KineJing-Pro on TriWorldBench: A multidimensional evaluation of tri-view robot video prediction

YiCheng Pan | KineJing Research Team

Benchmark report | 25 October 2026

Abstract

Robot video predictors must preserve scene structure across cameras while representing the intended manipulation. We report the complete TriWorldBench profile of KineJing-Pro, a system producing synchronized head and wrist videos. In the 25 October 2026 public snapshot, the submission achieves TWB-Score 54.96 among 60 published entries. Perspective scores 91.43, ranked first; physical/3D coherence and temporal consistency each rank sixth. Subject consistency ranks fifth and interaction quality eighth. These strengths coexist with substantially weaker task alignment and motion rankings. We analyze all six categories and nineteen component metrics, verify the published extraction, and distinguish camera-compatible appearance from task- and motion-sensitive prediction. The resulting profile identifies concrete development priorities without attributing aggregate scores to an untested architectural mechanism. This benchmark report concerns the scored submission; it establishes neither real-robot performance nor general-purpose planning capability.

A robot observes one manipulation through cameras with different visibility, scale and occlusion. A generated rollout can retain an apparently stable scene while representing the wrong arm movement or an incorrect interaction. Evaluating the resulting system requires examining what each measurement establishes, rather than treating a single favorable visual score as complete evidence of useful prediction.

KineJing-Pro predicts synchronized head, left-wrist and right-wrist videos from initial visual observations and supplied robot trajectories. This report characterizes its submitted outputs using the published TriWorldBench evaluation [1]. The question is specific: which aspects of the scored system are well preserved, and which aspects remain limiting? The unit of analysis is the dated aggregate benchmark entry. We analyze the submission as evaluated, without combining its score with results from later local models.

The main observation is an uneven but informative performance profile. Perspective and selected temporal measurements occupy strong positions within the published entry cohort, whereas several task and motion measurements occupy much weaker positions. Our contribution is a complete, source-checked characterization of this profile, a component-level analysis of its internal differences, and an evaluation agenda tied to the observed weaknesses. The report does not infer that a particular undisclosed component caused any favorable score.

System and evaluation setting

Submitted prediction interface

The evaluated artifact consists of predicted video triplets covering the same manipulation interval. The initial head and two wrist images provide the visual starting point; robot trajectory inputs condition the subsequent rollout. The generation implementation builds on publicly released BWM pretrained components [2]. KineJing coordinates the prediction and submission workflow, including synchronized camera outputs. Core implementation and model assets remain private; the present report documents the interface and evaluation evidence rather than releasing a new pretrained architecture.

The retained submission description states that prediction uses initial observations and supplied actions without future test-reference images. This is the documented submission contract. The public aggregate record does not expose a checkpoint hash, a full execution configuration or per-episode inference traces, so the report cannot independently audit every aspect of the original generation run. Later adaptation

experiments are outside the evaluated version described here.

Benchmark scope and recorded evidence

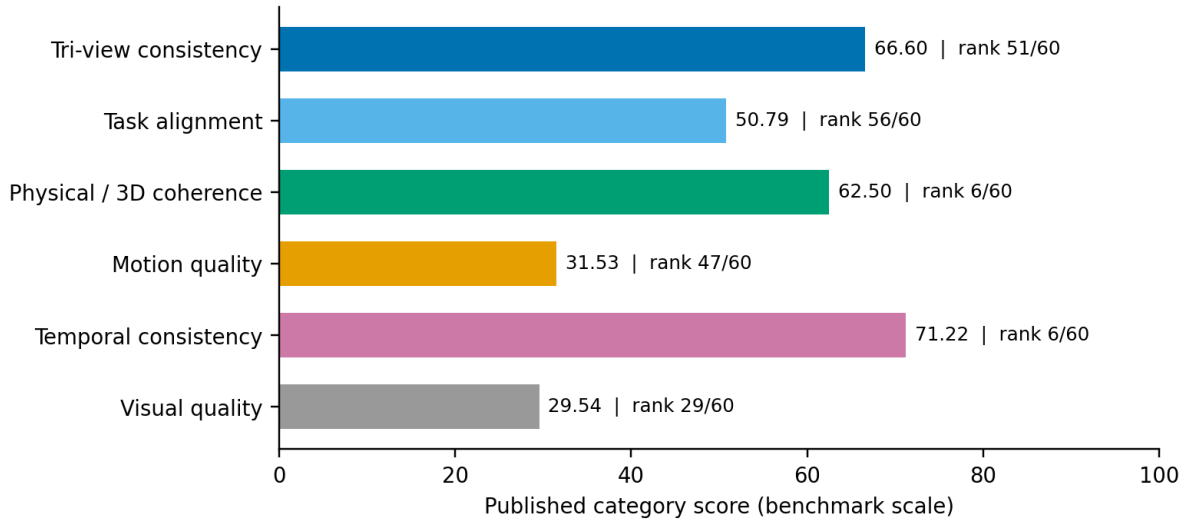
TriWorldBench covers 500 synchronized episodes over 50 bimanual manipulation tasks, with nineteen component metrics organized into six dimensions [1]. Its camera-aware protocol combines reference comparisons, cross-view judgments and task-sensitive measurements. We use the public KineJing-Pro entry captured on 25 October 2026 and retain the published values and rankings. The cohort contains sixty model entries; this count does not denote sixty independent institutions or architectures.

