WORLDLINEBENCH: TOWARD PROCESS-ORIENTED EVALUATION OF VIDEO-GROUNDED STREAMING INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming video generation endows video, for the first time, with a true worldline: a continuously unfolding visual trajectory that remains open to real-time user intervention. This transforms video generation into a fundamentally new interactive task that requires timely responses while preserving an already-established visual history, posing new challenges to existing evaluation paradigms centered on endpoint quality. To this end, we present WorldLineBench, to our knowledge the first large-scale video-grounded benchmark for streaming interaction, comprising 1,000 synthetic and 1,000 real videos spanning five content contexts and seven interaction types. Synthetic videos provide controlled instruction coverage while real videos probe whether interactive capability generalizes to natural scenarios. We adopt a 45-second long-horizon design supporting multiple rounds of history-dependent interaction, constructing branching worldlines from different interactions at designated nexus points. Under a process-oriented evaluation covering four metrics across two capability dimensions, WorldLineBench uniformly evaluates seven representative models from three paradigms, revealing consistent gaps across responsiveness, semantic fulfillment, and history-dependent preservation.

1 INTRODUCTION

Streaming models (Yin et al., 2025; Yang et al., 2026), overcoming the one-shot limitation of conventional video models (Ho et al., 2022; Bar-Tal et al., 2024; Yang et al., 2025), have established themselves as a promising new paradigm for generation. It enables users to re-enter the process at any time, issue new instructions based on the unfolded state, and redirect the following trajectory. By treating the generation process as a worldline[1] with an established history and an open future, it opens up a vast new field of research, streaming video interaction (Jiao et al., 2026), which grounds each instruction in the current visual state and requires the model to respond promptly and accordingly, preserve content outside the intended intervention across successive rounds of interaction.

However, a specialized benchmark for streaming video interaction remains lacking, forcing the community to using existing benchmarks centered on one-shot generation (Liu et al., 2023) and offline editing (Chen et al., 2025; Li et al., 2025) as a substitute, which cannot adequately or comprehensively evaluate model capability in this new task. Therefore, we seek to redesign a new specialized benchmark aligned with the requirements of streaming interaction. This vision faces three core challenges: First, user intent in streaming scenarios is inherently tied to the established visual state, while previous benchmarks (Ying et al., 2026; Duan et al., 2025) typically rely on predefined prompts or instructions that are weakly correlated with, or even irrelevant to, most of the video state, falling short of genuine video-grounded interaction. Second, most evaluations operate within relatively short windows (Feng et al., 2025; Wu et al., 2026b), making it difficult to capture accumulated effects across multiple interactions or to assess long-term stability after each intervention. Third, previous evaluations remain predominantly outcome-oriented (Zheng et al., 2025), focusing on the final output rather than the process, including response onset, instruction fulfillment, and post-interaction stability. Together, these challenges call for a fundamental redesign of streaming interaction evaluation, spanning assessment methodology, interaction grounding, and temporal coverage.

[1] We borrow this term from relativity to describe the evolving trajectory of visual states in a video.

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.

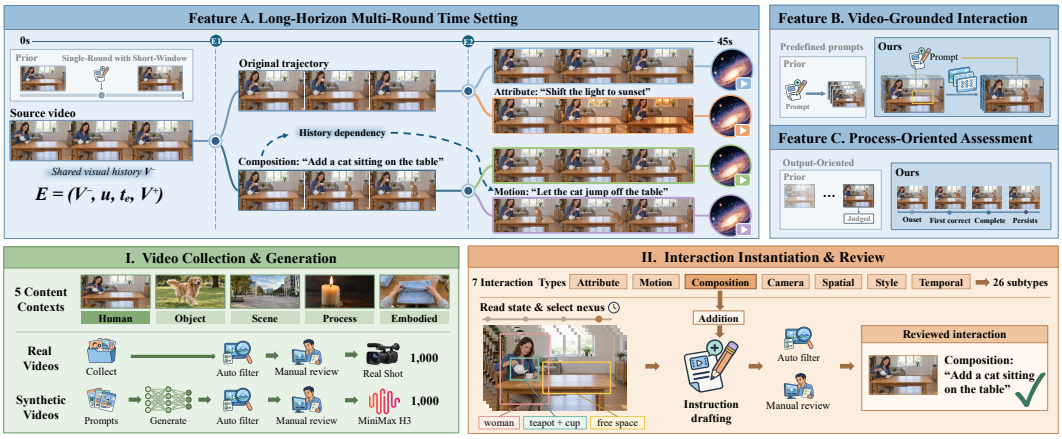


Figure 1: **Overview of WorldLineBench.** The upper panel illustrates three key features : a long-horizon multi-round time setting, video-grounded interaction, and process-oriented assessment, all jointly integrated into WorldLineBench. The lower panel shows the two-stage construction pipeline, from video collection and generation to interaction instantiation and review.

To overcome these limitations, we introduce WorldLineBench, the first benchmark for streaming interaction that integrates video-grounded interaction, long-horizon multi-round settings, and process-oriented assessment, a joint treatment that prior benchmarks have yet to offer. WorldLineBench begins by grounding each interaction in an already-unfolding video worldline, with instructions constructed from the visual state at the intervention time. Building on this grounding, WorldLineBench adopts a 45-second long-horizon, multi-round protocol to capture both post-interaction evolution and the accumulation of history across interventions, where the state produced by one intervention becomes part of the history for subsequent interactions, providing sufficient temporal scope to examine long-term preservation, drift, and accumulated errors. Such a multi-round protocol demands an evaluation approach that goes beyond final outcomes. WorldLineBench therefore anchors each interaction at a nexus point and defines dedicated observation windows to systematically track response onset, instruction fulfillment, and post-interaction stability along the temporal axis. Accordingly, four evaluation dimensions are assessed over the corresponding windows, grouped into two capability: Responsiveness and Semantic Fulfillment under Real-time Interaction Capability, and Preservation and Video Quality under Worldline Generation Capability. Altogether, WorldLineBench contains 1,000 real and 1,000 synthetic videos spanning five content contexts and seven interaction semantics, where real videos probe generalization to natural scenarios and synthetic videos ensure controllable interaction coverage, yielding 6,000 interaction opportunities for systematic evaluation.

We systematically evaluate seven representative models across three paradigms on WorldLineBench. Extensive results reveal that pervasive capability imbalances exist across four evaluation dimensions, with a severe gap observed between timely responses and successful instruction execution. Fine-grained analysis further reveals substantially different capability profiles across interaction conditions. As interaction history accumulates in long-horizon multi-round settings, models grow prone to semantic degradation, content drift, and stability decay. These findings indicate that current models remain far from meeting the demands of real-world streaming interaction, underscoring the value of WorldLineBench, which exposes and tracks these gaps.

Together, we summarize our contributions as follows:

1. We identify a critical gap, as video generation shifts toward streaming interaction. Current benchmarks remain confined to outcome-oriented, single-round settings with predefined instructions, leaving this emerging paradigm without adequate evaluation support.
2. We introduce WorldLineBench, the first process-oriented benchmark for streaming interaction, comprising 2,000 videos spanning five content contexts and seven interaction types, with 6,000 video-grounded interactions under a 45-second multi-round protocol.
3. We conduct extensive experiments on seven representative models across three paradigms, revealing pervasive capability gaps across four evaluation dimensions, with fine-grained analysis exposing model-specific boundaries and long-horizon degradation patterns.

Table 1: **Comparison with representative benchmarks.** We compare interaction settings, evaluation coverage, and dataset scale across existing benchmarks and WorldLineBench.

Benchmark	Interaction		General Evaluation			Process Evaluation				Scale		
	Multi-turn	Branch.	Quality	Adher.	Preserv.	Hist. cons.	Resp. time	Stages	Persist.	Cases	Instr.	Horizon (s)
General video generation												
EvalCrafter (Liu et al., 2024)	✗	✗	✓	✓	✗	✗	✗	✗	✗	700	700	1–4
VBench++ (Huang et al., 2026)	✗	✗	✓	✓	✗	✗	✗	✗	✗	946	946	≈2–4
VBench-2.0 (Zheng et al., 2025)	✗	✗	✗	✓	✗	✗	✗	✗	✗	1,230	1,230	5–10.1
Instruction-guided video editing												
IVEBench (Chen et al., 2026)	✗	✗	✓	✓	✓	✗	✗	✗	✗	600	600	≈1–43
OmniEdit-Bench (Miao et al., 2026a)	✗	✗	✓	✓	✓	✗	✗	✗	✗	790	790	≈3–10
CoVEBench (Wu et al., 2026a)	✗	✗	✓	✓	✓	✗	✗	✗	✗	416	626	3–21
Temporal and interactive evaluation												
NarraStream-Bench (Liu et al., 2026a)	✓	✗	✓	✓	✗	✓	✗	✗	✗	324	1,944	60
EgoPlay (Mai et al., 2026) benchmark	✗	✗	✓	✓	✓	✗	✓	✗	✗	150	600	5
InfinityEdit (Tong et al., 2026) benchmark	✓	✗	✓	✓	✓	✓	✗	✗	✗	200	600	≈25
StateBench (Miao et al., 2026b)	✗	✗	✗	✓	✗	✓	✗	✗	✗	200	200	≈5–10
WBench (Ying et al., 2026)	✓	✗	✓	✓	✗	✓	✗	✗	✗	289	1,058	≈10–45
What-If World (Cai et al., 2026)	✗	✓	✓	✓	✓	✗	✗	✗	✗	319	638	≈4–10
StreamAV-Bench (Liu et al., 2026b)	✓	✗	✓	✓	✓	✓	✓	✗	✗	160	960	180
WorldLineBench (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	2,000	6,000	45

Branch.: alternative paths from the same visual history, evaluated jointly. **Preserv.**: non-target content quality; **Persist.**: interaction effects afterwards; **Hist. cons.**: visual state retention across turns. **Resp. time**: when the change first appears in video time; **Stages**: scored separately at interaction onset, first correct change, and full completion. ≈ marks estimated durations; our 45 s horizon includes the given history. **Ours**: 2,000 source videos × 3 unique interactions across branches; at most two successive interactions per path.

2 Related Work

2.1 From Video Generation to Real-Time Interaction

With substantial progress achieved in visual fidelity (Chen et al., 2024; Team Wan et al., 2025), temporal coherence (Blattmann et al., 2023; Guo et al., 2024), and generation duration (Qiu et al., 2024; Zhang et al., 2025), real-time interaction capability has emerged as a new frontier in video generation as corresponding real-world demands increase, giving rise to two lines of research. The first line pursues real-time generation by reducing latency (Yin et al., 2025; Huang et al., 2025) and extending generation length (Liu et al., 2026c; Li et al., 2026b), then introduces interaction capability through dynamic prompt updates (Yang et al., 2026) and history-conditioned instruction following (Tong et al., 2026). Conversely, the second line starts from video editing, exploring interaction guided by text, images, and other multimodal signals (Geyer et al., 2024; Ku et al., 2024; Jiang et al., 2025), while later efforts improve editing efficiency through temporal reuse (Liang et al., 2025) and causal processing (Xing et al., 2024; Wang et al., 2026). It further achieves real-time performance through autoregressive modeling and distillation (Li et al., 2026a; Xiao et al., 2026). Despite these advances, both lines remain susceptible to failure modes under real-world scenarios that are insufficiently captured by current evaluation protocols, which still predominantly assess one-shot instructions against final outputs. Specifically, streaming generation models lose consistency once prompts change mid-generation (Tan et al., 2026), while real-time editing models face persistent error accumulation over successive instructions (Tong et al., 2026). Interactive streaming video models therefore still require further improvement to meet the demands of real-world interactive scenarios.

2.2 From Single-Shot Evaluation to Streaming Interactive Benchmarking

As video models advance, principled benchmarks become essential for tracking progress, exposing model limitations, and guiding future improvement (He et al., 2024). Early benchmarks focus on visual quality (Liu et al., 2024; Huang et al., 2024), prompt following (Huang et al., 2026; Sun et al., 2025), and temporal coherence (Yuan et al., 2024). As interaction complexity grows, the evaluation focus shifts from generation quality toward instruction execution: IVEBench (Chen et al., 2026), OmniEdit-Bench (Miao et al., 2026a), and CoVEBench (Wu et al., 2026a) provide testbeds covering diverse instructions for evaluating semantic compliance and content preservation. This shift further deepens as streaming scenarios demand continuous, responsive interaction. Targeting

this need, NarraStream-Bench (Liu et al., 2026a), EgoPlay (Mai et al., 2026), and InfinityEdit (Tong et al., 2026) introduce multi-turn updates, history consistency, and content preservation from streaming and continuation settings, while StreamAV-Bench (Liu et al., 2026b), StateBench (Miao et al., 2026b), WBench (Ying et al., 2026), and What-If World (Cai et al., 2026) extend into audio-video and world-model scenarios respectively. Yet existing benchmarks shown in Tab. 1 are built around a fundamentally different assumption: interaction is a one-shot, offline event, rather than a continuous process grounded in visual state. This mismatch motivates the design of WorldLineBench, which treats streaming interaction as a continuous, video-grounded, and process-oriented evaluation target.

3 Dataset Construction

3.1 Video Worldline Corpus Construction

A video, in its natural form, is itself a worldline: a continuously unfolding trajectory of visual states. When a user intervenes, it branches, with the established history remaining shared while the future diverges along the direction of the new instruction, giving rise to multiple coexisting worldlines. Real-world streaming interaction is fundamentally shaped by this structure, as users track the ongoing worldline and issue instructions grounded in its accumulated history, rather than predefined inputs. WorldLineBench constructs its source corpus accordingly, selecting worldlines with clearly traceable state evolution to ground every interaction in an established visual history.

Constructing such a corpus of worldlines, however, requires balancing two complementary goals: natural visual grounding and systematic scenario coverage. Real videos inherently satisfy the former, providing authentic visual histories with genuine physical complexity but with limited distributional control. Synthetic videos address the latter by enabling controlled construction of diverse states. WorldLineBench therefore combines both sources and further organizes the corpus around five content contexts: Human, Object, Scene, Process, and Embodied. This organization provides balanced semantic coverage as well as diversity in subject composition, environmental complexity, and temporal dynamics, providing a broad basis for constructing diverse subsequent video-grounded instructions. The resulting corpus contains 2,000 worldlines, evenly split between real and synthetic videos and systematically distributed across all five contexts. Details of data sources, generation configurations, and quality control procedures are provided in the Appendix A.

3.2 Interaction Taxonomy and Video-Grounded Instantiation

To systematically cover the range of possible user interactions, we organize them into seven primary categories: Attribute, Motion, Composition, Camera, Spatial, Style, and Temporal. Though comprehensive in scope, these categories are too abstract to translate directly into concrete, executable instructions. Moreover, possible subcategories derived top-down from them alone may be semantically coherent yet lack grounding in real worldlines. We therefore define finer-grained, video-grounded interaction subtypes by jointly considering seven primary categories above and the five content contexts introduced in Sec. 3.1. As a result, the same primary category unfolds differently across contexts. Motion, for instance, yields local and full-body subtypes in Human videos, while manifesting as object movement or interaction in Object videos.

Building on this fine-grained taxonomy, we derive each specific instruction and its intervention timing from the current worldline, based on the established visual states, including the subject's condition, spatial relations, and process stage. In this way, the taxonomy provides structured interaction categorization, while the video history constrains the applicable content and the nexus point for intervention, yielding interactions that are both systematically organized and genuinely state-dependent. Detailed constructions and applicability rules are provided in the Appendix B.

3.3 Multi-Round Worldline Bifurcation for Long-Horizon Evaluation

Rather than merely triggering an isolated short-term response, each applied intervention marks a nexus point that redirects the video's subsequent trajectory. Afterward, it branches into two worldlines: one continues the original trajectory, the other evolves under the new instruction, depending on whether the intervention is applied. Both share the same past but diverge into different futures.

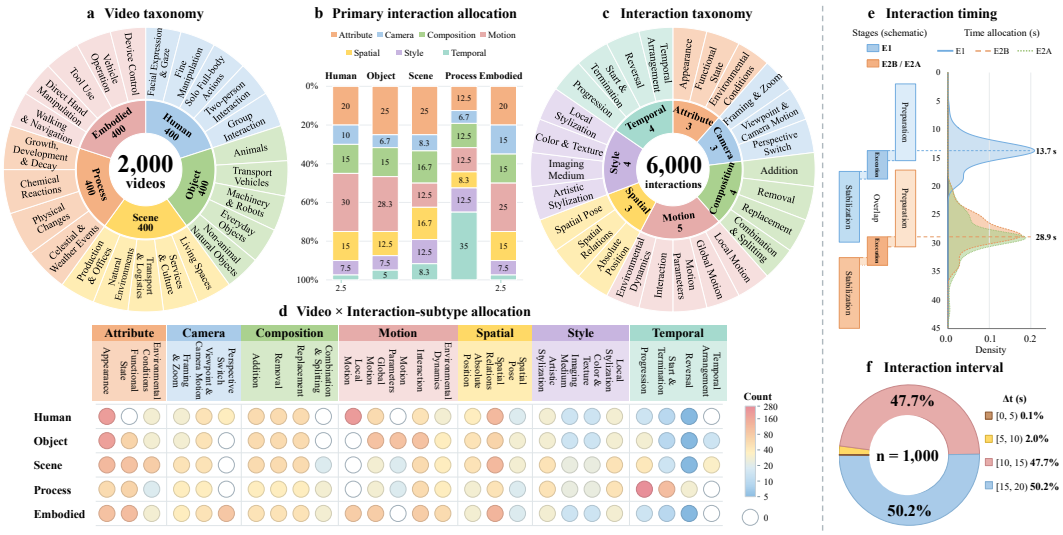


Figure 2: **Dataset statistics of WorldLineBench.** The dataset comprises 2,000 worldlines uniformly distributed across five content contexts and real and synthetic sources, yielding 6,000 interactions across 7 interaction categories and 26 subtypes whose distribution reflects context-dependent constraints rather than uniform allocation. Nexus points concentrate at approximately 13.7 s and 28.9 s across the two rounds, with 97.9% of inter-round intervals Δt falling within $[10, 20)$ s.

Yet this divergence does not settle instantaneously, as the resulting changes typically need several following frames to stabilize into a new visual state. Determining whether an intervention truly fulfills user expectations, and whether its result persists over time, therefore necessitates an extended observation window. Within this window, new content, relations, and states continue to accumulate and evolve differently on each worldline, forming the basis for subsequent interactions. Meanwhile, in real streaming scenarios, users rarely stop at a single interaction. They continue to issue new requests based on these newly formed states, motivating a multi-round design that captures how user intent and video state co-evolve over successive interventions. A new nexus point then naturally emerges on both diverged worldlines, forming a second round of interaction that, similarly as the previous one, branches each into a new future, together yielding four interrelated worldlines.

Together, each interaction round follows the same three-step rhythm: understanding the accumulated history, executing the intervention, and observing until the resulting state stabilizes. Each step takes time, and across two rounds, their combined span substantially exceeds what short-window evaluation can accommodate. WorldLineBench therefore adopts a 45-second long horizon, supporting the measurement of both post-intervention state persistence and subsequent interactions on each branch, while retaining the accumulated differences across intervention paths for cross-round analysis of history dependence and state drift. Further details are provided in the Appendix C.

