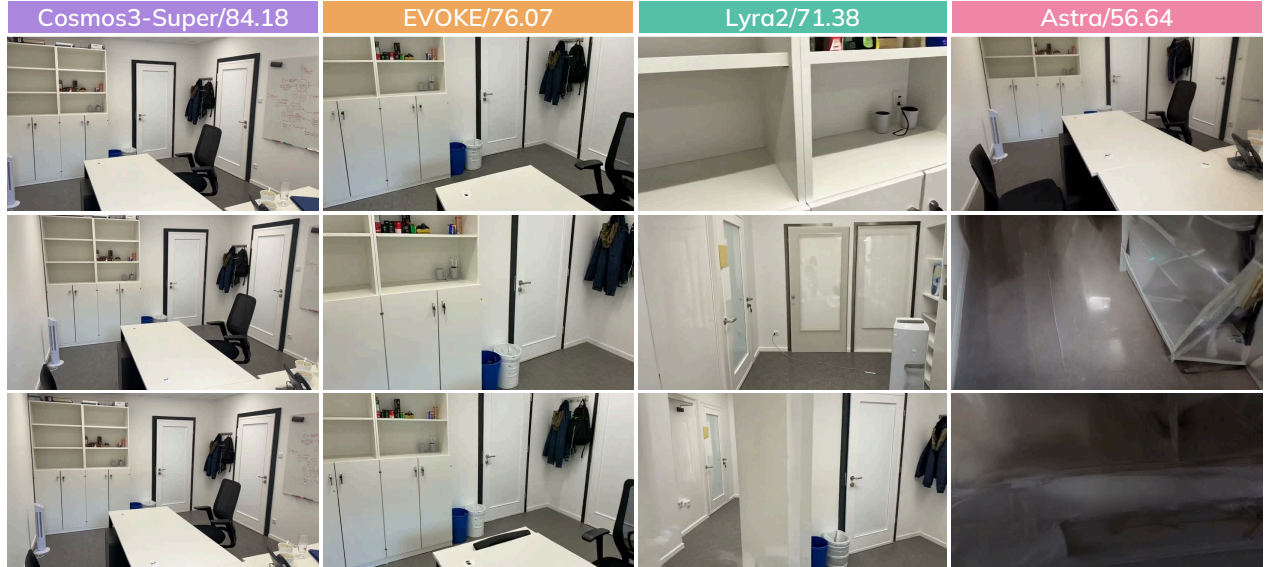


Figure 27. Visual quality on a ScanNet++ forward-backward example. Rows show three matched timestamps, with models arranged in columns. Displayed scores are for this case.

for each successive camera window. The first request uses I_0 , and each continuation uses the preceding generated last frame. We assemble the outputs by removing shared boundary frames and trimming the tail.

 **Cosmos3-Super** (NVIDIA, 2026) uses the 4-Step Image2Video checkpoint with guidance scale 1 and 832×480 output. The scene caption is combined with the motion text for each generation window. Short cases generate 357 frames in one call. Long cases use windows of 357, 357, and 249 frames, covering requested indices 0–356, 356–712, and 712–960. The 2 repeated boundary frames are removed to obtain 961 frames. Prompt times follow the native 24-FPS clock, and the assembled frames are published at 16 FPS.

 **HunyuanVideo1.5** (Tencent Hunyuan, 2025) uses the 480p image-to-video step-distilled checkpoint, 12 steps, and 848×480 output. Each window generates 121 frames at native 24 FPS, with the preceding generated last frame



Figure 28. Visual quality on a SpatialVID lateral-return example. Rows show three matched timestamps, with models arranged in columns. Displayed scores are for this case.



Figure 29. Generated-content persistence on a Sekai yaw-return example. Each column compares a model’s first visit (top) and revisit (bottom) at registered matching requested poses. Displayed scores are for this case.

conditioning the next window. Thus, 3 windows produce 361 frames for short cases, cropped to 357; 8 windows produce exactly 961 frames for long cases. Each window receives a scene description based on its actual conditioning image, followed by the window-specific camera script. The camera script preserves the timing, turning angles, and motion reversals specified by the requested trajectory within each generation window.

MiniMax-H3 (MiniMax, 2026) uses the Base FL2VA variant with 50 sampling time points, corresponding to 49 denoising iterations, at 768p. It receives the conditioning image and the scene caption combined with its window-specific motion script. Short requests produce 362 frames, cropped to 357. Long generation uses 362, 362, and 243 frames, with each later segment initialized from the preceding generated last frame. Removing the 2 repeated boundary frames leaves 965 frames, from which the first 961 are retained. The motion prompts are constructed for these specific windows at native 24 FPS before output conversion to 16 FPS.

⊗ **LTX2.5** (Lightricks, 2026) uses the 22B distilled checkpoint and its 2-stage 8 + 3-step schedule. The prompt describes the scene and the complete requested trajectory. Both short and long sequences are generated in a single call from I_0 : the short request is padded to 361 frames and trimmed to 357, while the long request produces 961 frames. Images are padded from 1280×720 to 1280×768 for inference and cropped back afterward. Prompt timings use the

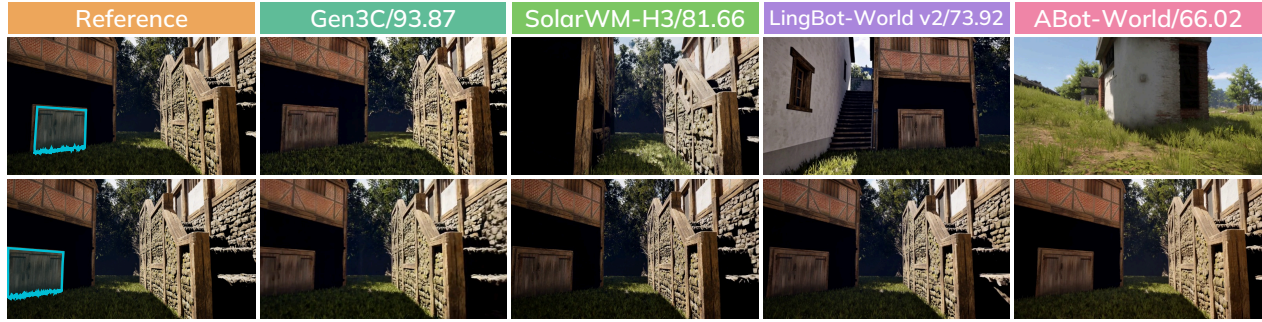


Figure 30. Input preservation on a UE yaw-return example. The left column shows the input and registered query references. The other columns show model returns (top) and queries (bottom). Ideally, Return reproduces the input, while Query preserves the reference anchor’s position and appearance. Cyan contours mark the wooden-crate anchor in the input and its input-visible query region. Displayed scores are for this case.



Figure 31. Input preservation on a UE lateral-return example. The left column shows the input and registered query references. The other columns show model returns (top) and queries (bottom). Ideally, Return reproduces the input, while Query preserves the reference anchor’s position and appearance. Cyan contours mark the input anchor and its input-visible query region. Displayed scores are for this case.

native 24-FPS clock, and output frames are published at 16 FPS.

Wan2.7 (Wan Team et al., 2025) uses the image-to-video API at 720p and native 30 FPS. Short cases request 12 seconds; long cases request 15, 15, and 3 seconds. Each request receives the relevant portion of the motion script with local times expressed at 30 FPS. Continuation requests use the generated last frame, and the assembled sequence removes repeated boundary frames before terminal trimming to 357 or 961 frames. The retained frame sequence is played at 16 FPS.

LongCat (Meituan LongCat Team, 2025) uses the base model with its step-distillation LoRA, 16 steps, guidance scale 1, and 832×480 output. Each segment contains 93 frames. The first is initialized from I_0 ; subsequent segments use 13 generated context frames and add 80 new frames through the native video-continuation interface. Consequently, 5 segments yield 413 frames for short cases and 12 yield 973 for long cases, followed by terminal trimming to 357 and 961. The scene caption is preserved and combined with window-specific English motion text. Window starts advance by 80 frames, and local times use the native 15-FPS clock. The motion script continues from the retained context and describes only the new portion of the trajectory. The model carries its native continuation cache from one segment to the next throughout the same evaluation case, for short and long videos.

E Extended Results and Analysis

We examine the capabilities behind the overall scores through complete component tables and qualitative examples. Cross-tier relationships and analyses across trajectories, horizons, and data sources appear in Sections 5.3–5.6.

Table 11. Tier 5: Detailed 3D self-consistency results. Scores use a 0–100 scale (↑). Models are ordered by Tier Score within each group. Apricot, gold, and gray mark the three highest scores per column and group.

	3D	Depth		Reprojection		Regional	GS reconstruction			Tier
Model	Repr.	Short	Medium	Short	Medium	scale	PSNR	SSIM	LPIPS	Score
World Models										
HY-WorldPlay	×	79.83	78.33	78.19	68.54	84.41	63.44	40.03	30.00	70.34
Cosmos3-Nano	×	75.79	73.59	81.96	73.17	79.05	75.89	41.88	29.52	70.10
SolarWM-H3	×	77.99	76.28	75.09	62.06	83.82	64.79	42.79	29.89	68.84
EVOKE	✓	75.88	72.93	75.21	66.40	80.33	64.55	44.38	30.70	68.02
Echo-WM	×	72.47	70.36	73.15	59.97	78.92	64.80	40.57	27.01	65.26
FantasyWorld	✓	71.51	67.24	78.78	69.42	72.42	50.22	39.13	25.25	63.53
Lyra2	✓	66.65	61.92	71.02	56.12	68.81	70.86	45.14	32.10	61.51
Genie3	–	66.75	63.53	67.53	49.72	70.05	80.60	43.12	29.89	61.25
Gen3C	✓	68.23	60.72	72.33	58.19	65.98	66.47	40.98	27.36	60.16
Astra	×	61.70	56.07	71.87	57.95	59.08	88.38	47.54	34.36	59.91
HappyOyster	–	60.16	55.32	71.49	55.32	63.24	83.19	42.01	31.73	59.17
LingBot-World v1	×	64.85	60.56	71.63	56.02	69.40	52.06	41.66	25.71	58.93
AlayaWorld	✓	65.25	60.96	73.08	59.87	67.55	54.46	36.96	24.25	58.92
LingBot-World v2	×	62.13	59.45	71.07	56.73	68.00	53.32	33.28	20.66	57.11
Matrix game 3.5	✓	55.41	53.19	70.35	56.77	60.45	65.77	31.52	23.05	54.61
SANA-WM	×	52.45	45.98	70.75	57.50	59.96	72.98	34.47	21.87	54.10
DreamX-World-5B	×	43.80	41.53	70.84	58.46	51.84	60.42	27.75	19.07	48.73
ABot-World	×	38.91	39.34	67.12	53.75	42.44	88.46	38.03	27.33	48.32
Voyager	✓	47.61	44.22	67.26	53.08	50.26	32.86	37.63	16.27	46.32
TI2V Models										
Cosmos3-Super	–	84.31	83.62	80.87	72.45	90.61	61.72	46.26	32.85	74.55
Seedance2.5	–	82.91	81.78	77.31	66.06	88.92	63.94	45.10	32.60	72.54
LTX2.5	–	80.81	79.33	75.69	63.74	87.27	69.66	46.01	32.86	71.64
HunyuanVideo1.5	×	83.92	82.65	78.37	70.65	89.25	48.35	38.54	27.68	71.31
Kling3.0	–	81.55	80.15	77.79	65.80	87.78	61.44	40.97	31.89	71.30
MiniMax-H3	–	80.29	78.47	74.64	61.82	87.09	53.05	43.59	32.31	69.42
Wan2.7	–	72.65	69.57	75.24	62.71	77.70	49.85	39.28	27.80	64.19
LongCat	×	72.77	69.39	71.82	59.83	79.07	56.88	38.62	25.83	64.11

E.1 Detailed Results by Evaluation Tier

E.1.1 Tier 1: Control Alignment

Direction following and path accuracy distinguish different control skills.

