SEKAI2: FROM WORLD EXPLORATION TO INTERACTIVE WORLD MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale exploration videos provide rich observations of the real world, but interactive world modeling demands more than diverse visual coverage: models must preserve scene state over long trajectories, respond consistently to camera motion, and remain coherent when previously observed regions are revisited. Existing video datasets rarely combine minute-scale continuity, camera trajectories, temporally grounded semantics, and explicit revisit structure. Building on the broad real-world exploration coverage of Sekai, we introduce **Sekai2**, extending the data paradigm from *world exploration* to *interactive world modeling*. Rather than simply expanding the original corpus, Sekai2 re-curates selected Sekai footage and combines it with newly collected perspective and panoramic videos under a unified construction and annotation pipeline. **Sekai2 contains 128,892 clips totaling 2,826 hours** across 113 countries or regions, including **1,453 hours of continuous two-minute sequences**. Each released clip is paired with a camera trajectory and hierarchical clip- and segment-level annotations that disentangle subject motion, environmental dynamics, static scene content, and camera behavior, yielding 649,597 temporally grounded segments. To strengthen long-range spatial supervision, we further include 982 panoramic sequences totaling 119 hours that preserve non-linear routes, loops, and revisits. Fine-tuning experiments show consistent improvements across multiple world-model benchmarks, while extensive data-quality evaluations further validate the reliability of Sekai2 for long-horizon and camera-controllable generation. The complete data and annotations will be released at https://anonymous.4open.science/r/Sekai2.

1 INTRODUCTION

Recent advances in video generation are pushing generative models beyond short visual synthesis toward *long-horizon world modeling*, where generated observations must remain spatially and temporally coherent as the camera moves through an evolving environment (Bruce et al., 2024; Sun et al., 2025; Gao et al., 2026b; Mao et al., 2026; Team et al., 2026b). Such a setting requires richer supervision than conventional video generation. A model must not only capture appearance and short-term motion, but also associate visual changes with camera movement and preserve scene structure over extended trajectories. This makes *long temporal continuity*, *camera geometry*, and *temporally grounded semantics* particularly important for interactive world modeling.

Yet these forms of supervision are rarely available together in existing datasets, as summarized in Table 1. Large-scale video–text collections such as WebVid (Bain et al., 2021), Panda-70M (Chen et al., 2024b), OpenVid-1M (Nan et al., 2025), and MiraData (Ju et al., 2024) provide broad visual diversity, but generally lack explicit camera trajectories and fine-grained temporal annotations. Datasets with stronger geometric supervision, including RealEstate10K (Zhou et al., 2018), DL3DV-10K (Ling et al., 2024), SpatialVID (Wang et al., 2026b), and OmniWorld (Zhou et al., 2025), primarily target reconstruction, novel-view synthesis, or short-range camera control. Action-centric datasets and interactive pipelines (Chen et al., 2026; Che et al., 2025; He et al., 2025b; Wang et al., 2026c) introduce richer behavioral signals, but often focus on local interactions or game and simulated environments. Explicit loop and revisit trajectories are also rarely represented. **Thus, existing real-world datasets still lack unified supervision over long temporal continuity, camera motion, temporal semantics, and revisits.**

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

Table 1: Comparison with representative video datasets. "Pose" indicates camera trajectories, while "Long obs." denotes minute-scale continuous observations. "Temp. sem." refers to segment-level temporal annotations aligned with temporal intervals, and "C+S" indicates annotations at both clip and segment levels. "Revisit" denotes explicit loop or revisit trajectories. Symbols indicate whether each property is available (✓), partially available or uncertain (∼), or absent (✗).

Dataset	Units	Hours	Avg. (s)	Countries	Real	Pose	Long obs.	Temp. sem.	G+S	Revisit
WebVid (Bain et al., 2021)	10.7M	52K	18.0	–	✓	✗	✗	✗	✗	✗
Panda-70M (Chen et al., 2024b)	70.8M	167K	8.5	–	✓	✗	✗	✗	✗	✗
OpenVid-1M (Nan et al., 2025)	1.0M	2.1K	7.2	–	✓	✗	✗	✗	✗	✗
MiraData (Ju et al., 2024)	330K	–	72.1	–	∼	✗	✓	∼	∼	✗
Ego4D (Grauman et al., 2022)	–	3.7K	–	9	✓	∼	✓	✓	∼	✗
Action100M (Chen et al., 2026)	∼100M seg.	∼128K	–	–	✓	✗	∼	✓	✓	✗
SpatialVID (Wang et al., 2026b)	2.7M	7.1K	9.5	–	✓	✓	✗	∼	✗	✗
OmniWorld (Zhou et al., 2025)	96K	214	8.0	–	✗	✓	✗	✓	∼	✗
Sekai (Li et al., 2026b)	–	5K+	60	100+	✓	∼	✓	✗	✗	✗
Sekai2 (ours)	**128,892**	**2,826**	**78.9**	**113**	✓	✓	✓	✓	✓	✓

To bridge these gaps, we introduce **Sekai2**. We re-curate selected videos from Sekai (Li et al., 2026b) and process them together with newly collected real-world videos under a unified construction and annotation pipeline, providing consistent representations of visual content, camera motion, and temporal semantics. The resulting dataset contains **128,892 clips** totaling **2,826 hours** across 113 countries or regions and diverse real-world environments. Among them, **43,594 segments reach the full two-minute duration, contributing 1,453 hours**, while segments of at least one minute account for 92.3% of the total footage. To complement the limited long-range revisit supervision in ordinary perspective videos, we further include 982 panoramic sequences totaling 119 hours, explicitly preserving non-linear routes, loops, and repeated visits. Each released clip is paired with a camera trajectory and hierarchical semantic annotations, including a clip-level global annotation and segment-level temporal annotations that factorize *subject motion*, *environmental dynamics*, *static scene content*, and *camera behavior*, yielding 649,597 temporally grounded segments.

To our knowledge, **Sekai2 is the first large-scale minute-level real-world video dataset** that jointly provides camera poses, hierarchical temporal semantics, and explicit revisit trajectories. Models fine-tuned on Sekai2 achieve consistent gains across iWorld-Bench, WorldScore, and panoramic-generation metrics, demonstrating its effectiveness for long-horizon and camera-controllable generation. Extensive corpus-level evaluations further confirm the overall quality and reliability of Sekai2. Our main contributions are summarized as follows:

- **A large-scale real-world video corpus with extensive minute-scale continuity.** Sekai2 contains 128,892 clips totaling 2,826 hours across 113 countries or regions, including 43,594 continuous two-minute segments spanning 1,453 hours, providing substantial long-horizon real-world observations for world-model training.
- **Joint camera-trajectory and hierarchical temporal-semantic supervision.** Each released clip is paired with camera-pose annotations and hierarchical semantics at clip and segment levels, separating subject motion, environmental dynamics, static scene content, and camera behavior.
- **Revisit-rich panoramic trajectories for long-range spatial consistency.** We provide 982 panoramic sequences totaling 119 hours, with non-linear trajectories, loops, and repeated visits that provide supervision for spatial consistency over extended temporal and viewpoint changes.
- **Empirical validation of both data quality and training effectiveness.** Corpus-scale evaluations verify data quality, while world-model fine-tuning demonstrates consistent gains on panoramic and long-horizon generation benchmarks.

2 Related Work

Video Data for World Modeling. Large-scale video–text datasets such as WebVid (Bain et al., 2021), InternVid (Wang et al., 2024a), Panda-70M (Chen et al., 2024b), and OpenVid-1M (Nan et al., 2025) provide broad visual diversity, while ShareGPT4Video (Chen et al., 2024a), Vript (Yang et al., 2024), MiraData (Ju et al., 2024), and Action100M (Chen et al., 2026) enrich semantic and

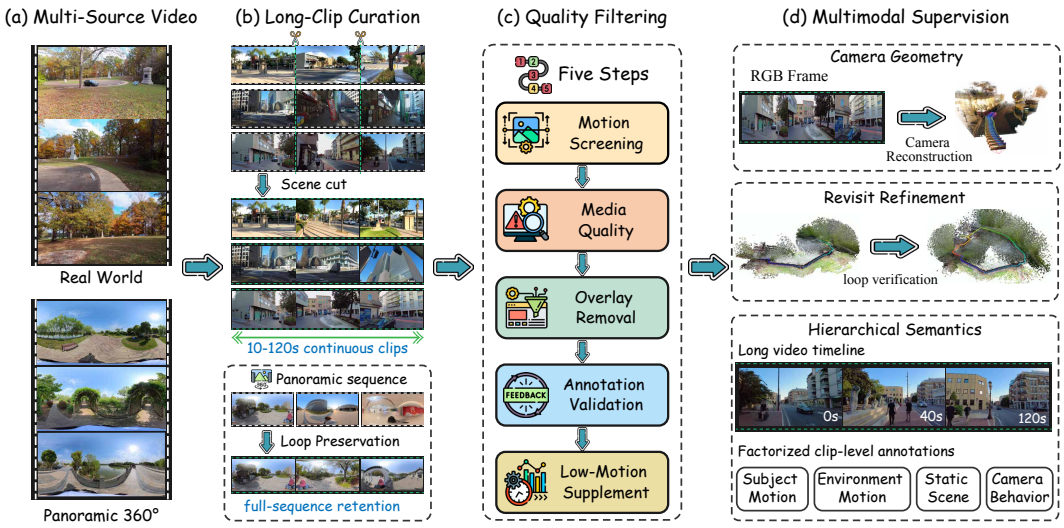


Figure 1: Overview of the Sekai2 pipeline. Multi-source videos are curated into long, scene-consistent clips, followed by quality filtering and multimodal supervision with camera trajectories, revisit refinement, and hierarchical semantic annotations.

temporal supervision. Complementary spatial datasets, including RealEstate10K (Zhou et al., 2018), DL3DV-10K (Ling et al., 2024), SpatialVID (Wang et al., 2026b), and OmniWorld (Zhou et al., 2025), provide camera poses, depth, or geometric annotations. Camera-controllable video generation further relies on poses, rays, or Plücker representations (He et al., 2024; Wang et al., 2024b; He et al., 2025a; Xu et al., 2024; Ren et al., 2025; Yu et al., 2025). Despite these advances, temporal extent, camera geometry, and fine-grained semantics are rarely available together at scale, with most large video corpora still dominated by relatively short observations. Sekai (Li et al., 2026b), our closest predecessor, provides geographically diverse real-world exploration videos. Sekai2 instead re-curates selected Sekai footage and combines it with newly collected traversal videos under a unified pipeline, extending supervision toward minute-scale and up-to-120-second continuous observations with camera trajectories and temporally grounded annotations.

Long-Horizon Interactive World Models. Interactive world models have progressed from action-conditioned simulation in driving and robotics (Yang et al., 2023; Hu et al., 2023) to controllable generation in games and open environments (Bruce et al., 2024; Valevski et al., 2025; Decart et al., 2024; He et al., 2025b; Wang et al., 2026c; Che et al., 2025; Gao et al., 2026b). Recent work increasingly emphasizes long-term spatial consistency and memory: WorldMem (Xiao et al., 2026) and long-term-memory video world models (Wu et al., 2026b) study scene persistence across large temporal gaps, while Infinite-World (Wu et al., 2026a) and PanoWorld (Li et al., 2026a) highlight revisits and panoramic context for maintaining spatial coherence. These developments place increasing demands on training data that preserve coherent observations over longer temporal and spatial horizons. Sekai2 complements this direction with sustained real-world trajectories and a dedicated revisit-rich panoramic subset, while aligning camera motion with fine-grained temporal semantics for long-horizon world modeling.

3 Dataset Construction

We construct Sekai2 through a unified pipeline that converts heterogeneous real-world videos into long-horizon training data with aligned camera trajectories and temporally grounded semantics. As illustrated in Figure 7, the pipeline consists of multi-source video collection, scene-consistent long-clip construction, quality filtering, camera-pose reconstruction, and hierarchical semantic annotation. Starting from more than 9,100 hours of candidate footage, the final release contains **128,892 clips totaling 2,826 hours**, including **1,453 hours of complete 120-second segments**. Implementation details, filtering thresholds, and annotation prompts are provided in Appendix B.

3.1 Multi-Source Video Collection

We assemble the source pool from three complementary data sources, covering diverse environments, camera motions, and trajectory structures. **(1) Sekai videos.** We incorporate the 5,000+-hour Sekai collection (Li et al., 2026b) as a large-scale source of real-world exploration footage. **(2) Newly collected perspective videos.** We collect long-form walking, traversal, driving, rail, drone, and FPV videos from YouTube. After removing 110 overlaps with Sekai, 5,089 source videos remain, producing 173,303 candidate clips totaling 3,957.4 hours before filtering. **(3) Revisit-rich panoramic videos.** We additionally capture 1,283 panoramic videos (approximately 160 hours) using Insta360 cameras. These recordings follow non-linear routes with turns, loops, and revisits, preserving surrounding context and repeated observations over extended trajectories.

Together, these sources provide more than 9,100 hours of candidate footage for subsequent long-clip construction, quality filtering, camera-pose reconstruction, and semantic annotation.

3.2 Scene-Consistent Long-Clip Construction

Long-horizon training requires temporal continuity that reflects physical scene evolution rather than editing artifacts. For perspective videos, we therefore use OmniShotCut (Wang et al., 2026a) to detect hard cuts and gradual transitions, remove temporal margins around detected boundaries, and partition each continuous shot into clips of at most 120 seconds. Residual segments shorter than 10 seconds are discarded to avoid fragmented or semantically incomplete observations. Additional segmentation and filtering details are provided in Appendix B.1.

Panoramic recordings are retained as complete sequences. Because they are directly captured along continuous trajectories, preserving the full recording maintains both the original route topology and repeated observations associated with turns, loops, and revisits. This provides uninterrupted long-range context for subsequent trajectory reconstruction and consistency analysis.