3.4 Dataset Statistics and Coverage

Following automatic quality filtering and manual review, we construct WorldLineBench along three dimensions: video corpus, interaction taxonomy, and temporal structure.

The dataset comprises 2,000 45-second videos, uniformly distributed across five content contexts and evenly split between real and synthetic source, as visualized in Fig. 2(a). Interactions within these worldlines are organized into 7 primary categories, further divided into 26 subtypes, whose distribution across different content reflects the constraints each context places on executable interactions, as shown in Fig. 2(b,c,d). For the temporal structure, each worldline is annotated with three nexus points across two rounds of interaction, one first-layer and two second-layer, yielding three evaluable interaction events per video and 6,000 in total. The first-layer nexus point is concentrated at roughly 13.7 s and the second-layer ones at approximately 28.9 s. Both distributions exhibit non-fixed timing, with the two peaks separated by close to 15 s, providing sufficient observation space for semantic response and state preservation after each intervention, as depicted in Fig. 2(e,f).

4 METRICS

4.1 EVALUATION TASK FORMULATION

WorldLineBench formalizes each user intervention as an interaction event as:

$$\mathcal{E} = (V^-, u, t_e, V^+), \qquad (V^-, u, t_e) \to V^+, \tag{1}$$

where V^- denotes the visual history established before the intervention, u denotes the instruction user issued at the nexus point t_e, categorized as *bounded instructions*, requiring completion within a finite interval, and *persistent instructions*, demanding sustained maintenance and V^+ denotes the worldline branching afterward. Start from t_e, we define two time windows that standardize the computation scope of all metrics: an evaluation window $W_u = [t_e, t_e + 15\text{s}]$, used to assess whether the instructed change occurs and whether non-target states are preserved, and a response window $W_r = [t_e, t_e + 5\text{s}]$, used to examine whether the response emerges promptly for streaming task. Built around these two windows, WorldLineBench adopts a process-oriented view that evaluates model performance along two capability dimensions comprising four metrics, reported separately:

Real-time Interaction Capability assesses whether the model executes the intervention both correctly and promptly. *Semantic Fulfillment* judges whether the response semantically aligns with the instruction demands, based on different criteria of bounded or persistent instructions.*Responsiveness* measures how promptly the model produces observable responses to the instruction.

Worldline Generation Capability further evaluates the model's ongoing generative capacity during and after the intervention. *Preservation* examines whether non-target states remain intact within W_u, while *Video Quality* measures the visual and temporal quality of the subsequent worldline itself.

4.2 REAL-TIME INTERACTION CAPABILITY

This dimension examines the model's ability to dynamically respond to user interactions, jointly measured by semantic fulfillment and responsiveness metrics. All instructions are decomposed into atomic sub-clauses, on which the metrics below are built. More details can be find in Appendix D.

Semantic Fulfillment Overall Semantic Consistency (OSC) measures the holistic alignment between the visual content and the expected description implied by the instruction. Instruction Satisfaction – Global (IS-g) and Edit Fulfillment Accuracy (EFA), both judged by Gemini 3.8 Flash, respectively rate the overall degree of instruction satisfaction on a predefined 1–5 scale and determine whether each atomic sub-clause is satisfied in V^+ , with the fulfillment ratio across all sub-clauses reported as the EFA score. For persistent instructions, Semantic Effective Occupancy (SEO) measures the proportion of frames within W_u in which all atomic sub-clauses hold simultaneously.

Responsiveness Time to First Frame (TTFF) records the timestamp of the first frame generated in response to the instruction. Semantic Onset Latency (SOL) marks the earliest moment an atomic sub-clause becomes satisfied, and once all sub-clauses are met, Semantic Completion Latency (SCL) settles, indicating successful instruction completion. Deadline Success Rate (DSR) determines whether this completion timestamp falls within the timeliness window W_r.

4.3 WORLDLINE GENERATION CAPABILITY

This dimension examines whether the model can sustain a stable, high-quality worldline after intervention, jointly measured by preservation and video quality metrics. Similarly, each instruction is associated with a predefined set of protected atomic states based on our seven primary interaction categories, from which non-target states are derived. Further details are provided in Appendix E.

Preservation Content Fidelity – Global (CF-g) and Protected-State Retention (PSR) are both judged by Gemini 3.8 Flash. The former rates the overall content preservation on a predefined 1-5 scale, while the latter assess whether each protected atomic state is retained in V^+, with the retention ratio across all protected states reported as the PSR score. Background Consistency (BC) and Subject Consistency (SC) follow the protocol of VBench (Huang et al., 2024) to measure the frame-to-frame consistency of the background and subject. Coverage-Adjusted History-Dependency Following (CA-HDF) is specifically designed for multi-round evaluation, restricted to samples exhibiting

Table 2: **Performance of seven models on WorldLineBench. Bold** / underlined respectively denote the first/second-best result per metric across all models while ↑ and ↓ indicate higher / lower is better.

Model	Real-time interaction capability								Worldline generation capability							
	Semantic fulfillment				*Responsiveness*				*Preservation*					*Video quality*		
	OSC↑	EFA↑	IS-g↑	SEO↑	TTFF↓	SOL↓	SCL↓	DSR↑	CF-g↑	PSR↑	CA-HDF↑	SC↑	BC↑	NRQ↑	TC↓	AE↑
Streaming-native models																
StreamEdit	0.184	21.17	1.75	11.59	71.33	3.40	13.28	9.20	2.99	58.72	19.68	0.887	0.917	56.61	**0.0106**	-0.015
StreamDiffusionV2	0.184	20.86	1.72	12.12	**2.30**	**2.08**	13.11	10.13	1.81	19.32	10.46	0.854	0.915	50.02	0.0142	0.003
Real-time video editors																
LiveEdit	**0.186**	32.19	2.17	24.65	26.24	2.57	11.99	18.67	2.23	35.24	20.57	0.871	0.918	**63.26**	0.0142	**0.037**
SANA-Streaming	0.176	19.11	1.68	9.52	14.41	3.78	13.42	7.73	**3.71**	**71.95**	17.91	0.896	0.926	41.13	0.0129	-0.060
JoyAI	0.182	**37.92**	**2.46**	**31.36**	2.54	2.26	**11.23**	**22.67**	3.48	68.46	**39.72**	0.890	0.920	55.63	0.0153	-0.026
Full-context controllable models																
Lucy Edit 1.1	0.182	16.41	1.56	8.03	62.37	4.80	13.79	5.17	1.78	18.83	3.19	0.823	0.901	45.57	0.0174	-0.037
Wan2.1-VACE	0.179	22.22	1.79	10.96	199.54	4.19	13.26	8.77	3.40	66.21	15.07	**0.901**	**0.929**	60.89	0.0135	0.016

cross-round interaction between interactions, measuring whether an earlier interction result is retained through a later round, which other single-round evaluations entirely overlook.

Video Quality No-Reference Quality (NRQ), computed with MUSIQ (Ke et al., 2021), measures the perceptual quality of individual frames without references. Temporal Continuity (TC), computed following VBench (Huang et al., 2024) Temporal Flickering and Motion Smoothness, jointly captures frame-level flickering and artifacts along with motion smoothness. Aesthetic Evaluation (AE), computed with Dover (Wu et al., 2023), measures the aesthetic quality and naturalness of the video.

5 Experiment

5.1 Experiment Setup

We evaluate seven representative models on WorldLineBench, organized into three groups according to their underlying technical paradigm. **Streaming-native Models**, comprising StreamEdit (Jiao et al., 2026) on LongLive (Yang et al., 2026) and StreamDiffusionV2 (Feng et al., 2026) , build real-time interaction capabilities from causal streaming generation architectures. **Real-time Video Editors**, namely LiveEdit (Wang et al., 2026) , SANA-Streaming (Zhao et al., 2026), and JoyAI-Video (Xiao et al., 2026) , start from conditional video editing and push toward real-time response through architectural and training optimizations. **Full-context Controllable Models** include Lucy Edit 1.1 (DecartAI Team, 2025) and Wan2.1-VACE (Jiang et al., 2025), which maintain full-sequence bidirectional temporal modeling and drive interaction through explicit source video conditioning. Together, these three groups represent the primary technical paradigms in current interactive video generation, enabling a systematic examination of their capability boundaries. Evaluation protocol, hardware environment, and per-model configuration details are provided in the Appendix F.

5.2 Overall Interactive Performance

Table 2 summarizes the results of seven models across two capability dimensions in four evaluation aspects, with different models demonstrating distinct strengths and weaknesses across metrics.

Within Real-time Interaction Capability, JoyAI and LiveEdit lead on EFA, IS-g, and SEO, with EFA gaps exceeding 15% over the weakest models, reflecting stronger semantic execution. TTFF also spans widely, from 2–3 s to nearly 200 s, showing the differences of generation paradigms. Beyond this initial response gap, SOL and SCL show relatively limited variation across models, while DSR remains consistently low, exposing a gap between initial response and full instruction completion."

Turning to Worldline Generation Capability, SANA-Streaming excels at CF-g and PSR, with Wan2.1-VACE following closely, it further achieves the highest SC and BC scores, suggesting that full-context modeling confers a certain advantage in content preservation. However, CA-HDF remains low across all models, indicating that cross-round history-dependent preservation is a systemic challenge yet to be addressed. On video quality, LiveEdit achieves the highest perceptual quality and aesthetic naturalness, while StreamEdit attains the lowest temporal continuity error.

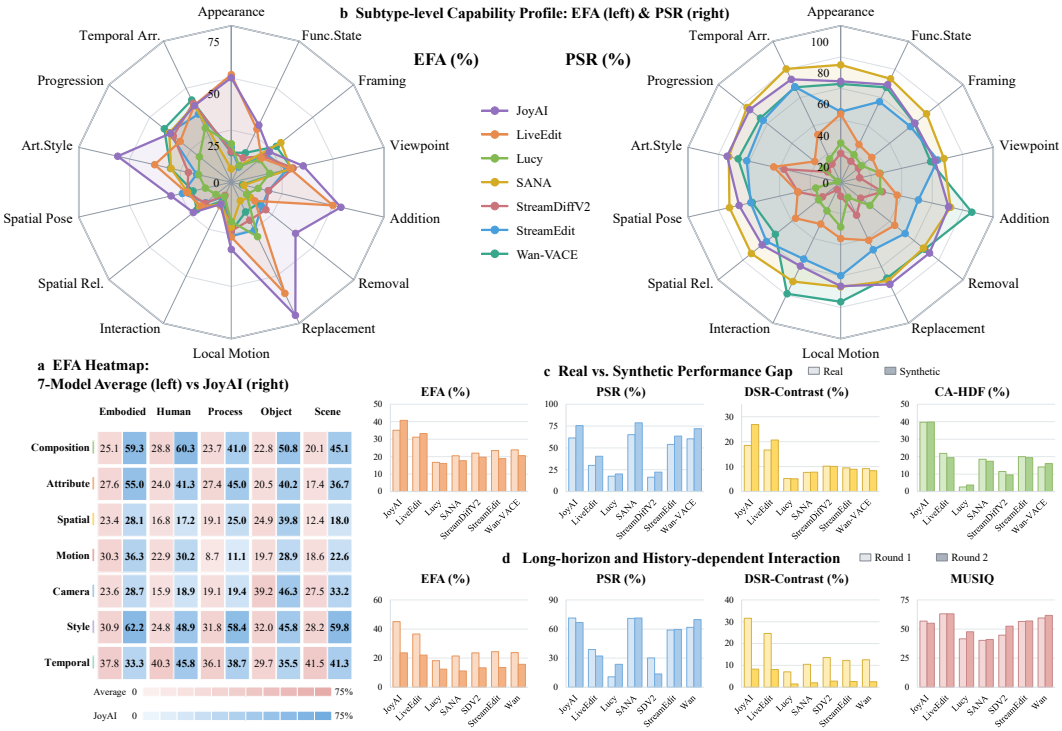


Figure 3: **Fine-grained analysis and long-horizon evaluation on WorldLineBench.** The difficulty landscape varies across interaction types and content contexts, and subtype-level profiles reveal divergent capability shapes across models. Real vs. Synthetic shows asymmetric effects across capability dimensions, while round-wise results under multi-round interaction reveal consistent degradation in semantic fulfillment and responsiveness, with preservation effects varying by model.

Taken together, current models face two pervasive gaps: in Real-time Interaction Capability, the transition from producing an initial response to fully and promptly realizing the instructed semantic change remains broadly unresolved, while in Worldline Generation Capability, cross-round history inheritance consistently fails even when single-round performance is adequate, a limitation that only becomes visible under WorldLineBench's shared-history branching design and real scenarios.

5.3 Fine-grained Capability Analysis

Interaction-type and Content-Context Difficulty Analysis. Fig. 3(a) presents both the seven-model average EFA as a difficulty landscape and JoyAI's individual performance as as a leading reference. Among interaction types, Temporal achieves the highest EFA across all content contexts, as its atomic conditions are relatively straightforward to satisfy. Spatial and Motion remains consistently challenging, with Motion interactions under Process context yield the lowest EFA at 8.65%, possibly due to the difficulty of precise motion modification in structured procedural scenes. Within content contexts, Scene consistently underperforms across most interaction types, especially on Spatial at 12.40%, consistent with the challenge of spatial manipulation in complex scene layouts. Comparing JoyAI against the seven-model average performance, gains are most pronounced on Style and Composition, but remain limited on Spatial and Motion, both of which require sustained global control throughout generation that causal streaming architectures may struggle to maintain.

Subtype-level Capability Profile Fig. 3(b) presents the EFA and PSR profiles of seven models across 14 representative interaction subtypes. LiveEdit and SANA sit at opposite ends of the EFA-PSR spectrum, as LiveEdit achieves strong EFA but consistently low PSR, while SANA shows the reverse, with an expansive PSR contour and depressed EFA, reaching as low as 6.41% on Appearance. JoyAI, in contrast, strikes a balance between the two, performing well on both dimensions across several subtypes. At the subtype level, Interaction registers the lowest EFA across all models,

with a mean of 9.73% and a narrow 7.05% to 12.49% range,, suggesting that multi-agent coordinated motion poses a difficulty intrinsic to the task rather than specific to any model architecture.

Real and Synthetic Video Source Analysis Fig. 3(c) compares seven models on Real and Synthetic videos across four representative metrics, EFA, PSR, CA-HDF, and DSR, with complete results provided in the Appendix. G. Overall, the two video sources impose asymmetric challenges on models across different capability dimensions. On preservation, CF-g and PSR are consistently higher on Synthetic across all seven models, with average gains of 0.33 points and 9.70% respectively, likely because synthetic videos tend to have more regular textures and cleaner object boundaries, making unintended modifications to non-target regions less likely. However, EFA is slightly lower on Synthetic overall, with an average decline of 0.84%, revealing an important divergence: while synthetic videos facilitate non-target preservation, they do not consistently improve instruction-following compliance, with most models showing reduced compliance on synthetic videos. Model-level adaptability varies in direction across metrics and models, indicating that the effect of video source is model-dependent rather than a uniform shift. CA-HDF shows no consistent directional difference between Real and Synthetic, suggesting that cross-round history-dependent preservation remains a shared challenge regardless of video source. Taken together, Real and Synthetic videos elicit different capability profiles, supporting WorldLineBench's design decision to include both.

5.4 Long-horizon and History-dependent Interaction

Fig. 3(d) compares seven models across the first and second interaction rounds on key metrics, revealing systematic capability changes under multi-round interaction.

On semantic fulfillment and responsiveness, all models show consistent degradation in the second round: EFA drops from 27.59% to 16.00% on average and DSR collapses from 16.03% to 3.97%. Alongside this, response timing slows as SOL rises from 3.10 s to 4.83 s and SCL from 12.22 s to 14.34 s, collectively indicating that executing new instructions on top of an established interaction imposes a substantial additional burden that current models struggle to absorb. For preservation, models diverge in the second round: JoyAI, LiveEdit, and StreamDiffusionV2 decline on PSR, while Lucy-Edit and Wan2.1-VACE improve. However, these PSR gains does not necessarily reflect better history-dependent preservation, as CA-HDF remains consistently low across all models, suggesting that retaining prior interaction results across rounds remains challenging independent of strong single-round preservation performance. Regarding video quality, multi-round interaction introduces pervasive degradation: across all paired samples, 71.5% decline on at least two quality metrics and 31.3% degrade across all three simultaneously. At the dimension level, however, temporal continuity and aesthetics do not degrade in tandem, as StreamDiffusionV2 and StreamEdit improve temporal stability while aesthetics worsen, whereas most other models decline on both. This comparison is nevertheless confounded by differing inputs between rounds.

6 Conclusion

We present WorldLineBench, the first benchmark to jointly integrate video-grounded interaction, long-horizon multi-round settings, and process-oriented assessment for streaming interaction. As video generation shifts toward streaming interaction, existing benchmarks remain confined to outcome-oriented, single-round settings with predefined instructions, leaving this emerging paradigm without dedicated evaluation support. WorldLineBench fills this gap by grounding each interaction in an already-unfolding worldline, organizing 1,000 real and 1,000 synthetic videos under a 45-second multi-round protocol with 6,000 video-grounded interactions, and measuring real-time interaction and worldline generation capabilities across five content contexts and seven interaction semantics. Systematic evaluation of representative models reveals that current models remain far from what real-world streaming interaction requires: DSR is consistently low across all models, indicating that fully executing instructions within a user-acceptable time window remains a broadly unresolved challenge. As visual history accumulates, models exhibit consistent degradation in semantic fulfillment, preservation, and video quality, exposing stability challenges particularly under multi-round history-dependent settings. Together, these results provide a capability baseline for streaming interaction and highlight concrete directions for improvement, which we hope will serve the open community in building more reliable streaming video interaction systems.

AI USE STATEMENT

Generative AI tools were used in this work for paper writing, literature retrieval, and synthetic data construction. During paper preparation, AI tools were primarily used for English language polishing, sentence rewriting, and improving clarity of expression, rather than independently drafting paper content. During literature review, AI tools were used to assist in discovering related work, expanding search keywords, and organizing candidate references. All cited works were subsequently verified by the authors against the original sources. Generative AI was also used in constructing the synthetic portion of WorldLineBench, with the detailed generation pipeline and quality-control procedures described in the main paper and supplementary materials. All AI-assisted content was reviewed and confirmed by the authors, who take full responsibility for the final text, data, experimental results, and conclusions.