Lyra2 leads translation and Move Accuracy, while SolarWM-H3 leads rotation and Look Accuracy among world models (Table 9). LingBot-World v2 illustrates the gap between the two types of measurement: its Move Accuracy is 79.45, but its translation score is 50.94. **Moving in the right direction** does not ensure the correct *distance, timing, or camera trajectory*. The strongest TI2V model, Seedance2.5, also remains well below the leading world models across control components.

Turning is a major source of control differences.

Look Accuracy scores vary the most across camera-conditioned models, and all three action-conditioned models score lower on Look than Move. HunyuanVideo1.5 leads TI2V Move Accuracy but ranks last on Look, showing how differently the two commands can be executed. This distinction matters for yaw sweeps, curved paths, and orbits, which repeatedly change the viewing direction. In several inspected Genie3 cases, lateral requests produce turns instead of translation. Separately, its lower Look Accuracy indicates difficulty executing requested turns; the lateral-motion errors instead reflect confusion between requested translation and turning.

For ABot-World, Genie3, and HappyOyster, Tier Score uses Move and Look Accuracy. Their parenthesized translation

Table 12. Tier 6: Detailed temporal stability results. Short- and long-video means receive weights of 30% and 70%, respectively. Scores use a 0–100 scale (↑). Models are ordered by Tier Score within each group. Apricot, gold, and gray mark the three highest scores per column and group.

3D		Visual quality		Control		Geometry		Tier
Model	Repr.	Stability Score	Mean Performance	Stability Score	Mean Performance	Stability Score	Mean Performance	Score
World Models								
SolarWM-H3	×	83.16	60.29	80.01	76.78	88.96	75.85	84.05
EVOKE	✓	87.61	64.01	74.09	66.73	87.83	73.33	83.18
Genie3	–	84.70	61.43	82.72	59.83	82.04	64.91	83.15
Echo-WM	×	85.64	61.61	74.40	68.87	88.02	72.64	82.69
Lyra2	✓	81.12	54.94	81.97	78.47	83.83	64.49	82.31
Gen3C	✓	76.12	46.67	75.17	70.61	85.58	66.40	78.96
HappyOyster	–	76.13	49.28	79.66	59.98	78.97	63.57	78.25
Voyager	✓	77.53	49.49	74.48	65.08	82.12	56.86	78.04
HY-WorldPlay	×	77.23	52.96	66.05	46.37	88.70	77.46	77.32
LingBot-World v2	×	84.74	62.71	65.08	58.41	82.02	66.08	77.28
Matrix game 3.5	✓	82.35	59.11	69.21	62.75	77.68	60.96	76.41
ABot-World	×	73.87	46.83	78.08	53.59	77.01	52.80	76.32
AlayaWorld	✓	81.48	54.81	62.86	54.32	84.59	67.16	76.31
LingBot-World v1	×	81.74	58.20	61.36	50.08	83.01	64.77	75.37
SANA-WM	×	83.19	57.21	64.72	58.19	77.19	56.98	75.03
DreamX-World-5B	×	83.21	60.44	67.91	51.56	73.34	54.14	74.82
Cosmos3-Nano	×	78.45	55.77	59.33	34.93	86.36	76.22	74.71
FantasyWorld	✓	74.85	50.26	56.42	40.73	86.26	72.02	72.51
Astra	×	64.70	35.58	53.97	33.04	83.26	64.17	67.31
TI2V Models								
Seedance2.5	–	86.92	63.81	62.88	43.55	91.38	80.15	80.39
Cosmos3-Super	–	89.24	67.43	57.27	33.42	92.69	82.80	79.73
Kling3.0	–	87.31	65.18	61.29	38.15	89.31	79.52	79.30
HunyuanVideo1.5	×	85.37	63.09	59.88	37.05	91.62	81.08	78.96
MiniMax-H3	–	88.35	66.75	55.62	34.32	88.56	76.51	77.51
LTX2.5	–	88.03	66.00	54.52	31.90	89.35	78.63	77.30
Wan2.7	–	86.62	64.66	56.90	35.02	83.85	72.37	75.79
LongCat	×	86.03	64.93	50.29	27.88	86.33	70.89	74.22

and rotation scores provide additional context. Figures 24–25 compare the requested camera trajectories with those recovered by DA3 from the generated videos for the same cases.

E.1.2 Tier 2: Visual Quality

TI2V models achieve more uniformly high visual quality.

Seven of the eight TI2V models exceed the best world-model quality aggregate (Table 10). Their advantage spans aesthetic, imaging, and human-preference scores, with a smaller spread than among world models. Cosmos3-Super leads the TI2V group, while EVOKE leads world models. Figures 26–28 show the corresponding differences in detail, clarity, and scene appearance.

Perceptual quality shows larger differences between models than temporal smoothness.

HPS scores vary the most across models in both groups, and world models also differ in imaging quality. Motion-smoothness scores range only from 97.05 to 99.40, while flicker scores are also high and similar across models. These temporal measures reveal broadly smooth outputs; perceptual quality and preference account for more of the remaining differences between models. Cosmos3-Nano leads world models in flicker and smoothness, whereas EVOKE’s overall quality advantage comes from its perceptual scores.

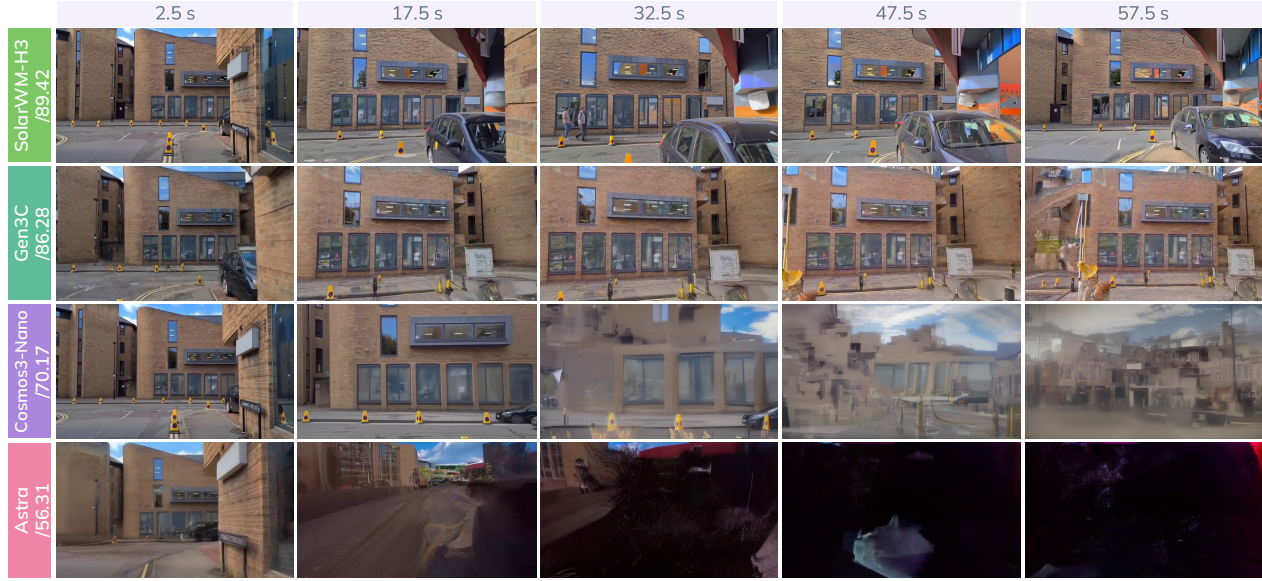


Figure 32. Temporal stability on a Sekai forward-backward example. Rows show models and columns show five timestamps from the minute-long sequence. Displayed scores are for this case.

E.1.3 Tier 3: Input Preservation

Recovering the initial scene remains harder than retaining recognizable anchors.

Return scores vary the most across models in both groups, while **Identity** scores are generally high and more similar across models (Table 3). **Identity** is evaluated during the input-to-query interval, when the initial observation is more recent. **Return** requires recovery of the starting scene after the full exploration. The contrast points to a longer-term memory challenge: recognizable objects can survive early exploration even when the original view is not faithfully recovered after completing the full exploration trajectory and returning to the initial viewpoint.

Stored scene evidence can support recovery after exploration.

Gen3C leads world-model **Return** and overall input preservation despite ranking 17th in visual quality. Its *persistent 3D cache* supplies target-view renderings during generation (Ren et al., 2025), keeping earlier scene evidence accessible when the camera returns. This cache may help it recover the original view after exploration. SolarWM-H3 instead leads **Query**, and Cosmos3-Nano leads **Identity**, showing that different models lead recovery of the starting view, fidelity at the requested query viewpoint, and recognition of the original anchors.

T12V models have higher mean **Identity** and **Return** scores but lower mean **Query** fidelity than world models. Seedance2.5 leads both checkpoints within its group; other models retain recognizable content with weaker viewpoint-specific fidelity. Figures 30 and 31 compare these checkpoints with their registered references.

E.1.4 Tier 4: Generated-Content Persistence

Recognizable revisits can still lose detailed appearance.

Different-view global DINO similarity remains relatively high for many models, while same-view perceptual fidelity exposes larger failures (Table 4). MiniMax-H3, for example, scores 77.27 on global DINO but 17.32 on normalized same-view LPIPS. A scene can retain its *semantic character* while its **textures, objects, or layout change**. The matched views in Figures 3 and 29 make these failures visible.

Appearance preservation and spatial agreement favor different mechanisms.

Lyra2 leads world-model same-view LPIPS, patch DINO, and reprojected DINO. Its stored geometry and dense



Figure 33. Temporal stability on a UE arc-return example. Rows show models and columns show five timestamps from the minute-long sequence. Displayed scores are for this case.



Figure 34. Temporal stability on a SpatialVID arc-return example. Rows show models and columns show five timestamps from the minute-long sequence. Displayed scores are for this case.

correspondences retrieve historical frames (Shen et al., 2026), which may help preserve appearance across revisits. HY-WorldPlay leads reprojected depth consistency, while SolarWM-H3 leads surface F-score. Depth compares distances on overlapping surfaces; F-score measures agreement between recovered point sets. These distinct strengths bring SolarWM-H3 and Lyr2 to **nearly identical tier scores**.

Cosmos3-Super leads TI2V persistence through strong appearance and surface agreement, while Seedance2.5 leads reprojected depth. The best models in the two groups are close, but their within-group orderings differ from visual quality: MiniMax-H3 is second in TI2V quality and sixth in persistence.

E.1.5 Tier 5: 3D Self-Consistency

Wider temporal separation exposes inconsistencies hidden between nearby views.