3.3 Multi-Stage Quality Filtering

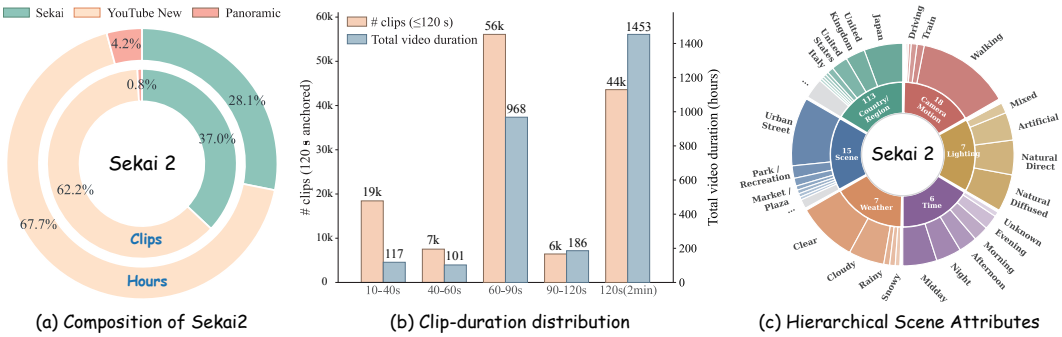
We apply a four-stage filtering procedure to ensure visual quality, temporal dynamics, and annotation reliability. First, *motion screening* uses RAFT-Small (Teed & Deng, 2020) optical flow over uniformly sampled frames to remove clips with insufficient visual variation. Second, *media-integrity filtering* checks decoding, resolution, frame rate, exposure, and border artifacts to exclude corrupted or visually degraded videos. Third, *overlay filtering* removes persistent HUDs, subtitles, watermarks, and excessive on-screen text that may interfere with scene modeling. Finally, *annotation validation* verifies the reliability of camera trajectories and semantic annotations before release.

We additionally retain 5,608 *static supplement* clips, where the camera remains largely stationary while scene dynamics are still present. These examples provide supervision for stationary camera states, helping the model capture environmental changes without introducing unnecessary viewpoint drift. Detailed filtering thresholds and stage-wise statistics are provided in Appendix B.1.

3.4 Camera-Pose Annotation

We reconstruct frame-wise camera trajectories using ViPE (Huang et al., 2025). To improve robustness in unconstrained real-world videos, GeoCalib (Veicht et al., 2024) provides camera intrinsics, while dynamic foreground regions are masked before DROID-SLAM optimization (Teed & Deng, 2021). A monocular depth prior further regularizes reconstruction over extended sequences. For each released clip, we provide frame-aligned camera-to-world poses with estimated intrinsics, which can be converted into camera rays, Plücker coordinates, or other geometric conditioning representations.

For panoramic sequences, we further exploit long-range revisits to refine accumulated trajectory drift. Loop candidates are retrieved between temporally distant keyframes, verified geometrically on the sphere, and incorporated into pose-graph optimization. This procedure successfully refines 1,051 of 1,283 panoramic trajectories (81.9%); sequences without reliable loop closures retain their original estimates. Further details are provided in Appendix B.2 and B.3.



(a) Composition of Sekai2 (b) Clip-duration distribution (c) Hierarchical Scene Attributes

Figure 2: Scale, temporal extent, and scene diversity of **Sekai2.** (a) Source composition by clip count (inner) and duration (outer). (b) Distribution of non-overlapping segments capped at 120 s, showing segment count and total duration. (c) Distribution of six clip-level attributes; inner labels indicate category counts and outer sectors show empirical frequencies.

3.5 Hierarchical Semantic Annotation

Whole-clip captions provide limited temporal specificity and often conflate scene dynamics with viewpoint changes. We therefore use Kimi-K2.6 (Team et al., 2026c) to construct hierarchical semantic annotations at two temporal scales: a ***clip-level global annotation*** summarizing the entire clip and ***segment-level temporal annotations*** describing localized content over time.

The clip-level annotation includes structured scene attributes and four factorized semantic components: ***subject motion*** for salient actions, interactions, and state changes of foreground entities; ***environmental dynamics*** for changes in surrounding objects, conditions, and ambient activity; ***static scene content*** for persistent objects, scene layout, and spatial context; and ***camera behavior*** for viewpoint evolution, camera direction, and motion patterns. This factorization separates persistent scene structure and environmental dynamics from camera-induced visual changes, providing more explicit supervision for both scene understanding and camera control.

For temporal grounding, each clip is further partitioned into segments according to changes in scene content, illumination, surface type, environmental state, or camera behavior. Each segment is annotated with timestamps, the same four semantic components, a discrete `camera_path` label, a camera-aware `full_prompt`, and a camera-agnostic `short_prompt`. This design provides fine-grained temporal supervision while explicitly separating scene content from camera information, yielding **649,597 temporally grounded semantic segments** in total. The full annotation schema and prompting details are provided in Appendix B.4.

After annotation, we perform a final verification of the released geometric and semantic supervision. Camera trajectories are checked for validity and temporal stability, while semantic annotations are examined for temporal and content consistency. We further compare pose-derived motion with the annotated `camera_path` labels as a cross-modal consistency check. Detailed quantitative evaluations are presented in Sec. 5.3 and Appendix B.5.

4 Dataset Statistics and Analysis

4.1 Scale, Composition, and Temporal Extent

Sekai2 contains 128,892 clips totaling 2,826 hours from three complementary sources. As shown in Figure 2(a), newly collected perspective videos constitute the largest portion, with 80,211 clips and 1,912 hours, followed by 47,699 re-curated Sekai clips totaling 795 hours. The panoramic subset contains only 982 sequences but contributes 119 hours; despite accounting for just 0.8% of the clips, it represents 4.2% of the total duration, reflecting its substantially longer trajectories.

Long temporal continuity is another defining property of Sekai2. Figure 2(b) shows that 43,594 segments reach the full 120-second duration, totaling **1,453 hours and 51.4% of the corpus**. Moreover, segments of at least 60 seconds account for 92.3% of the total footage. These statistics

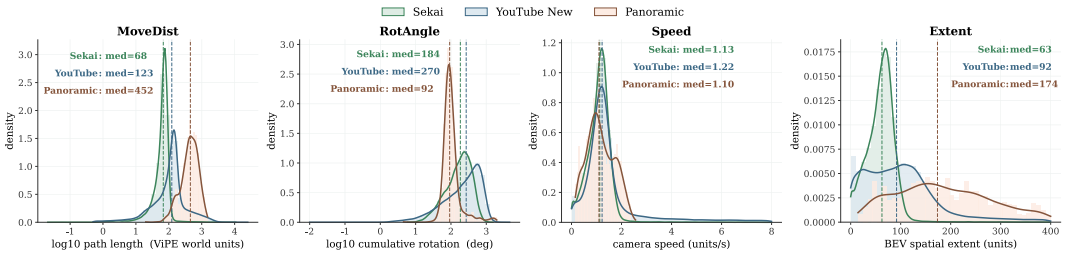


Figure 3: Camera-motion statistics across data sources. Distributions of path length, cumulative rotation, camera speed, and BEV extent; dashed lines indicate medians. Translation-related quantities use clip-wise ViPE world units and reflect distributional differences rather than metric distances.

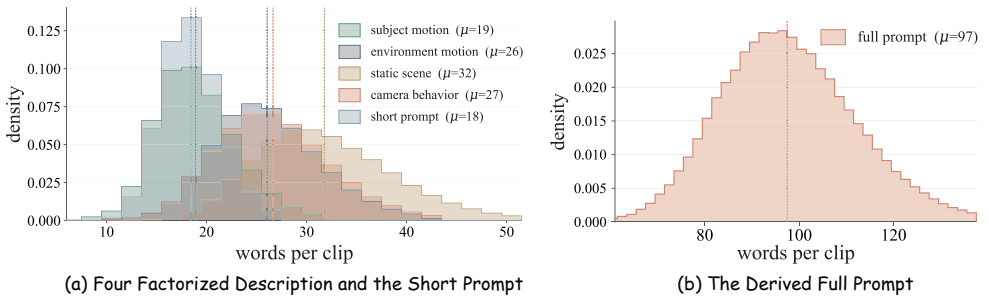


Figure 4: Length distributions of hierarchical semantic annotations. (a) Four factorized descriptions and the `short_prompt`. (b) Integrated `full_prompt`. **Dashed lines indicate mean word counts.**

indicate that minute-scale continuity is a dominant characteristic of Sekai2 rather than being limited to a small subset. Additional duration statistics are provided in Appendix A.1.

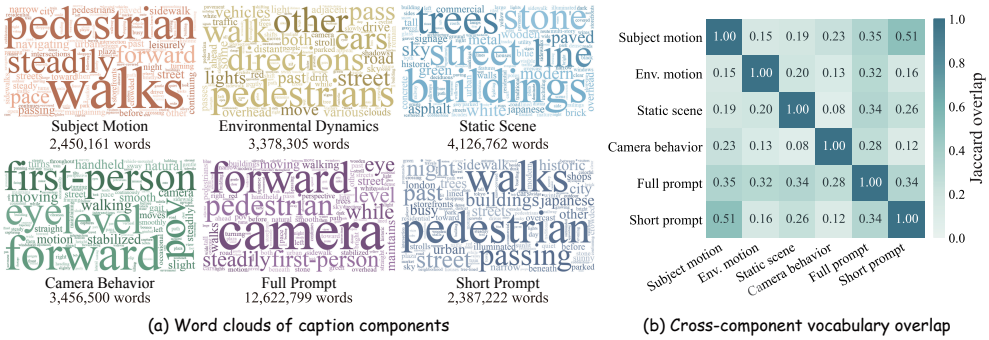
4.2 Environment and Scene Diversity

Beyond temporal scale, Sekai2 provides broad geographic and environmental coverage across 113 countries or regions. Its geographic distribution includes major sources such as Japan, the United States, the United Kingdom, China, and South Korea, together with a long tail of locations worldwide. More importantly, the diversity extends beyond geography. As shown in Figure 2(c), the annotations span 15 scene types, 18 camera-motion categories, 7 lighting conditions, 6 times of day, and 7 weather conditions, covering variations from urban streets and public spaces to natural environments, as well as different motion, illumination, and weather patterns.

The newly collected videos further complement the distribution inherited from Sekai. In particular, natural environments increase from 10.8% to 28.7%, while urban-street scenes decrease from 70.3% to 56.9%, yielding a more balanced mixture of real-world environments. Figure 8 further illustrates the global coverage and its naturally long-tailed distribution. Additional geographic and attribute statistics are provided in Appendix A.3.

4.3 Camera-Trajectory Diversity

Figure 3 compares four pose-derived trajectory descriptors across the three data sources. Compared with the re-curated Sekai subset (Li et al., 2026b), the newly collected perspective videos exhibit larger trajectory variation, with higher median path length (123 vs. 68 ViPE units) and cumulative rotation (270° vs. 184°), indicating richer translational and rotational camera motion. The panoramic subset shows a complementary profile, with the largest median path length (452 units) and bird's-eye-view extent (174 units), but substantially lower cumulative rotation (92°), consistent with omnidirectional acquisition. Median camera speed remains similar across all three sources (1.10–1.22 units/s), suggesting that their main differences lie in trajectory structure and spatial coverage rather than traversal pace. Together, the three sources provide complementary motion patterns for camera-controllable long-horizon world modeling.



(a) Word clouds of caption components (b) Cross-component vocabulary overlap

Figure 5: Lexical specialization across hierarchical annotations. (a) Word clouds of six caption components. (b) Pairwise vocabulary overlap measured by Jaccard similarity, showing distinct yet complementary lexical patterns across components.

4.4 Caption Density and Lexical Structure

Sekai2 provides dense semantic supervision at both clip and temporal-segment levels. The annotation corpus contains approximately 135 million words and 649,597 temporally grounded segments, averaging 5.04 segments per clip with a median duration of 15 seconds. Figure 4 shows that the factorized descriptions remain concise, while the integrated full prompt provides substantially richer context, averaging 97 words at the clip level and 79.6 words at the segment level. This hierarchy provides semantic supervision at complementary levels of temporal and descriptive granularity.

Figure 5(a) illustrates the lexical structure of the clip-level annotations. The factorized components exhibit semantically distinct vocabularies: subject motion emphasizes actions and movement, static scene descriptions focus on persistent scene elements, and camera behavior centers on viewpoint and motion terms. The full and short prompts integrate these complementary cues at different levels of detail, reflecting the structured design of the hierarchical annotation schema. A corresponding segment-level lexical analysis is provided in Appendix C.4.

5 Experiments

5.1 Experimental Setup

World-model fine-tuning. To validate Sekai2 as training data for world modeling, we conduct fine-tuning experiments in two settings. For ***panoramic world modeling***, we fine-tune PanoWorld (Li et al., 2026a) on approximately 120 hours of panoramic videos from Sekai2 to obtain **PanoWorld-FT**. We evaluate it on 50 panoramic test sequences, each generated as an 81-frame 960×480 equirectangular video conditioned on the initial frame, camera trajectory, and caption, against PanoWorld, HY-World 1.5 (HunyuanWorld, 2025), LingBot-World 1.5 (Gao et al., 2026a), and AlayaWorld (Team et al., 2026a). For ***long-horizon world modeling***, we fine-tune HY-World 1.5 and AlayaWorld on approximately 800 hours of newly collected perspective videos from Sekai2, training separate 40-second and 60-second variants to obtain **HYWorld-FT** and **AlayaWorld-FT**. We evaluate the corresponding variants on 100 test cases against their respective base models. Further details on the evaluation settings and protocol are provided in Appendix D.1.

Evaluation Metrics. For ***panoramic generation***, we report PSNR, FID, FAED, NIQE, and Q-Align to assess reconstruction fidelity, distributional similarity, and perceptual quality. For ***long-horizon generation***, all methods are evaluated on the same 100-case benchmark using metrics from **WorldScore** (Duan et al., 2025), **VBench** (Huang et al., 2024), and **iWorld-Bench** (Fang et al., 2026). WorldScore measures controllability and scene consistency, VBench evaluates visual and temporal quality, and iWorld-Bench assesses camera-trajectory following and auxiliary visual quality. Detailed definitions of the panoramic and long-horizon metrics are provided in Appendices D.2 and D.3, respectively. The complete per-dimension results are reported in Appendix D.4.

Table 2: **Panoramic image-to-video generation** on 50 held-out Sekai2 panoramic cases. All methods are evaluated under identical inputs. PSNR is reported over four temporal windows, while FID is computed overall and over polar and equatorial regions. Best results are shown in **bold**.

Method	PSNR ↑				FID ↓			FAED ↓	NIQE ↓	Q-Align ↑	
	20–25	50–55	70–75	75–80	Overall	Pole	Equator			Quality	Aesthetic
HY-World 1.5	15.45	14.99	14.81	14.78	76.29	109.23	48.72	1.72	2.93	4.03	3.11
LingBot-World	14.20	12.91	12.36	12.32	73.95	94.20	59.52	1.41	2.81	3.99	3.09
AlayaWorld	15.81	15.29	15.16	15.07	83.43	109.13	61.18	3.11	2.75	3.67	2.88
PanoWorld	16.69	15.54	15.01	14.91	61.88	78.35	41.02	1.19	**2.74**	4.01	**3.31**
PanoWorld-FT	**18.66**	**17.13**	**16.62**	**16.50**	**46.75**	**56.55**	**35.84**	**1.06**	2.75	**4.17**	3.26

Table 3: **Long-horizon generation** on the 100-case benchmark at 40s and 60s. A representative subset of WorldScore, VBench, and iWorld-Bench metrics is reported; values are higher-is-better.