ETHICS STATEMENT

WorldLineBench contains both real and synthetic video data. Real videos are collected from publicly available datasets under their respective licenses, without additional collection of personally identifiable information, and samples involving apparent privacy concerns, sensitive content, or inappropriate material are excluded during curation. Synthetic videos and interaction instructions may inherit biases from underlying generative models and data distributions, so we apply automated filtering together with manual review to reduce low-quality, inappropriate, or clearly biased samples, while acknowledging that residual biases may remain. WorldLineBench is intended for research on interactive streaming video generation, and its results should not be interpreted as fully representative of all real-world scenarios or populations. Because related technologies may also be misused to create misleading or harmful visual content, we encourage lawful and responsible use and discourage applications involving privacy violations, impersonation, deception, or other real-world harm. Detailed data sources and licensing information are provided in the supplementary material.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide a complete description of the WorldLineBench data construction and evaluation pipeline in the main paper and supplementary material, including data sources and filtering criteria, synthetic data generation, the interaction taxonomy, branching protocol, quality-control procedures, metric definitions, evaluation protocols, and model inference settings. For key steps in data construction and automated evaluation, we further provide the corresponding construction rules, implementation details, and evaluation configurations.

REFERENCES

Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A Space-Time Diffusion Model for Video Generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11. Association for Computing Machinery, 2024. doi: 10.1145/3680528.3687614.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align Your Latents: High-Resolution Video Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575, 2023. doi: 10.1109/CVPR52729.2023.02161.

Kunlin Cai, Rui Song, Jinghuai Zhang, Kaiyuan Zhang, Pranav Bodapati, Alicia Yu, Fnu Suya, Mohammad Rostami, Jiaqi Ma, and Yuan Tian. What-If World: A Causal Benchmark for General World Models in Embodied Scenarios. arXiv preprint arXiv:2605.27589, 2026.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7310–7320, 2024. doi: 10.1109/CVPR52733.2024.00698.

Yinan Chen, Jiangning Zhang, Teng Hu, Yuxiang Zeng, Zhucun Xue, Qingdong He, Chengjie Wang, Yong Liu, Xiaobin Hu, and Shuicheng Yan. IVEBench: Modern Benchmark Suite for Instruction-Guided Video Editing Assessment. In *International Conference on Learning Representations (ICLR)*, 2026.

Yupeng Chen, Penglin Chen, Xiaoyu Zhang, Yixian Huang, and Qian Xie. EditBoard: Towards a Comprehensive Evaluation Benchmark for Text-Based Video Editing Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(15):15975–15983, 2025. doi: 10.1609/aaai.v39i15.33754.

DecartAI Team. Lucy Edit: Open-Weight Text-Guided Video Editing. Technical report, 2025.

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 (ICCV)*, pp. 27713–27724, 2025. doi: 10.1109/ICCV51701.2025.02573.

Tianrui Feng, Zhi Li, Shuo Yang, Haocheng Xi, Muyang Li, Xiuyu Li, Lvmin Zhang, Keting Yang, Kelly Peng, Song Han, Maneesh Agrawala, Kurt Keutzer, Akio Kodaira, and Chenfeng Xu. StreamDiffusionV2: A Streaming System for Dynamic and Interactive Video Generation. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 8, pp. 1656–1669, 2026.

Weixi Feng, Jiachen Li, Michael Saxon, Tsu-Jui Fu, Wenhu Chen, and William Yang Wang. TC-Bench: Benchmarking Temporal Compositionality in Conditional Video Generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4638–4662, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.241.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. TokenFlow: Consistent Diffusion Features for Consistent Video Editing. In *International Conference on Learning Representations (ICLR)*, 2024.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *International Conference on Learning Representations (ICLR)*, 2024.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: Building Automatic Metrics to Simulate Fine-grained Human Feedback for Video Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2105–2123. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.127.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen Video: High Definition Video Generation with Diffusion Models. arXiv preprint arXiv:2210.02303, 2022.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 167283–167308, 2025. doi: 10.52202/085713-5576.

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 (CVPR)*, pp. 21807–21818, 2024. doi: 10.1109/CVPR52733.2024.02060.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3268–3285, 2026. doi: 10.1109/TPAMI.2025.3633890.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. VACE: All-in-One Video Creation and Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17191–17202, 2025. doi: 10.1109/ICCV51701.2025.01597.

Guanlong Jiao, Chenyangguang Zhang, Jia Jun Cheng Xian, Zewei Zhang, and Renjie Liao. StreamEdit: Training-Free Video Editing via Few-Step Streaming Video Generation. In *European Conference on Computer Vision (ECCV)*, volume 17057 of *Lecture Notes in Computer Science*, pp. 1–20. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-37335-9_1.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-Scale Image Quality Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5148–5157, 2021. doi: 10.1109/ICCV48922.2021.00510.

Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. AnyV2V: A Tuning-Free Framework For Any Video-to-Video Editing Tasks. *Transactions on Machine Learning Research*, 2024.

Minghan Li, Chenxi Xie, Yichen Wu, Lei Zhang, and Mengyu Wang. FiVE-Bench: A Fine-grained Video Editing Benchmark for Evaluating Emerging Diffusion and Rectified Flow Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16672–16681, 2025. doi: 10.1109/ICCV51701.2025.01548.

Runjia Li, Moayed Haji-Ali, Ashkan Mirzaei, Chaoyang Wang, Arpit Sahni, Ivan Skorokhodov, Aliaksandr Siarohin, Tomas Jakab, Junlin Han, Sergey Tulyakov, Philip Torr, and Willi Menapace. EgoEdit: Dataset, Real-Time Streaming Model, and Benchmark for Egocentric Video Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16042–16053, 2026a.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable Video Infinity: Infinite-Length Video Generation with Error Recycling. In *International Conference on Learning Representations (ICLR)*, 2026b.

Feng Liang, Akio Kodaira, Chenfeng Xu, Masayoshi Tomizuka, Kurt Keutzer, and Diana Marculescu. Looking Backward: Streaming Video-to-Video Translation with Feature Banks. In *International Conference on Learning Representations (ICLR)*, 2025.

Jinzhuo Liu, Jiangning Zhang, Wencan Jiang, Yabiao Wang, Dingkang Liang, Zhucun Xue, Ran Yi, and Yong Liu. Advancing Narrative Long Video Generation via Training-Free Identity-Aware Memory. arXiv preprint arXiv:2605.18733, 2026a.

Kaiqi Liu, Haoxuan Zeng, Jingqi Liu, Jiacong Fang, Ziqi Cai, Yunyao Mao, Henglin Liu, Yu Sheng, Shuchen Weng, and Boxin Shi. StreamAV-Bench: A Comprehensive Benchmark for Streaming Audio-Video Generation. arXiv preprint arXiv:2608.26336, 2026b.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling Forcing: Autoregressive Long Video Diffusion in Real Time. In *International Conference on Learning Representations (ICLR)*, 2026c.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and Evaluating Large Video Generation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22139–22149, 2024. doi: 10.1109/CVPR52733.2024.02090.

Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. FETV: A Benchmark for Fine-Grained Evaluation of Open-Domain Text-to-Video Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 62352–62387, 2023. doi: 10.52202/075280-2723.

Jinjie Mai, Gordon Guocheng Qian, Willi Menapace, Arpit Sahni, Chaoyang Wang, Ashkan Mirzaei, Runjia Li, Sergey Tulyakov, Bernard Ghanem, Peter Wonka, and Rameen Abdal. EgoPlay: Event-Triggered Video Editing for Egocentric Streams. In *SIGGRAPH Asia 2026 Conference Papers*, 2026. Accepted for publication.

Chenxuan Miao, Yutong Feng, Yi Lu, Yunfeng Yan, Donglian Qi, Shiwei Zhang, Yu Liu, Xi Chen, and Hengshuang Zhao. OmniEdit-Bench: A Comprehensive Benchmark for Instruction-based Video Editing. arXiv preprint arXiv:2608.05049, 2026a.

Yingmao Miao, Pengfei Zhang, Chaoran Xu, Meng Yu, Jing Tang, Xiangxiang Chu, Chao Shen, and Chenhao Lin. Do Video Generators Track the World Across Segments? A Benchmark and Method for World-State Reasoning in Video Continuation. arXiv preprint arXiv:2609.03673, 2026b.

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. FreeNoise: Tuning-Free Longer Video Diffusion via Noise Rescheduling. In *International Conference on Learning Representations (ICLR)*, 2024.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8406–8416, 2025. doi: 10.1109/CVPR52734.2025.00787.

Shanwen Tan, Hao Li, Jingtao Zhang, Xiaosong Jia, Xue Yang, Shaofeng Zhang, and Yanyong Zhang. SWIFT: Prompt-Adaptive Memory for Efficient Interactive Long Video Generation. arXiv preprint arXiv:2605.09442, 2026.

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

Yunze Tong, Mushui Liu, Canyu Zhao, Shiyi Zhang, Didi Zhu, Peng Zhang, Wanggui He, Jinlong Liu, Ying Chen, Hao Jiang, Pipei Huang, and Bo Zheng. InfinityEdit: Infinite Video Editing with a Lightweight Edit-Ignition Adapter. arXiv preprint arXiv:2608.20910, 2026.

Xinyu Wang, Chongbo Zhao, Fangneng Zhan, and Yue Ma. LiveEdit: Towards Real-Time Diffusion-Based Streaming Video Editing. In *European Conference on Computer Vision (ECCV)*, volume 17021 of *Lecture Notes in Computer Science*, pp. 631–649. Springer Nature Switzerland, 2026. doi: 10.1007/978-3-032-37574-2_35.

Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring Video Quality Assessment on User Generated Contents from Aesthetic and Technical Perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20144–20154, 2023. doi: 10.1109/ICCV51070.2023.01843.

Jiangtao Wu, Jiaming Wang, Yiwen He, Yuanxing Zhang, Shihao Li, Dunyuan Liu, Xuedong Zhao, Jialu Chen, Zekun Moore Wang, and Jiaheng Liu. CoVEBench: Can Video Editing Models Handle Complex Instructions? arXiv preprint arXiv:2606.08415, 2026a.

Meiqi Wu, Zhixin Cai, Fufangchen Zhao, Xiaokun Feng, Rujing Dang, Bingze Song, Ruitian Tian, Jiashu Zhu, Jiachen Lei, Hao Dou, Jing Tang, Lei Sun, Jiahong Wu, Xiangxiang Chu, Zeming Liu, and Kaiqi Huang. Omni-WorldBench: Towards a Comprehensive Interaction-Centric Evaluation for World Models. arXiv preprint arXiv:2603.22212, 2026b.

Yicheng Xiao, Wenxun Dai, Xinran Qin, Lin Song, Maoquan Zhang, Hang Xu, Yukang Chen, Yitong Li, Guohui Zhang, Yuan Zhang, Xuying Zhang, Tommy Zhang, Jianlong Yuan, Peihao Li, Shuai Lu, Siming Fu, Chuyang Zhao, Xin Han, Jie Huang, Wenbo Li, Guoqing Ma, Wei Huang, Xiaojuan Qi, Haoyang Huang, and Nan Duan. JoyAI-Video-Edit: Real-Time Open-Ended Video Editing with Autoregressive Diffusion. arXiv preprint arXiv:2608.03974, 2026.

Zhening Xing, Gereon Fox, Yanhong Zeng, Xingang Pan, Mohamed Elgharib, Christian Theobalt, and Kai Chen. Live2Diff: Live Stream Translation via Uni-directional Attention in Video Diffusion Models. arXiv preprint arXiv:2407.08701, 2024.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time Interactive Long Video Generation. In *International Conference on Learning Representations (ICLR)*, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *International Conference on Learning Representations (ICLR)*, 2025.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From Slow Bidirectional to Fast Autoregressive Video Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974, 2025. doi: 10.1109/CVPR52734.2025.02138.

Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Henghui Ding. WBench: A Comprehensive Multi-turn Benchmark for Interactive Video World Model Evaluation. arXiv preprint arXiv:2605.25874, 2026.

Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Ruijie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. ChronoMagic-Bench: A Benchmark for Metamorphic Evaluation of Text-to-Time-lapse Video Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 21236–21270, 2024. doi: 10.52202/079017-0669.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame Context Packing and Drift Prevention in Next-Frame-Prediction Video Diffusion Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 30546–30566, 2025. doi: 10.52202/085713-1024.

Yuyang Zhao, Yicheng Pan, Qiyuan He, Jincheng Yu, Junsong Chen, Tian Ye, Haozhe Liu, Enze Xie, and Song Han. SANA-Streaming: Real-time Streaming Video Editing with Hybrid Diffusion Transformer. arXiv preprint arXiv:2605.30409, 2026.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing Video Generation Benchmark Suite for Intrinsic Faithfulness. arXiv preprint arXiv:2503.21755, 2025.

APPENDIX FOR WORLDLINEBENCH

A CONSTRUCTION OF THE VIDEO CORPUS

A.1 VIDEO CONTENT TAXONOMY

A.1.1 PRIMARY CATEGORIES AND THEIR BOUNDARIES

Streaming interaction is grounded in an evolving visual history, whose organization varies across videos. WorldLineBench therefore groups both real and synthetic videos into five primary categories: Human, Object, Scene, Process, and Embodied. Primary categories describe the dominant observational focus and temporal organization of a clip, while secondary categories provide finer content distinctions.

Human. Videos centered on people and their appearance, expressions, gaze, posture, actions, or interpersonal relations, ranging from fine manipulation to full-body and multi-person activities.

Object. Videos centered on an identifiable non-human subject and its appearance, structure, state, motion, or interaction. Subjects include animals, vehicles, machinery, everyday objects, and natural objects.

Scene. Videos centered on an environment as a whole, including its spatial layout, distributed entities, and ongoing activities. People and objects may be present, but collectively characterize the place rather than serving as a single dominant subject.

Process. Videos centered on an event with clear *semantic progression*, where successive states indicate how the event develops, which stage it has reached, and how later evolution follows from earlier progress.

Embodied. Videos organized around an acting observer and the observer's interaction with the environment, including navigation, manipulation, tool use, vehicle operation, and device control.

Boundaries between primary categories. The categories are distinguished by how the clip organizes its principal content rather than by the mere presence of particular entities. Human, Object, and Scene primarily differ in whether observation is centered on a person, a non-human subject, or the broader environment. The less direct boundaries concern Process and Embodied.

1. **Process versus other content categories.** Motion or state change alone does not define Process. Human, Object, and Scene may all contain substantial temporal variation. A clip is categorized as Process when the progression of an event itself becomes the primary object of observation, such that successive states encode its stage and development.
2. **Embodied versus external observation.** Embodied is not determined by first-person viewpoint alone. It requires the clip to be organized around the observer's actions and their consequences in the environment. If a first-person or moving camera merely observes an external subject, scene, or process, the corresponding non-Embodied category is retained.

Object count, viewpoint, shot scale, and motion magnitude therefore serve only as supporting cues rather than independent classification criteria.

A.1.2 CONTENT OF THE SECONDARY CATEGORIES

The five primary categories are further divided into 24 secondary categories, providing finer distinctions according to the observed subject, environment, process, or mode of embodied interaction.

Human. *Facial Expression & Gaze* covers facial and gaze behavior; *Fine Manipulation* focuses on hand–object coordination; *Solo Full-body Actions* covers posture and whole-body activity; and *Two-person Interaction* and *Group Interaction* capture interpersonal behavior involving two and three-or-more participants, respectively.

Object. The five subclasses distinguish *Animals*, *Transport Vehicles*, *Machinery & Robots*, *Everyday Objects*, and *Non-animal Natural Objects*, separating non-human subjects primarily by their nature and functional role.

Table 3: Primary video categories, their secondary categories, and the number of secondary categories. Counts refer to category types rather than video samples.

Primary	Secondary categories	Count
Human	Facial Expression & Gaze; Fine Manipulation; Solo Full-body Actions; Two-person Interaction; Group Interaction	5
Object	Animals; Transport Vehicles; Machinery & Robots; Everyday Objects; Non-animal Natural Objects	5
Scene	Living Spaces; Services & Culture; Transport & Logistics; Natural Environments; Production & Offices	5
Process	Celestial & Weather Events; Physical Changes; Chemical Reactions; Growth, Development & Decay	4
Embodied	Walking & Navigation; Direct Hand Manipulation; Tool Use; Vehicle Operation; Device Control	5
Total	**5 primary categories and 24 secondary categories**	**24**

Scene. Scene subclasses describe environments by function: *Living Spaces*, *Services & Culture*, *Transport & Logistics*, *Natural Environments*, and *Production & Offices*.

Process. Process is divided into *Celestial & Weather Events*, *Physical Changes*, *Chemical Reactions*, and *Growth, Development & Decay*, covering major forms of temporally evolving phenomena.

Embodied. Embodied subclasses distinguish *Walking & Navigation*, *Direct Hand Manipulation*, *Tool Use*, *Vehicle Operation*, and *Device Control* according to how the acting observer interacts with the environment.

A.2 Source Datasets and Their Licenses

RealVideo's 1,000 clips are derived from six publicly released research datasets. In selecting these sources, we required each to have been released to the research community and to provide an explicit license, terms of use, or official usage statement. The number of clips contributed by each source was not determined by a fixed per-dataset quota. Instead, the final source composition reflects the clips retained after our curation procedure, together with consideration of source diversity and coverage across the five content contexts. The detailed filtering and quality-control criteria are described separately in Sec. A.4.1. Table 4 summarizes the resulting source composition, together with the corresponding licenses and main obligations.

Table 4: Source datasets of RealVideo and their licenses.

Source	Original scale	Clips used	Categories	License	Main obligations
LaSOT	1,400 tracking sequences, 70 object classes	359	Object, Scene, Embodied, Process	Free for personal, academic, or educational use (LaSOT official site)	Cite original paper; use for non-commercial research purposes only
FineVideo	Over 43,000 YouTube videos	336	All five	CC-BY, under FineVideo Terms of Use	Attribute per-video license; redistribution requires including ToU and user agreement
ChronoMagic-Bench	1,649 time-lapse reference videos	169	Process	Apache 2.0 (as declared by authors)	Preserve license/attribution notices; cite original paper
Walking Tours	Long single-take egocentric city walks	65	Scene	CC-BY	Attribution; indicate changes
EPIC-KITCHENS-100	100 hours, 700 egocentric videos, 45 kitchens	47	Embodied	CC BY-NC 4.0	Attribution; indicate changes; non-commercial only
MEVA	About 328 hours, 29 camera views	24	Scene	CC BY 4.0	Attribution; indicate changes

LaSOT (359 clips). LaSOT is a large-scale single-object tracking benchmark. Its original authors searched YouTube for candidate videos, collecting over 5,000 and filtering them down to 1,400

sequences, each following a single target over a long, continuous span (**?**). Of the 359 clips we use, 188 are assigned to Object, 88 to Scene, 62 to Embodied, and 21 to Process. The LaSOT paper does not state that the underlying YouTube videos carry a Creative Commons license; our usage basis is instead the dataset's official download page, which states that LaSOT can be freely used for personal, academic, or educational purposes. We use it accordingly for non-commercial research and cite the original paper.

FineVideo (336 clips). FineVideo, released by Hugging Face, comprises over 43,000 YouTube videos licensed under CC-BY. These videos were obtained through the YouTube-Commons project, and the dataset provides provenance information for each video. Our 336 clips span all five categories: 200 Human, 91 Embodied, 23 Scene, 12 Object, and 10 Process. FineVideo's terms of use require that any use of all or part of the videos comply with their original licenses, including attribution, and that anyone who hosts or redistributes the data include the FineVideo Terms of Use and require users to agree to them.