The analysis records three levels: the published overall score, all six dimension scores and all nineteen component scores. Category and component rankings describe positions in this particular cohort. Scores retain their original benchmark scale: for example, normalized PSNR is not a PSNR value in decibels, and JEPA similarity is not a task-success percentage. A rank indicates ordering, not the magnitude or statistical significance of a performance difference.

Results

Complete category profile

The published TWB-Score is 54.96, with overall ordinal position 44 of 60. Temporal consistency is 71.22 and physical/3D coherence 62.50; each ranks sixth. Tri-view consistency is 66.60, ranked 51st; task alignment is 50.79, ranked 56th; motion quality is 31.53, ranked 47th; and visual quality is 29.54, ranked 29th. Reporting all six categories preserves the distinction between the system's localized strengths and its overall position.



KineJing-Pro category profile in the 25 October 2026 public snapshot. Bars show the six published scores; annotations give each category rank among sixty model entries. Categories use different evaluation constructions, so bar length is not a common physical-error or task-success scale. No uncertainty estimates are available from the aggregate entry.

Category	Published score	Rank / 60
Tri-view consistency	66.60	51
Task alignment	50.79	56
Physical / 3D coherence	62.50	6
Motion quality	31.53	47
Temporal consistency	71.22	6
Visual quality	29.54	29

Complete dimension results. The separately published overall TWB-Score is 54.96 (ordinal position 44/60); it is not reconstructed by taking the arithmetic mean of these six displayed values.

Geometric and interaction measurements

Perspective is 91.43 and ranks first in the snapshot. Interaction Quality is 33.57 and ranks eighth; together their published arithmetic mean gives the physical/3D category score of 62.50. These results support a favorable benchmark profile for the appearance and interaction properties assessed by those metrics. They do not establish calibrated 3D reconstruction, measured contact forces or correct physical state transitions in every episode.

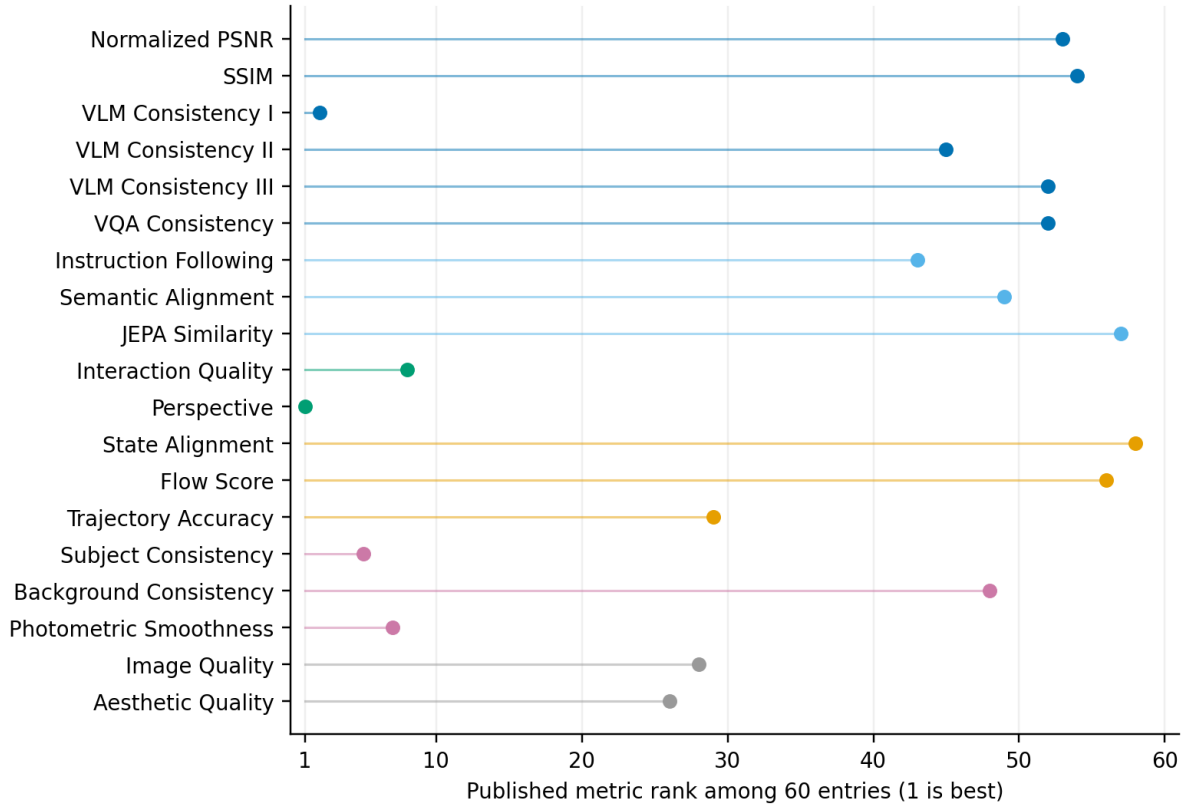
The component distinction matters because perspective regularity and interaction quality are not the same property. A high perspective score can coexist with an incorrect requested manipulation. The physical/3D category therefore identifies a promising aspect of the submitted video behavior while leaving task execution as a separate empirical question. No per-task distribution is available to determine whether the favorable aggregate is broad or concentrated in particular scenes.