Every model has lower feature consistency at *medium* than at *short* temporal offsets, with greater between-model variation at the medium interval (Table 11). Depth and regional scale also separate world models more strongly than short-range features. Together, these results locate the harder part of 3D consistency: maintaining corresponding content and spatial relationships as the camera moves farther through the scene.

High reconstruction scores can coexist with weak depth and scale agreement.

HY-WorldPlay leads depth and regional scale. Its Reconstituted Context Memory retrieves geometrically relevant historical frames, and Context Forcing preserves memory use during distillation (Sun et al., 2025), consistent with its local geometric strength. Astra achieves strong reconstruction scores but weaker depth and regional-scale agreement. Reconstruction measures how well a shared 3D representation reproduces the generated views; low-detail outputs may make this task easier without yielding equally strong geometric agreement. ABot-World also combines high reconstruction PSNR with low depth and scale consistency.

Five TI2V models exceed the best world-model self-consistency aggregate, with their largest group advantages in depth and regional scale. Their weaker control results indicate that internally compatible generated views can still follow a different path from the requested cameras.

E.1.6 Tier 6: Temporal Stability

Models differ most in temporal control.

Temporal control scores vary more across models than quality or geometry scores, in both average performance and stability (Table 12). TI2V models have higher group means in quality and local geometry but weaker control, bringing their aggregate Tier 6 performance close to that of world models. Seedance2.5 leads its group overall through stronger control, while Cosmos3-Super leads quality and geometry.

Different models lead in average performance and temporal stability.

Genie3 leads the *temporal control score*, while Lyra2 has the highest *mean control level*. EVOKE leads world-model temporal quality; SolarWM-H3 leads geometric stability and the overall tier, while HY-WorldPlay has the highest mean geometry level. SolarWM’s camera-conditioned autoregressive training, distribution-matching distillation (Huang et al., 2026), and retained rollout history offer a possible explanation for its sustained performance. Figures 32–34 show how the generated views evolve; Section 5.5 examines the corresponding temporal performance.

E.2 Additional Annotated Examples and Review Interface

These examples extend the four source-group figures in Appendix A.3 with additional exploration templates. The final figure shows the manual review interface for a Sekai orbit case.

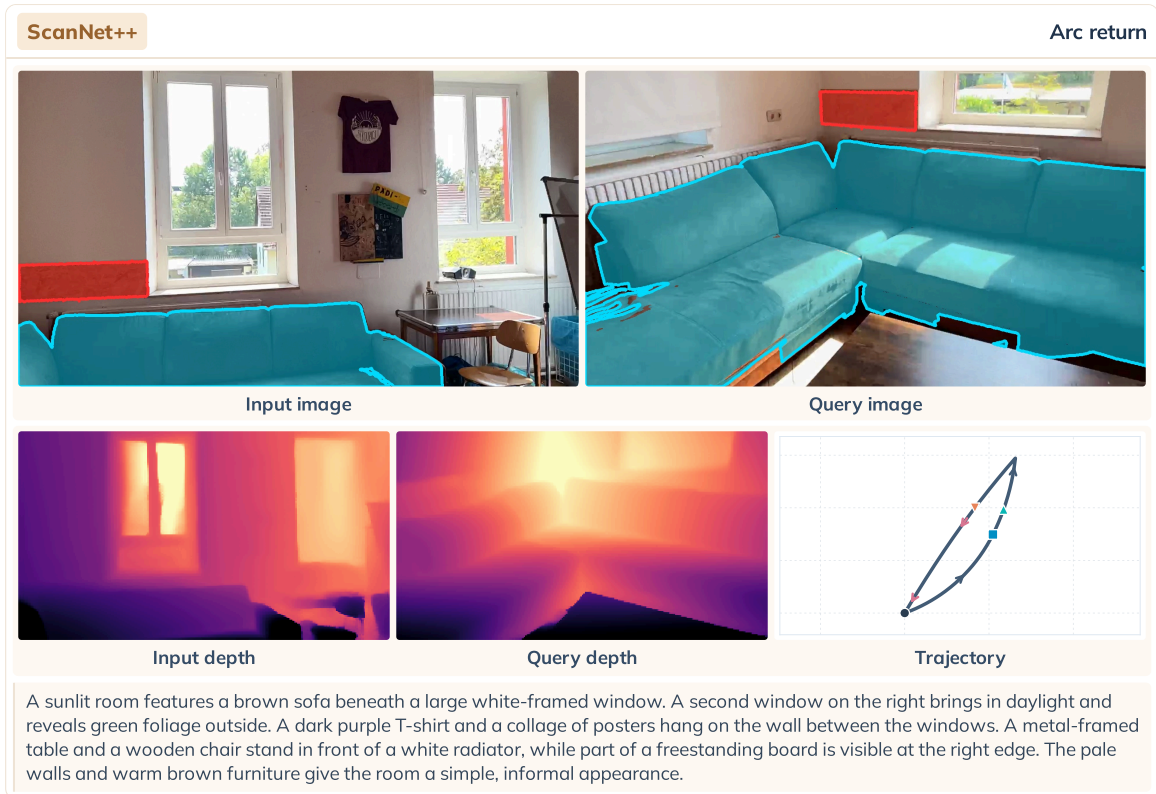
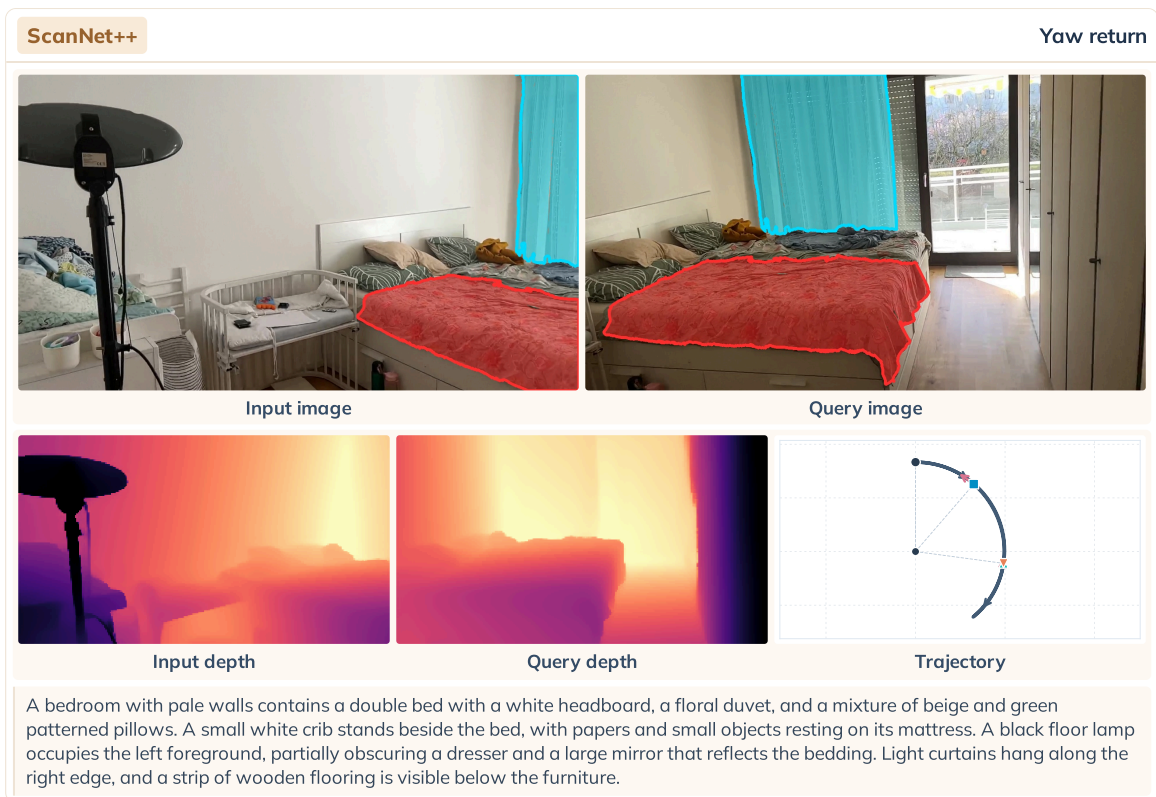


Figure 35. Examples of the yaw return (top) and arc return (bottom) templates from ScanNet++.

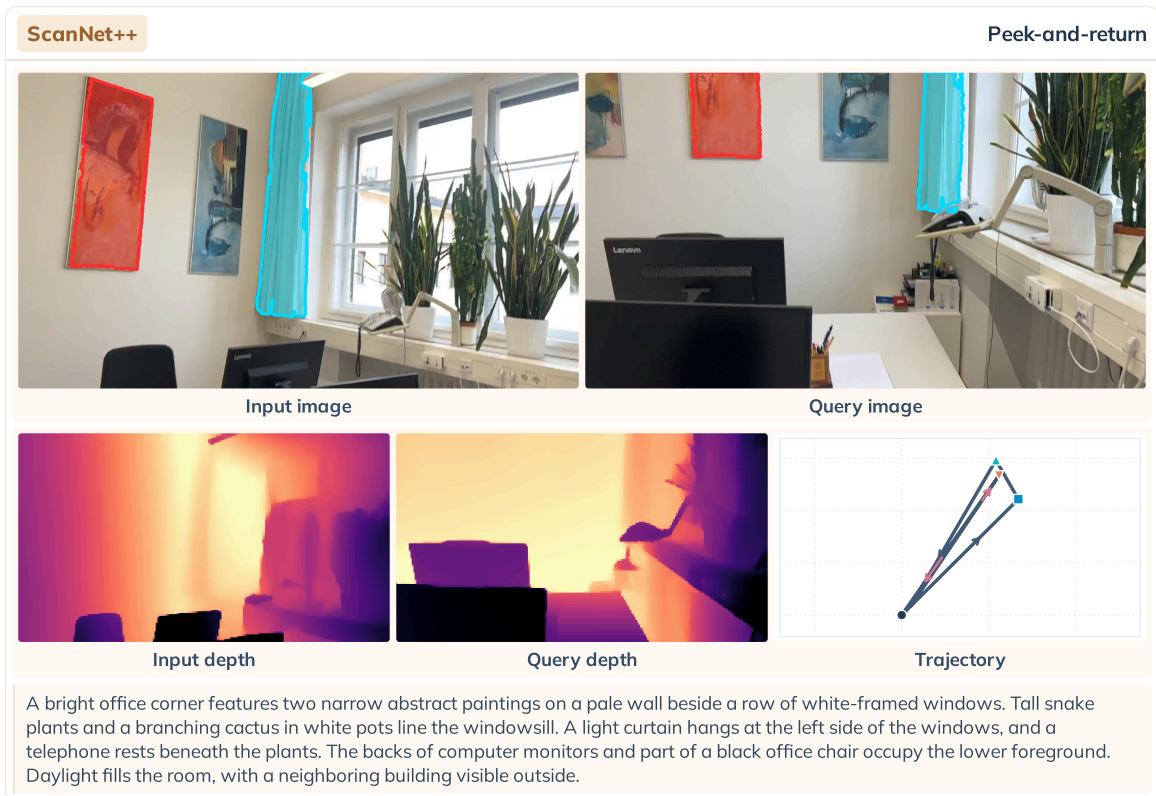
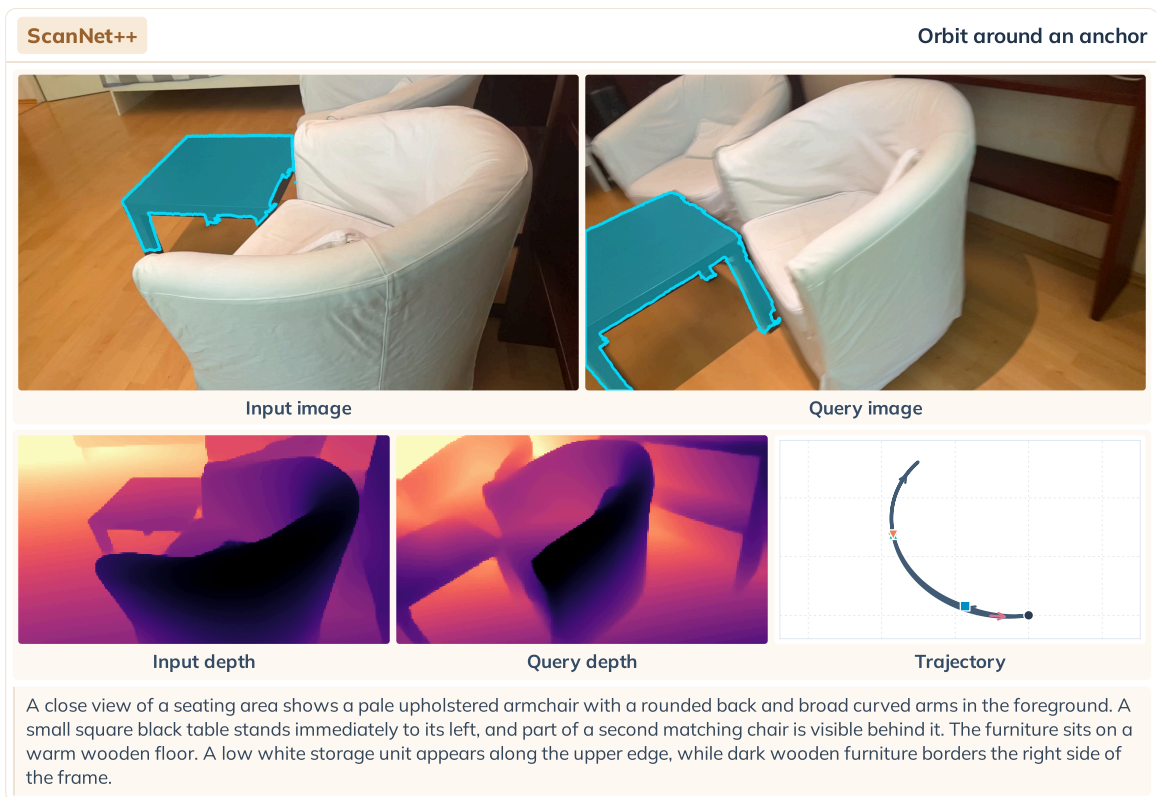