ChronoMagic-Bench (169 clips). ChronoMagic-Bench is a benchmark for text-to-time-lapse generation whose reference set consists of real time-lapse videos depicting physical, biological, and environmental processes. Our 169 clips are all assigned to Process. The dataset is released by its authors on Hugging Face under the Apache 2.0 license; we use its videos as distributed and preserve the corresponding license and attribution notices.

Walking Tours (65 clips). Walking Tours consists of hours-long, single-take, first-person videos of city walks, captured in 4K at 60 fps and released on YouTube under a CC-BY license. Our 65 clips are all taken from the videos officially released as part of this dataset and are assigned to Scene. In accordance with CC-BY, we attribute the source and describe the changes we made.

EPIC-KITCHENS-100 (47 clips). EPIC-KITCHENS-100 contains 100 hours of unscripted, first-person daily kitchen activity recorded by voluntary participants with head-mounted cameras in their own homes. Our 47 clips span 34 different participants and are all assigned to Embodied. The dataset is released under CC BY-NC 4.0, which requires attribution, indication of changes, and non-commercial use only; this is the most restrictive license among our six sources and is why RealVideo as a whole is limited to non-commercial research.

MEVA (24 clips). MEVA is a multi-view video dataset in which roughly 100 actors performed scripted and unscripted activities at an access-controlled facility, recorded simultaneously by 29 cameras. Data collection was overseen by an Institutional Review Board (IRB), and all participants gave informed consent. The dataset is released under CC BY 4.0. Our 24 clips come from 24 distinct camera views and are all assigned to Scene.

Compliance. Given these licenses, we use the source material as follows. First, attribution: all six datasets are cited in this paper, and the provenance table released with the data records, for every clip, its source dataset, source video identifier, and the start and end times of the trimmed segment. Second, indication of changes: our processing of the original videos is limited to temporal trimming, spatial resizing, re-encoding, and, where applicable, removal of the audio track, and these changes are listed in the data documentation. Third, benchmark-level release policy: to provide a uniform usage policy across the benchmark, we release RealVideo for non-commercial research use only, consistent with the most restrictive source license included in the collection (CC BY-NC 4.0 for EPIC-KITCHENS-100). Each clip nevertheless retains the license and attribution requirements of its original source. Fourth, redistribution: on release, access to the data will require users to accept the FineVideo Terms of Use before it is provided.

A.3 Synthetic Video Generation

WorldLineBench asks whether a model, after watching a video history, can follow a subsequent editing instruction while preserving the people, objects, and scene states established so far. For this question to be answerable, the source video itself must stand up to re-examination: a person must remain recognizable, object relations must stay visible, and the ongoing activity must still be underway when the instruction arrives. Without these conditions, even a plausible-looking continuation tells us little, since we cannot tell whether the model inherited the past or simply repainted a similar scene. We therefore design synthetic sources around two goals—traceable states and continuing activity—so that every later instruction has a well-defined target. WorldLineBench contains 1,000

real and 1,000 synthetic videos, with 200 synthetic videos in each of Human, Object, Scene, Process, and Embodied. This section describes how each synthetic video goes from a content specification to two generation prompts, and how the two clips are joined into a 45-second source.

A.3.1 Source Content and Prompt Design

Designing a video begins with the opening frame: who and what is visible, and where each entity stands relative to the others. We then separate the states that may change from those that must persist throughout the clip. A person's clothing and identity, an object's count and support, and a scene's landmarks and camera position form the history that can later be checked; position, pose, local motion, and the stage of an event leave room for subsequent interaction. A prompt therefore does more than describe an action—it spells out the conditions of that action: where the participants start, what each one holds, how they connect to the environment, and what must remain visible once the change has taken place. In multi-person scenes, we favor clear and independent motion paths, avoiding the identity and handoff ambiguities that crossings and repeated exchanges create.

The five categories rely on different observable evidence to support this design. Human preserves identity, body motion, and relations between people; Object preserves a subject's silhouette, parts, support, and motion pattern; and Embodied ties first-person action to the environment's response. Process organizes its timeline around the semantic progress of a single event—how the current stage came about, what changes within it, and how that stage constrains what comes next—whether the event is carried by people, objects, or the environment. Scene keeps the place itself at the center: the local layout and landmarks are readable from the opening, while pedestrians passing by, animals pausing, foliage swaying, or ripples on water unfold concurrently at a natural pace. A scene video does not have to synthesize one event from beginning to end; this keeps it temporally alive without reducing it to a static portrait of a single object.

Each specification is finally written as two prompts that join head to tail. Both keep the same setting, subject scale, and camera constraints. Activities with meaningful stages are narrated in temporal order, while simple concurrent activities within a place follow their natural rhythm. Every prompt also states the visible content, ambient sound, and music requirement, with no music by default. The continuation may be sketched in advance, but the version actually submitted must be finalized against what the opening clip truly generated.

A.3.2 Full-Model Generation and Continuation

The second half of a 45-second video should inherit the midpoint state the model actually produced, not the one we expected. We therefore split generation into two requests. Full, unquantized MiniMax H3 is served through SGLang, and each request runs jointly on eight A100 40GB GPUs. The opening clip is generated from text for 22.5 seconds. We review the entire clip, confirm the motion and relations at its end, and then take the actually delivered last frame as the image condition at the start of a second, 22.5-second generation. The continuation prompt retains the people, objects, positions, and activity stage established in the opening, and describes only what can plausibly develop from that state.

The last frame tells the model the visible layout and appearance, but it cannot say who an occluded person is, who held what earlier, or in which direction something was moving; these evidence-backed historical conditions must still be stated in text. If the opening itself has lost a critical identity or relation, we revise the opening first, rather than letting the continuation presuppose an end state that was never reached.

Let R_1 and R_2 be the zero-indexed raw outputs (549 frames each, 1344×768 at 24 fps), with $K = R_1[539]$ the continuation's sole reference image. Using half-open frame intervals, we assemble

$$V = R_1[0{:}540) \,\|\, R_2[1{:}541),$$
$$|V| = 1{,}080, \qquad T = \frac{|V|}{24} = 45.000\,\text{s}.$$

Here $\|$ denotes direct temporal concatenation; the second slice omits its conditioned first frame. The 22.5-second join uses no transition, interpolation, or speed adjustment. Audio is trimmed at the same boundary, with 50-ms fades on both sides.

A.3.3 Revision Grounded in Generated Evidence

The two prompts are not simply written in advance and submitted back to back. The opening first undergoes technical checks and visual review, which record both its expected and its observed end state; only an acceptable opening is used to finalize the continuation. The continuation and the assembled video then go through a further round of checks: whether subjects are retained, relations remain consistent, the join is clean, and the ongoing activity is suitable as a source for later editing. When a deviation occurs, the revision targets the particular action or dependency that failed, rather than piling on constraints indiscriminately. Section A.4 describes the video-quality checks in detail.

The recorded revision of `human_104`, reproduced in Box A.3.3, makes this principle concrete. Its accepted opening establishes two gardeners working the same green hose in a fixed wide shot. The initial continuation asked the yellow-jacket gardener to attach the rear coupling and open the valve, and review flagged the coupling and water-flow stages as failed. Rather than adding constraints everywhere, the revision kept the accepted opening, the terminal-frame condition, and the continuation seed, and changed only the description of the ensuing activity. In other words, the edit pulled the continuation back inside the two-person, single-hose relation that the opening had established and that the frame actually shows. The final review accepted the revised full video, noting that identities, hose ownership, spatial relations, and the join all remained usable.

Revision Record: `human_104`

Opening (accepted). A yellow-jacket gardener stands beside a hose reel, and a blue-jacket gardener stands at the far end of the same green hose along a garden path, with a fixed wide shot covering both people and the full hose route. Review of the opening confirmed that the yellow-jacket gardener remained at the reel and the blue-jacket gardener stood at the far end of the extended hose, with no handoff.

Initial continuation (flagged). The yellow-jacket gardener is asked to "connect the already-present rear coupling to the faucet," and then to "slowly open the valve so that water begins to flow from the far end." The revision record flagged the coupling and water-flow stages as failed, without pinpointing a specific frame or geometric error.

Revised continuation (accepted). The yellow-jacket gardener "continues to work the reel slowly," the blue-jacket gardener "takes two small steps forward so the same hose straightens slightly," and the two then make small adjustments at their respective ends. Coupling, valve operation, water flow, and handoff are explicitly excluded.

Videos that pass the full-clip review are archived as synthetic sources and selected for SyntheticVideo according to category coverage. For each video, the archive keeps the two submitted prompts, the image condition, the revision records, the review decisions, and their correspondence to the final media. Descriptions used for subsequent editing are written from what is actually visible in the accepted video, not copied from the outcome the generation prompt intended.

A.4 Quality Control and Data Curation

A.4.1 Real-Video Quality Screening

We adopt a three-stage quality-control pipeline for real videos, consisting of automatic technical screening, VLM-based semantic review, and human verification. The goal is not only to remove low-quality videos, but also to ensure that each retained clip forms a continuous, stable, and interaction-ready visual history. We first construct candidate pools from the six source datasets and, when structural metadata such as object categories, participants, camera views, or other annotations are available, sample across these dimensions to reduce over-representation of any single content pattern. We then perform automatic technical screening and retain only videos from which at least 45.5 seconds of continuous footage can be extracted. Candidate segments are checked for shot changes, scene transitions, obvious temporal discontinuities, large text or graphical overlays, black frames, severe blur, and prolonged static content. Segments containing artifacts that disrupt the continuity of the visual history or substantially hinder subsequent state understanding are removed at this stage.

Candidates that pass the technical screening are further evaluated by a VLM at the semantic level. Unlike the previous stage, which focuses on measurable technical constraints, this stage assesses whether the complete clip constitutes a valid dynamic worldline for benchmark construction. We

provide the full candidate video to the VLM and use a fixed prompt shown in Box A.4.1 that asks it to examine a predefined set of criteria, including whether the main subjects and scene remain continuous throughout the clip, whether unexplained semantic discontinuities occur, whether the video contains clear and sustained state evolution, whether the principal activity or process can be consistently understood over time, and whether the established visual history can support subsequent interactions that are reasonable and verifiable. For each criterion, the model returns `Yes`, `No`, or `Uncertain`, together with a brief justification, and assigns the sample to one of three outcomes: `Pass`, `Reject`, or `Review`. Samples for which all required criteria are satisfied are admitted to the candidate pool, samples with any clearly violated criterion are rejected, and samples containing uncertain judgments or conflicting evidence are forwarded to human review.

VLM Prompt

Watch the full video and judge whether it is suitable as a real source worldline for an interactive video benchmark, based on: continuity of subject and scene, presence of clear state evolution, temporal coherence of the main activity, and whether the visual history supports a verifiable follow-up interaction.

Output `Pass`, `Reject`, or `Review` with a brief justification, based on the full video rather than a single frame.

Human verification is used both to resolve borderline cases and to detect potential failures among automatically accepted samples. All candidates labeled `Review` are inspected manually, and we additionally audit a random 10% subset of samples labeled `Pass`. Two Human reviewers verify temporal continuity, the trackability of subjects, scenes, and key states, the sufficiency of dynamic state evolution, and the suitability of the clip as the visual history for subsequent grounded interaction. If the selected temporal interval contains a local defect but a more suitable continuous interval exists within the same source video, the clip may be re-trimmed and re-enter the corresponding screening stages. When the automatic judgment and human review disagree, the human decision is used as the final outcome.

After this three-stage procedure, only clips that satisfy the duration requirement, basic visual-quality constraints, temporal and semantic continuity, sufficient state evolution, and interaction suitability are retained in the final RealVideo subset. This design separates low-level technical defects, semantic validity as a worldline, and borderline cases that may be missed by automated screening, rather than relying on a single automatic model for all quality decisions.

A.4.2 Synthetic-Video Quality Screening

SyntheticVideo follows the same three-stage quality-control framework as RealVideo, consisting of automatic technical screening, VLM-based semantic review, and human verification. We therefore omit the shared low-level screening details here and focus on the criteria specific to generated videos. Unlike RealVideo, where the main objective is to identify a continuous and interaction-ready source worldline from existing footage, SyntheticVideo must additionally verify whether the generated result faithfully realizes the predefined generation prompt and forms a stable dynamic trajectory suitable for subsequent interaction construction.

During VLM-based review as shown in Box A.4.2, the model is provided with both the original generation prompt and the complete generated video. The review focuses on prompt fulfillment and temporal validity: whether the main subject, scene, and overall content are consistent with the prompt; whether the required action, process, or state evolution is correctly realized; whether the subject, background, and key visual attributes remain stable over time; and whether the resulting video exhibits a clear and sustained state evolution that can support a verifiable follow-up interaction. Videos with evident prompt mismatch, identity drift, unexplained object disappearance or replacement, structural distortion, abrupt state changes, or other generation failures are rejected even when their overall visual appearance is otherwise plausible.

VLM Prompt

Given the original generation prompt and the full generated video, judge whether it is suitable as a synthetic source worldline for an interactive video benchmark, based on: consistency with the generation prompt, correct realization of the required action or state evolution, temporal stability of the subject and scene, and whether the resulting visual history supports a verifiable follow-up interaction.

Output `Pass`, `Reject`, or `Review` with a brief justification, based on the full video rather than a single frame.

Human verification follows the same protocol as for RealVideo. All samples assigned to `Review` are manually inspected, and an additional random 10% of automatically accepted samples are audited to identify potential false acceptances or overlooked generation artifacts. Human reviewers compare the generated video against its original prompt and verify semantic fulfillment, temporal stability, state evolution, and suitability for downstream grounded interaction. When automatic and human judgments disagree, the human decision is used as the final outcome.

Any synthetic sample rejected during automatic screening, VLM review, or human verification is regenerated under the same content configuration and re-enters the complete quality-control pipeline. The replacement preserves the original content context, target category, and predefined coverage setting, preventing quality filtering from altering the intended distribution of SyntheticVideo. Only samples that satisfy prompt consistency, visual and temporal stability, sufficient state evolution, and interaction suitability are retained in the final subset.

B Construction of the Interaction Instructions

B.1 Interaction Types and Instruction Allocation

Every interaction in WorldLineBench is characterized by two taxonomies: the video content context describes the visual history already established at the moment of intervention, and the interaction type describes what the user wants to change in the subsequent generation. The former supplies the context, the latter the goal; the same interaction can act on different content contexts, and the same context can host many kinds of interactions. On top of the five content contexts, we define seven interaction types: Attribute, Camera, Composition, Motion, Spatial, Style, and Temporal. An instruction is classified by the main object it asks to change and the verifiable outcome it requires, not by the verb it happens to contain, nor by incidental changes in the frame. When a single instruction involves several changes, we identify one primary interaction target for counting and treat the remaining requirements as auxiliary constraints, so that nothing is counted twice. This section explains what each interaction type changes, how neighboring types are separated, and how instructions are allocated across the five content contexts; the 26 interaction subtypes derived from them are described in Section B.2.

Attribute changes a persistent attribute or state of an entity or the environment already present in the frame, such as appearance, functional state, or environmental conditions. After the interaction, the existing content persists with its new attribute, so the test is whether the target state is established and maintained. **Composition**, by contrast, changes which entities and parts the frame contains: adding, removing, replacing, or altering part–whole relations. Both can change a local region of the picture, and the distinction lies here: Attribute asks the existing object to persist with new attributes, whereas Composition asks the existence, identity, or composition of objects themselves to change.

Motion changes how entities or environmental elements subsequently move and interact, including initiating and switching actions, adjusting motion parameters, and coordinating dynamics among multiple entities. **Spatial** changes where existing entities stand in the world, their relative layout, or their spatial pose, aiming at a specified spatial configuration. Both can involve displacement; the test is what the instruction asks to verify. If it asks how the motion unfolds, it belongs to Motion; if it asks whether the target position, orientation, or inter-entity relation holds, it belongs to Spatial.

Camera changes the way the visual world is observed, including framing, viewpoint, camera movement, and observing identity, so that the same world is presented differently. A change of perspective alters the apparent size and on-screen position of objects, but the target of Camera is the manner of observation, not the physical size or spatial configuration of entities in the world. **Style** changes how

Table 5: Planned instruction counts per video content context and interaction type. Values are numbers of instructions; column order matches the legend of the statistics figure in the main text.

Content context	Attribute	Camera	Composition	Motion	Spatial	Style	Temporal	Total
Human	240	120	180	360	180	90	30	1,200
Object	300	80	180	340	150	90	60	1,200
Scene	300	100	200	150	200	150	100	1,200
Process	150	80	150	150	100	150	420	1,200
Embodied	240	180	180	300	180	90	30	1,200
Total	**1,230**	**560**	**890**	**1,300**	**810**	**570**	**640**	**6,000**

the same content is visually rendered, applied either to the whole frame or to a designated region. It differs from Attribute in that its target is the presentation, such as rendering, medium, and tonality, rather than the properties of objects or the environment themselves; and it differs from Camera in that it leaves the geometry of observation untouched.

Temporal targets activities or processes with semantic progress, changing how they advance, start or stop, reverse, or align with one another in time. It is concerned with how an event develops along its semantic progress, not merely with how fast objects move in the frame. Adjusting the direction, magnitude, or period of a motion therefore belongs to Motion; adjusting when a staged process starts, pauses, resumes, or reaches its next stage belongs to Temporal. The applicability of Temporal is decided by the specific instruction and by a recognizable process in the video; it is not restricted to the Process context.

Table 5 reports the planned instruction counts of the seven interaction types across the five video content contexts, consistent with the interaction-allocation panel of the statistics figure in the main text. The full configuration assigns 400 videos to each content context and three planned interaction events per video, giving 1,200 planned instructions per context and 6,000 in total. All values are *planned* instruction counts, not collected samples; dividing a cell by its row's 1,200 yields the percentage of the corresponding bar segment in the figure. Column totals give the planned count of each interaction type across the five contexts.

The allocation reflects the design emphasis of different content contexts. Process receives 420 Temporal instructions, 35% of its quota, echoing its design around the semantic progress of events. Human, Object, and Embodied receive 360, 340, and 300 Motion instructions, accounting for 30.0%, 28.3%, and 25.0% of their respective rows; Scene is dominated by Attribute (300, 25.0%) together with Composition and Spatial (200 each, 16.7%). These proportions describe the planned structure of instruction coverage. How concrete instructions are grounded in the visible state of each video, and which effective sample denominators apply during model evaluation, are deferred to Section B.3 and the experiment sections.

B.2 Interaction Subtypes and Instruction Allocation

Section B.1 assigns each instruction to one of seven interaction types according to its main target; here we refine each type into verifiable operations, yielding 26 interaction subtypes. A subtype states what must change in the continuation, while the five content contexts, namely Human, Object, Scene, Process, and Embodied, state the source-video context in which that change takes place. Crossing the two taxonomies yields planning units for instructions, not new interaction categories. So that counts remain additive, each instruction is counted under exactly one subtype by its main target; any other simultaneous requirements are kept as constraints rather than counted again.