Dur.	Method	WorldScore Metrics ↑				VBench Metrics ↑			iWorld-Bench Metrics ↑		
		Camera	3D	Photo.	Static	Subject	Imaging	Overall	Traj. Acc.	Traj. Tol.	Bright.
40s	LingBot-World	42.33	66.07	34.42	53.44	77.18	65.18	19.73	**83.46**	**64.31**	48.20
	HYWorld 1.5	43.41	77.76	81.79	58.85	70.91	59.99	17.54	78.78	59.95	38.40
	HYWorld-FT	**44.10**	83.32	89.83	59.64	72.14	62.10	19.37	80.04	60.72	46.50
	AlayaWorld	41.32	91.56	94.45	64.50	86.96	65.71	20.64	81.38	61.77	64.60
	AlayaWorld-FT	42.20	**93.04**	**95.36**	**66.11**	**88.10**	**67.54**	**23.01**	81.69	62.54	**65.30**
60s	LingBot-World	**43.67**	67.38	32.32	53.31	76.03	63.12	19.61	**82.29**	**63.06**	43.70
	HYWorld 1.5	41.63	75.12	83.09	60.24	65.41	57.40	15.89	76.27	57.84	37.40
	HYWorld-FT	42.35	79.65	84.94	62.15	67.80	59.22	16.29	77.30	58.99	40.20
	AlayaWorld	39.59	91.78	94.67	64.44	83.58	64.55	20.75	80.48	61.87	57.20
	AlayaWorld-FT	40.06	**93.30**	**94.95**	**65.98**	**85.03**	**66.35**	**22.34**	81.06	62.54	**58.10**

5.2 Quantitative Results

Panoramic World Modeling. As shown in Table 2, fine-tuning PanoWorld on the panoramic subset of Sekai2 yields consistent improvements in reconstruction quality and distributional consistency. PanoWorld-FT achieves the highest PSNR across all temporal windows, outperforming the base model by 1.59–1.97 dB; notably, PSNR remains higher in the final 75–80 frame window (16.50 vs. 14.91), indicating better preservation of visual fidelity over extended generation. Meanwhile, overall FID decreases from 61.88 to 46.75, with clear improvements in both polar and equatorial regions; the reduction is particularly pronounced at the poles, from 78.35 to 56.55, suggesting improved distributional consistency in regions that are more sensitive to equirectangular distortion. FAED further decreases from 1.19 to 1.06, while Q-Align Quality increases from 4.01 to 4.17; NIQE remains nearly unchanged and the Aesthetic score varies only marginally. Taken together, these improvements in temporal fidelity, spatial distributional consistency, and perceptual quality indicate that the panoramic subset of Sekai2 provides effective supervision for panoramic world modeling, thereby validating its utility as training data for long-horizon panoramic generation.

Long-Horizon World Modeling. To assess the training contribution of Sekai2 while controlling for differences between base models, we keep the model architecture and pretrained initialization fixed and compare performance before and after fine-tuning. As shown in Table 3, both HY-World 1.5 and AlayaWorld consistently benefit from fine-tuning on Sekai2 at 40- and 60-second horizons, outperforming their corresponding base models across multiple evaluation dimensions. For HY-World 1.5, the 40s 3D consistency and photorealism scores increase from 77.76 to 83.32 and from 81.79 to 89.83, respectively; similar gains are observed on VBench and iWorld-Bench and largely persist under the more challenging 60-second setting. AlayaWorld-FT exhibits the same overall trend, indicating that the observed benefits are not specific to a particular model architecture but generalize across different world models. These results further validate Sekai2 as effective training data for long-horizon, camera-controllable world modeling. Additional analysis are provided in Appendix D.

5.3 Dataset Quality Evaluation

Visual Quality and Temporal Dynamics. We compare 1,200 perspective clips from Sekai2 with equal-sized subsets of MiraData (Ju et al., 2024), RealEstate10K (Zhou et al., 2018), and SpatialVID (Wang et al., 2026b) under an identical evaluation protocol. As shown in Figure 6(a),

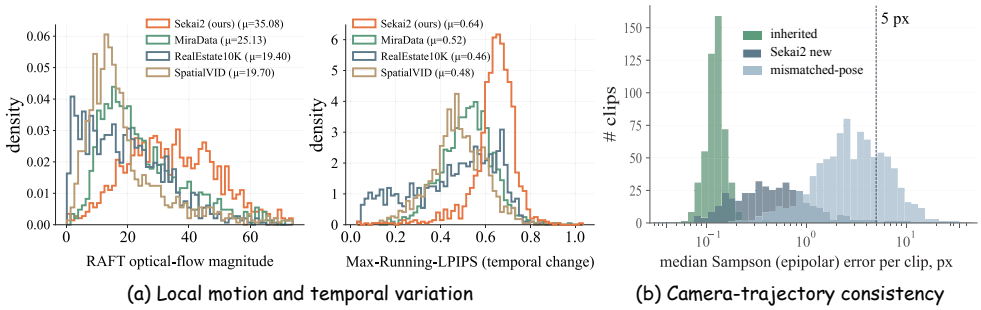


Figure 6: Dataset quality evaluation. (a) Local motion and temporal variation across Sekai2 and representative video datasets. (b) Camera-trajectory consistency measured by pose-induced epipolar error with a mismatched pose–frame control.

Sekai2 exhibits substantially richer local dynamics, with a mean RAFT optical-flow magnitude of 35.08 and Max-Running-LPIPS of 0.64, both clearly higher than the compared datasets. Meanwhile, its LAION aesthetic score remains competitive at 5.23, close to RealEstate10K (5.20) and SpatialVID (5.18). Together, these results show that Sekai2 preserves visual quality while providing richer local motion and temporal variation. Further visual-quality analysis is provided in Appendix C.1.

Camera-Trajectory Quality. We assess the geometric reliability of the released camera trajectories by comparing pose-induced epipolar geometry with image-side SIFT correspondences. As shown in Figure 6(b), the inherited Sekai and newly collected subsets achieve low median Sampson errors of 0.12 and 0.43 px, respectively, with nearly all clips below the 5 px threshold. To verify that this consistency arises from correct pose–frame alignment, we deliberately mismatch poses and frames while keeping the image correspondences and evaluation protocol unchanged. The median error increases to 2.83 px, and the fraction below 5 px drops to 74%. This degradation provides direct evidence that *the released trajectories are geometrically consistent with the observed videos*. Additional pose-quality analyses are provided in Appendix C.2.

Semantic Annotation Quality. We assess the hierarchical annotations in terms of semantic fidelity, cross-component complementarity, and temporal grounding. The frame-aware VLM judge yields consistently high scores above $4.0/5$ across coverage, scene, motion, and camera correctness (Detailed in Appendix C.3). As shown in Figure 5(b), the four factorized clip-level components exhibit limited lexical overlap (Jaccard 0.08–0.23), indicating **clear semantic specialization across annotation views**. At the segment level, independently estimated camera motion shows a consistent ordering across `camera_path` labels, with substantially larger yaw changes for turning than stationary segments. The agreement across semantic assessment, lexical structure, and pose-derived motion provides complementary evidence for the quality and reliability of our hierarchical annotations. Additional evidence on annotation quality appears in Appendix C.3 and Appendix C.4.

Case Studies. Qualitative cases further show that Sekai2 preserves long-form temporal continuity across diverse camera motions, scene transitions, and trajectory structures. Across representative sequences, visual observations align with camera trajectories and temporally grounded annotations, highlighting the dataset's geometric and semantic richness. Additional videos, annotations, trajectories, and panoramic reconstructions are provided in Appendix E.

6 Conclusion

We introduce **Sekai2**, a real-world video dataset designed to bridge *world exploration* and *interactive world modeling*. Sekai2 unifies long-form videos, camera trajectories, and hierarchical temporal annotations, providing aligned supervision for scene content, dynamic events, and viewpoint evolution. A dedicated panoramic subset further captures long non-linear trajectories, loops, and revisits, enabling richer supervision for long-range spatial and revisit consistency. Extensive fine-tuning experiments and quality evaluations validate the reliability and training value of Sekai2, supporting long-horizon and camera-controllable world modeling.

AI USE STATEMENT

We used generative AI tools to assist with hierarchical semantic annotation, selected components of the dataset construction and processing pipeline, and manuscript drafting, language editing, and organization. Generative AI was not used to produce or modify any reported quantitative results, which were obtained through our evaluation pipelines. All AI-assisted annotations, pipeline components, analyses, and manuscript text were manually reviewed and verified by the authors. The authors take full responsibility for the correctness of the final content, experimental claims, reported results, and released artifacts.

REPRODUCIBILITY STATEMENT

Detailed dataset construction, preprocessing, filtering, annotation, and quality-control procedures are provided in Appendix B, while evaluation protocols, metric definitions, and experimental details are documented in the corresponding appendix sections. To facilitate reproducibility, we will release the processing code, configurations, annotations, camera trajectories, and structured dataset indices in a format similar to UltraVideo-Long. Together with the released metadata, these resources support reconstruction of the dataset processing pipeline and reproduction of the main experimental procedures across the full workflow.

ETHICS STATEMENT

For web-sourced videos, we will provide the original source URLs and precise temporal boundaries rather than redistributing third-party video files. This release strategy preserves source attribution while minimizing unnecessary redistribution of externally hosted content.

REFERENCES

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1708–1718. IEEE, 2021.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*, 2024.

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. In *International Conference on Learning Representations*, volume 2025, pp. 37546–37593, 2025.

Delong Chen, Tejaswi Kasarla, Yejin Bang, Mustafa Shukor, Willy Chung, Jade Yu, Allen Bolourchi, Theo Moutakanni, and Pascale Fung. Action100m: A large-scale video action dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 8832–8842, June 2026.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024a.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13320–13331. IEEE, 2024b.

Etched Decart, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. *URL: https://oasis-model. github. io*, 2(3):6, 2024.

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*, pp. 27713–27724, 2025.

Jianjie Fang, Yingshan Lei, Qin Wan, Ziyou Wang, Yuchao Huang, Yongyan Xu, Baining Zhao, Weichen Zhang, Chen Gao, Xinlei Chen, et al. iworld-bench: A benchmark for interactive world models with a unified action generation framework. *arXiv preprint arXiv:2605.03941*, 2026.

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

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, et al. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026b.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Giorgio Grisetti, Rainer Kümmerle, Cyrill Stachniss, and Wolfram Burgard. A tutorial on graph-based slam. *IEEE Intelligent Transportation Systems Magazine*, 2(4):31–43, 2010.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13416–13426. IEEE, 2025a.

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

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

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Team HunyuanWorld. Hy-world 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. *arXiv preprint*, 2025.

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. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2024.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Foundations of Crystallography*, 32(5):922–923, 1976.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Rainer Kümmerle, Giorgio Grisetti, Hauke Strasdat, Kurt Konolige, and Wolfram Burgard. g 2 o: A general framework for graph optimization. In *2011 IEEE international conference on robotics and automation*, pp. 3607–3613. IEEE, 2011.

Haoyuan Li, Dizhe Zhang, Yuemei Zhou, Xiangkai Zhang, Haoran Feng, Xiaofan Lin, Wenjie Jiang, Bo Du, Ming-Hsuan Yang, and Lu Qi. Panoworld: Real-world panoramic generation. *arXiv preprint arXiv:2607.09661*, 2026a.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. *Advances in Neural Information Processing Systems*, 38, 2026b.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169. IEEE, 2024.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55, 2024.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, and Kaipeng Zhang. Yume1. 5: A text-controlled interactive world generation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7752–7761, 2026.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *International conference on learning representations*, volume 2025, pp. 1045–1064, 2025.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132. IEEE, 2025.

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pp. 2564–2571. Ieee, 2011.

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.

AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Mingliang Zhai, et al. Alayaworld: Interactive long-horizon world modeling–full technical report. *arXiv preprint arXiv:2607.18367*, 2026a.

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*, 2026b.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026c.

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

Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations*, volume 2025, pp. 73754–73776, 2025.

Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. Geocalib: Learning single-image calibration with geometric optimization. In *European Conference on Computer Vision*, pp. 1–20. Springer, 2024.

Boyang Wang, Guangyi Xu, Jiahui Zhang, Zhipeng Tang, and Zezhou Cheng. Omnishotcut: Holistic relational shot boundary detection with shot-query transformer. *arXiv preprint arXiv:2604.24762*, 2026a.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, et al. Spatialvid: A large-scale video dataset with spatial annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42592–42603, 2026b.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2024, pp. 42055–42079, 2024a.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024b.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026c.

Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, et al. Infinite-world: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. *arXiv preprint arXiv:2602.02393*, 2026a.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *Advances in Neural Information Processing Systems*, 38:49371–49393, 2026b.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *Advances in Neural Information Processing Systems*, 38:49632–49652, 2026.

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

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. *Advances in Neural Information Processing Systems*, 37:57240–57261, 2024.

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.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 100–111. IEEE, 2025.

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

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, et al. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. *arXiv preprint arXiv:2509.12201*, 2025.



Figure 7: Composition and visual coverage of **Sekai2**. Sekai2 contains 128,892 clips and 2,826 hours across 113 countries or regions, each with camera trajectories and hierarchical annotations.

A Detailed Dataset Statistics

A.1 Release and Temporal Statistics

The newly collected perspective-video pipeline begins with 173,303 segmented candidates. After motion, media-quality, OCR, pose, and caption filtering, 74,603 dynamic clips remain. A separately curated low-motion branch contributes another 5,608 clips, resulting in 80,211 newly collected perspective clips. Together with 47,699 re-curated Sekai clips and 982 panoramic sequences, the final processed corpus contains 128,892 video units totaling 2,826 hours.

For temporal analysis, all videos are decomposed into non-overlapping analysis segments capped at 120 seconds. This representation is used only for duration statistics and does not alter the released video units. Among them, 43,594 segments reach the full 120-second duration, contributing 1,453 hours and 51.4% of the corpus. Segments of at least 60 seconds account for 92.3% of the total footage. These statistics confirm that the temporal scale of Sekai2 is dominated by sustained observations rather than short video fragments.