Figure 36. Examples of the orbit around an anchor (top) and peek-and-return (bottom) templates from ScanNet++.

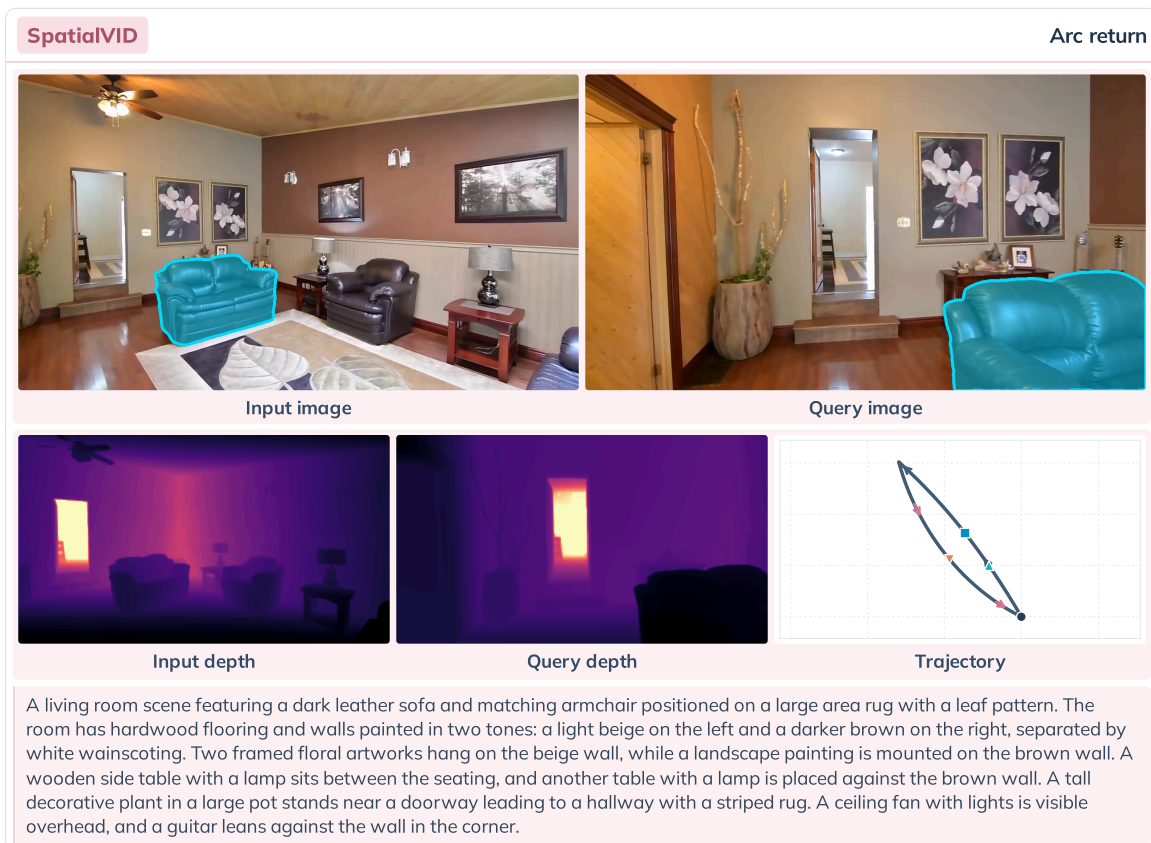


Figure 37. Examples of the yaw return template from Sekai (top) and the arc return template from SpatialVID (bottom).

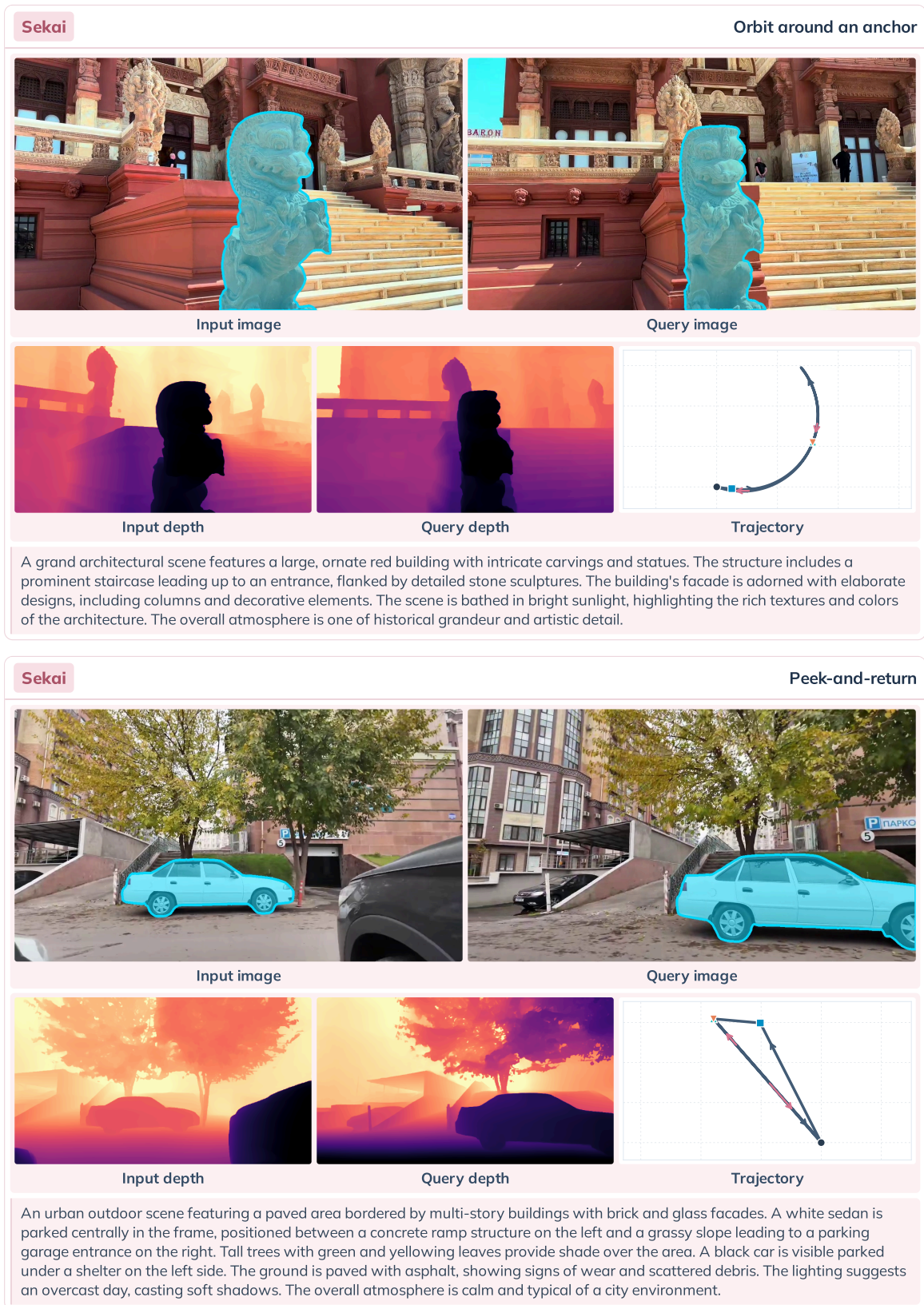


Figure 38. Examples of the orbit around an anchor (top) and peek-and-return (bottom) templates from Sekai.