Attribute (3 subtypes). *Appearance* changes a persistent visible property of an existing entity, such as its color, material, or form, while the entity's identity is preserved. *Functional State* specifies an operating, usage, or otherwise functionally meaningful sustained state of an existing entity; what is checked is whether the target state is established after the interaction, not how far an entire event has progressed. *Environmental Conditions* changes scene conditions such as lighting or weather. These conditions act on the environment itself, rather than applying an overall tonal treatment to the rendered image.

Camera (3 subtypes). *Framing & Zoom* changes the field of view and the apparent scale of a target in the frame. *Viewpoint & Camera Motion* changes the observation position, direction, or

camera trajectory. *Perspective Switch* changes the observing identity, letting the same scene alternate between an actor's first-person view and an external third-person view. All three share the same target, namely how the world is observed: none of them requires changing an entity's physical size or position in the world to produce a similar-looking frame.

Composition (4 subtypes). *Addition* requires an entity or component absent from the established scene to appear in the continuation; *Removal* requires specified existing content to disappear; *Replacement* substitutes a new entity for an existing one. *Combination & Splitting* instead changes part–whole structure, combining components into a whole or separating a whole into components. The first three concern the appearance, disappearance, and replacement of content; the fourth concerns how content is constituted as a whole.

Motion (5 subtypes). *Local Motion* controls the action of a local region of an entity, while *Global Motion* controls the action or motion state of the entity as a whole. *Motion Parameters* adjusts speed, direction, trajectory, amplitude, or periodicity while preserving the basic action type. *Interaction* requires dynamic action or coordination among multiple entities, not merely a static spatial arrangement. *Environmental Dynamics* controls the motion of background objects or environmental elements; sustained lighting and weather conditions remain under Attribute.

Spatial (3 subtypes). *Absolute Position* specifies the target position of an existing entity relative to the scene or the frame. *Spatial Relations* specifies distances, order, or layout among entities. *Spatial Pose* specifies an entity's own orientation or pose in space. What is verified is, respectively, the location of a single entity, the configuration among several entities, and the entity's own spatial orientation; how the entity travels to that configuration is not the question this subtype asks.

Style (4 subtypes). *Artistic Stylization* changes the overall artistic expression or rendering system, *Imaging Medium* changes the overall medium and its imaging characteristics, and *Color & Texture* changes overall color, tonal contrast, or image-level texture. *Local Stylization* changes the visual presentation of only a designated object or region, leaving the rest untouched. To avoid double counting, every local-style instruction is assigned to Local Stylization by its local target, with the target region and the chosen style, medium, or tonality recorded only as further descriptors rather than counted again under the three global subtypes. Likewise, a change in image-level texture does not by itself imply a change in an entity's material or in the scene's lighting.

Temporal (4 subtypes). *Progression* controls how an event with semantic progress advances toward a target stage, covering its rate, duration, stage, and degree of completion. *Start & Termination* controls whether and when a process starts, pauses, resumes, or terminates. *Reversal* requires future states to evolve along the opposite direction of semantic progress, without rewriting frames that have already been generated. *Temporal Arrangement* controls the order, synchronization, or overlap of multiple processes. Whether Temporal applies is decided by a recognizable semantic progress and inter-process relations, so these instructions are not confined to the Process context; conversely, merely adjusting the speed or trajectory of a motion remains Motion Parameters.

Table 6 expands the context-by-subtype allocation shown in Fig. 2(d) into exact counts. Each content context plans 1,200 instructions, 6,000 in total. Each row gives the planned count of one interaction subtype across the five content contexts; the rows within a type sum to its subtotal, and the seven subtotals agree item by item with the allocation in Table 5. A zero, shown as a white circle in the figure, only means that the plan allocated no instruction to that cell; it says nothing about whether the interaction is semantically applicable there, nor about any model's capability. All values are *planned instruction counts*, not counts of collected or evaluated samples.

The same subtype can appear in several content contexts, though with different weights. In Process, for example, 280 of the 420 Temporal instructions target Progression, 110 target Start & Termination, and 30 target Reversal, echoing this context's design emphasis on semantic progress. Human receives 220 Local Motion instructions and Embodied receives 90 Perspective Switch instructions, corresponding to the local actions and the changes of observing identity available in their respective content. These differences describe planned coverage in task design, not an inherent ordering of content contexts, and not observed model performance.

Table 6: Planned instruction counts by interaction subtype and video content context. English subtype labels match the statistics figure. Group subtotals agree with Table 5.

Interaction subtype	Human	Object	Scene	Process	Embodied	Total
Attribute (3)						
A01 Appearance	210	200	120	60	100	690
A02 Functional State	0	70	100	70	110	350
A03 Environmental Conditions	30	30	80	20	30	190
Attribute subtotal	**240**	**300**	**300**	**150**	**240**	**1,230**
Camera (3)						
V01 Framing & Zoom	30	30	50	40	40	190
V02 Viewpoint & Camera Motion	50	50	50	40	50	240
V03 Perspective Switch	40	0	0	0	90	130
Camera subtotal	**120**	**80**	**100**	**80**	**180**	**560**
Composition (4)						
C01 Addition	60	60	60	40	50	270
C02 Removal	60	60	60	40	50	270
C03 Replacement	60	60	60	40	50	270
C04 Combination & Splitting	0	0	20	30	30	80
Composition subtotal	**180**	**180**	**200**	**150**	**180**	**890**
Motion (5)						
M01 Local Motion	220	0	0	0	80	300
M02 Global Motion	60	100	30	30	80	300
M03 Motion Parameters	0	100	20	20	0	140
M04 Interaction	50	100	60	60	80	350
M05 Environmental Dynamics	30	40	40	40	60	210
Motion subtotal	**360**	**340**	**150**	**150**	**300**	**1,300**
Spatial (3)						
S01 Absolute Position	40	50	50	30	30	200
S02 Spatial Relations	120	70	120	50	130	490
S03 Spatial Pose	20	30	30	20	20	120
Spatial subtotal	**180**	**150**	**200**	**100**	**180**	**810**
Style (4)						
Y01 Artistic Stylization	30	30	50	50	30	190
Y02 Imaging Medium	15	15	25	25	15	95
Y03 Color & Texture	15	15	25	25	15	95
Y04 Local Stylization	30	30	50	50	30	190
Style subtotal	**90**	**90**	**150**	**150**	**90**	**570**
Temporal (4)						
T01 Progression	15	30	45	280	15	385
T02 Start & Termination	10	10	15	110	10	155
T03 Reversal	5	5	5	30	5	50
T04 Temporal Arrangement	0	15	35	0	0	50
Temporal subtotal	**30**	**60**	**100**	**420**	**30**	**640**
Total (26 subtypes)	**1,200**	**1,200**	**1,200**	**1,200**	**1,200**	**6,000**

B.3 Grounding Interaction Content and Intervention Times

Sections B.1 and B.2 fix the benchmark's coverage of interaction types. This section explains how a concrete instruction is grounded in a particular worldline. An interaction is valid only for what is visible at its nexus point: the target must be present and identifiable, its current state known, and the requested outcome verifiable in the continuation. A vision language model (VLM) first proposes candidate nexus points from the video, and human reviewers check them against the moving image before an instruction is finalized. This procedure is applied to the source A for the two interactions that branch from it and, after E_1 has been executed, to each evaluated model's generated branch B for E_{2B}. Across the 2,000 source worldlines spanning five video content contexts, each evaluation has three interaction events per video, giving 6,000 events in total.

B.3.1 Constructing Candidates from Visual History

For each 45-second source video, the VLM examines its visual history and proposes candidate nexus points. Human reviewers then inspect the full video and the continuous interval around each proposed time, identifying the target and current state and correcting or rejecting points where the target is absent, ambiguous, or about to leave the frame. The review follows a content timeline of persistent entities and identifying features, parts and support relations, spatial and contact changes, event stages, camera cuts, and occlusions. Object ownership inferred from a single frame, fully hidden parts, and unobserved causal states are not treated as known conditions.

An interaction candidate is defined by the observed state of its parent video at the nexus point, a uniquely locatable target, one principal operation, and an observable desired outcome. A color change, for example, requires an existing object with a known appearance; an action change requires a visible actor and current motion; process control requires a recognizable stage of the same event. The evidence differs by video content context: Human requires stable person identity and visible body regions; Object requires real parts and support; Scene requires a readable layout and distributed activity; Process requires semantic progress; and Embodied requires a coherent relation among viewpoint, body motion, tools, and contact. These checks determine whether an interaction is applicable without restricting any interaction type to one content context.

Human reviewers discard a candidate anchored at a VLM proposed point if its target or reference cannot be distinguished, the requested result already holds, the operation contradicts known contact or geometry, or the outcome cannot be judged in the ensuing generation. Each instruction specifies one main change; entities outside the target, protected text, and relations to preserve are recorded separately. Only visually verified candidates enter selection guided by the interaction type and subtype coverage targets of Sections B.1 and B.2. This ordering prevents a quota from creating an unsupported target.

B.3.2 Choosing Times and Branches

Let A denote the original source worldline and m an evaluated model. The first interaction event E_1 produces that model's branch $B^{(m)}$. In the second round, E_{2A} branches independently from A to $D^{(m)}$. Only after $B^{(m)}$ has been generated do we select $E_{2B}^{(m)}$ from its observed content and use it to produce $C^{(m)}$:

$$B^{(m)} = E_1^{(m)}(A), \qquad D^{(m)} = E_{2A}^{(m)}(A), \qquad C^{(m)} = E_{2B}^{(m)}\big(B^{(m)}\big).$$

The superscript indicates model execution; the E_1 and E_{2A} instructions are grounded in the shared source A, whereas the $E_{2B}^{(m)}$ instruction is grounded in the realized $B^{(m)}$. Thus its target, content, and time may differ across models. The three events may have different interaction types or subtypes, and branch identity is not an interaction type.

The VLM proposes a nexus point at time t_e when the target appears stable and identifiable, its current state is clear, and the requested result has not already arisen. A reviewer determines the final time after checking cuts, occlusions, contact transitions, and process boundaries around the proposal. The record distinguishes intervention time from expected effect onset and completion; a delayed effect may begin after the instruction is issued. We favor placing the first interaction earlier and the second-round interaction later, while constraining each branch through $B^{(m)}$ to $0 < t_{2B}^{(m)} - t_1 \leq 15\,\mathrm{s}$. This interval does not pin every intervention to the same second. The video interval inspected to justify a candidate is distinct from the standard metric windows defined in Appendix C.

B.3.3 Conditioning the Second Interaction on the First

The second interaction on the B branch is selected only after the evaluated model has completed the generation of $B^{(m)}$ under E_1. The VLM proposes nexus points in this realized video, and a human reviewer checks the actual target, state, and timing before choosing the $E_{2B}^{(m)}$ instruction. Its content is thus determined by the model's observed continuation. Whether E_1 succeeded is assessed separately from whether the subsequent instruction is fulfilled.

Reviewers first distinguish a persistent result that is visible in $B^{(m)}$ from a bounded action that has ended without a lasting state. They then check whether the proposed $E_{2B}^{(m)}$ target exists in $B^{(m)}$, remains visible at its nexus point, and is compatible with the observed identities, count, space, support, contact, and event stage. If an expected E_1 result did not appear, a candidate that requires it is rejected. An object absent from the realized $B^{(m)}$ cannot serve as a target; a part actually introduced or exposed there can.

This check separates semantic dependence from mere compatibility. If the target, location, and evaluation of $E_{2B}^{(m)}$ would remain unchanged without the observed E_1 result, the two interactions are independent but compatible. If the target or its evaluation is defined by a persistent result visible in $B^{(m)}$, $E_{2B}^{(m)}$ is semantically dependent. In either case, the instruction acts on $B^{(m)}$ and must

preserve any E_1 result that actually persists. For example, a request to pull the active loop on the right panel of a split crochet fabric is eligible only if the generated $B^{(m)}$ visibly contains that panel and loop. Otherwise, the proposed target is absent and that instruction cannot be selected. The procedure and human review criteria are shared across models, but the resulting $E_{2B}^{(m)}$ instructions need not be identical.

B.3.4 Finalizing the Instruction Record

Each selected interaction is checked against its actual parent video: A for E_1 and E_{2A}, and the generated $B^{(m)}$ for $E_{2B}^{(m)}$. An invalid VLM proposal is corrected or replaced after human review rather than meeting a quota by changing what the video visibly contains. The final instruction remains a short request specifying when, which target, and what principal operation. A separate structured record retains the observed parent state, desired outcome, effect timing, visual evidence, preservation constraints for other entities, and the observed basis on which $E_{2B}^{(m)}$ depends on or is compatible with E_1. This keeps the request executable while allowing fulfillment and preservation to be evaluated independently. Appendix B.4 describes the structural checks and visual review used before records are finalized.

B.4 Interaction Instruction Quality Control

For each interaction instruction, we adopt a three-stage quality-control procedure consisting of rule-based checking against the video content description, VLM-based visual verification, and human review. The objective is not to evaluate the instruction text in isolation, but to verify that the interaction type, instruction content, and nexus point are jointly consistent with the visual state already established in the current worldline. A valid interaction event should therefore be visually grounded, executable from the current state, and verifiable in the subsequent observation window.

The first stage performs rule-based checking using the content description associated with each video. This description records the major subjects, objects, scene context, ongoing actions, state changes, and relevant relations, and is used to eliminate clearly incompatible candidate instructions. Specifically, we verify that the target entities referenced by the instruction are present, that the corresponding attributes, actions, or relations have already been established in the current visual history, and that the instruction is consistent with its assigned primary and secondary interaction types. Instructions containing multiple requirements are further decomposed into atomic sub-clauses, each of which is checked against the established state. Candidates that introduce nonexistent entities, assume incorrect preconditions, contain conflicting requirements, or specify unverifiable targets are removed at this stage. This rule-based stage serves as an efficient consistency filter rather than the final determination of visual grounding.

Candidates that pass the rule-based check are further reviewed visually using GPT-5.6 Sol. The model directly observes the pre-interaction video history together with the candidate interaction type, instruction, and corresponding nexus point. It verifies whether the subject, attribute, action, or relation to be modified is actually observable in the video, whether the requested change is grounded in the currently established state, whether the target state is sufficiently specific, and whether its outcome can be reliably evaluated within the subsequent observation window. For second-round interactions, this verification is performed against the visual history of the corresponding branch rather than the shared history before the first intervention. In particular, instructions at E_{2B} must be grounded in the state that actually emerges after the first intervention on branch B, rather than inherited from the original source description.

Human review forms the final stage of quality control. Candidates for which the VLM identifies ambiguous grounding, insufficient visual evidence, or conflicting conditions are manually inspected, and an additional subset of automatically accepted interactions is randomly audited. Reviewers jointly examine the complete video history, content description, interaction type, instruction, and nexus point, with particular attention to whether the target state is genuinely present, whether the instruction defines a natural and unambiguous intervention, whether sufficient pre-interaction evidence has been established at the selected time point, and whether the resulting change can be evaluated reliably. When the textual description, automatic visual judgment, and actual video content disagree, the visible video state and human review determine the final decision.

Through this three-stage procedure, the retained samples are not merely linguistically valid instructions, but complete interaction events whose instruction content, interaction type, and intervention time are jointly grounded in the evolving visual history. If any component of the interaction type, instruction content, or nexus point is modified during review, the revised event re-enters the corresponding quality-control stages to ensure that the final (V^-, u, t_e) remains consistent at the semantic, visual, and temporal levels.

C Temporal Design of Multi-Round Interaction

C.1 45-Second Horizon and Observation Windows

WorldLineBench uses a 45-second source worldline to support two rounds of history-dependent interaction. Each sample contains a first interaction E_1, followed by second-round interactions E_{2A} and E_{2B} on the two resulting branches. The 45-second horizon is not obtained by mechanically dividing the video into several fixed segments. Instead, it is chosen to provide enough temporal context for each interaction to be grounded in an already established visual history, while still leaving sufficient time to observe how the requested change emerges and evolves afterward. This is particularly important for the second round, whose validity depends on the state produced by the first intervention rather than on the original source video alone.

For each interaction round, we characterize the local temporal process using three functional stages: Preparation, Execution, and Stabilization. Preparation corresponds to the period in which the relevant visual preconditions become sufficiently established, such as the identity of a subject, an ongoing action, an object state, a spatial relation, or the current stage of a process. Execution begins once the interaction is issued and covers the period in which the requested semantic change starts to appear and develops toward the target state. Stabilization follows the emergence of the target change and reflects whether the new state can persist under continued generation rather than immediately reverting, drifting, or becoming inconsistent.

These three stages are conceptual rather than fixed temporal partitions. Their onset and duration vary across videos and interaction types because different worldlines establish usable states at different rates, while different interventions may require different response and settling times. A simple motion change may become grounded and observable quickly, whereas a process-level or multi-subject interaction may require a longer Preparation interval and a slower transition toward a stable outcome. WorldLineBench therefore does not enforce a universal duration for any of the three stages, but instead uses them to describe the functional structure of each interaction round.

The two interaction rounds are temporally coupled. After E_1, the edited continuation forms a new branch-specific visual history. The resulting state first enters Stabilization for the first interaction and then progressively becomes the Preparation context for the second round. Consequently, E_{2A} and E_{2B} are determined from the actual states reached on their respective branches rather than inherited from the shared history before E_1. In our data, the first interaction is typically located around 13.7 s, while second-round interactions are concentrated around 28.9 s. These values describe the empirical distribution of selected nexus points rather than fixed intervention timestamps.

Around each nexus point t_e, we define two complementary observation windows. The semantic-effect window

$$W_u = [t_e, t_e + 15\text{s}]$$

is used to evaluate whether the instructed state is established and whether it remains valid during subsequent generation. This window spans the main Execution and Stabilization behavior after an interaction and therefore captures more than a transient post-edit change. In parallel, a shorter response window W_r, also anchored at t_e, focuses on the early emergence of the semantic response and is used by the responsiveness metrics. The detailed duration and computation of W_r are specified together with those metrics.

The 45-second horizon, the three-stage interaction process, and the two observation windows serve different purposes. The 45-second duration provides the overall temporal budget for two-round interaction, Preparation–Execution–Stabilization describes how each interaction unfolds within that budget, and W_r and W_u define consistent ranges for measuring early response and subsequent semantic effectiveness. Together, they allow WorldLineBench to evaluate not only whether a model

reacts to an instruction, but also whether the resulting state becomes part of a coherent visual history that can support the next interaction.

C.2 Nexus-Point Selection and Intervention Timing

Interaction times are not predetermined absolute timestamps. Instead, each nexus point is selected according to the visual state already established in the current worldline. Before an interaction can be introduced, the relevant preconditions must be sufficiently observable, such as a stable subject identity, an ongoing action, a visible object attribute, a clear spatial relation, or a process that has entered an appropriate stage for intervention. At the same time, the selected point must leave enough subsequent video for evaluating responsiveness, semantic fulfillment, and state preservation. Nexus-point selection is therefore jointly constrained by the maturity of the pre-interaction state and the remaining evaluation horizon.