A.2 Geographic Coverage

As shown in Figure 8, Sekai2 spans 113 countries or regions, with substantial coverage across East Asia, Europe, North America, and a long tail of locations worldwide. Japan contributes the largest number of clips, followed by the United Kingdom and the United States, while Italy and South Korea are also among the major contributors. Beyond these high-frequency regions, the dataset includes geographically diverse samples across South America, Southeast Asia, Oceania, and parts of Africa and the Middle East. The resulting distribution is naturally long-tailed rather than geographically uniform. This reflects the availability of long-form exploration videos while preserving broad coverage across distinct urban, natural, transportation, and mixed-use environments. The dedicated panoramic subset further complements the geographically distributed perspective videos with revisit-rich trajectories collected under a controlled acquisition protocol.

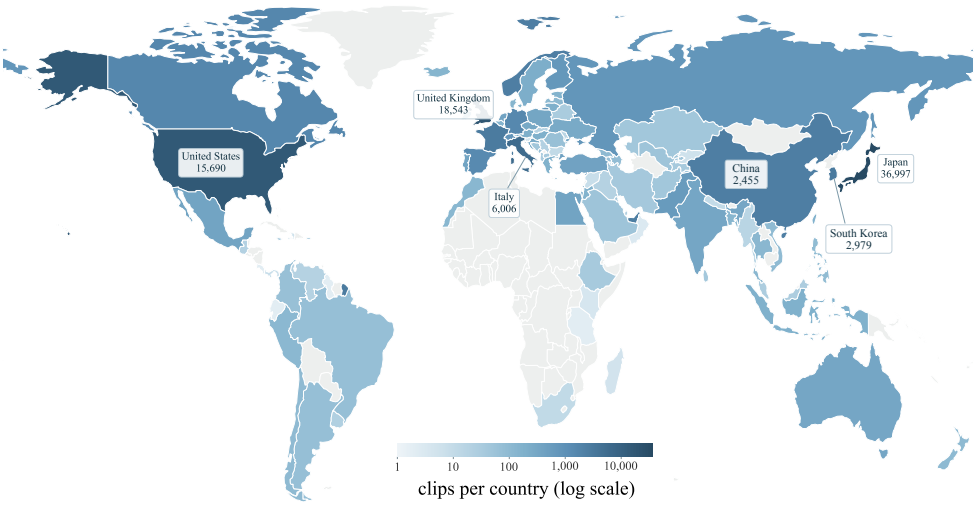


Figure 8: **Geographic coverage of Sekai2.** The map shows the number of clips per country or region on a $\log_{10}$ scale, with grey indicating no available clips. The five largest contributors are labelled, together with China, where the panoramic subset was captured.

A.3 Attribute Dependence

To complement the marginal diversity statistics in Sec. 4.2, we measure normalized mutual information (NMI) among the ten clip-level attributes. As shown in Figure 9, most attribute pairs exhibit weak dependence: across the 45 pairs, the mean NMI is 0.13 and the median is 0.05, with 69% below 0.1. This indicates that most scene, environmental, and acquisition attributes vary largely independently rather than collapsing into a small set of recurring combinations.

Stronger dependencies occur primarily among semantically related attributes. Camera motion is correlated with perspective (0.66), video style (0.55), and camera height (0.45), while lighting is associated with time of day (0.43) and weather (0.33). In contrast, dependencies across these viewpoint- and illumination-related groups remain consistently low. These results suggest that Sekai2 retains expected correlations within related attributes while preserving broad compositional diversity across scene, appearance, and acquisition conditions.

B Dataset Construction Details

This section provides the implementation details of the construction pipeline summarized in Sec. 3. We describe temporal segmentation, filtering, camera-pose reconstruction, panoramic trajectory refinement, hierarchical captioning, and release validation.

B.1 Temporal Segmentation and Filtering

Perspective videos are first checked for valid media metadata and segmented using OmniShotCut (Wang et al., 2026a). We remove a 6-second margin around detected shot boundaries, retain continuous intervals between 10 and 120 seconds, and preserve panoramic recordings as complete sequences so that loops and revisits remain intact.

We apply a multi-stage filtering pipeline covering motion, media integrity, persistent overlays, OCR-detected text, pose quality, and annotation validity. For motion filtering, at most 30 frames are uniformly sampled over the complete clip and processed with RAFT-Small (Teed & Deng, 2020). For adjacent sampled frames, we compute the mean optical-flow magnitude m_i and define

$$r_{\text{static}} = \frac{1}{T-1} \sum_{i=1}^{T-1} \mathbb{I}[m_i < 1.0]. \tag{1}$$

Clips with $r_{\text{static}} \geq 0.5$ fail the strict motion gate. Media checks further remove clips with abnormal resolution, frame rate, exposure, or black borders. Persistent HUD elements are detected from

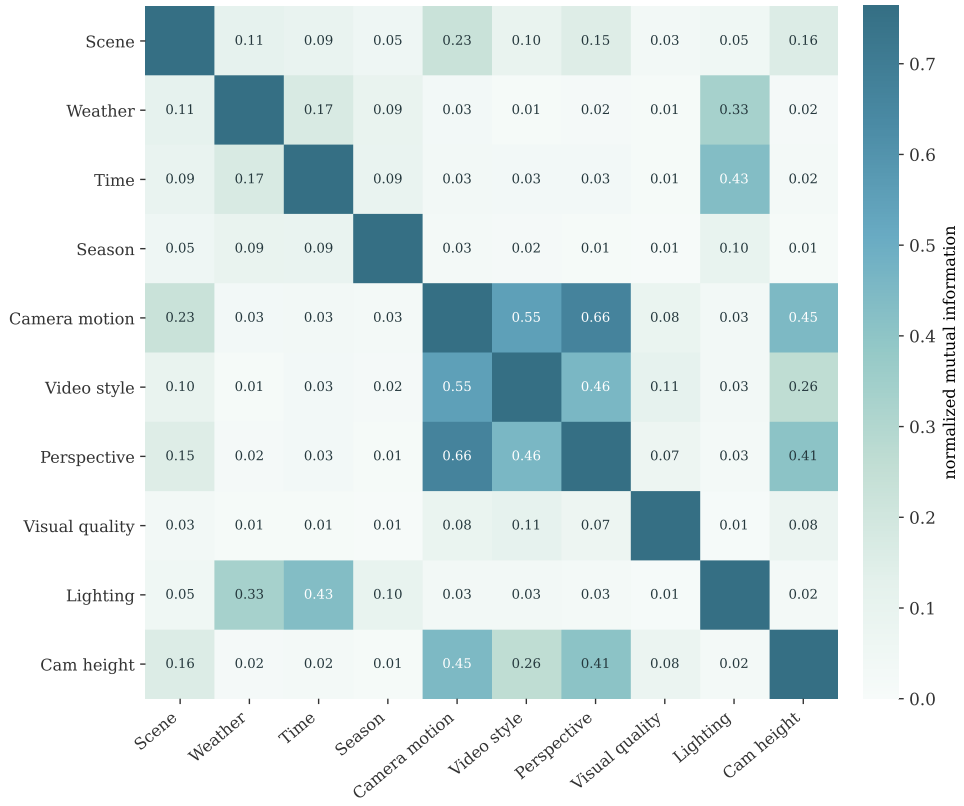


Figure 9: **Attribute dependence.** Normalized mutual information between the ten clip-level attributes. The diagonal denotes self-information and is saturated for visualization.

temporally stable high-gradient regions, while EasyOCR is used to identify subtitles, watermarks, and excessive on-screen text.

To retain meaningful scene dynamics under limited camera motion, clips whose only failed criterion is the strict motion gate are considered for a low-motion recovery branch. After passing the remaining quality checks, a duration-balanced subset is retained, contributing 5,608 clips to the final corpus. The complete release thresholds are summarized in Table 4.

B.2 Camera Pose Generation

Perspective clips are processed using the pose-only configuration of ViPE (Huang et al., 2025), which retains camera reconstruction while omitting dense outputs unnecessary for camera-conditioned modeling. GeoCalib (Veicht et al., 2024) provides initial camera intrinsics, reducing sensitivity to unknown focal length and field of view. To improve robustness in unconstrained videos, dynamic foreground regions are removed before geometric optimization. Grounding DINO (Liu et al., 2024) and SAM (Kirillov et al., 2023) detect and segment common moving objects, including people, animals, and vehicles, reducing violations of the static-scene assumption.

DROID-SLAM (Teed & Deng, 2021) performs sequential tracking, keyframe selection, local refinement, and global trajectory optimization. A monocular depth prior regularizes long-sequence geometry and improves reconstruction stability over extended traversal. The deployed configuration uses an eight-frame warm-up, a keyframe threshold of 4.0, a 25-frame frontend window, and 24 global backend iterations. Videos above 40 fps are processed with stride two and the remaining videos with stride one, yielding trajectories at approximately 30 fps. This configuration balances temporal resolution and computational cost while preserving dense trajectories for geometric supervision.

Each successful reconstruction stores frame-aligned camera-to-world transformations, source-frame indices, and estimated intrinsics. Because the trajectories are reconstructed monocularly, their translation coordinates represent relative geometry rather than metric scale.

B.3 Panoramic Captures and Trajectory Refinement

Data collection. We collect 1,283 panoramic videos using Insta360 cameras, totaling approximately 160 hours. The recordings span diverse indoor and outdoor environments and deliberately include non-linear traversal, sharp turns, elevation changes, loops, and revisits. Whenever feasible, the operator returns to a previously observed region after extended traversal, providing explicit long-range spatial constraints and richer revisit supervision.

Initial trajectory estimation. The equirectangular videos are resized to 960×480 and processed using the panorama mode of ViPE (Huang et al., 2025). The resulting trajectories generally preserve local camera motion and short-range geometric structure. However, accumulated drift over several-minute sequences can gradually introduce global inconsistency, particularly along long non-linear routes and when the camera revisits previously observed regions. Such drift may cause spatially identical locations to become misaligned in the reconstructed trajectory. We therefore apply an additional loop-closure refinement stage to improve long-range geometric consistency.

Loop-closure refinement. We retrieve candidate loop closures between temporally distant keyframes using ORB correspondences (Rublee et al., 2011) and verify them geometrically in spherical coordinates. Matched equirectangular features are converted to unit viewing rays $(\mathbf{r}_k, \mathbf{r}'_k)$, and the relative rotation is estimated by

$$\mathbf{R}^* = \arg\min_{\mathbf{R}\in \mathrm{SO}(3)} \sum_k \|\mathbf{r}'_k - \mathbf{R}\mathbf{r}_k\|_2^2 . \tag{2}$$

The rotational alignment admits a closed-form solution (Kabsch, 1976). A candidate is retained only when sufficient correspondences are geometrically consistent and the median angular residual remains low, suppressing visually similar but spatially distinct locations.

Verified loop closures are incorporated into pose-graph optimization (Kümmerle et al., 2011; Grisetti et al., 2010). Sequential ViPE motions provide local odometry constraints, while loop factors impose long-range constraints on revisited camera centers without requiring identical viewing directions. The optimized corrections are propagated through the sequence via interpolation on $\mathrm{SE}(3)$ to preserve dense frame-level trajectories. This procedure refines 1,051 of 1,283 trajectories, corresponding to 81.9%; sequences without reliable loop closures retain the original ViPE estimates.

B.4 Hierarchical Captioning

Annotations are generated with Kimi-K2.6 (Team et al., 2026c) using a unified inference request that jointly produces clip-level attributes and temporally grounded segments. Videos are resized to 360p and sampled at approximately 1 fps, with at most 64 timestamped frames distributed across the complete clip.

At the clip level, the model predicts structured scene and camera attributes together with six text fields: `subject_motion`, `environment_motion`, `static_scene`, `camera_description`, `full_prompt`, and `short_prompt`. The first four factorize subject dynamics, environmental dynamics, persistent scene content, and camera behavior. The `full_prompt` integrates these components into a detailed camera-aware description, whereas the `short_prompt` provides a compact camera-agnostic description.

Temporal boundaries are introduced when salient scene content, illumination, surface type, environmental state, or dominant camera behavior changes. Each segment contains timestamps, the same four factorized descriptions, a discrete `camera_path`, and full/short prompts. The 16 camera-path labels cover translation, rotation, coupled translation–rotation, stationary behavior, and mixed trajectories.

Model outputs are serialized as structured JSON and validated for schema completeness, continuous temporal coverage, legal camera-path labels, balanced camera tags, and separation between camera-aware and camera-free descriptions. The summarized production prompt is shown in Figure 10; the complete prompt and annotation code will be released for reproducibility.

```
ROLE
You are annotating a video clip for a world-model training dataset.
Watch the complete video and return:
(1) clip-level attributes and descriptions;
(2) temporally grounded scene segments.

INPUT
Video duration: [DURATION] seconds.
Sampled frames: [TIMESTAMPED_FRAMES].

CLIP-LEVEL ANNOTATION
Return controlled attributes for scene, environment, and camera properties.
Also return:
subject_motion,
environment_motion,
static_scene,
camera_description,
full_prompt,
short_prompt.

TEMPORAL SEGMENTS
Create a boundary when salient scene content, illumination, surface,
environmental state, or dominant camera behavior changes.
Segments must tile the complete video without gaps.

For each segment return:
segment_index,
time_range_s,
subject_motion,
environment_motion,
static_scene,
camera_description,
camera_path,
full_prompt,
short_prompt.

SEMANTIC FACTORIZATION
subject_motion: salient foreground actions and state changes.
environment_motion: other moving entities and dynamic conditions.
static_scene: persistent layout, objects, materials, and appearance.
camera_description: viewpoint, framing, path, and stability.

camera_path in:
{straight, reverse, strafe_left, strafe_right,
 ascend, descend, turn_left, turn_right,
 orbit_left, orbit_right, pan_left, pan_right,
 tilt_up, tilt_down, stationary, mixed}

TRAINING TEXT
full_prompt: coherent camera-aware description.
short_prompt: compact scene-centric description without camera language.

SELF-CHECK
Use legal enums, distinguish subject from camera motion,
cover the complete video, and return valid JSON.
```

Figure 10: **Hierarchical annotation prompt.** The summarized prompt jointly produces clip-level attributes and temporally grounded segments while factorizing subject motion, environmental dynamics, static scene content, and camera behavior.