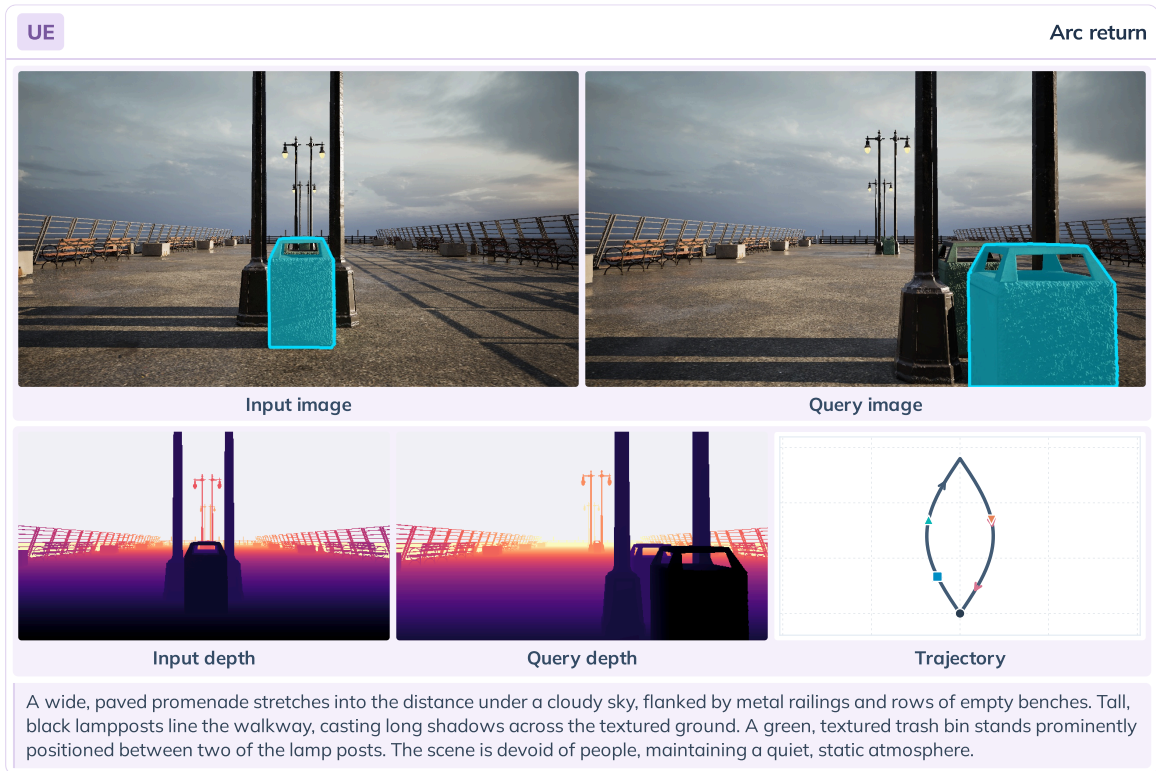
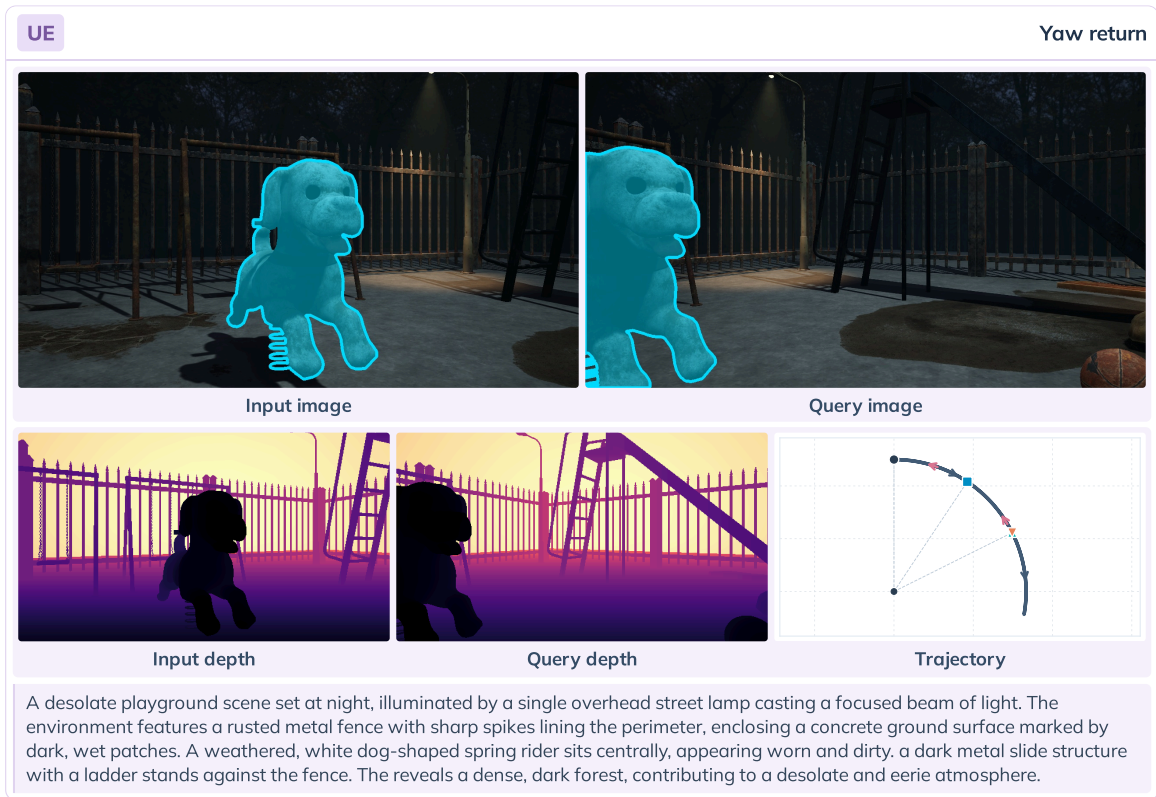


Figure 39. Examples of the yaw return (top) and arc return (bottom) templates from UE.

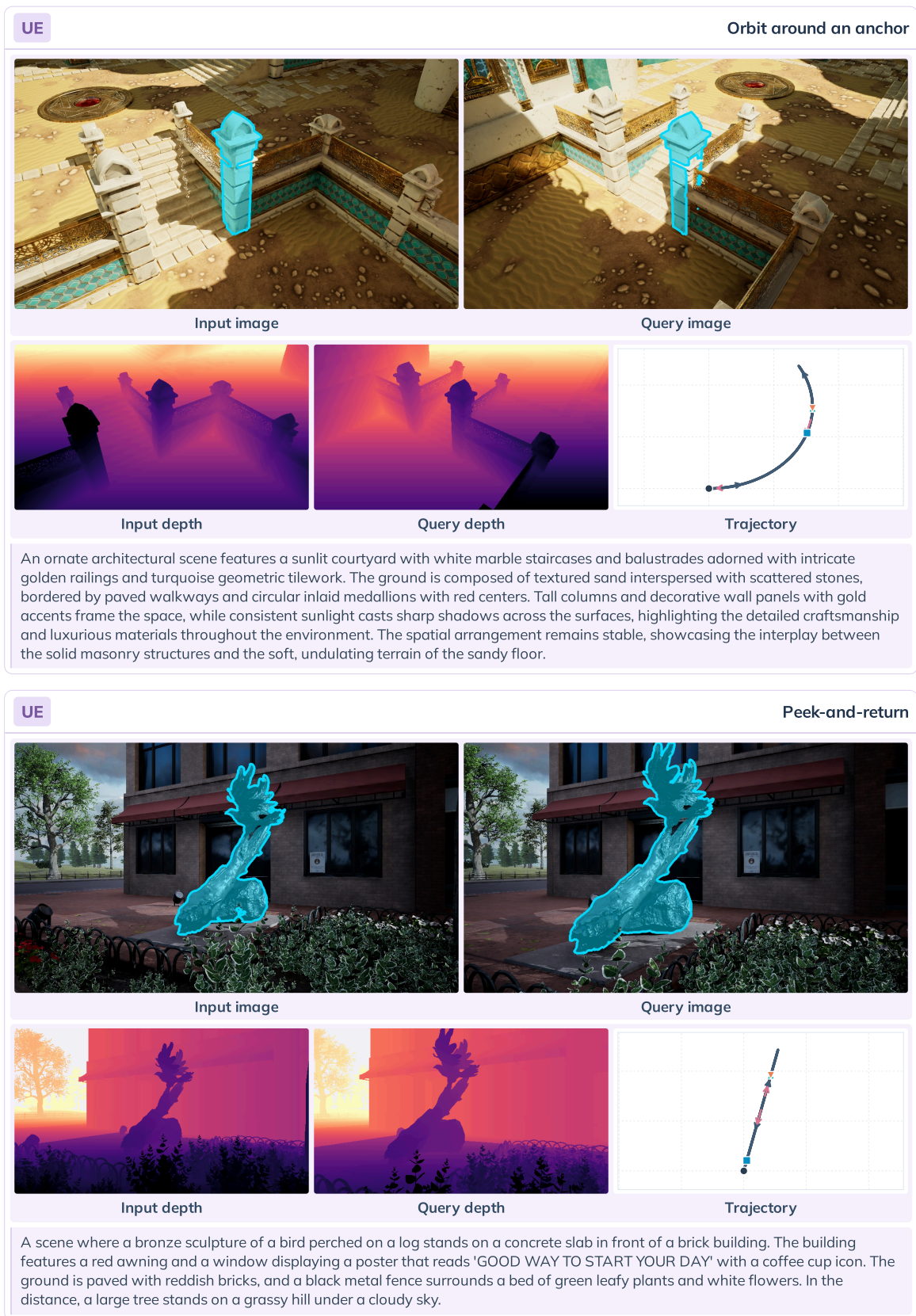


Figure 40. Examples of the orbit around an anchor (top) and peek-and-return (bottom) templates from UE.