The first interaction E_1 is selected on the original worldline A. We follow the temporal evolution of A and identify a candidate point once the intended interaction type and instruction have acquired sufficient visual grounding. Different interaction types require different preconditions. For example, a Motion interaction typically requires the target action to be clearly established, an Attribute interaction requires the relevant property to be visible, a Spatial interaction requires the corresponding relation to be identifiable, and Temporal or process-related interactions require the underlying process to have already started. Consequently, even videos within the same content category may have different E_1 locations.

The second round contains two interactions applied to different worldlines. E_{2A} is selected directly from the later evolution of the original worldline A, once a new interaction opportunity becomes sufficiently grounded. In contrast, E_{2B} is defined on branch B, which is produced only after applying E_1. Its timing therefore cannot be fixed in advance. We first execute E_1, observe the actual state reached on branch B, and then determine both the applicable interaction content and the corresponding nexus point from the visible subject state, action progress, spatial relations, or process stage. Thus, E_{2A} depends on the natural continuation of A, whereas E_{2B} is explicitly conditioned on the edited history produced by the first intervention.

In practice, we avoid selecting nexus points either too early or too late. An early intervention may lack sufficient visual grounding, while a late intervention leaves too little time for subsequent Execution and Stabilization. We therefore choose t_e after the relevant preconditions have been established but while sufficient evaluation time remains. For successive interactions on the same evolving worldline, the two nexus points are typically separated by approximately 10–20 seconds. This interval provides enough time for the preceding state to develop and stabilize while still allowing the second interaction and its observation window to fit within the 45-second horizon. The 10–20 second range is an empirical guideline rather than a fixed temporal constraint.

Finally, nexus-point selection is reviewed jointly with the interaction type and instruction rather than in isolation. Candidate points are first checked against the video content description to verify that the required preconditions have already been established. A VLM then inspects the pre-interaction video history to confirm that the selected time has sufficient visual evidence and that the instruction is natural, grounded, and executable at that point. Ambiguous or inconsistent cases are forwarded to human review. If the nexus point is modified, the corresponding interaction event is re-evaluated to ensure consistency among V^-, u, and t_e.

Under this state-driven protocol, a nexus point is more than a temporal coordinate. It marks a moment at which the accumulated visual history supports a meaningful intervention. E_1 creates the first divergence from worldline A, E_{2A} continues to intervene on the original evolution of A, and E_{2B} operates on the edited history of branch B, enabling the benchmark to capture both natural-history and intervention-conditioned second-round interactions.

D METRICS FOR REAL-TIME INTERACTION CAPABILITY

D.1 SEMANTIC FULFILLMENT METRICS

Semantic fulfillment evaluates whether an interaction makes the requested change to the correct target and how long the resulting state remains valid. Four metrics provide complementary evidence: OSC measures feature alignment with the target description, IS-g rates the complete instruction, EFA checks individual requirements, and SEO measures the occupancy of a persistent target state; higher values are better for all four. For interaction i, we retain the main-text notation $\mathcal{E}_i = (V_i^-, u_i, t_{e,i}, V_i^+)$, local media time $\tau = t - t_{e,i}$, and the 15-second window W_u; shorter outputs use their actual duration $T_i = \min(15\,\mathrm{s}, T_i^{\mathrm{avail}})$. The judge is Gemini 3.8 Flash, with six reference frames from the second before the interaction and post-interaction output sampled at 3 FPS. E1 and E2A use A as reference; E2B uses the same model's actual B.

Overall Semantic Consistency (OSC). OSC measures feature alignment between the post-interaction content and its target semantics, using the common-window, phase-aware variant OSC-CW-PA. Windows are chosen by the interaction time, effect start and end times, and video duration; each spans 2 seconds and supplies 8 uniformly sampled frames (RGB, resized to 224×224, ImageNet-normalized), encoded by VideoCLIP-XL-v2. Each window is matched against a phase-specific target text rather than a single final-state description: pre-onset windows describe an effect not yet present, transition windows its formation, boundary windows onset or termination, and post-completion windows the persistent result. For model m and window w, with clip G_{miw}, phase text d_{iw}, and encoders f_v, f_t,

$$s_{miw} = \frac{f_v(G_{miw})^\top f_t(d_{iw})}{\|f_v(G_{miw})\|_2 \, \|f_t(d_{iw})\|_2}. \tag{2}$$

With $\mathcal{W}_{mi}$ the windows available to model m for interaction i, the common support is $\mathcal{W}_i^\cap = \bigcap_{m \in \mathcal{M}} \mathcal{W}_{mi}$. We average within each interaction's common windows and then equally over interactions with nonempty support:

$$\mathrm{OSC}_{mi} = \frac{1}{|\mathcal{W}_i^\cap|} \sum_{w \in \mathcal{W}_i^\cap} s_{miw}, \qquad \mathrm{OSC}_m = \frac{1}{|\mathcal{I}_{\mathrm{OSC}}|} \sum_{i \in \mathcal{I}_{\mathrm{OSC}}} \mathrm{OSC}_{mi}. \tag{3}$$

The per-interaction value is stored as `osc_cw_pa`. OSC retains the raw cosine scale: a larger value indicates stronger feature alignment, not a probability of success; actual fulfillment is checked by IS-g and EFA.

Instruction Satisfaction—Global (IS-g). The judge directly rates the complete instruction's implementation on the correct target from 1 to 5, considering explicit counts, ordering, durations, and entity relations:

Score	Criterion
1	No correct execution, an incorrect target, an opposite change, or insufficient visual evidence.
2	A small relevant change on the correct target, without a substantial requested outcome.
3	Substantial partial fulfillment, with a major outcome or essential condition still missing.
4	All major outcomes and essential conditions hold, with only an identifiable minor discrepancy.
5	All explicit requirements are fully satisfied on the correct target.

Adjacent grades are separated by what is still missing:

Boundary	Lower grade	Higher grade
1 vs 2: any correct change?	No correct relevant change, or unjudgeable.	A clear but small target-relevant change on the correct object.
2 vs 3: a start or partial result?	Only scattered changes, no confirmable sub-outcome.	One substantial sub-outcome or necessary stage is done, but the whole is not.
3 vs 4: missing key requirement or only a flaw?	A major outcome, required count, necessary stage, or key relation is still missing.	All major outcomes and essential conditions hold, with only a local minor discrepancy.
4 vs 5: can a specific gap be named?	An identifiable minor discrepancy related to the request.	All requirements are supported by sufficient evidence, with no nameable semantic gap.

For example, for "make the cup entirely blue": unchanged or unjudgeable is grade 1, a few blue spots 2, clearly half blue 3, entirely blue with slight edge leakage 4, and fully blue 5. The example only illustrates the grades and adds neither atoms nor area thresholds. For directly assigned grades $g_i \in \{1, 2, 3, 4, 5\}$,

$$\text{IS-g} = \frac{1}{|\mathcal{I}_g|} \sum_{i \in \mathcal{I}_g} g_i. \tag{4}$$

The holistic grade is independent of EFA: satisfying several auxiliary conditions can still leave the principal outcome unfulfilled. Readable outputs with insufficient evidence receive grade 1; API or decoding errors are recorded separately.

Edit Fulfillment Accuracy (EFA). Each instruction is decomposed before evaluation into K_i required atoms covering the target entity, the requested change, and necessary count, spatial, ordering, or duration constraints. The shared E1 and E2A instructions use one list across models; for E2B, whose instruction is selected per model from its own generated branch (Appendix B.3), the list derives from that model's instruction. No list is revised to match an output. Atom k receives $a_{ik} = 1$ only when visual evidence establishes its acceptance condition on the correct target, else 0. We compute

$$\text{EFA}_i = \frac{1}{K_i} \sum_{k=1}^{K_i} a_{ik}, \qquad \text{EFA} = \frac{1}{|\mathcal{I}_{\text{EFA}}|} \sum_{i \in \mathcal{I}_{\text{EFA}}} \text{EFA}_i, \tag{5}$$

averaging atoms within each interaction and then weighting interactions equally. Acceptance follows each atom's semantics: a state must be established, a bounded action must complete, and a sequence must satisfy its events and their order. Complete interaction success is

$$S_i = \prod_{k=1}^{K_i} a_{ik} = \mathbf{1}\{\text{EFA}_i = 1\}. \tag{6}$$

A positive EFA can indicate partial fulfillment; only $S_i = 1$ establishes all required conditions and supplies the complete-success gate used in Section D.2.

Semantic Effective Occupancy (SEO). SEO applies to instructions with a persistent target state and measures its occupancy within the observed interval. At the regular sampling times $\tau_{i\ell} = \ell/3 < T_i$ (L_i observations), let $z_i(\tau_{i\ell}) = 1$ when all applicable target states hold simultaneously, else 0. Then

$$\text{SEO}_i = \frac{1}{L_i} \sum_{\ell=0}^{L_i-1} z_i(\tau_{i\ell}), \qquad \text{SEO} = \frac{1}{|\mathcal{I}_{\text{SEO}}|} \sum_{i \in \mathcal{I}_{\text{SEO}}} \text{SEO}_i. \tag{7}$$

The indicator is instantaneous, not a cumulative "ever satisfied" flag, so a briefly appearing and then reverting target lowers SEO. Instantaneous or finite-count events without a persistent postcondition are not applicable and receive no zero score. EFA and SEO are computed as proportions and displayed as percentages, with the effective denominators retained alongside the aggregates.

D.2 Responsiveness Metrics

Responsiveness distinguishes the actual wait for a generated frame from the progression of target semantics within the video: TTFF uses the generation system's wall clock, while SOL, SCL, and DSR use media time after the interaction, and the two axes are not added together. We retain the main-text windows $W_u = [t_e, t_e + 15\,\mathrm{s}]$ and $W_r = [t_e, t_e + 5\,\mathrm{s}]$, with $T_i = \min(15\,\mathrm{s}, T_i^{\mathrm{avail}})$ the observed duration, $\mathcal{I}$ the valid evaluation set, and $N = |\mathcal{I}|$. Target atoms and acceptance criteria follow Section D.1; $S_i = 1$ exactly when all required atoms pass. Gemini 3.8 Flash judges semantic timing from the regular 3 FPS observations.

Time to First Frame (TTFF). TTFF measures system time from instruction-triggered generation to the availability of the first output frame. With w_i^{start} the recorded generation-timing origin and w_i^{first} the wall-clock time of first-frame availability,

$$\mathrm{TTFF}_i = w_i^{\mathrm{first}} - w_i^{\mathrm{start}}, \qquad \mathrm{TTFF} = \frac{1}{|\mathcal{I}_{\mathrm{TTFF}}|} \sum_{i \in \mathcal{I}_{\mathrm{TTFF}}} \mathrm{TTFF}_i. \tag{8}$$

TTFF comes from generation-run timing, in seconds, and is not recovered from the final video's frame timestamps or replaced by a judge's inference time. A runner returning a chunk or a full sequence must wait until its first output unit containing generated frames is available; dividing this wait by the unit's frame count is an amortized cost, not the first-output wait, and is reported separately. First-frame availability does not require correct semantic execution, which SOL evaluates; system latency is not capped by the 15-second window. Instrumentation, hardware, and execution settings belong to Appendix F.

Semantic Onset Latency (SOL). SOL is the first media-time instant at which any required atom is fulfilled on the correct target. With $b_{ik}(\tau)$ indicating that the evidence up to τ satisfies atom k,

$$\mathrm{SOL}_i = \min\{\tau \in [0, T_i] : \exists k,\ b_{ik}(\tau) = 1\}. \tag{9}$$

Recognizing the correct object without executing the change, changing an incorrect object, or unrelated natural motion does not trigger SOL; for a bounded action or a required count, the atom must complete, not merely begin. Instructions with a wait or trigger condition keep the interaction as the timing origin. SOL is unobserved when no required atom passes in the window. For the set $\mathcal{O}$ of interactions with at least one fulfilled atom,

$$\mathrm{SOL} = \frac{1}{|\mathcal{O}|} \sum_{i \in \mathcal{O}} \mathrm{SOL}_i, \qquad C_{\mathrm{SOL}} = \frac{|\mathcal{O}|}{N}. \tag{10}$$

Unobserved onsets are neither filled with zero nor replaced by the window endpoint; coverage C_{SOL} accompanies the mean so that a short SOL is read against how many interactions initiated at all.

Semantic Completion Latency (SCL). SCL localizes the first time the complete instruction satisfies all acceptance conditions. With $Q_i(\tau) = 1$ when the correct target, all required outcomes, and any count, order, or duration conditions are verified by τ, the completion time is $t_i^* = \min\{\tau \in [0, T_i] : Q_i(\tau) = 1\}$. The primary statistic caps unsuccessful interactions at their observed duration:

$$\mathrm{SCL}_i = \begin{cases} t_i^*, & S_i = 1, \\ T_i, & S_i = 0, \end{cases} \qquad \mathrm{SCL} = \frac{1}{N} \sum_{i \in \mathcal{I}} \mathrm{SCL}_i. \tag{11}$$

This mean combines observed completion times with observation limits for incomplete interactions, so it is not the mean completion time of successful interactions alone; a shorter output's smaller cap does not imply faster success.

Deadline Success Rate (DSR). DSR tests whether the entire interaction completes within the common 5-second budget W_r, combining complete success with SCL:

$$\mathrm{DSR}_5 = \frac{1}{N} \sum_{i \in \mathcal{I}} S_i\, \mathbf{1}\{\mathrm{SCL}_i \le 5\,\mathrm{s}\}. \tag{12}$$

The denominator covers all valid evaluations and completion exactly at 5 seconds counts. The success gate is essential: a failed four-second output has a capped SCL of four seconds but still contributes zero. DSR is displayed as a percentage under one common budget; it does not replace analyses stratified by temporal requirements.

E METRICS FOR WORLDLINE GENERATION CAPABILITY

E.1 PRESERVATION METRICS

Preservation evaluates whether an interaction retains content outside its authorized scope and earlier results that remain valid. Five metrics, all higher-is-better, provide complementary evidence: CF-g gives a holistic fidelity grade, PSR checks protected states, CA-HDF evaluates the handling of historical dependencies, and SC and BC are feature-based continuity measures. The protected scope follows the content description, the current and historical instructions, and pre-interaction visual evidence, excluding authorized changes and their necessary consequences. The observed window is W_u with duration $T_i = \min(15\,\mathrm{s}, T_i^{\mathrm{avail}})$; the reference is the second before the interaction, and the branches are E1: A→B, E2A: A→D, E2B: B→C (E2B uses the same model's actual B). For E2B, protection conditions follow each model's realized B and its selected instruction; what is shared across models is the protection rule, not identical protected facts. CF-g, PSR, and CA-HDF use Gemini 3.8 Flash; SC and BC use the feature protocol below.

Content Fidelity—Global (CF-g). CF-g directly assigns a 1–5 grade to non-target preservation across the observed window, weighing the importance and extent of losses to protected identities, structures, relations, actions, and states. Changes required by an authorized edit are not damage. The rubric is:

Score	Criterion
1	Almost no applicable content is retained, or the output cannot be matched to the reference; also used when required evidence is insufficient.
2	Some content survives, but a core identity, structure, action, or state is substantially damaged.
3	Core content survives, but an important non-core protected fact changes substantially.
4	Core and important facts are retained, with only minor local detail drift.
5	Important facts and verifiable details have sufficient preservation evidence, with no confirmed unauthorized change.

Adjacent grades are separated as follows:

Boundary	Distinguishing evidence
1 vs 2	Applicable content cannot be matched to the reference overall (1); some content is still clearly retained (2).
2 vs 3	A core element is damaged, such as the main subject being replaced, a key support relation broken, or a key process reset (2); the core survives and the substantive loss is confined to important non-core content (3).
3 vs 4	An important non-core fact changes, such as a striped shirt turning solid, a secondary action reversed, or a required shot mode or style switched (3); all important facts hold and only local texture or similar detail drifts (4).
4 vs 5	Some applicable category shows a specific minor drift (4); every applicable category has no confirmed drift and sufficient evidence (5).

For global grades f_i over the applicable, valid set $\mathcal{I}_f$,

$$\text{CF-g} = \frac{1}{|\mathcal{I}_f|} \sum_{i \in \mathcal{I}_f} f_i. \tag{13}$$

Confirmed damage is not erased by later recovery, and CF-g is assigned holistically rather than derived from PSR or interaction success. An interaction with no applicable protected content is not applicable.

Protected-State Retention (PSR). PSR organizes non-target requirements into seven protection categories, with at most one binary atom per applicable category:

Category	Content to preserve unless a change is authorized
Motion	Actions, direction, speed trends, amplitude, and interactions; a stationary state can also require protection.
Attribute	Entity or environmental color, material, shape, clothing, lighting, weather, and physical state.
Spatial	World-space position, orientation, distance, support, contact, and layout relations.
Composition	Entity identity, instance count, existence, and part composition.
Camera	Continuation of viewpoint, shot scale, and camera-motion mode.
Style	Valid global or local visual style, medium, and rendering.
Temporal	Event order, process stage, and established historical facts that remain valid.

With P_i the applicable categories, $p_{ik} = 1$ when all protected facts in category k are retained, else 0; the whole window is considered and later recovery does not cancel a violation. The category of the current interaction is not automatically removed; a category is excluded only when no requirement remains. The per-interaction score and the model score, a macro-average over the seven primary interaction types $\mathcal{T}$, are

$$\mathrm{PSR}_i = \frac{1}{|P_i|} \sum_{k \in P_i} p_{ik}, \qquad \mathrm{PSR} = \frac{1}{7} \sum_{c \in \mathcal{T}} \left(\frac{1}{N_c} \sum_{i \in \mathcal{I}_c} \mathrm{PSR}_i \right), \tag{14}$$

with $\mathcal{I}_c = \{i : y_i = c,\ |P_i| > 0\}$ and $N_c = |\mathcal{I}_c|$; $|P_i| = 0$ makes PSR not applicable. If a required type has no valid records, its missing coverage is reported rather than silently reweighted.

Coverage-Adjusted History-Dependency Following (CA-HDF). CA-HDF evaluates the chain $A \xrightarrow{\mathrm{E1}} B \xrightarrow{\mathrm{E2B}} C$ on history-dependent E2B interactions. Each model m has its own set $\mathcal{D}_m$, constructed from its realized branch $B^{(m)}$ under Appendix B.3 and finalized before $C^{(m)}$ is generated and scored. Each applicable interaction has four binary gates: u_i (the prerequisite E1 object or state exists in the actual B), c_i (the new E2B interaction is completed in C), d_i (E2B correctly acts on the corresponding E1 result), and h_i (earlier facts that remain valid are retained). An interaction scores one only if all four pass:

$$\text{CA-HDF}_i = u_i c_i d_i h_i, \qquad \text{CA-HDF}_m = \frac{1}{|\mathcal{D}_m|} \sum_{i \in \mathcal{D}_m} u_i c_i d_i h_i. \tag{15}$$

For example, adding a hat and then recoloring that hat red requires the hat's establishment in B, successful recoloring, binding to the same hat, and retention of its identity; a merely present red hat does not suffice. Under Appendix B.3, a dependent candidate whose prerequisite is absent from the realized B is rejected at selection and never enters $\mathcal{D}_m$. The gate u_i re-checks the prerequisite at scoring time: a selected case whose prerequisite is not confirmed remains in the denominator with zero credit rather than being dropped, so the metric does not condition on judge-confirmed histories. E1, E2A, and inapplicable E2B interactions are excluded. We report $|\mathcal{D}_m|$ per model; a model with an empty set receives a not-applicable result rather than a zero.