B.5 Quality Assurance and Release Configuration

Camera trajectories undergo structural and temporal consistency checks before release. Each pose file must contain a finite, non-degenerate trajectory with valid frame-wise transformations. For pose $(\mathbf{R}_t, \mathbf{t}_t)$, we define a forward reference point to characterize its local viewing direction:

$$\mathbf{p}_t = \mathbf{R}_t[0, 0, 1]^\mathsf{T} + \mathbf{t}_t, \tag{3}$$

and measure its deviation from a temporally median-filtered trajectory. Peak residual, local jitter, median residual, and long-range drift are normalized into a pose-quality score,

$$q_{\text{pose}} = 0.3q_{\text{peak}} + 0.3q_{\text{jitter}} + 0.2q_{\text{median}} + 0.2q_{\text{drift}}. \tag{4}$$

Table 4: **Release configuration for perspective-video curation.** "Aux." denotes a recorded descriptor that is not itself a hard gate.

Component	Configuration	Decision
Segmentation	OmniShotCut; 6 s boundary buffer	retain 10–120 s intervals
Motion	RAFT-Small; 30 uniform frames; long side 256	reject $r_{\text{static}} \geq 0.5$
Format	height ≥ 720; fps 24–65; H.264/HEVC/AV1	hard gate
Exposure	luminance ≤ 16 or ≥ 240	reject run > 15 samples
Black border	low-mean, low-variance rows/columns	reject ratio > 0.10
HUD	12 frames at 640×360	reject stable-edge area > 0.04
OCR	8 frames; EasyOCR	subtitle $\geq$ 2 / watermark $\geq$ 3 frames
Flow descriptors	direction variance, saliency, temporal variance	Aux.
Static recovery	duration budget; eight motion bins	target 6% of final hours
Pose score	peak/jitter/median/drift residuals	reject $q_{\text{pose}} \geq 0.5$
Caption	Kimi-K2.6; 360p; up to 64 frames	schema + valid segments

Clips with $q_{\text{pose}} \geq 0.5$ are rejected, together with trajectories that trigger hard failure conditions for extreme residuals, jitter, or insufficient path length.

Caption outputs undergo both online validation and a corpus-level structural audit. We verify required fields, continuous and increasing temporal ranges, legal camera-path values, non-empty factorized descriptions, balanced camera tags, and the absence of camera leakage into `short_prompt`. Operational errors and validation flags remain recorded in the annotation metadata.

As an additional cross-modal diagnostic, coarse camera motion derived independently from the pose trajectories is compared with the VLM-produced `camera_path` labels. This diagnostic is used to identify systematic caption or trajectory failures rather than as external geometric ground truth.

B.6 Data Organization and Reproducibility

The release is organized around a canonical manifest with one row per final video unit. Each row stores source identifiers, precise temporal boundaries, split information, and references to the corresponding pose and caption artifacts. Dataset statistics are generated directly from this manifest to avoid including stale or intermediate processing outputs.

Pose files contain frame-aligned $N \times 4 \times 4$ camera-to-world transformations and source-frame indices. Caption files contain clip-level attributes, temporally grounded segments, factorized descriptions, training prompts, and annotation metadata. Processing stages record their configurations, versions, per-clip quality scores, and decision reasons, enabling reconstruction of the final manifest from the released source references and configurations.

The annotation corpus was generated using distributed VLM inference. All processing stages support deterministic sharding and resumable execution, while cached intermediate scores allow filtering thresholds to be re-applied without recomputing expensive visual features.

B.7 Known Failure Modes

The construction pipeline reduces common artifacts but does not eliminate all residual noise. Automatically generated annotations may miss fine temporal events or contain uncertain semantic attributes, while monocular camera reconstruction remains sensitive to low texture, reflections, dynamic foregrounds, rapid rotation, and long-range drift. The trajectories also retain global scale ambiguity. These limitations motivate the independent semantic and geometric evaluations in Sec. 5.3 and the additional analyses below.

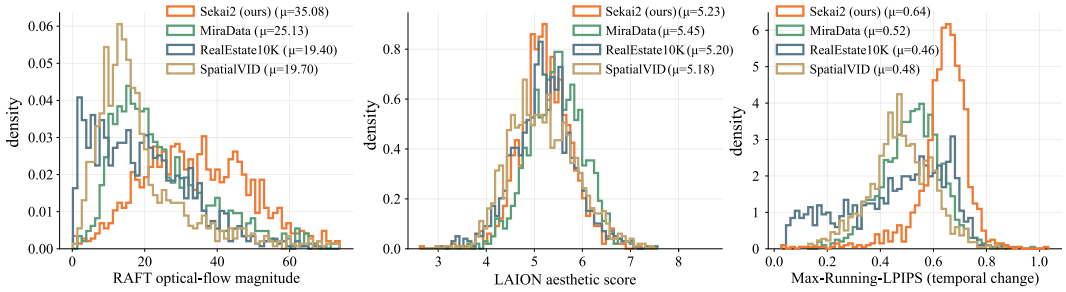


Figure 11: Intrinsic visual quality and temporal dynamics. Distributions of RAFT optical-flow magnitude, LAION aesthetic score, and Max-Running-LPIPS for Sekai2 and three video datasets.

C Additional Dataset Quality Evaluations

This section provides additional analyses supporting the compact dataset-quality evaluation in Sec. 5.3. We report distribution-level visual statistics, corpus-wide and cross-run trajectory diagnostics, extended caption-quality assessment, and segment-level pose–caption grounding.

C.1 Additional Visual-Quality Analysis

We extend the visual-quality evaluation in Sec. 5.3 with distribution-level comparisons. Following the same protocol, we sample 1,200 perspective clips from Sekai2, MiraData (Ju et al., 2024), RealEstate10K (Zhou et al., 2018), and SpatialVID (Wang et al., 2026b). All subsets are evaluated with identical models, preprocessing, resolution, and sample size, while motion statistics are computed over adjacent frame pairs to avoid duration-dependent bias.

As shown in Figure 11, Sekai2 maintains comparable aesthetic quality while exhibiting substantially stronger local temporal variation. Its mean RAFT optical-flow magnitude is 35.08, compared with 25.13, 19.40, and 19.70 for MiraData, RealEstate10K, and SpatialVID, respectively; Max-Running-LPIPS is 0.64, compared with 0.52, 0.46, and 0.48. The separation is distribution-wide rather than driven by a small highly dynamic subset: 81% of Sekai2 clips exceed the median optical-flow magnitude of MiraData, and 91% exceed its median Max-Running-LPIPS. Meanwhile, the mean LAION aesthetic score remains competitive at 5.23, compared with 5.45, 5.20, and 5.18 for the three baselines. These results support the main-paper observation that richer local dynamics are obtained without a substantial loss in frame-level visual quality.

The comparison is restricted to perspective videos because equirectangular panoramic frames are not directly comparable under the adopted visual metrics.

C.2 Additional Camera-Trajectory Analysis

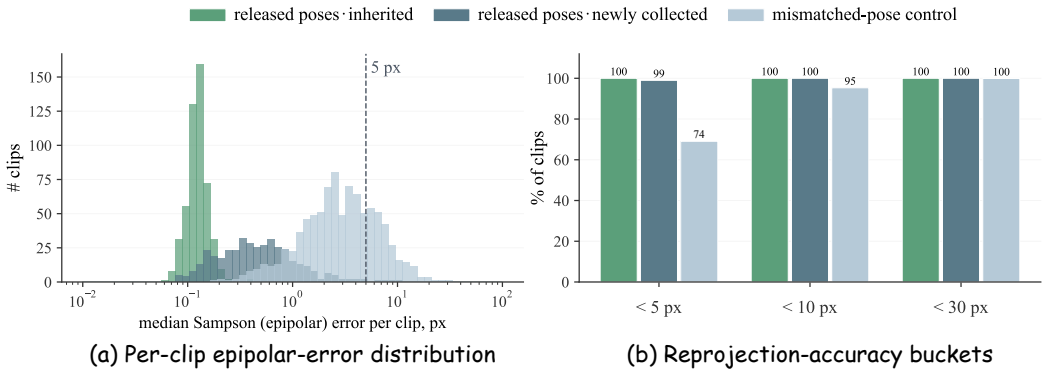
Corpus-wide validity and smoothness. We first audit all released trajectories using the valid-pose ratio, the rate of robust inter-frame translation outliers, and the median second difference of inter-frame rotation. As shown in Table 5, all three data sources achieve a median valid-pose ratio of 1.0, together with near-zero translation-jump rates and low rotational jerk. This corpus-wide analysis complements the sampled geometric evaluation in the main paper by verifying numerical completeness and local smoothness over the full dataset.

C.2.1 Ground-Truth-Free Epipolar Validation

Evaluation protocol. For each sampled perspective clip, we extract SIFT correspondences between frame pairs separated by 0.5 s and construct the fundamental matrix induced by the released camera poses and intrinsics. We then compute the Sampson error for each correspondence and discard the largest 20% of errors within each frame pair, reducing sensitivity to feature mismatches, dynamic foreground objects, and other local violations of rigid-scene geometry.

Table 5: **Validity and smoothness of all released trajectories.** Values are medians over clips. A jump denotes an inter-frame translation step exceeding the clip-specific median by five robust standard deviations. Lower is better except for validity.

Source	Clips	Valid ↑	Jump rate ↓	Rot. jerk ↓
Re-curated Sekai	47,699	1.000	0.00000	0.00318
Newly collected	80,211	1.000	0.00028	0.01052
Panoramic	982	1.000	0.00010	0.00427
All	**128,892**	**1.000**	**0.00000**	**0.00725**



(a) Per-clip epipolar-error distribution (b) Reprojection-accuracy buckets

Figure 12: **Ground-truth-free epipolar validation.** (a) Per-clip median Sampson-error distributions for the inherited and newly collected subsets, together with a mismatched-pose negative control that shifts the pose pair along the same trajectory while preserving all image-side inputs. (b) Fraction of clips below the 5, 10, and 30 px thresholds.

The re-curated Sekai subset is evaluated using its released intrinsics. For the newly collected subset, we adopt a fixed prior focal length as the primary protocol rather than directly reusing the same estimated calibration employed during pose reconstruction.

Mismatched-pose control. To test whether low errors depend on correct pose–frame correspondence, we replace each pose pair with another pair shifted by one third of the same trajectory. For an image pair indexed by (a, b), the control uses

$$a' = (a + \lfloor N/3 \rfloor) \bmod (N-1), \qquad b' = \min\left(a' + (b-a), N-1\right). \tag{5}$$

The image pair, SIFT correspondences, temporal separation, intrinsic protocol, and robust trimming remain unchanged; only the pose–frame association is disrupted.

As shown in Fig. 12(a), the re-curated Sekai subset achieves a median per-clip Sampson error of 0.12 px over 500 clips. Under the fixed-focal protocol, 399 newly collected clips achieve a median error of 0.43 px. Figure 12(b) further shows that 99.0% of these clips remain below the 5 px threshold. In contrast, the mismatched-pose control shifts the error distribution substantially, increasing the median error to 2.83 px and reducing the fraction below 5 px to 74.0%. Because all image-side inputs remain unchanged, this degradation directly demonstrates sensitivity to correct pose–frame alignment.

Allowing the focal length to vary over a predefined per-clip sweep reduces the median error of the newly collected subset to 0.14 px. We treat this result only as a sensitivity analysis and retain the fixed-prior result in the main paper as the more conservative protocol. The epipolar evaluation measures local relative geometry and synchronization; it does not constrain monocular translation scale or directly measure long-range accumulated drift.

C.2.2 Cross-Run Geometric Consistency

As a complementary diagnostic, we compare the released ViPE trajectories with separate DROID-SLAM reconstructions (Teed & Deng, 2021) under matched GeoCalib intrinsics and identical

Table 6: **Cross-run geometric consistency under matched intrinsics.** Results are source–motion-weighted per-clip medians. Rotation RPE and translation-direction errors are reported in degrees, while ATE is normalized by trajectory path length. Lower is better except for reconstruction success.

Source	n	Success ↑	Rot. 1s ↓	Rot. 5s ↓	T-dir. 1s ↓	T-dir. 5s ↓	ATE_{norm} ↓
Re-curated Sekai	395	99.9%	0.072°	0.281°	1.09°	0.86°	0.0100
Newly collected, main	480	100.0%	0.134°	0.541°	2.57°	2.19°	0.0281
Newly collected, static	125	89.3%	0.088°	0.313°	13.14°	7.26°	0.0257
All (weighted)	**1,000**	**99.5%**	**0.103°**	**0.414°**	**1.88°**	**1.48°**	**0.0176**

reconstruction settings. We attempt reconstruction on 1,040 perspective clips stratified by source and motion regime and obtain 1,000 successfully paired trajectories. Panoramic sequences are excluded because the reconstruction configuration assumes pinhole imagery.

Because monocular trajectories are defined only up to a global similarity transform, each DROID trajectory is aligned to its ViPE counterpart by solving

$$\min_{s,\mathbf{Q},\mathbf{b}} \sum_{i=1}^{N} \left\| s\mathbf{Q}\mathbf{p}_i^{\mathrm{D}} + \mathbf{b} - \mathbf{p}_i^{\mathrm{V}} \right\|_2^2, \tag{6}$$

where s is a global scale factor, $\mathbf{Q} \in \mathrm{SO}(3)$ a global rotation, and $\mathbf{b} \in \mathbb{R}^3$ a translation. We evaluate rotation relative pose error (RPE), translation-direction error, and ATE normalized by ViPE path length:

$$\text{ATE}_{\text{norm}} = \frac{\text{ATE}\left(\mathbf{p}^{\mathrm{D}}, \mathbf{p}^{\mathrm{V}}\right)}{\sum_{i=2}^{N} \left\| \mathbf{p}_i^{\mathrm{V}} - \mathbf{p}_{i-1}^{\mathrm{V}} \right\|_2}. \tag{7}$$

As summarized in Table 6, the reconstructed trajectories exhibit strong cross-run agreement across data sources, with a weighted success rate of 99.5%, rotation RPE of 0.103° at one second and 0.414° at five seconds, and normalized ATE of 0.0176. Translation-direction error is substantially larger for the static supplement because directional estimates become poorly conditioned as displacement approaches zero. Straight trajectories are generally more stable than curved or winding trajectories, which tend to accumulate greater geometric and calibration uncertainty over longer temporal spans.

Because ViPE itself contains a DROID-based geometric component, this experiment should be interpreted as cross-run geometric consistency rather than accuracy against independent metric ground truth. Together with the image-based epipolar analysis in Fig. 12, it provides complementary evidence for the geometric reliability of the released trajectories.