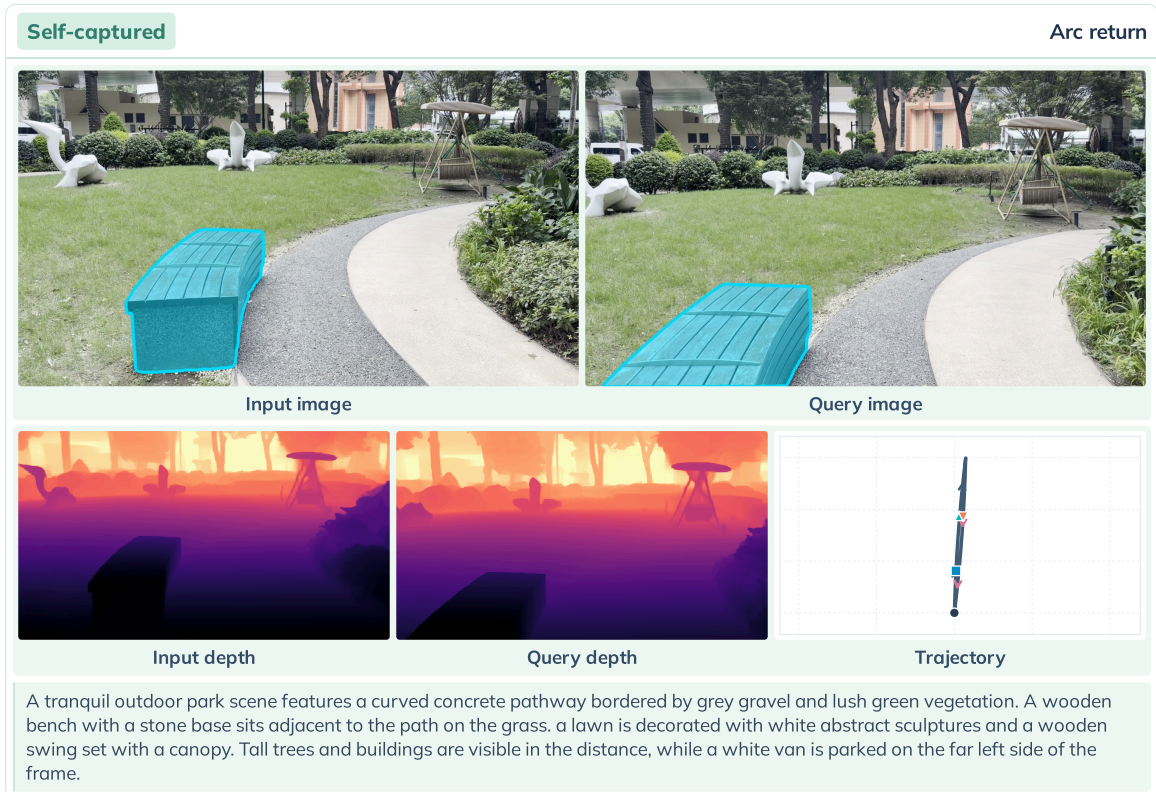
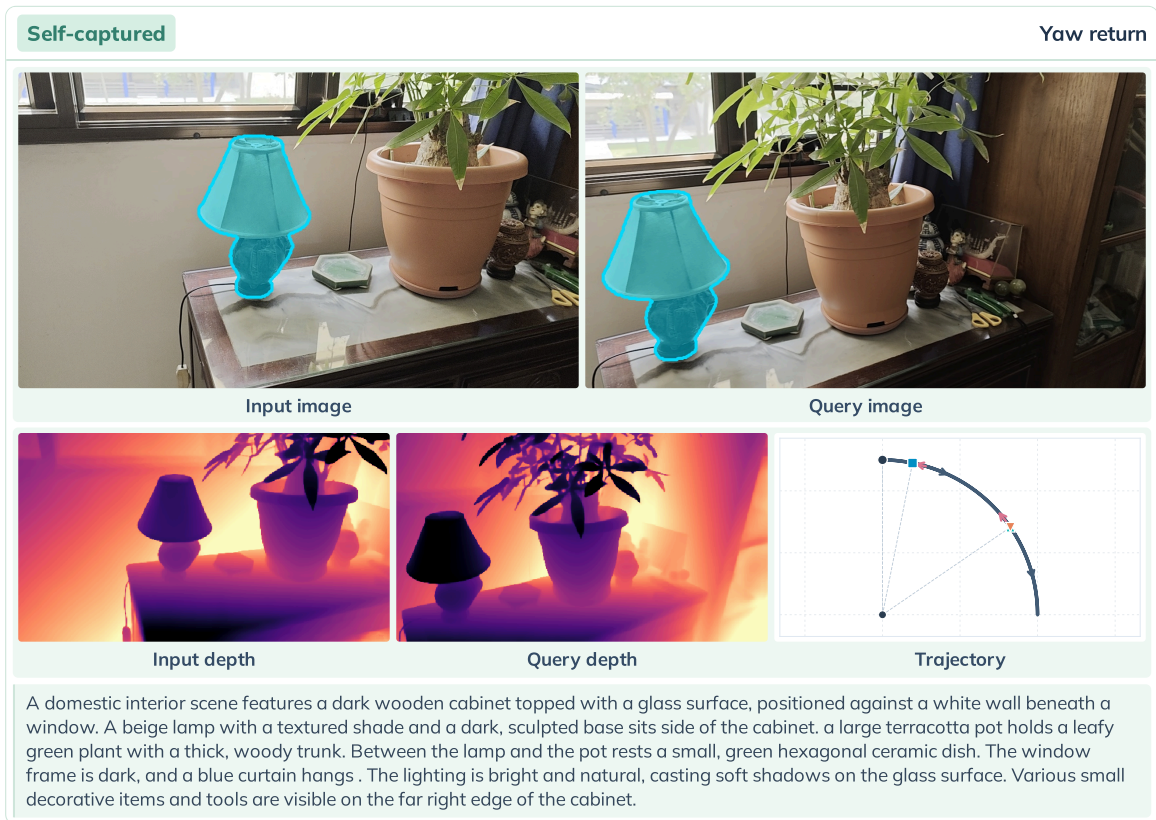


Figure 41. Examples of the yaw return (top) and arc return (bottom) templates from self-captured videos.

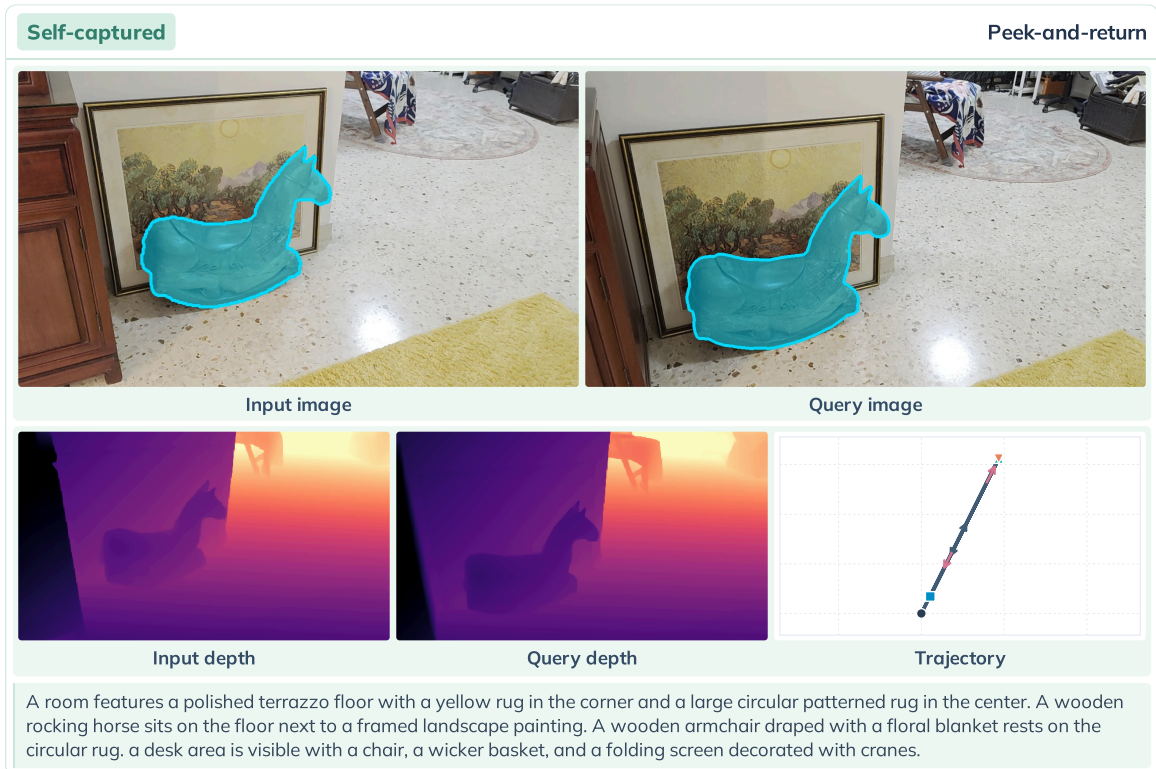
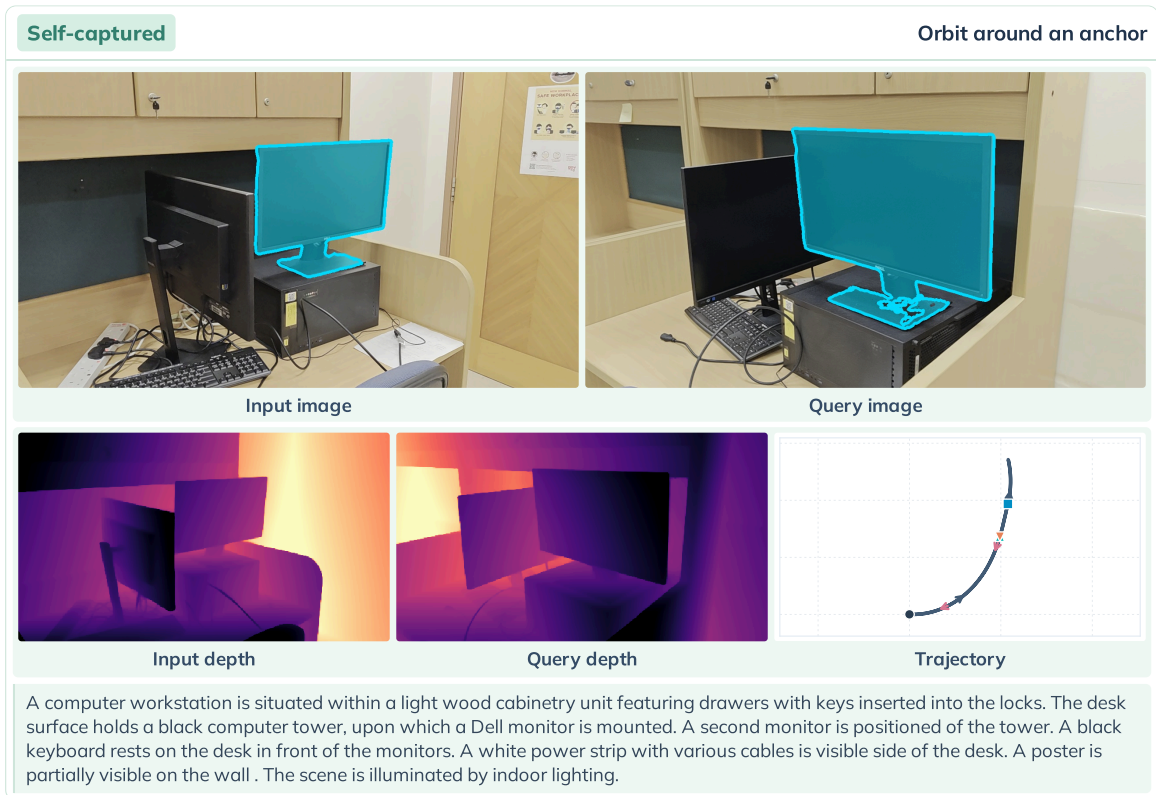


Figure 42. Examples of the orbit around an anchor (top) and peek-and-return (bottom) templates from self-captured videos.

#2 Kept by reviewer 57.6

Scene VLM PASS · The scene is a continuous outdoor shot of an ornate red building with stone carvings and stairs. People are small, peripheral, and not moving through the main scene. No overlays, weather issues, or editing boundaries are present.
No scene hazards detected

Image 5.0/5 · Anchor 29.0/35 · Trajectory 18.5/20 · Completion 5.1/10 · Total score 57.6/70

Q0/QE sharpness 1906/1827 · Completion moderate

Manual review

Keep scene and all anchors Keep scene; choose anchors Reject scene Undecided Keep

Review individual anchors

A1 building steps Reject

Area / placement 8.4/13 · Integrity 6.8/12 · depth 3.0/10

Keep anchor Reject anchor Undecided

A2 ornate stone guardian Keep

Area / placement 13.0/13 · Integrity 7.1/12 · depth 8.9/10

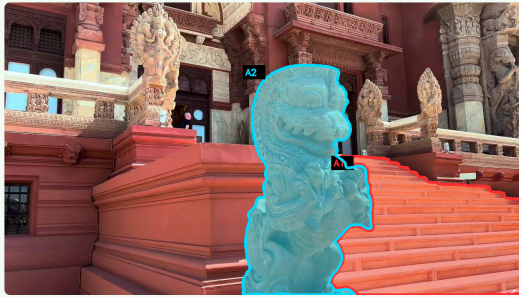
Keep anchor Reject anchor Undecided

Notes

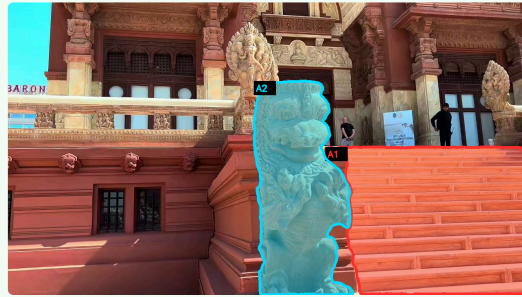
e.g., fragmented mask, cropped object, keep A2, or other observations...