Subject Consistency (SC). SC combines adjacent-frame continuity and pre-interaction reference similarity using global DINO v1 ViT-B/16 features (encoder and transforms follow VBench (Huang et al., 2024); sampling and six-reference fusion follow the `wl-scbc-adjref6-v1` protocol), with no tracking, segmentation, or ROI crop. For interaction i, reference frames $r_{i1}, \ldots, r_{i6}$ sit at local times $-1, -5/6, \ldots, -1/6$ s, and output frames $g_{i1}, \ldots, g_{in_i}$ are sampled at $k/6 < T_i$. With encoder f, normalized features $\widehat{f}(x) = f(x)/\|f(x)\|_2$, and clipped cosine $s_f(x, y) = \mathrm{clip}_{[0,1]}(\widehat{f}(x)^\top \widehat{f}(y))$,

$$R_{ij}^f = \frac{1}{6} \sum_{\ell=1}^{6} s_f(g_{ij}, r_{i\ell}), \qquad A_{ij}^f = \begin{cases} s_f(g_{i1}, r_{i6}), & j = 1, \\ s_f(g_{ij}, g_{i,j-1}), & j > 1, \end{cases} \tag{16}$$

$$C_i^f = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{A_{ij}^f + R_{ij}^f}{2}, \qquad \mathrm{SC}_i = C_i^{\mathrm{DINO}}, \qquad \mathrm{SC} = \frac{1}{|\mathcal{I}_{\mathrm{SC}}|} \sum_{i \in \mathcal{I}_{\mathrm{SC}}} \mathrm{SC}_i, \tag{17}$$

giving adjacent and six-reference terms weight $1/2$ each. DINO uses `dino_vitb16` with short-side-224 resizing (no center crop), official normalization, and FP32 features.

Background Consistency (BC). BC applies the same C_i^f to global OpenAI CLIP ViT-B/32 features from the same frames:

$$\mathrm{BC}_i = C_i^{\mathrm{CLIP}}, \qquad \mathrm{BC} = \frac{1}{|\mathcal{I}_{\mathrm{BC}}|} \sum_{i \in \mathcal{I}_{\mathrm{BC}}} \mathrm{BC}_i. \tag{18}$$

CLIP uses `clip_transform(224)` (short-side resize plus 224×224 center crop); the center crop is encoder preprocessing, not background segmentation, so BC is likewise a global-feature proxy. SC and BC lie in $[0, 1]$; authorized changes can lower similarity while static or unexecuted outputs can score highly, so both are read together with CF-g, PSR, and EFA.

E.2 Video Quality Metrics

Video quality covers imaging quality, temporal stability, and aesthetic presentation through NRQ, TC, and AE, respectively. These metrics score the generated video itself, without requiring an ideal reference or judging instruction fulfillment. With G_i the video input of interaction i, each metric first yields a per-interaction score and then averages equally over its valid set $\mathcal{I}_M$:

$$M = \frac{1}{|\mathcal{I}_M|} \sum_{i \in \mathcal{I}_M} M_i, \qquad M \in \{\mathrm{NRQ}, \mathrm{TC}, \mathrm{AE}\}. \tag{19}$$

The three metrics keep their native scales; NRQ and AE are higher-is-better, TC is an error for which lower is better.

No-Reference Quality (NRQ). NRQ uses MUSIQ-KonIQ-10k (Ke et al., 2021) to estimate perceptual imaging quality (blur, noise, compression, and other degradations) from 16 frames uniformly sampled over the video's valid interval, $\mathcal{F}_i = \{x_{i1}, \ldots, x_{i16}\}$. With raw MUSIQ output $q_{\mathrm{MUSIQ}}(x)$,

$$\mathrm{NRQ}_i = \frac{1}{16} \sum_{\ell=1}^{16} q_{\mathrm{MUSIQ}}(x_{i\ell}). \tag{20}$$

The primary result uses the native-input setting `NRQ_native`; scores retain the raw MUSIQ scale. NRQ measures perceptual quality, not whether the correct object was changed.

Temporal Continuity (TC). TC is expressed as an error, following VBench's distinction between Temporal Flickering and Motion Smoothness (Huang et al., 2024): the flickering error $E_{\mathrm{TF},i}$ is the residual difference between adjacent frames after optical-flow motion compensation, and the smoothness error $E_{\mathrm{MS},i}$ is the difference between RIFE-interpolated frames and the actual intermediate frames. The combined score is

$$\mathrm{TC}_i = \frac{1}{2} E_{\mathrm{TF},i} + \frac{1}{2} E_{\mathrm{MS},i}, \tag{21}$$

the `equal_weight_raw_error_v1` definition, for which lower is better. A frozen video can yield a low error, so TC is read alongside EFA for Motion and Temporal instructions and the semantic timing metrics.

Aesthetic Evaluation (AE). AE uses DOVER's aesthetic branch (Wu et al., 2023) to assess aesthetic presentation and overall visual impression. With $f_{\mathrm{DOVER,aesthetic}}(G_i)$ its raw output,

$$\mathrm{AE}_i = f_{\mathrm{DOVER,aesthetic}}(G_i), \tag{22}$$

using the native-input setting `dover_aesthetic_native_v1`. AE retains the branch's raw scalar (negative values are valid) without remapping to $[0, 1]$ or percentages. It is a learned proxy for composition and appearance, not a test of physical correctness or instruction fulfillment.

F Experimental Setup and Evaluation Protocol

F.1 Implementation and Model Inference Details

All evaluated models are run in a unified server environment using NVIDIA H20 GPUs. We use publicly released official checkpoints or the corresponding versions provided by the model authors

and follow their released inference implementations. Except for quantization, parallelism, caching, or other mechanisms required by the original implementations, we do not modify the model architectures or perform benchmark-specific retraining or parameter fine-tuning. Inference configurations, random seeds, and input preprocessing are kept fixed throughout evaluation, while model-specific settings are described separately below.

Because the seven models differ substantially in their native generation paradigms, we preserve their original inference mechanisms rather than forcing all systems into a single internal generation pipeline. This choice avoids introducing adaptation-specific bias: Streaming-native Models, Real-time Video Editors, and Full-context Controllable Models differ structurally in their intended usage, input organization, and generation granularity, and forcing them into the same execution pipeline could make performance depend on the quality of the adaptation itself rather than on the underlying model capability. We therefore preserve model-level differences while enforcing a common benchmark-level task definition and evaluation protocol.

Specifically, Streaming-native Models retain their native frame-wise or chunk-wise streaming generation procedures; Real-time Video Editors use their released editing interfaces and source-conditioning mechanisms; and Full-context Controllable Models retain their sequence-level inference procedures. These paradigms differ in how inputs are organized, how inference is executed, and how outputs are produced, but they follow the same interaction taxonomy, construction rules, and evaluation protocol. For interactions defined on the shared source worldline, including E_1 and E_{2A}, the corresponding instruction and nexus point are shared across models. In contrast, E_{2B} is constructed after each model has produced its edited branch B, and its instruction and nexus point are therefore determined from that model's actual branch state under the same construction rules.

The unified evaluation window and the subsequent time normalization operate at different levels. The evaluation window defines the temporal range over which responsiveness, semantic fulfillment, and preservation are measured after an interaction. Time normalization, in contrast, aligns runtime statistics across models with different native output granularities, including frame-wise, chunk-wise, and sequence-level generation. The visual conditions available to each model, the mapping between its official interface and the WorldLineBench input protocol, and the corresponding latency normalization rules are specified in the Unified Evaluation Protocol.

We also distinguish system-level failures from model-level failures during inference. System-level failures include out-of-memory errors, abnormal process termination, decoding failures, or other execution errors unrelated to the semantic quality of the generated result. Such cases are rerun with the same input and inference configuration. Model-level failures, by contrast, refer to runs that complete normally but fail to satisfy the interaction target or produce semantic or visual defects. These outputs are not resampled because of poor quality and are retained for evaluation. In practice, the large majority of samples complete inference normally, with system-level failures occurring only occasionally.

F.2 Unified Evaluation Protocol

WorldLineBench represents every evaluated interaction using

$$(V^-, u, t_e) \to V^+,$$

where V^- denotes the visual history established before the interaction, u is the instruction issued at nexus point t_e, and V^+ is the post-interaction worldline used for evaluation. The unified protocol standardizes how an interaction event is constructed and evaluated rather than requiring every model to receive an identical event at every branch. All models follow the same interaction taxonomy, grounding criteria, temporal constraints, and evaluation windows.

For E_1 and E_{2A}, which are defined on the original source worldline A, the visual history is shared across models; consequently, the corresponding instruction and nexus point can be kept identical. E_{2B}, however, is inherently model-dependent. It is defined only after E_1 has been executed and branch B has been generated by the evaluated model. Because different models may produce different post-E_1 states, the valid instruction and nexus point for E_{2B} are reconstructed from each model's actual branch history using the same grounding and timing rules described in Sec. **??**. Thus, comparability at E_{2B} is established through a shared construction protocol rather than through an identical instruction or timestamp.

The three model families also differ in their accessible visual conditions, and we therefore do not assume an identical causal information constraint. Streaming-native Models operate on causal visual history and use only states available up to the current time. Real-time Video Editors receive source conditions according to their released online editing interfaces. After an instruction is issued at t_e, new source chunks may continue to arrive during the response delay and can be edited once available, but unavailable future content is not exposed in advance. This setting reflects a practical streaming system in which the source stream may continue while the edit response is being established.

Full-context Controllable Models follow a different formulation based on sequence-level source conditioning. We retain this native full-context inference mechanism and treat these systems as non-causal reference baselines rather than as models operating under the same causal information constraint as Streaming-native Models. They follow the same interaction-construction rules and are evaluated using the same metric definitions and observation windows, while their broader source context is explicitly reported as an additional information condition. Their results therefore characterize full-context controllable generation under the same benchmark objectives without implying identical deployment conditions across model families.

Regardless of whether a model produces outputs frame-wise, chunk-wise, or sequence-wise, its generated result is mapped to the corresponding post-interaction timeline and evaluated within the same W_r and W_u. These windows standardize the temporal scope of metric computation for each constructed interaction event. Runtime normalization serves a separate purpose: it aligns computational cost across models with different native output granularities. The evaluation windows determine *what temporal content is evaluated*, whereas runtime normalization determines *how inference cost is reported*.

For runtime comparison, we use 81 frames as a common normalization window. This benchmark-level hyperparameter is chosen because most evaluated models can natively accept or stably generate video segments of approximately this scale, allowing a common computational unit without excessive cropping, padding, or non-native adaptation. The 81-frame window is used only for runtime reporting and cross-model cost alignment and is independent of the semantic evaluation windows W_r and W_u.

For native frame-wise systems, including JoyAI and StreamDiffusionV2, we retain their measured per-frame runtime without 81-frame normalization. For block-wise or sequence-level systems, we instead report the total cost corresponding to an equivalent 81-frame output, including the preparation cost associated with that output scale and the time required to generate the 81 frames. This normalization provides a common runtime reporting scale without implying that non-frame-wise models possess native frame-level streaming capability.

Overall, WorldLineBench enforces comparability through shared task definitions, interaction-construction rules, and evaluation criteria rather than by requiring every branch to contain an identical instruction event. Interactions on the shared source worldline are directly aligned across models, while branch-dependent interactions such as E_{2B} are reconstructed from each model's realized visual history under the same protocol. Native generation mechanisms and information-access conditions are preserved and reported explicitly, allowing heterogeneous systems to be compared without discarding the history dependence that defines multi-round interactive video generation.

G Additional Results and Evaluation Analysis

G.1 Complete Numerical Results for the Main-Text Analyses

This section provides the complete numerical results for Fig. 3(a–d), covering the interaction-type and content-context heatmap, subtype-level radar profiles, Real–Synthetic comparison, and long-horizon round-wise comparison. EFA, PSR, CA-HDF, and DSR are reported as percentages; NRQ uses the native MUSIQ score. Higher is better for all metrics reported here. In model-comparison tables, **bold** and underlined values denote the best and second-best results among the seven models under the same condition. The heatmap table presents the seven-model mean alongside JoyAI.

Interaction types and content contexts. Table 7 expands Fig. 3(a) into all 35 combinations of seven primary interaction types and five content contexts. For each combination, Mean is obtained by averaging sample-level EFA within each model and then averaging the seven models equally; JoyAI reports that model on the same combination. Row and column orders follow the heatmap.

Table 7: **EFA across interaction types and content contexts (%, ↑).** Full values for Fig. 3(a). Each context has a seven-model Mean column and a JoyAI column.

Interaction type	**Embodied**		**Human**		**Process**		**Object**		**Scene**	
	Mean	JoyAI	Mean	JoyAI	Mean	JoyAI	Mean	JoyAI	Mean	JoyAI
Composition	25.06	59.28	28.80	60.30	23.69	40.99	22.80	50.84	20.06	45.09
Attribute	27.57	55.04	24.03	41.33	27.39	44.96	20.46	40.17	17.42	36.68
Spatial	23.43	28.08	16.82	17.21	19.14	25.00	24.95	39.77	12.40	18.00
Motion	30.31	36.25	22.88	30.23	8.65	11.07	19.70	28.89	18.55	22.64
Camera	23.64	28.68	15.87	18.88	19.11	19.38	39.17	46.25	27.52	33.18
Style	30.90	62.20	24.76	48.86	31.77	58.42	32.04	45.76	28.18	59.80
Temporal	37.83	33.27	40.30	45.81	36.11	38.71	29.70	35.53	41.52	41.33

Subtype-level capability profiles. Tables 8 and 9 report all seven models on the 14 representative interaction subtypes in the EFA and PSR radar plots of Fig. 3(b). Subtype labels and their order follow the radar axes, starting with Appearance. Model columns follow the main results table. SDV2, SANA, Lucy, and Wan denote StreamDiffusionV2, SANA-Streaming, Lucy Edit 1.1, and Wan2.1-VACE, respectively. Best and second-best results are compared across the seven models within each row.

Table 8: **EFA by interaction subtype (%, ↑).** Full values for the left radar plot in Fig. 3(b). Highlighting compares models within each row.

	Streaming-native models		Real-time video editors			Full-context controllable models	
Subtype	StreamEdit	SDV2	LiveEdit	SANA	JoyAI	Lucy	Wan
Appearance	16.69	14.68	**51.53**	6.41	<u>50.04</u>	18.45	14.22
Func.State	8.10	13.00	<u>28.19</u>	8.91	**30.38**	8.76	15.58
Framing	21.48	19.92	18.59	**30.36**	23.33	17.38	<u>27.56</u>
Viewpoint	27.35	<u>30.36</u>	28.89	29.30	**35.33**	18.90	29.77
Addition	18.27	18.21	<u>49.90</u>	6.19	**53.87**	13.18	5.46
Removal	18.60	<u>21.16</u>	14.37	11.68	**39.37**	9.37	16.42
Replacement	25.49	20.26	<u>59.14</u>	9.85	**70.82**	28.96	15.82
Local Motion	25.90	23.38	<u>26.22</u>	21.56	**32.19**	18.88	23.66
Interaction	7.77	7.05	**12.49**	11.58	<u>11.65</u>	7.60	9.94
Spatial Rel.	15.76	16.12	18.39	19.29	**23.28**	9.42	<u>21.20</u>
Spatial Pose	<u>24.07</u>	21.56	22.00	21.46	**29.62**	12.75	18.84
Art.Style	17.04	21.14	<u>37.72</u>	29.67	**55.89**	16.08	29.88
Progression	<u>38.06</u>	35.57	31.39	38.01	37.45	19.60	**41.11**
Temporal Arr.	36.22	<u>42.07</u>	40.95	40.92	40.73	28.81	**43.72**

Real and synthetic video sources. Table 10 reports EFA, PSR, CA-HDF, and DSR for each model on Real and Synthetic videos, expanding Fig. 3(c). Columns within each metric are ordered as Real, then Synthetic. PSR retains the equal-weight macro-average over the seven primary interaction types, and CA-HDF is evaluated on applicable samples with cross-round dependencies. DSR-Contrast in the figure denotes the five-second DSR used in the metric definitions and this table.

Long-horizon and history-dependent interaction. Table 11 expands Fig. 3(d), with Round 1 denoted by E1 and Round 2 by E2B. E1 is the first interaction, and E2B is the second interaction conditioned on the first interaction output. Each metric is ordered as E1, then E2B. We report EFA,

Table 9: **PSR by interaction subtype (%, ↑).** Full values for the right radar plot in Fig. 3(b). Highlighting compares models within each row.

	Streaming-native models		Real-time video editors			Full-context controllable models	
Subtype	StreamEdit	SDV2	LiveEdit	SANA	JoyAI	Lucy	Wan
Appearance	45.02	18.42	43.59	**74.99**	64.49	25.04	62.81
Func.State	57.06	14.94	26.70	**73.29**	69.19	19.41	67.25
Framing	57.02	14.44	25.43	**70.15**	60.54	17.08	60.32
Viewpoint	63.62	12.45	25.12	**67.68**	61.71	25.84	58.95
Addition	50.85	27.27	36.95	71.99	70.68	25.93	**85.91**
Removal	52.64	16.33	44.15	67.57	**72.63**	23.74	68.68
Replacement	47.92	23.42	41.17	69.99	**72.48**	10.72	68.08
Local Motion	59.77	8.95	36.07	66.92	66.44	28.60	**76.42**
Interaction	54.45	5.26	29.66	70.43	59.43	20.38	**79.09**
Spatial Rel.	60.75	14.95	36.95	**73.14**	64.31	18.02	53.39
Spatial Pose	58.93	28.01	28.21	**72.93**	66.64	16.42	58.00
Art.Style	61.61	37.22	43.98	73.38	**74.49**	2.30	67.24
Progression	63.44	10.64	21.34	**76.62**	74.41	14.43	65.68
Temporal Arr.	67.73	12.91	33.70	**80.36**	72.96	16.60	67.37

Table 10: **Real and Synthetic results (%, ↑).** Full values for Fig. 3(c). Best and second-best values are identified independently within each column.

Model	**EFA ↑**		**PSR ↑**		**CA-HDF ↑**		**DSR ↑**	
	Real	Synthetic	Real	Synthetic	Real	Synthetic	Real	Synthetic
Streaming-native models								
StreamEdit	23.45	18.86	53.90	63.52	20.00	19.39	9.47	8.93
StreamDiffusionV2	22.02	19.68	16.44	22.22	11.48	9.52	10.20	10.07
Real-time video editors								
LiveEdit	31.17	33.22	29.89	40.42	21.85	19.39	16.67	20.67
SANA-Streaming	20.51	17.68	**65.19**	**78.69**	18.52	17.35	7.67	7.80
JoyAI	**35.09**	**40.74**	61.42	75.53	**39.63**	**39.80**	**18.47**	**26.87**
Full-context controllable models								
Lucy Edit 1.1	16.71	16.11	17.52	20.19	2.59	3.74	5.20	5.13
Wan2.1-VACE	23.82	20.60	60.34	72.03	14.07	15.99	9.20	8.33

DSR, PSR, and NRQ for every model; NRQ corresponds to MUSIQ in the figure. Under this round correspondence, mean EFA across the seven models decreases from 27.59% to 16.00%, and mean DSR from 16.03% to 3.97%. PSR decreases for JoyAI, LiveEdit, and StreamDiffusionV2, while increasing for Lucy Edit 1.1 and Wan2.1-VACE.