C.3 Additional Semantic-Annotation Analysis

Judge-based assessment. We report the complete frame-aware VLM evaluation summarized in Sec. 5.3. For each of 1,500 sampled clips, the judge receives 12 uniformly sampled frames together with the corresponding hierarchical annotation and scores four aspects on a 1–5 scale. As shown in Table 7, all four dimensions exceed 4.0, with scores of 4.28 for coverage, 4.31 for scene correctness, 4.14 for motion correctness, and 4.01 for camera correctness. The slightly lower motion and camera scores reflect the difficulty of assessing temporally extended dynamics from sparse frame samples.

Pairwise comparison. We further compare our structured annotations with conventional one-shot captions on 800 clips. For each clip, both captions are generated from the same visual input using the same VLM, isolating the annotation protocol from differences in model capacity or visual evidence. The judge then selects the preferred and more informative caption.

We evaluate both presentation orders and observe a pronounced positional bias: the first candidate is selected in 71.6% of comparisons, with no ties. To avoid benefiting from this bias, Table 7 reports the conservative setting in which our annotation is always presented second. Even under this unfavorable ordering, **our annotations are preferred in 60.9% of cases and judged more informative in 68.4%**. These results indicate that hierarchical annotation provides richer and more useful descriptions than conventional one-shot captioning under controlled conditions.

Table 7: **Judge-based annotation assessment.** Absolute VLM scores are averaged over 1,500 clips on a 1–5 scale. Pairwise results compare our structured annotation with one-shot captions generated by the same VLM from identical visual inputs over 800 clips. To account for the judge's first-position bias, we report the conservative ordering with our annotation shown second.

Protocol	Metric	Value
VLM Judge	Coverage	4.28
	Scene Correctness	4.31
	Motion Correctness	4.14
	Camera Correctness	4.01
Pairwise	Ours Preferred (shown second)	60.9%
	Ours More Informative (shown second)	68.4%
	First-position Preference	71.6%

Table 8: **Clip-level annotation statistics.** Mean length denotes words per released clip.

Annotation field	Mean words	Total words
`subject_motion`	19	2,450,161
`environment_motion`	26	3,378,305
`static_scene`	32	4,126,762
`camera_description`	27	3,456,500
`short_prompt`	18	2,387,222
`full_prompt`	97	12,622,799

C.4 Additional Caption Statistics

Annotation density. The annotation corpus contains approximately 135 million words and 649,597 temporally grounded segments, corresponding to 5.04 segments per clip on average. The median segment duration is 15 seconds. Table 8 summarizes the clip-level annotation fields.

The four factorized descriptions remain relatively concise, whereas `full_prompt` integrates them into a richer camera-aware description. The `short_prompt` provides a compact camera-agnostic alternative for settings in which camera information is controlled separately.

Description diversity. The annotations exhibit little exact repetition. Among approximately 650K segment-level short prompts, 99.86% are unique, while the exact duplicate rate across the factorized description fields is 0.02%. This suggests that the annotation corpus captures substantial linguistic variation rather than repeatedly applying a small set of templates.

Segment-level lexical complementarity. Figure 13 extends the clip-level lexical analysis in Figure 5(b) to temporally grounded segments. The same qualitative structure is preserved at this finer temporal granularity: the four factorized descriptions remain lexically distinct, whereas the full prompt exhibits broader overlap because it integrates complementary information from the individual fields. Compared with clip-level annotations, segment-level descriptions share somewhat more spatial, motion-related, and connective vocabulary. This is expected because temporally adjacent segments often describe successive states within the same local scene while differing primarily in camera behavior, subject motion, or short-term environmental changes.

C.5 Pose–Caption Temporal Grounding

Camera trajectories and semantic annotations are produced independently, enabling end-to-end cross-modal consistency analysis. For 620,487 segments, covering 95.5% of all annotations, we align each semantic interval with its camera trajectory and measure pose-derived yaw change.

As shown in Figure 14, median yaw change increases from 1.3° for `stationary` segments to 8.5° for `straight`, 20.6° for `mixed`, and 27.9° for `turn`. Along the same ordering, the fraction of segments with less than 5° rotation decreases from 72% to 14%, whereas the fraction above 45° increases from 4% to 33%. Using yaw change to distinguish turning from non-turning segments yields an AUC of 0.709.

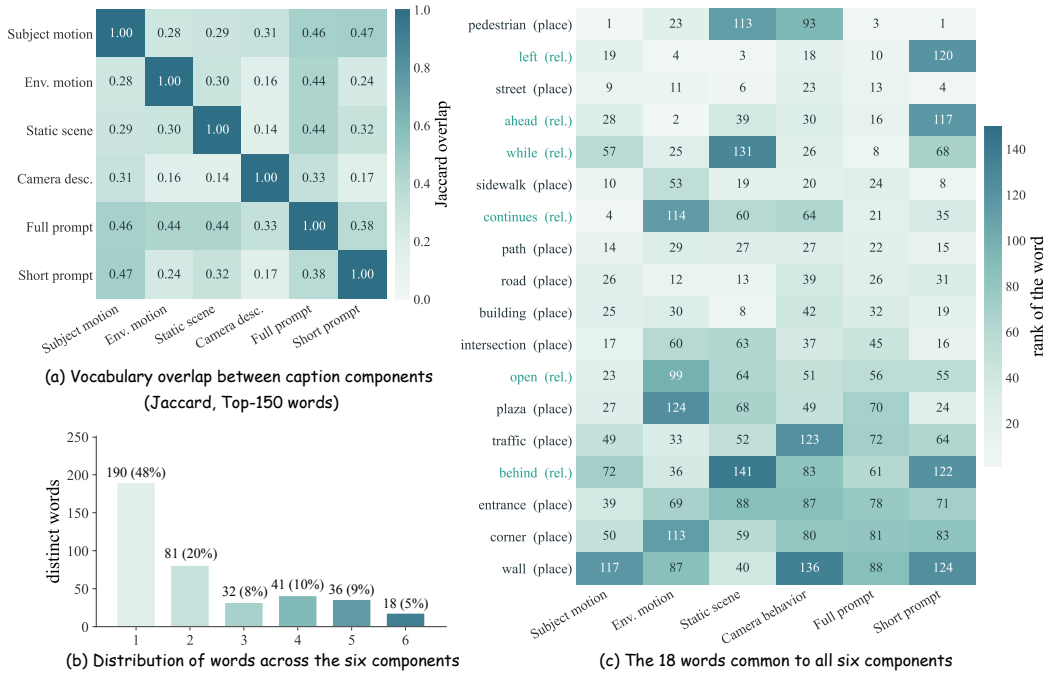


Figure 13: **Segment-level lexical complementarity.** (a) Pairwise Jaccard overlap among the top-150 content words of the six annotation components. (b) Number of components in which each word appears. (c) Shared words and their within-component ranks.

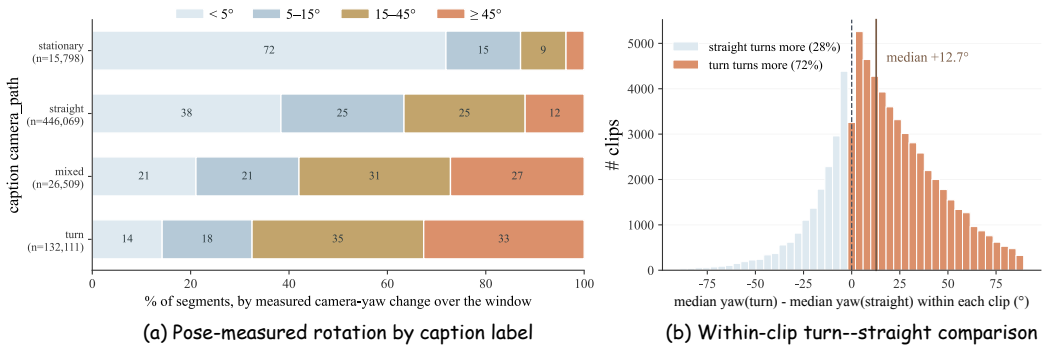


Figure 14: **Segment-level pose–caption grounding.** (a) Distribution of pose-derived yaw change across semantic `camera_path` categories. (b) Within-clip difference between median yaw change of `turn` and `straight` segments.

To control for clip-level differences in motion scale and capture style, we further compare `turn` and `straight` segments within the same video. Among 74,437 clips containing both labels, `turn` segments exhibit larger median yaw change in 71.8% of clips, with a median within-clip difference of 12.7°. The remaining overlap is expected because `camera_path` represents coarse semantic camera behavior rather than a discretization of geometric rotation magnitude.

D Generation Evaluation Details

This section provides additional evaluation protocols, metric definitions, and complete per-dimension results for the world-model experiments in Sec. 5, together with further details on the panoramic and long-horizon evaluation settings.

Table 9: **WorldScore evaluation on long-horizon generation.** We report the static branch at 40- and 60-second horizons. Dynamic-branch metrics are omitted because they were not reliably produced in our evaluation. All values are higher-is-better.

Dur.	Method	Camera	Object	Content	3D	Photo.	Style	Subj.	Static
40s	LingBot-World 1.5	42.33	45.12	68.51	66.07	34.42	82.17	35.45	53.44
	HY-World 1.5	43.41	41.18	71.05	77.76	81.79	84.98	28.49	58.85
	HYWorld-FT	44.10	42.45	75.15	83.32	89.83	83.02	28.15	59.64
	AlayaWorld	41.32	43.62	75.85	91.56	94.45	85.50	19.17	64.50
	AlayaWorld-FT	42.20	43.97	77.24	93.04	95.36	86.23	21.30	66.11
60s	LingBot-World 1.5	43.67	43.99	68.61	67.38	32.32	82.10	35.10	53.31
	HY-World 1.5	41.63	39.79	70.97	75.12	83.09	81.71	29.14	60.24
	HYWorld-FT	42.35	40.20	74.55	79.65	84.94	82.08	29.90	62.15
	AlayaWorld	39.59	42.48	77.75	91.78	94.67	85.47	19.33	64.44
	AlayaWorld-FT	40.06	43.21	79.52	93.30	94.95	86.09	20.57	65.98

D.1 Evaluation Protocol

Panoramic world modeling. We evaluate panoramic generation on 50 panoramic test sequences. Each model receives the same initial frame, camera trajectory, and semantic caption and generates an 81-frame equirectangular video at 960×480 resolution. **PanoWorld-FT** is compared with PanoWorld (Li et al., 2026a), HY-World 1.5 (HunyuanWorld, 2025), LingBot-World 1.5 (Gao et al., 2026a), and AlayaWorld (Team et al., 2026a) under matched conditioning.

Long-horizon world modeling. We evaluate 100 perspective test cases at both 40- and 60-second horizons to assess model behavior under extended generation. HY-World 1.5 and AlayaWorld are separately fine-tuned on the newly collected perspective subset of Sekai2 for each duration, yielding 40- and 60-second variants of **HYWorld-FT** and **AlayaWorld-FT**. All variants are evaluated under the same test cases and conditioning protocol. Each fine-tuned model is compared with its corresponding base model, while LingBot-World 1.5 serves as an additional reference.

Reporting protocol. For panoramic world modeling, PanoWorld is fine-tuned on approximately 120 hours of panoramic videos from Sekai2. For long-horizon world modeling, HY-World 1.5 and AlayaWorld are fine-tuned on approximately 800 hours of newly collected Sekai2 videos, with separate variants trained for the 40- and 60-second settings. All long-horizon methods are evaluated on the same 100-case test set using metrics from **WorldScore**, **VBench**, and **iWorld-Bench**, ensuring a consistent comparison across models and generation horizons. The main paper reports representative dimensions for compact comparison in Sec 5.2.

D.2 Panoramic Generation Metrics

We evaluate panoramic generation from complementary perspectives of reconstruction fidelity, distributional similarity, and perceptual quality. **PSNR** measures pixel-level reconstruction fidelity and is reported over four temporal windows—frames 20–25, 50–55, 70–75, and 75–80—to characterize how fidelity evolves over the generation horizon. **FID** measures distributional similarity between generated and reference frames and is computed both globally and separately over polar and equatorial regions, where equirectangular distortion differs substantially. **FAED** provides an additional feature-space measure of distributional discrepancy. **NIQE** evaluates no-reference perceptual image quality, while **Q-Align** reports learned Quality and **Aesthetic** scores aligned with human perceptual judgments. Together, these metrics characterize temporal fidelity, spatial distributional consistency, and perceptual quality of panoramic generation.

D.3 Long-Horizon Evaluation Metrics

WorldScore. We use the static branch of WorldScore (Duan et al., 2025) to evaluate controllability and long-range scene consistency. **Camera** measures whether the generated viewpoint evolution follows the prescribed camera motion, while **Object** and **Content** assess object-level controllability

Table 10: **VBench evaluation on long-horizon generation.** All values are higher-is-better.

Dur.	Method	Subject Cons.	BG Cons.	Motion Smth.	Aesthetic	Imaging	Overall Cons.
40s	LingBot-World 1.5	77.18	87.31	97.17	52.08	65.18	19.73
	HY-World 1.5	70.91	82.98	96.17	43.00	59.99	17.54
	HYWorld-FT	72.14	85.45	97.36	46.63	62.10	19.37
	AlayaWorld	86.96	92.78	98.85	49.47	65.71	20.64
	AlayaWorld-FT	88.10	93.55	99.07	52.32	67.54	23.01
60s	LingBot-World 1.5	76.03	86.26	97.30	51.49	63.12	19.61
	HY-World 1.5	65.41	80.37	94.51	40.19	57.40	15.89
	HYWorld-FT	67.80	82.28	94.88	42.56	59.22	16.29
	AlayaWorld	83.58	91.24	98.84	48.07	64.55	20.75
	AlayaWorld-FT	85.03	92.37	98.72	49.80	66.35	22.34

and semantic agreement with the conditioning input. The remaining dimensions characterize visual and structural consistency over time: **3D** measures cross-view geometric coherence, **Photometric** evaluates appearance consistency across frames, **Style** captures visual-style stability, and **Subjective** reflects overall perceptual quality. **WorldScore-Static** aggregates these static-world dimensions into an overall score. The dynamic branch is omitted because its motion-related metrics were not reliably produced in our evaluation.