Candidate anchors

A1 building steps A2 ornate stone guardian



Q0 selected anchors



QE selected anchors

Show score breakdown and supporting evidence

Anchor scores

- A1 building steps: Area / placement 8.4/13 · Integrity 6.8/12 · depth 3.0/10
- A2 ornate stone guardian: Area / placement 13.0/13 · Integrity 7.1/12 · depth 8.9/10

Trajectory quality

Motion extent 15.0/15 · Upright / vertical behavior 3.5/5

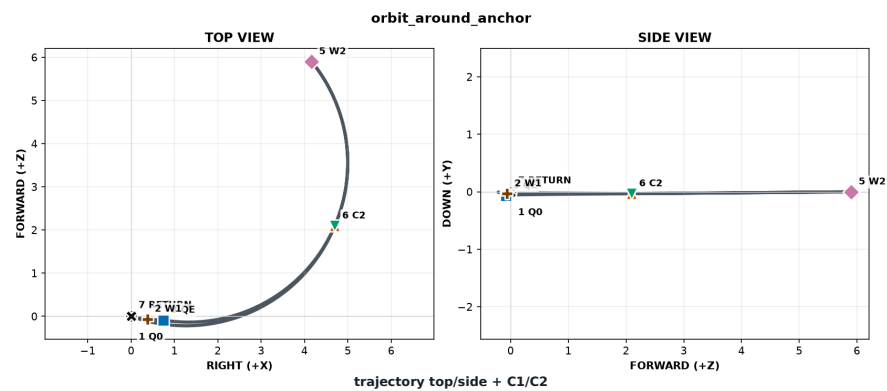


Figure 43. Manual review of a Sekai orbit case. The interface combines scene review, annotation scores, per-anchor decisions, reference overlays, sampled frames, and camera trajectories.

References

- AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, et al. AlayaWorld: Interactive Long-Horizon World Modeling – Full Technical Report. *arXiv preprint arXiv:2607.18367*, 2026. URL <https://arxiv.org/abs/2607.18367>.
- 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. *arXiv preprint arXiv:2406.03520*, 2024.
- Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI Technical Report, 2024. URL <https://openai.com/index/video-generation-models-as-world-simulators/>.
- Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning*, 2024.
- ByteDance Seed. Introducing Seedance 2.5. <https://seed.bytedance.com/en/blog/one-take-creation-flexible-referencing-introducing-seedance-2-5>, 2026.
- Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. URL <https://arxiv.org/abs/2511.16719>.
- Hanlin Chen, Jiaxin Wei, Xibin Song, Yifu Wang, Steve Wang, Hongdong Li, Pan Ji, and Gim Hee Lee. Robust dreamer: Deviation-aware latent gaussian memory for action-controlled AR video generation. *arXiv preprint arXiv:2605.30855*, 2026a.
- Haoyu Chen, Kaichen Zhou, Hang Hua, Kaile Zhang, Jingwen Qian, Wufei Ma, Haonan Chen, Chunjiang Liu, Yizhou Zhao, Xiaoyuan Wang, Weiye Li, Alan Yuille, Paul Pu Liang, and Yilun Du. MemoBench: Benchmarking world modeling in dynamically changing environments. In *European Conference on Computer Vision*, 2026b. URL <https://arxiv.org/abs/2606.27537>.
- Weiliang Chen, Haowen Sun, Jun Gao, Jiawei Chi, Hanyang Wang, Qiyu Dai, Yihao Li, Hao Li, Jingnan Gao, Yi-Hsin Hung, Xingzhuo Guo, Shangchen Miao, Zhiyuan Shi, Xiang Li, Fengrui Tian, Weihua Du, Ziqi Huang, Shenyuan Gao, Siqiao Huang, Mingyu Liu, Yifei Li, Shizun Wang, Xi Wang, Tianqi Zhang, Xue Luo, Xiyin Ren, Jinshan Ren, Xiaoyang Shen, Xiaobo Hu, Zhiyang Dou, Mingyu Ding, Yichao Yan, Xinchao Wang, Yizhou Wang, Shilong Liu, Wenzhao Zheng, Yueqi Duan, Yuan Gong, Ziwei Liu, Ming-Yu Liu, Jialong Wu, Jiangran Lyu, and Fangfu Liu. HarnessEval-W: Agentifying the Evaluation of Visual Worlds. *arXiv preprint arXiv:2608.16859*, 2026c. URL <https://arxiv.org/abs/2608.16859>.
- Zhifei Chen, Luozhou Wang, Guibao Shen, Dongyu Yan, Shuai Yang, Tianshuo Xu, Yihua Du, Wei Wang, Tianyi Gui, Lianghua Huang, and Yingcong Chen. ReWorld: An Interactive World Model with Long-Horizon Memory. *arXiv preprint arXiv:2608.23565*, 2026d. URL <https://arxiv.org/abs/2608.23565>.
- DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, et al. DreamX-World 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026. URL <https://arxiv.org/abs/2606.16993>.
- Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. WorldScore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL <https://arxiv.org/abs/2504.00983>.
- Nan Duan, Haoyang Huang, Weiyang Jin, Haoran Li, Yaowei Li, Yuming Li, Yijun Liu, Xin Lu, Xiaoxiao Ma, Yanwen Ma, Yaofeng Su, Yilang Sun, Haoyu Wang, Zeyue Xue, Songchun Zhang, and Junhao Zhuang. Long-Horizon Audio-Visual Generation for Persistent Stories and Interactive Worlds. *arXiv preprint arXiv:2608.23383*, 2026. URL <https://arxiv.org/abs/2608.23383>.
- Chenrui Fan and Paolo Favaro. GeoT2V-Bench: Benchmarking 3D Consistency in Text-to-Video Models via 3D Reconstruction. *arXiv preprint arXiv:2606.24829*, 2026. URL <https://arxiv.org/abs/2606.24829>.
- Jianjie Fang, Yingshan Lei, Qin Wan, Ziyu 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. *arXiv preprint arXiv:2605.03941*, 2026. URL <https://arxiv.org/abs/2605.03941>.
- FantasyWorld Team. FantasyWorld. <https://github.com/Fantasy-AMAP/fantasy-world>, 2025.
- Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, et al. Seedance 1.0: Exploring the Boundaries of Video Generation Models. *arXiv preprint arXiv:2506.09113*, 2025. URL <https://arxiv.org/abs/2506.09113>.

- Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026. URL <https://arxiv.org/abs/2607.07534>.
- Google DeepMind. Veo 3.1. Official model release, 2025. URL <https://deepmind.google/models/veo/>.
- Bohai Gu, Yueyang Yuan, Taiyi Wu, Dazhao Du, Jian Liu, Xiaoyi Pang, Jie Zhang, Xiaocheng Lu, Haobin Zhong, Xiaotong Zhao, Alan Zhao, and Song Guo. WorldCycle: Self-Verifiable Reinforcement Learning for Long-Horizon Video World Models. *arXiv preprint arXiv:2608.04964*, 2026a. URL <https://arxiv.org/abs/2608.04964>.
- Leslie Gu, Junhwa Hur, Charles Herrmann, Fangneng Zhan, Todd Zickler, Deqing Sun, and Hanspeter Pfister. GeCo: Evaluating Geometric Consistency for Video Generation via Motion and Structure. *arXiv preprint arXiv:2512.22274*, 2025. URL <https://arxiv.org/abs/2512.22274>.
- Qiwen Gu, Bingjie Gao, Rui Chen, Geng Li, Jifan Li, Qishuai Wen, Li Niu, Jing Tang, Xiangxiang Chu, and Junqiao Zhao. R2M-Bench: Evaluating revisit memory via relative consistency in interactive video world models. *arXiv preprint arXiv:2608.27328*, 2026b.
- 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. URL <https://arxiv.org/abs/2404.02101>.
- Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, et al. Matrix-Game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.
- Yue Hu, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. EWM-Bench: Evaluating Scene, Motion, and Semantic Quality in Embodied World Models. *arXiv preprint arXiv:2505.09694*, 2025. URL <https://arxiv.org/abs/2505.09694>.
- Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025a. URL <https://arxiv.org/abs/2508.10934>.
- Junchao Huang, Guian Fang, Shengju Qian, Xianghao Kong, Zhuoran Zhao, Wei Huang, Yihua Du, Zixin Zhang, Justin Cui, Yuchao Gu, Yukang Chen, Xinting Hu, Tianyu He, Shaoshuai Shi, Zhuotao Tian, Xin Wang, Mike Zheng Shou, and Li Jiang. SolarWM: Open data and scalable training for long-horizon video world models. *arXiv preprint arXiv:2609.02886*, 2026. URL <https://arxiv.org/abs/2609.02886>.
- Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson W. H. Lau, Wangmeng Zuo, and Chunchao Guo. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *arXiv preprint arXiv:2506.04225*, 2025b. URL <https://arxiv.org/abs/2506.04225>.
- Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.
- InSpatio Team, Donghui Shen, Guofeng Zhang, Haomin Liu, Haoyu Ji, Jialin Liu, Jing Guo, Nan Wang, Siji Pan, Weihong Pan, Weijian Xie, Xiaojun Xiang, Xiaoyu Zhang, Xianbin Liu, Yifu Wang, Yipeng Chen, Zhewen Le, Zhichao Ye, and Ziqiang Zhao. InSpatio-WorldFM: An open-source real-time generative frame model. *arXiv preprint arXiv:2603.11911*, 2026. URL <https://arxiv.org/abs/2603.11911>.
- Fan Jiang, Zhaoxu Sun, Mengchao Wang, Ziyu Zhu, Chiyu Wang, Yunpeng Zhang, et al. ABot-World-0: Infinite interactive world rollout on a single desktop gpu. *arXiv preprint arXiv:2607.19191*, 2026. URL <https://arxiv.org/abs/2607.19191>.
- Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *arXiv preprint arXiv:2407.06358*, 2024. URL <https://arxiv.org/abs/2407.06358>.
- Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5148–5157, 2021.
- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.
- Kling AI. Kling Video 3.0. <https://kling.ai/dev/model-release/video-3-omni>, 2026.