Temporal and cross-view evidence

Subject Consistency scores 84.96, ranked fifth, and Photometric Smoothness scores 59.12, ranked seventh. Background Consistency scores 69.59, ranked 48th. These values yield temporal consistency 71.22. The strong category rank consequently does not imply that all temporal submetrics perform uniformly well; it combines distinct identity, background and appearance signals.

VLM Consistency I scores 88.62, ranked second, whereas Consistency II scores 73.79, ranked 45th, and Consistency III scores 83.28, ranked 52nd. The benchmark distinguishes compatibility under occlusion, explicit verification across views and exemplar-calibrated judgments [1]. The disagreement shows that a strong position under one rubric does not establish uniformly strong cross-view verification. The scores also have different cohort distributions, so the numeric gaps between rubrics are not controlled effect sizes.

Normalized PSNR is 47.16, SSIM is 65.36 and VQA Consistency is 41.40, with ranks 53, 54 and 52 respectively. Averaging these three measurements with the three VLM values reproduces the published tri-view category score of 66.60. Thus the second-place compatibility result is one component of a substantially less favorable aggregate cross-view profile.



Ranks of all nineteen component metrics for the same KineJing-Pro submission. Smaller ranks are better; colors retain the six category groupings used in the preceding figure. All ranks use one plus the number of strictly greater published metric values. These positions are descriptive and depend on the entry cohort; they are not confidence intervals or calibrated distances between capabilities.

Task alignment and motion remain the limiting dimensions

Instruction Following scores 63.76, Semantic Alignment 88.14 and JEPA Similarity 0.46, with ranks 43, 49 and 57. Their published average yields task alignment 50.79. A high-looking semantic score on its own therefore does not establish strong task performance relative to the evaluated cohort. The different raw scales further make the category a benchmark-defined index rather than a direct success probability.

State Alignment is 58.64 (rank 58), Flow Score 4.59 (rank 56) and Trajectory Accuracy 31.37 (rank 29), yielding motion quality 31.53. These measurements motivate closer inspection of action phase, camera-specific motion and end-effector path agreement. The aggregate record cannot distinguish insufficient motion, excess motion, timing mismatch and object-response errors at the episode level. In particular, the flow value alone is not evidence that increasing generated motion would improve correctness.

Image Quality is 35.01 and Aesthetic Quality 24.07, ranked 28 and 26. Their combined visual category sits near the middle of the cohort. The resulting assessment is specific: geometric appearance and selected temporal properties are the strongest recorded aspects of this entry, while action- and task-sensitive prediction require further development.

Analysis protocol

Extraction and arithmetic checks

We preserve the public HTML response and independently parse its displayed table and embedded public records. All sixty names and their order are matched; 1,560 score/rank cells are checked against the records, allowing at most 0.005 for two-decimal score formatting. The two representations originate in one response, so agreement verifies extraction rather than independently replicating the benchmark evaluation. The snapshot SHA-256 is 9f9730288829c9d81ff494d2bc75b62a077584aed40f98cbf9cd96054ff6ca4d.

For component value x, its rank is one plus the number of strictly greater values in the sixty-entry snapshot; ties share a metric rank. The overall published ordinal ordering is retained. Using the metric groups given by the benchmark, we check that each of the six component means reproduces its category value to numerical tolerance. Their unweighted mean is 52.03, which does not reproduce the published overall 54.96; accordingly, we retain the official aggregate instead of inventing a replacement aggregation rule.

Scope of statistical interpretation

The available observations are aggregate category and metric values. Per-episode scores, task-wise dispersion and evaluator-repeat measurements are unavailable for this report. We therefore report no bootstrap intervals, significance tests, task success rate or episode-level failure frequency. Nineteen metrics are nineteen related measurements of one system, not nineteen independent replications. A rank change in a future leaderboard snapshot would also reflect changes to the comparison cohort.