VBench. We report six dimensions from VBench (Huang et al., 2024) to characterize complementary aspects of temporal and perceptual video quality. **Subject Consistency** measures whether foreground subjects preserve their visual identity over time, while **Background Consistency** evaluates the stability of the surrounding scene. **Motion Smoothness** captures temporal continuity and penalizes abrupt or irregular motion. **Aesthetic Quality** measures the visual appeal of individual frames, whereas **Imaging Quality** focuses on low-level perceptual fidelity and image degradation. Finally, **Overall Consistency** evaluates global agreement between the generated sequence and its conditioning information. Together, these dimensions provide a complementary view of appearance stability, temporal coherence, and visual quality over extended generation.

iWorld-Bench. We adopt the trajectory-following metrics from iWorld-Bench (Fang et al., 2026) to evaluate camera controllability on our long-horizon benchmark. **Trajectory Accuracy** measures how faithfully generated camera motion follows the prescribed camera-command trajectory. For each video, ViPE recovers the camera trajectory, which is aligned to the reference coordinate system; motion tangents are then compared with those of the command trajectory using a calibrated cosine-similarity score averaged over time. **Trajectory Tolerance** follows the same reconstruction and alignment procedure but compares the recovered trajectory with the corresponding reference trajectory, providing a complementary measure of trajectory fidelity. We additionally report **Brightness**, **Clarity**, and **Hue** for luminance stability, image sharpness, and color consistency. All iWorld-Bench scores range from 0 to 100, with higher values indicating better performance.

D.4 Full Long-Horizon Results

Tables 9–11 report the complete long-horizon evaluation corresponding to the compact comparison in the main paper. Across both 40- and 60-second settings, **fine-tuning on Sekai2 yields consistent improvements over the corresponding base models across multiple evaluation dimensions**. For HY-World 1.5, the gains are particularly evident in WorldScore 3D consistency and photometric quality, while VBench also shows improvements in subject consistency, background consistency, imaging quality, and overall consistency. The same trend is observed for AlayaWorld, whose fine-tuned variants improve across most WorldScore and VBench dimensions at both horizons.

The iWorld-Bench results show a similar pattern. **HYWorld-FT** improves over HY-World 1.5 in trajectory accuracy, trajectory tolerance, and the auxiliary visual-quality metrics at both durations, while **AlayaWorld-FT** also improves trajectory following and most appearance-related dimensions. Although the magnitude of improvement varies across metrics and base models, the overall trend remains stable as the generation horizon increases from 40 to 60 seconds. These full results support the main-paper observation that the benefit of Sekai2 is not confined to a small subset of reported

Table 11: **iWorld-Bench evaluation on long-horizon generation.** Trajectory metrics are computed over cases with valid ViPE-recovered trajectories. All metric scores are reported on a 0–100 scale.

Dur.	Method	Trajectory Following		Visual Quality		
		Traj. Acc. ↑	Traj. Tol. ↑	Bright. ↑	Clarity ↑	Hue ↑
40s	LingBot-World 1.5	83.46	64.31	48.20	50.40	31.00
	HY-World 1.5	78.78	59.95	38.40	47.60	13.30
	HYWorld-FT	80.04	60.72	46.50	52.90	24.60
	AlayaWorld	81.38	61.77	64.60	64.70	50.40
	AlayaWorld-FT	81.69	62.54	65.30	64.50	51.00
60s	LingBot-World 1.5	82.29	63.06	43.70	48.20	28.30
	HY-World 1.5	76.27	57.84	37.40	39.40	11.20
	HYWorld-FT	77.30	58.99	40.20	43.30	18.50
	AlayaWorld	80.48	61.87	57.20	59.30	42.30
	AlayaWorld-FT	81.06	62.54	58.10	60.50	43.60

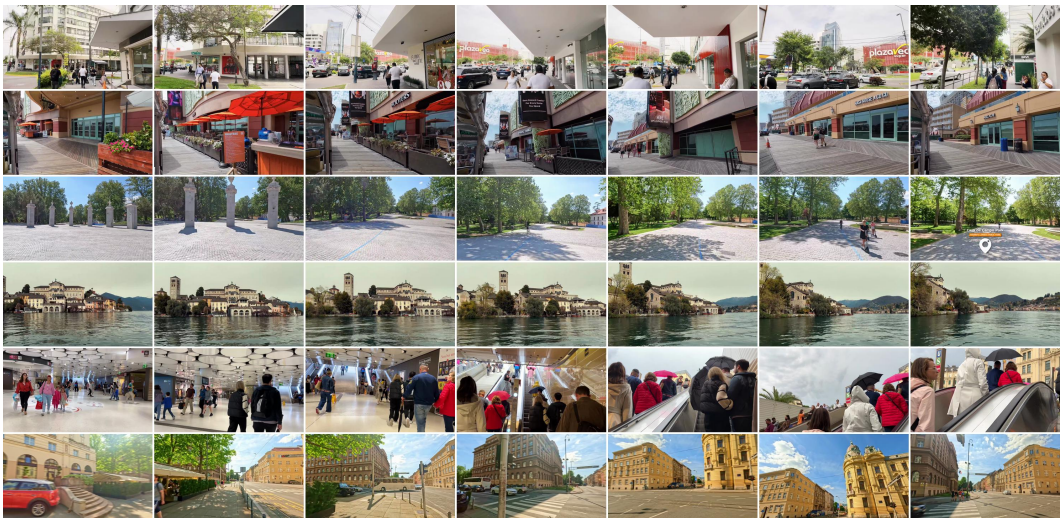


Figure 15: **Representative videos from the re-curated Sekai subset.** Each row shows one sequence at seven timestamps, covering diverse traversal modes and real-world environments.

metrics, but extends across complementary measures of geometric consistency, temporal quality, visual fidelity, and camera controllability.

E Qualitative Case Studies

This section provides qualitative examples of the released videos, hierarchical annotations, camera trajectories, and panoramic reconstruction results. The selected cases span diverse environments, motion patterns, and trajectory structures across the three data sources, and complement the quantitative evaluations by illustrating the temporal continuity, geometric characteristics, and annotation structure of Sekai2. Together, these examples provide a more intuitive view of the data modalities and their alignment over long video sequences.

E.1 Visual Cases Across Data Sources

Figures 15–17 show representative sequences from the re-curated Sekai, newly collected perspective, and panoramic subsets. The examples cover diverse environments and traversal patterns, including walking, driving, cycling, aerial motion, and non-linear panoramic trajectories. Frames sampled throughout each sequence illustrate sustained viewpoint and scene evolution rather than isolated keyframes, highlighting the long-form temporal continuity and complementary visual characteristics of the three data sources.

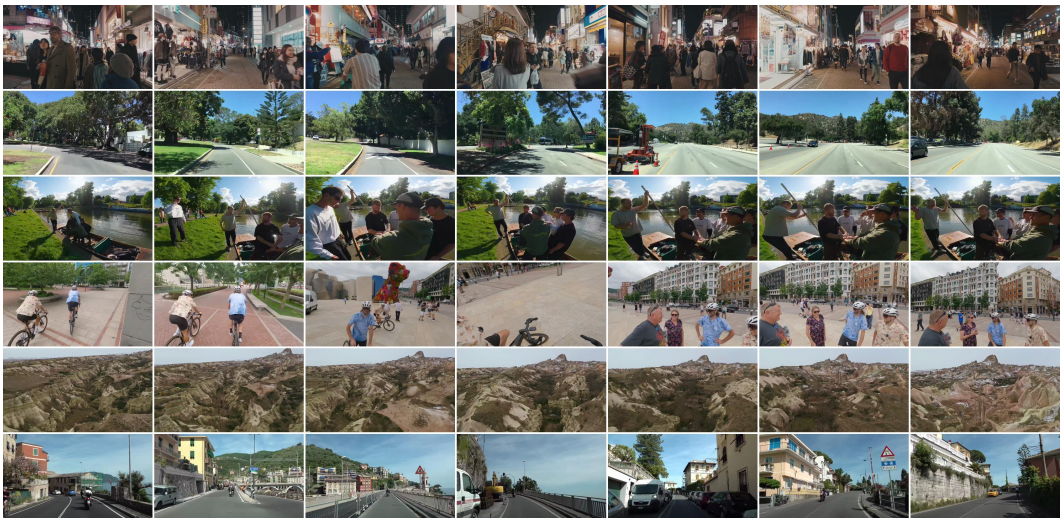


Figure 16: **Representative videos from the newly collected perspective subset.** The examples span walking, driving, cycling, boating, and aerial capture across diverse scenes.



Figure 17: **Representative panoramic videos.** Native 2:1 equirectangular observations preserve surrounding context during non-linear traversal, turns, stairs, and elevation changes.

E.2 Multimodal Long-Horizon Annotation

Figure 18 presents a 120-second first-person sequence covering the approach to, ascent of, observation from, and descent of a wooden tower. RGB observations, the camera trajectory, vertical profile, and temporally grounded annotations jointly characterize the evolving scene context and camera behavior throughout the sequence. The example highlights the limitation of a single clip-level description, which cannot precisely localize the distinct motion phases and corresponding scene transitions over an extended temporal horizon.

Figures 19–23 provide additional examples spanning direction reversals, vertical motion, loops, and orbital trajectories across different capture settings. Across these cases, the segment-level annotations remain aligned with changes in both camera behavior and scene state, while preserving the temporal ordering of successive motion phases. This finer-grained supervision provides more precise temporal grounding than a single global description, particularly for long sequences containing multiple camera transitions and scene changes.

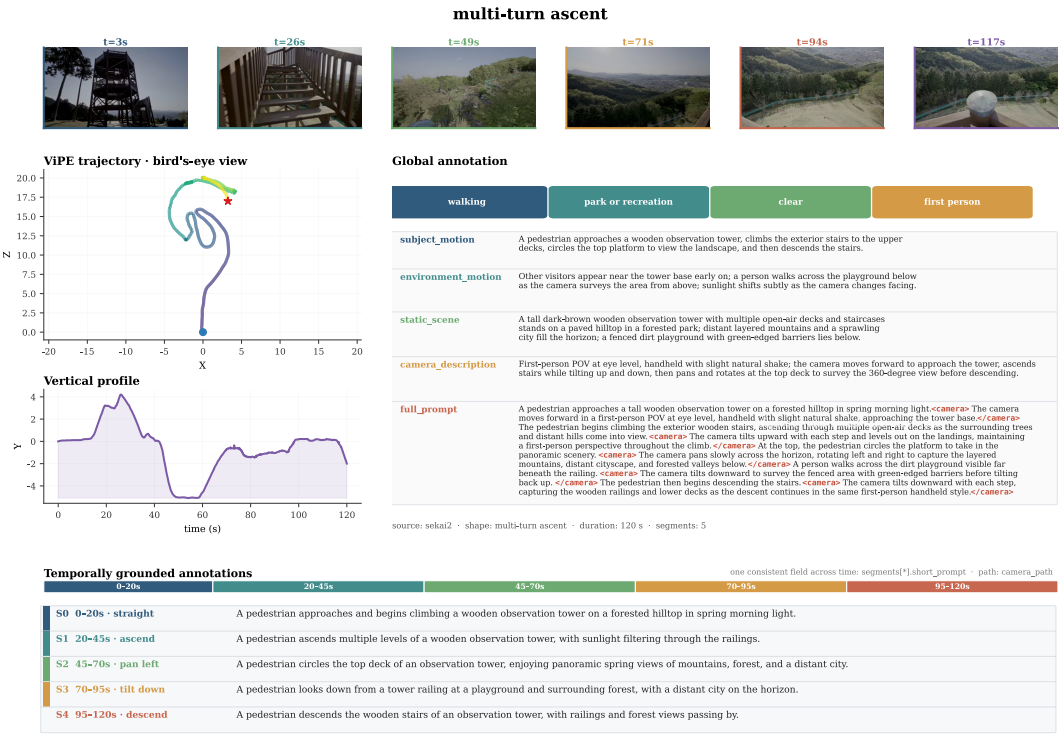


Figure 18: **Video–pose–caption case study.** A 120-second observation-tower sequence with RGB frames, ViPE trajectory, vertical profile, and temporally grounded annotations.



Figure 19: **U-turn case.** A 120s street-level traversal containing a mid-sequence trajectory reversal.



Figure 20: **L-turn case.** A re-curated Sekai walking sequence containing multiple direction changes.

E.3 Camera-Trajectory Cases

Figures 24 and 25 visualize representative ViPE trajectories from the re-curated Sekai and newly collected perspective subsets. The selected cases span diverse capture modes and trajectory structures, including straight traversal, gradual bends, sharp turns, returns, and winding paths. Despite this variation, the trajectories remain temporally smooth and preserve coherent global motion patterns over extended sequences. These examples complement the quantitative trajectory-quality evaluation by illustrating both the geometric diversity and continuity of the released camera trajectories.

E.4 Panoramic Full-Accumulation Reconstruction

Figures 26 and 27 visualize full-sequence accumulation using the loop-refined panoramic trajectories. Across non-linear paths, stairs, sharp turns, obstacle avoidance, and revisits, repeated structures remain spatially aligned and the reconstructed routes preserve their global topology. These visualizations serve as qualitative diagnostics of long-range trajectory consistency rather than additional released 3D supervision.



Figure 21: **Loop-and-return case.** A panoramic sequence that revisits its starting region while preserving complete surrounding context.

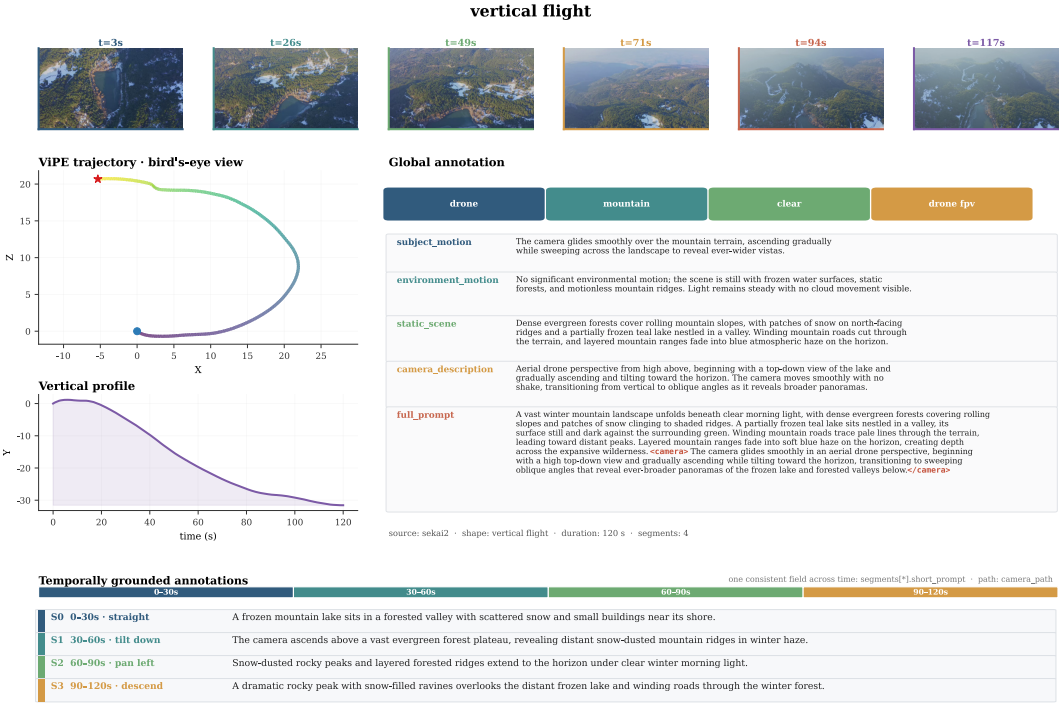


Figure 22: **Vertical-flight case.** An aerial sequence in which the vertical trajectory profile captures motion not apparent from the ground-plane projection alone.