Table 11: **Long-horizon results across two interaction rounds.** Full values for Fig. 3(d). EFA, DSR, and PSR are percentages; NRQ is the native MUSIQ score. Higher is better for all four metrics. E1 and E2B correspond to Round 1 and Round 2, respectively. Highlighting compares models within each column.

Model	EFA ↑		DSR ↑		PSR ↑		NRQ (MUSIQ) ↑	
	E1	E2B	E1	E2B	E1	E2B	E1	E2B
Streaming-native models								
StreamEdit	24.40	13.57	12.30	2.60	59.06	59.52	56.406	56.842
StreamDiffusionV2	23.63	13.32	13.60	2.80	30.15	13.68	44.670	52.392
Real-time video editors								
LiveEdit	36.59	22.12	24.60	8.10	38.96	32.17	**63.063**	**63.093**
SANA-Streaming	21.44	11.21	10.50	2.00	71.02	**71.36**	40.213	41.072
JoyAI	**45.08**	**23.59**	**31.60**	**8.30**	**71.35**	66.67	56.787	54.949
Full-context controllable models								
Lucy Edit 1.1	18.26	12.45	7.00	1.50	10.73	23.54	41.608	47.589
Wan2.1-VACE	23.76	15.73	12.60	2.50	61.71	69.72	59.446	61.748

G.2 Human Validation of VLM-Judged Metrics

Nine of the sixteen benchmark metrics rest on VLM judgments, so the credibility of the benchmark depends on a quantitative check of human–VLM agreement. We therefore audit these nine metrics against independent human review on 100 interaction outputs, comparing for each case the human-reference scores H with the scores V produced by the VLM judge. This section describes the sampling design, the blinded review protocol, the agreement statistics, and the audit results.

Scope and sampling design. The audit concerns nine metrics that use VLM judgments or are derived from them: EFA, IS-g, and SEO for fulfillment; SOL, SCL, and DSR for semantic responsiveness; and CF-g, PSR, and CA-HDF for preservation. The other seven metrics use different evidence: video–text features for OSC, system timing for TTFF, visual features for SC and BC, and quality, temporal, or aesthetic models for NRQ, TC, and AE. They are outside this agreement analysis.

The unit is one interaction output from one model. The audit contains 100 cases in total, rather than 100 per model. It allocates 50 cases to each source stratum, Real and Synthetic, with ten per content context within each stratum (Table 12). Seven anonymized model slots M1–M7 receive 15, 15, 14, 14, 14, 14, and 14 cases, respectively; every source–context stratum covers all seven slots.

Table 12: **Audit allocation.** E2B cases in this allocation are all dependency-applicable. SEO and CA-HDF columns describe eligible subsets, not additional cases.

Context	Real	Synthetic	E1	E2A	E2B / CA-HDF	SEO eligible
Human	10	10	6	6	8	14
Object	10	10	6	6	8	14
Scene	10	10	6	6	8	14
Process	10	10	6	6	8	14
Embodied	10	10	6	6	8	14
Total	50	50	30	30	40	70

The allocation contains 30 E1, 30 E2A, and 40 E2B cases; all 40 E2B cases are dependency-applicable, and SEO applies to 70 persistent-state cases. All seven primary interaction types are represented. The source, context, model, branch, and applicability manifest was frozen before scores were inspected, and each case carries a unique source-video identifier. These quotas serve validation only; they do not estimate dataset proportions or overall model performance.

Review protocol and metric alignment. Two independent reviewers scored each case, and a third adjudicated disagreements. Reviewers received the instruction, the prespecified target atoms and protection requirements, and the same visual evidence as the judge; model identity, VLM scores, and the other reviewer's scores remained hidden. Original judgments, timestamps, evidence notes, and adjudication reasons were retained. Pre-adjudication human–human agreement is reported separately from post-adjudication human–VLM agreement. The audit follows the paper's Gemini 3.8 Flash judge designation; the analysis below uses the two channels H and V.

Evidence and timing follow the metric definitions in Appendices D and E: six reference frames from the preceding second, post-interaction observations at 3 FPS, and an actual window $T_i = \min(15\,\mathrm{s}, T_i^{\mathrm{avail}})$. E1 and E2A use A as reference. E2B requires the same model's actual B, the subsequent C, and the relevant historical instructions, not an idealized B. The window allocation includes 90 cases of 15 seconds, five of 8 seconds, and five of 4 seconds to exercise short-output handling.

Reviewers supplied evidence-level judgments before aggregation. EFA averages the required binary target atoms within a case; IS-g independently applies the 1–5 holistic fulfillment rubric. SEO averages instantaneous joint target-state validity over applicable 3 FPS observations, including reversions. SOL is missing when no required atom is attained. SCL uses the actual completion time on success and T_i on failure. DSR requires both complete success and completion within five seconds, so a failed four-second output remains a failure. CF-g independently applies the 1–5 preservation rubric; PSR averages the applicable binary protection categories within each case. This audit compares paired case scores; the main benchmark PSR retains its macro-average over seven primary interaction types. CA-HDF multiplies source establishment u, current completion c, dependency binding d, and history preservation h on the fixed applicable E2B set, retaining $u = 0$ cases in its denominator.

Agreement statistics. Let $\mathcal{I}_m$ contain applicable cases with valid paired scores for metric m, with $n_m = |\mathcal{I}_m|$. For EFA, IS-g, SEO, SOL, SCL, CF-g, and PSR, Spearman correlation uses average ranks for ties. Mean bias and mean absolute error assess calibration:

$$\rho_m = \mathrm{Corr}(\mathrm{rank}(H_m), \mathrm{rank}(V_m)),$$
$$\Delta_m = \frac{1}{n_m}\sum_{i\in\mathcal{I}_m}(V_{im} - H_{im}), \qquad \mathrm{MAE}_m = \frac{1}{n_m}\sum_{i\in\mathcal{I}_m}|V_{im} - H_{im}|. \tag{23}$$

Binary DSR and CA-HDF use Cohen's $\kappa = (p_o - p_e)/(1 - p_e)$, where p_o is observed agreement and p_e is agreement under the two marginal distributions. IS-g and CF-g additionally use quadratic weighted $\kappa_w = 1 - D_o/D_e$, with disagreement weights $w_{ab} = (a - b)^2/16$ and observed and marginal-independent weighted disagreements D_o, D_e. Correlation evaluates ordering; bias and MAE evaluate score calibration.

We compute percentile 95% intervals from 2,000 paired bootstrap resamples within the ten source–context strata, resampling entire cases and retaining all metrics and missingness. Source-video identifiers are unique in this audit, so case-level resampling needs no source clustering; repeated sources would require source-cluster inference. Positive association is tested with 19,999 one-sided permutations of valid VLM scores against fixed human-reference scores, followed by Holm correction over nine primary tests. Rejecting the pairing-independence null does not by itself establish equivalence or case-level correctness.

Results. Tables 13–15 report the audit results.

Overall correlations for the seven score or timing metrics range from 0.873 to 0.992, and binary agreement is 0.767 for DSR and 0.875 for CA-HDF. All nine unadjusted permutation p values reach the Monte Carlo floor of 0.00005, giving adjusted $p = 0.00045$; equal reported p values do not imply equal evidential strength. Source-specific results vary: Synthetic SEO has $\rho = 0.726$ and

Table 13: **Human–VLM score agreement.** ρ: Spearman correlation; κ: Cohen agreement. Source-specific columns report descriptive coefficients and valid case counts.

Metric	Stat.	N	Overall [95% CI]	Real (N)	Synthetic (N)	Holm p
Fulfillment						
EFA	ρ	100	0.915 [0.867, 0.952]	0.965 (50)	0.859 (50)	< 0.001
IS-g	ρ	100	0.903 [0.858, 0.936]	0.921 (50)	0.888 (50)	< 0.001
SEO	ρ	70	0.873 [0.753, 0.960]	0.960 (35)	0.726 (35)	< 0.001
Semantic responsiveness						
SOL	ρ	79	0.992 [0.985, 0.995]	0.993 (40)	0.989 (39)	< 0.001
SCL	ρ	100	0.940 [0.881, 0.981]	0.978 (50)	0.885 (50)	< 0.001
DSR	κ	100	0.767 [0.621, 0.891]	1.000 (50)	0.464 (50)	< 0.001
Preservation						
CF-g	ρ	100	0.910 [0.854, 0.949]	0.919 (50)	0.899 (50)	< 0.001
PSR	ρ	100	0.971 [0.948, 0.988]	0.974 (50)	0.961 (50)	< 0.001
CA-HDF	κ	40	0.875 [0.481, 1.000]	1.000 (20)	0.828 (20)	< 0.001

Intervals use 2,000 paired bootstrap resamples within source–context strata. CA-HDF has 1,990 defined bootstrap coefficients; the other ten are undefined, not perfect agreement. The remaining metrics have 2,000 defined coefficients. Source-specific intervals are not estimated here. All nine adjusted p values equal 0.00045 at the Monte Carlo resolution used; this tests positive association, not equivalence or correctness.

Synthetic DSR has $\kappa = 0.464$, so agreement is not uniformly high in every subgroup. CA-HDF has only 40 eligible cases and four human-reference positives, yielding a wide interval of $[0.481, 1.000]$. A subgroup point estimate of 1.000 does not establish stable perfect agreement.

Table 14: **Score bias and absolute error.** H and V denote the human-reference and VLM score channels. Percentage-metric differences and MAE are in percentage points; ordinal and latency differences use grades and seconds.

Metric (mean unit)	N	H mean	V mean	$V - H$ [95% CI]	MAE
Fulfillment					
EFA (%)	100	57.92	62.92	+5.00 [+2.25, +7.83]	7.00
IS-g (grade)	100	2.91	3.10	+0.19 [+0.10, +0.29]	0.27
SEO (%)	70	21.01	26.06	+5.05 [+1.96, +8.95]	5.41
Semantic responsiveness					
SOL (s)	79	5.39	5.05	-0.34 [-0.40, -0.27]	0.36
SCL (s)	100	10.81	10.23	-0.59 [-1.00, -0.26]	0.61
DSR (%)	100	20.00	24.00	+4.00 [-1.00, +9.00]	8.00
Preservation					
CF-g (grade)	100	3.06	3.15	+0.09 [-0.01, +0.19]	0.23
PSR (%)	100	70.68	73.00	+2.31 [+0.60, +4.08]	3.81
CA-HDF (%)	40	10.00	12.50	+2.50 [+0.00, +8.11]	2.50

All entries use the same paired valid cases as Table 13; $V - H$ is computed from unrounded case-level scores. PSR here is the audit-case mean, not the benchmark macro-average over seven interaction types. For binary metrics, MAE in percentage points equals the percentage of disagreements.

VLM-channel scores exceed the human reference by 5.00 percentage points for EFA, 0.19 grades for IS-g, and 2.31 percentage points for PSR. SOL and SCL are earlier by 0.34 and 0.59 seconds, reflecting a stricter H channel relative to V. Not every bias is distinguishable from zero: the CF-g and DSR intervals cross zero, and the CA-HDF interval includes zero at its lower endpoint. The PSR audit mean in Table 14 does not replace the benchmark model-level macro-average.

Disagreement diagnostics. Table 15 treats H as reference. TP means both channels pass; FP means only V passes; FN means only H passes; TN means both fail.

DSR agrees exactly on 92.00% of cases but has six FP and two FN decisions. CA-HDF agrees on 97.50% but has one FP. IS-g and CF-g have quadratic weighted κ of 0.867 and 0.901, while exact-

Table 15: **Disagreement diagnostics.** With H as reference, FP means V passes while H fails; FN is the reverse.

Check	N	TP	FP	FN	TN	Agree. (%)	κ
Case-level decisions							
DSR	100	18	6	2	74	92.00	0.767
CA-HDF	40	4	1	0	35	97.50	0.875
SOL presence	100	79	6	0	15	94.00	0.798
Full completion	100	34	6	1	59	93.00	0.851
Atomic decisions (descriptive counts)							
EFA atom	296	171	20	4	101	91.89	0.829
PSR atom	501	346	19	5	131	95.21	0.883
History-dependency gates							
CA-HDF u	40	30	2	0	8	95.00	0.857
CA-HDF c	40	11	1	0	28	97.50	0.939
CA-HDF d	40	23	1	0	16	97.50	0.948
CA-HDF h	40	21	2	1	16	92.50	0.848

Atoms from the same case are correlated. The 296 target and 501 protection atoms are not additional independent audit cases. Primary confidence intervals resample cases, not atoms. CA-HDF gate c is current-round completion; u, d, and h denote source establishment, dependency binding, and history preservation.

grade agreement is only 74.00% and 79.00%. Good aggregate association can therefore coexist with individual disagreements.

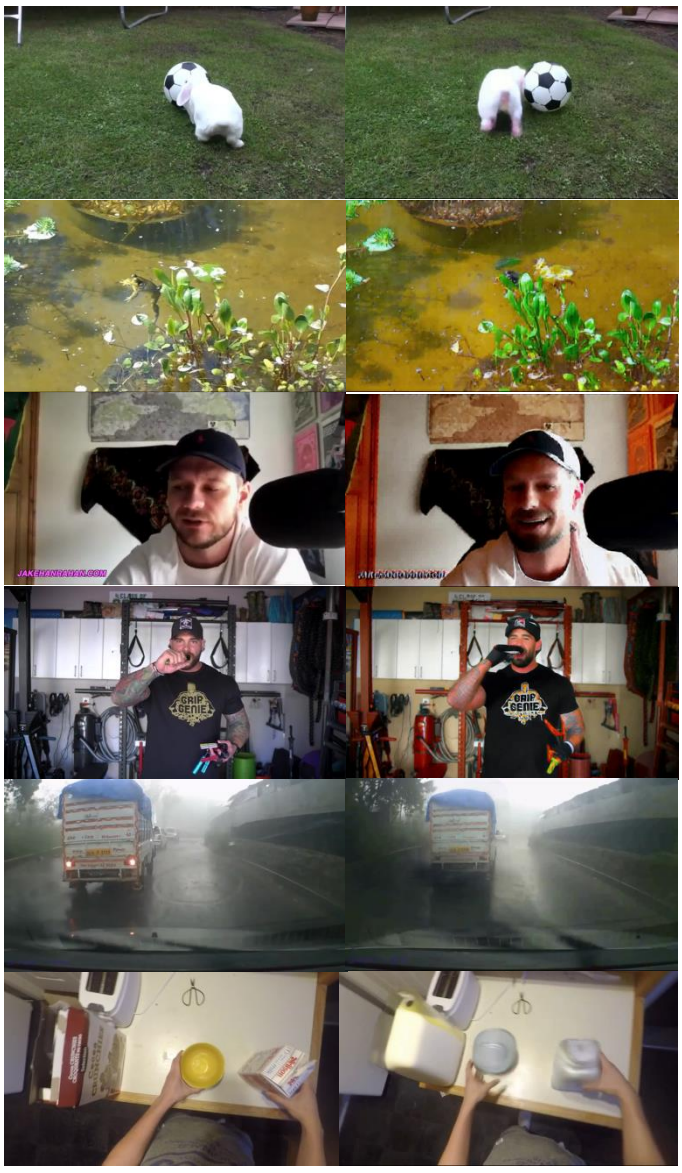
Timing validation must distinguish occurrence from timestamp agreement. SOL is observed in 79 human-reference and 85 VLM-channel cases, with 79 jointly observed cases used for paired timing statistics. The additional six occurrence disagreements remain in Table 15. Full-completion decisions agree on 93.00% of cases. Overall SCL correlation includes equal caps for jointly failed cases; among the 34 jointly completed cases, completion-time MAE is 0.37 seconds. Capped SCL correlation alone is thus insufficient evidence of accurate temporal localization. The 296 target and 501 protection atoms are descriptive judgment counts, not additional independent audit cases.

Interpretation. The audit shows high overall rank correlation and above-chance binary agreement between human review and the VLM judge, with a slightly stricter H channel relative to V, together with measurable score bias, individual disagreements, and subgroup differences. Because ordering, calibration, valid coverage, and disagreement patterns were assessed jointly, these results support the use of VLM-judged scoring for the nine audited metrics: the judge preserves the ordering of model outputs with small absolute error relative to the model differences the benchmark reports, and its mild leniency is quantified in Table 14. The support is not unbounded. Agreement is weakest on short-horizon binary decisions in the Synthetic stratum (DSR $\kappa = 0.464$), and CA-HDF rests on 40 eligible cases; conclusions that depend on these subgroups should be read with the corresponding intervals, and case-level disagreements of the kind in Table 15 remain possible even where aggregate association is high.

G.3 Failure Case Analysis

Figure 4 presents representative failure cases observed across different interaction types. These examples reveal that current models may fail at different stages of interactive generation rather than exhibiting a single dominant error pattern. Some models successfully execute the requested semantic change but fail to preserve unrelated visual content, as illustrated by the rabbit example, where the instructed motion is achieved while the subject identity drifts. Other cases show incomplete or incorrect semantic fulfillment, including failed viewpoint switching, unsuccessful object removal, and ambiguous object replacement. We also observe unintended spillover, where editing one target alters nearby objects or introduces spurious content, as well as substantial visual degradation after an otherwise detectable edit.

These cases highlight an important distinction captured by WorldLineBench: successful interaction requires more than producing a visible change. A model must simultaneously realize the intended



Motion succeeds, but identity drifts.
The rabbit develops abnormal pink ears and feet.

Instruction applied, but quality fail.
The pond shifts to an unintended look with severe visual degradation.

The requested viewpoint switch fails completely.
The video remains in the original third-person view.

Target changes fails with spurious additions.
The red gripper is unchanged, while extra grippers appear and quality degrades.

The requested removal fails completely.
The truck remains unchanged in the scene.

Replacement is ambiguous and affects nearby objects.
The bowl is not clearly turned into a silver can, and the milk carton is unintentionally changed.

Figure 4: Representative failure cases in WorldLineBench. From top to bottom, the examples illustrate identity drift despite successful motion editing, incorrect appearance transformation with severe quality degradation, failed viewpoint switching, spurious object additions, failed object removal, and ambiguous replacement with unintended changes to nearby content. These examples expose complementary failures in semantic fulfillment, preservation, and visual quality.

target, preserve content outside the requested modification, and maintain acceptable visual quality throughout the edited continuation. The observed failures therefore motivate evaluating interaction consistency, preservation, and video quality jointly rather than treating semantic fulfillment alone as sufficient evidence of successful interaction.