- Lingdong Kong, Yu Yang, Jianbiao Mei, Youquan Liu, Ao Liang, Dekai Zhu, Dongyue Lu, Wei Yin, Xiaotao Hu, Mingkai Jia, Junyuan Deng, Kaiwen Zhang, Yang Wu, Tianyi Yan, Shen Yuan Gao, Song Wang, Linfeng Li, Liang Pan, Yong Liu, Jianke Zhu, Wei Tsang Ooi, Steven C. H. Hoi, and Ziwei Liu. 3d and 4d world modeling: A survey. *arXiv preprint arXiv:2509.07996*, 2025. URL <https://arxiv.org/abs/2509.07996>.
- Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging Video Generation Models As World Models. *arXiv preprint arXiv:2502.20694*, 2025a. URL <https://arxiv.org/abs/2502.20694>.
- 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*, 2025b. URL <https://arxiv.org/abs/2506.17201>.
- 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*, 2025c. URL <https://arxiv.org/abs/2506.18903>.
- Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, Zizhen Li, Fanrui Zhang, Jiaxin Ai, Zhixiang Wang, Yuwei Wu, Tong He, Yunde Jia, and Kaipeng Zhang. Sekai: A video dataset towards world exploration. In *Advances in Neural Information Processing Systems*, 2025d. URL https://papers.neurips.cc/paper_files/paper/2025/hash/2368d4a3f24122b9f3779aaffa261719-Abstract-Datasets_and_Benchmarks_Track.html.
- Ao Liang, Lingdong Kong, Tianyi Yan, Hongsi Liu, Wesley Yang, Ziqi Huang, Wei Yin, et al. WorldLens: Full-spectrum evaluations of driving world models in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL <https://arxiv.org/abs/2512.10958>.
- 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. *arXiv preprint arXiv:2412.12091*, 2024. URL <https://arxiv.org/abs/2412.12091>.
- Lightricks. LTX-2.5. <https://huggingface.co/Lightricks/LTX-2.5>, 2026.
- Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e4cd50120b6d7e8daff1749d6bbaa889-Abstract-Conference.html.
- 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024.
- Dongyue Lu, Ao Liang, Tianxin Huang, Xiao Fu, Yuyang Zhao, Baorui Ma, Liang Pan, Wei Yin, Lingdong Kong, Wei Tsang Ooi, and Ziwei Liu. See4D: Pose-free 4d generation via auto-regressive video inpainting. In *Eurographics*, 2026a. URL <https://arxiv.org/abs/2510.26796>.
- Jinpeng Lu, Dexu Zhu, Haoyuan Shi, Linghan Cai, Guo Tang, Yinda Chen, Jie Cao, Duyu Tang, Yi Zhang, Yong Dai, and Xiaozhu Ju. Current world models lack a persistent state core. *arXiv preprint arXiv:2606.20545*, 2026b. URL <https://arxiv.org/abs/2606.20545>.
- Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A text-controlled interactive world generation model. *arXiv preprint arXiv:2512.22096*, 2025.
- Meituan LongCat Team. LongCat-Video. <https://github.com/Meituan-LongCat/LongCat-Video>, 2025.
- Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards World Simulator: Crafting Physical Commonsense-Based Benchmark for Video Generation. *arXiv preprint arXiv:2410.05363*, 2024. URL <https://arxiv.org/abs/2410.05363>.
- MiniMax. MiniMax-H3. <https://www.minimax.io/blog/minimax-h3>, 2026.
- NVIDIA. Cosmos 3 Model Reference. https://docs.nvidia.com/cosmos/latest/cosmos3/model_reference.html, 2026.
- NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, et al. Cosmos World Foundation Model Platform for Physical AI. *arXiv preprint arXiv:2501.03575*, 2025. URL <https://arxiv.org/abs/2501.03575>.

- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.
- Jack Parker-Holder and Shlomi Fruchter. Genie 3: A new frontier for world models. Google DeepMind Blog, 2025. URL <https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/>.
- Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, et al. Movie Gen: A Cast of Media Foundation Models. *arXiv preprint arXiv:2410.13720*, 2024. URL <https://arxiv.org/abs/2410.13720>.
- Runjia Qian, Zile Wang, Jihai Zhang, Kai Zou, Wei Yu, Jiaxing Li, Zexiang Liu, Yaokun Li, Fei Kang, Kaichen Huang, Mengyin An, Haobo Zhang, Biao Jiang, Jiahua Wang, Haofeng Sun, Yang Liu, and Yangguang Li. Matrix-Game 3.5: Enhancing real-time streaming interactive world models with patch memory. *arXiv preprint arXiv:2608.29910*, 2026. URL <https://arxiv.org/abs/2608.29910>.
- Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, Lei Bai, Wanli Ouyang, and Ruimao Zhang. WorldSimBench: Towards Video Generation Models as World Simulators. *arXiv preprint arXiv:2410.18072*, 2024. URL <https://arxiv.org/abs/2410.18072>.
- Qwen Team. Qwen3.5-9B. Official model card, 2026. URL <https://huggingface.co/Qwen/Qwen3.5-9B>.
- 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. *arXiv preprint arXiv:2503.03751*, 2025. URL <https://arxiv.org/abs/2503.03751>.
- Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294, 2022.
- Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, Chen Gao, Wei Wu, Xihui Liu, Dhruv Shah, Zhaoxiang Zhang, Zhibo Chen, Jun Zhu, Yonghong Tian, Tat-Seng Chua, Wenwu Zhu, and Yong Li. WorldArena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026. URL <https://arxiv.org/abs/2602.08971>.
- Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, Sanja Fidler, Jiahui Huang, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra 2.0: Explorable generative 3d worlds. *arXiv preprint arXiv:2604.13036*, 2026. URL <https://arxiv.org/abs/2604.13036>.
- Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. WorldPlay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025. URL <https://arxiv.org/abs/2512.14614>.
- Stanislaw Szymanowicz, Jason Y. Zhang, Pratul Srinivasan, Ruiqi Gao, Arthur Brussee, Aleksander Holynski, Ricardo Martin-Brualla, Jonathan T. Barron, and Philipp Henzler. Bolt3D: Generating 3D Scenes in Seconds. *arXiv preprint arXiv:2503.14445*, 2025. URL <https://arxiv.org/abs/2503.14445>.
- Junshu Tang, Jiacheng Liu, Jiaqi Li, Longhuang Wu, Haoyu Yang, Penghao Zhao, Siruis Gong, Xiang Yuan, Shuai Shao, Linfeng Zhang, and Qinglin Lu. Hunyuan-GameCraft-2: Instruction-following Interactive Game World Model. *arXiv preprint arXiv:2511.23429*, 2025. URL <https://arxiv.org/abs/2511.23429>.
- Tencent Hunyuan. HunyuanVideo 1.5. <https://github.com/Tencent-Hunyuan/HunyuanVideo-1.5>, 2025.
- Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion Models Are Real-Time Game Engines. *arXiv preprint arXiv:2408.14837*, 2024. URL <https://arxiv.org/abs/2408.14837>.
- Wan Team, Ang Wang, Baole Ai, Bin Wen, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

- Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, Hao Zhu, Zhaoxiang Zhang, Xun Cao, and Yao Yao. SpatialVID: A large-scale video dataset with spatial annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_SpatialVID_A_Large-Scale_Video_Dataset_with_Spatial_Annotations_CVPR_2026_paper.html.
- Tianyi Wang, Jiazhou Chen, Yiming Xu, Xiangyu Li, Tianyi Zeng, Chih-Hsien Chou, Ning Lu, Liang Peng, Junfeng Jiao, and Christian Claudel. RoboPhys-3D: A Comprehensive Embodied World Model Evaluation via 3D Reconstruction. *arXiv preprint arXiv:2608.28718*, 2026b. URL <https://arxiv.org/abs/2608.28718>.
- 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. *arXiv preprint arXiv:2312.03641*, 2023. URL <https://arxiv.org/abs/2312.03641>.
- World Labs Team. Marble: A multimodal world model. World Labs Blog, 2025. URL <https://www.worldlabs.ai/blog/marble-world-model>.
- World Labs Team. Atlas: A world model for spatial intelligence. World Labs Blog, 2026. URL <https://www.worldlabs.ai/blog/atlas>.
- Meiqi Wu, Zhixin Cai, Fufangchen Zhao, Xiaokun Feng, Rujing Dang, Bingze Song, Ruitian Tian, Jiashu Zhu, Jiachen Lei, Hao Dou, Jing Tang, Lei Sun, Jiahong Wu, Xiangxiang Chu, Zeming Liu, and Kaiqi Huang. Omni-WorldBench: Towards a Comprehensive Interaction-Centric Evaluation for World Models. *arXiv preprint arXiv:2603.22212*, 2026a. URL <https://arxiv.org/abs/2603.22212>.
- Xindi Wu, Sven Elflein, James Lucas, Olga Russakovsky, Laura Leal-Taixé, Despoina Paschalidou, Jonathan Lorraine, and Aljosa Osep. Addressable memory for video world models. *arXiv preprint arXiv:2608.07408*, 2026b.
- 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. *arXiv preprint arXiv:2606.31672*, 2026a.
- 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. *arXiv preprint arXiv:2604.21686*, 2026b. URL <https://arxiv.org/abs/2604.21686>.
- Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning Interactive Real-World Simulators. *arXiv preprint arXiv:2310.06114*, 2023. URL <https://arxiv.org/abs/2310.06114>.
- Yu Yang, Xuemeng Yang, Licheng Wen, Lingdong Kong, Xiaobin Hu, Dongyue Lu, Wei Chow, Xiyan Huang, Yuxiang Feng, Yue Liao, Jianbiao Mei, Daocheng Fu, Rong Wu, Pinlong Cai, Ran Yi, Ying Tai, Jiangning Zhang, Botian Shi, Yong Liu, and Shuicheng Yan. Quo vadis, world modeling? *arXiv preprint arXiv:2608.02713*, 2026a. URL <https://arxiv.org/abs/2608.02713>.
- Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. NeoVerse: Enhancing 4d world model with in-the-wild monocular videos. *arXiv preprint arXiv:2601.00393*, 2026b. URL <https://arxiv.org/abs/2601.00393>.
- Yixuan Ye, Xuanyu Lu, Yuxin Jiang, Yuchao Gu, Rui Zhao, Qiwei Liang, Jiachun Pan, Fengda Zhang, Weijia Wu, and Alex Jinpeng Wang. MIND: Benchmarking memory consistency and action control in world models. *arXiv preprint arXiv:2602.08025*, 2026.
- Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.
- Yuanyang Yin, Gongxuan Wang, Yifan Zhan, Chuanhao Li, Kaipeng Zhang, and Feng Zhao. Alaya-EVOKE: From Linear-Scaling Supervision to Endless World. *arXiv preprint arXiv:2608.13546*, 2026. URL <https://arxiv.org/abs/2608.13546>.
- 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. *arXiv preprint arXiv:2605.25874*, 2026.
- Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T. Freeman, and Jiajun Wu. WonderWorld: Interactive 3D Scene Generation from a Single Image. *arXiv preprint arXiv:2406.09394*, 2024a. URL <https://arxiv.org/abs/2406.09394>.
- 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*, 2024b. URL <https://arxiv.org/abs/2409.02048>.
- Songchun Zhang, Yaowei Li, Junhao Zhuang, et al. EchoWM: Open and Enterable Omnimodal World Models. <https://arxiv.org/abs/2608.23189>, 2026.
- Yuke Zhao, Wangbo Zhao, Weijie Wang, Zeyu Zhang, Dakai An, Akide Liu, Yinghao Yu, Jiasheng Tang, Fan Wang, Wei Wang, and Bohan Zhuang. WorldOlympiad: Can your world model survive a triathlon? *arXiv preprint arXiv:2606.11129*, 2026.

- Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yanan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. URL <https://arxiv.org/abs/2503.21755>.
- 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. *arXiv preprint arXiv:2605.15178*, 2026. URL <https://arxiv.org/abs/2605.15178>.
- Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jie Zhou, and Jiwen Lu. Astra: General Interactive World Model with Autoregressive Denoising. *arXiv preprint arXiv:2512.08931*, 2025. URL <https://arxiv.org/abs/2512.08931>.