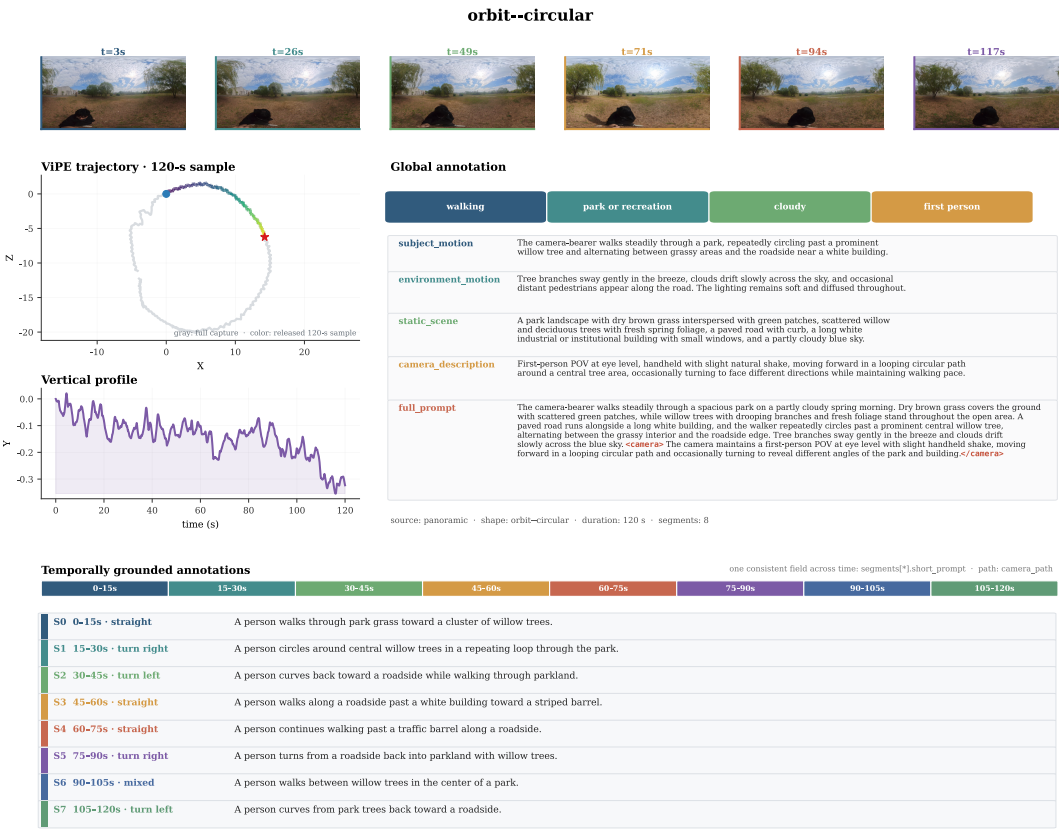


Figure 23: **Circular-orbit case.** A panoramic sequence with a strongly non-linear trajectory and repeated surrounding observations.

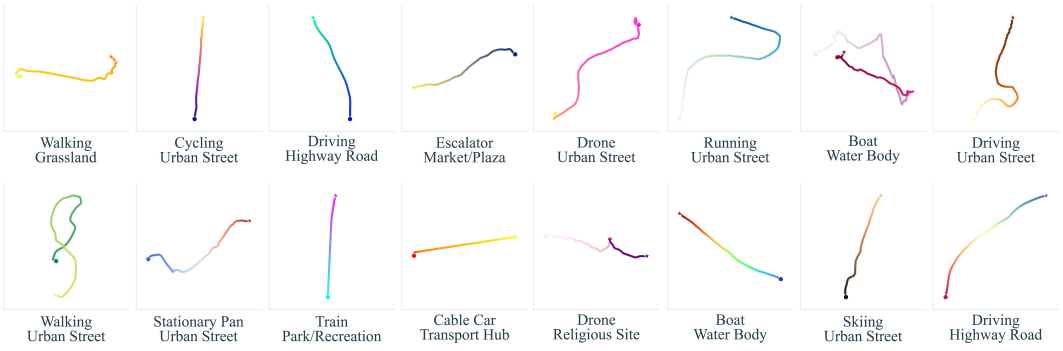


Figure 24: **Representative trajectories from the re-curated Sekai subset.** Camera centers are projected onto the ground plane and colored by time.

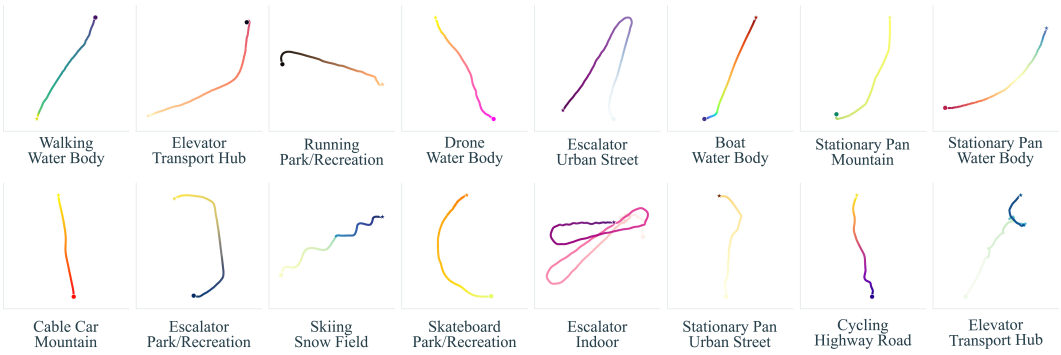


Figure 25: **Representative trajectories from the newly collected perspective subset.** The examples illustrate diverse camera-motion regimes while retaining smooth trajectory structure.

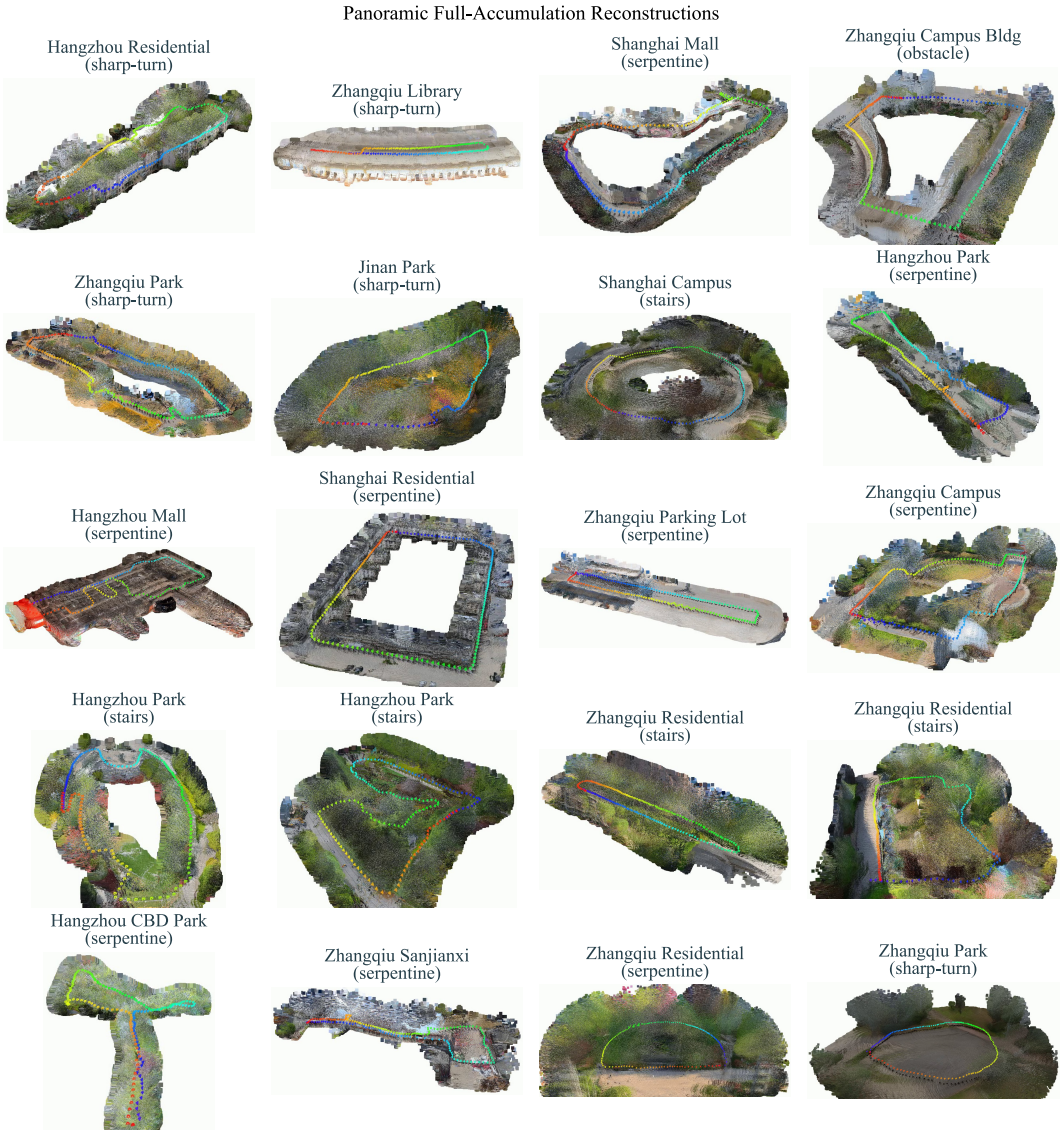


Figure 26: **Panoramic full-accumulation cases (part I).** The examples span diverse environments and non-linear camera trajectories.

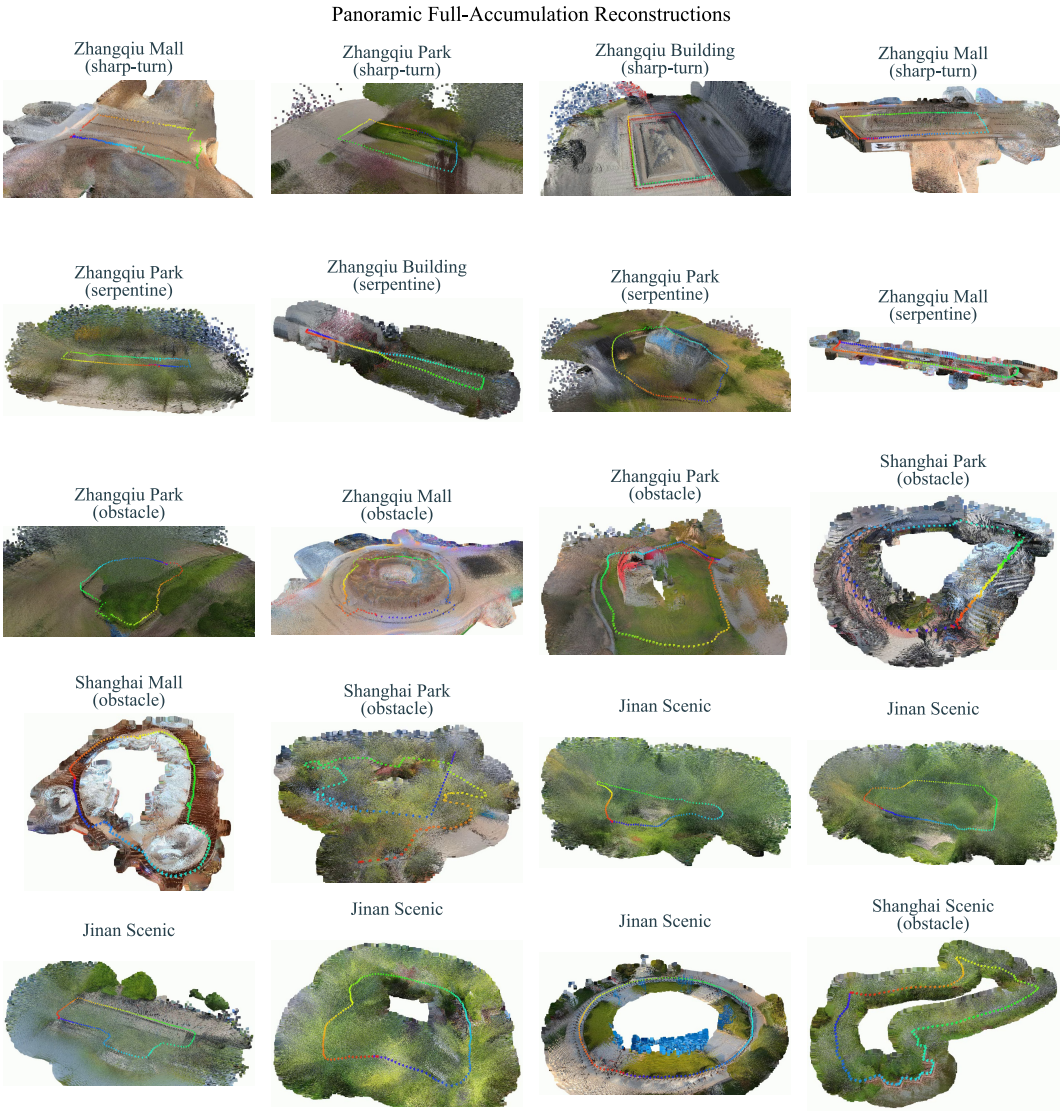


Figure 27: **Panoramic full-accumulation cases (part II).** Additional examples with long-range traversal and repeated viewpoints.