Discussion

The profile identifies a practical separation between visual regularity and action-faithful prediction. The evaluated system can occupy strong perspective and selected temporal positions while remaining weak in task alignment and state-conditioned motion. This is an observation about the joint pattern of aggregate measurements, not a causal explanation of the system's errors. It motivates evaluating progress across several dimensions simultaneously.

A targeted next evaluation should preserve the submitted model and input configuration while examining complete task episodes: synchronized arm motion, grasp and release transitions, object identity across cameras and the alignment of supplied trajectories with generated paths. Initial-frame persistence and action-removal controls would test whether apparent stability substitutes for useful dynamics. Such controls are proposed follow-up experiments; their outcomes are not inferred from the published aggregate.

The reporting choice is consistent with the broader distinction between perceptual video quality and functional predictive utility. VBench separates perceptual and temporal properties [3], EWMBench evaluates scene, motion and semantic aspects [4], and WorldArena connects visual assessment to downstream functions [5]. These frameworks motivate keeping geometric, temporal, task and motion evidence visible rather than treating a favorable component rank as a complete system claim.

The limitations are substantial. This is one scored system in a dated simulation benchmark snapshot. The complete original execution configuration and public checkpoint binding are unavailable; aggregate scores do not establish task-general consistency or a mechanism responsible for the observed strengths. The study includes no new blind submission, controlled architectural ablation, real-robot experiment, closed-loop planning result or hardware-independent deployment measurement. These boundaries define the scope of the reported evidence.

Conclusion

KineJing-Pro's TriWorldBench evaluation reveals first-ranked perspective, sixth-ranked physical/3D and temporal categories, and strong subject-consistency and interaction positions within the published cohort. The complete profile simultaneously exposes limitations in task alignment and motion. Preserving both aspects yields a concrete basis for further development: retain the recorded geometric and temporal strengths while testing whether action- and task-sensitive behavior improves under controlled evaluation.

Data and code availability

The source of the evaluation profile is the public TriWorldBench leaderboard at https://www.triworldbench.com/#leaderboard, captured on 25 October 2026. The accompanying report records all KineJing-Pro values, extraction checks and snapshot digest. Benchmark data and pretrained assets remain with their original providers under their terms. KineJing core source code and model weights are not publicly released. No customer data, private correspondence or model checkpoints are included in the manuscript source archive.

Author information and research assistance

YiCheng Pan is affiliated with KineJing Research Team, which develops KineJing-Pro; this affiliation is disclosed as a potential competing interest. AI tools assisted with source retrieval, artifact inspection, manuscript preparation and consistency checks, including Scientific Agent Skills procedures [6]. Reported benchmark measurements originate from the identified public evaluation record.

Complete component results

Component metric	Published value	Rank / 60
Normalized PSNR	47.16	53
SSIM	65.36	54
VLM Consistency I	88.62	2
VLM Consistency II	73.79	45
VLM Consistency III	83.28	52
VQA Consistency	41.40	52
Instruction Following	63.76	43
Semantic Alignment	88.14	49
JEPA Similarity	0.46	57
Interaction Quality	33.57	8
Perspective	91.43	1
State Alignment	58.64	58
Flow Score	4.59	56
Trajectory Accuracy	31.37	29
Subject Consistency	84.96	5
Background Consistency	69.59	48
Photometric Smoothness	59.12	7
Image Quality	35.01	28
Aesthetic Quality	24.07	26

All nineteen component measurements in benchmark group order: six tri-view, three task, two physical/3D, three motion, three temporal and two visual metrics. Values retain the published scale and are not converted to percentages of task success.

References

1. Liu, Xuanyi et al.. TriWorldBench: A Tri-View Consistency Perspective on Embodied World Models. arXiv:2609.26314 (2026). https://arxiv.org/abs/2609.26314

2. BWM Team. BWM: A Low-Cost High-Fidelity World Simulator for Robot Learning. arXiv:2607.29302 (2026). https://arxiv.org/abs/2607.29302

3. Huang, Ziqi et al.. VBench: Comprehensive Benchmark Suite for Video Generative Models. arXiv:2311.17982 (2023). https://arxiv.org/abs/2311.17982

4. Yue, Hu et al.. EWMBench: Evaluating Scene, Motion, and Semantic Quality in Embodied World Models. arXiv:2505.09694 (2025). https://arxiv.org/abs/2505.09694

5. Shang, Yu et al.. WorldArena: A Unified Benchmark for Evaluating Perception and Functional Utility of Embodied World Models. arXiv:2602.08971 (2026). https://arxiv.org/abs/2602.08971

6. Kassis, Timothy et al.. Scientific Agent Skills: A Library of Procedural Knowledge for Research Agents. arXiv:2609.00065 (2026). https://arxiv.org/abs/2609.00065