OOPSWORLD! Operation-Grounded Seamless World Generation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study seamless world generation, where a model must update a persistent visual world according to online user operations rather than full next-scene descriptions. To make this task well-defined, we introduce an operation-centric taxonomy that represents each scene with a global-local schema and each boundary with a typed operator program specifying change and preservation. This taxonomy enables SEAMLESSWORLD, a synthetic data and benchmark pipeline with 8 operation families, 23 atomic operators, dual prompt surfaces, and schema-derived evaluation probes. We then develop OOPSWORLD, a causal video generator that learns from SEAMLESSWORLD using operation-conditioned decoupled distillation: the student is driven by operation prompts and visual history, while scene-conditioned teachers provide target-state supervision. We further apply dynamics curriculum learning to progressively compose simple edits into high-dynamic world transitions. OOPSWORLD substantially improves dynamic responsiveness and long-video quality, achieving 91.67 VBench Dynamics and 80.42 Long-Quality with 100K samples. A systematic explicit–implicit study further reveals a persistent operation-prompting gap in current systems, especially for preserving visual anchors, calling for operation-driven world models beyond next-scene prompt switching.

1 Introduction

Real-time world generation aims to turn video synthesis into an interactive visual system: a model continuously renders a persistent world, receives user controls online, and updates the world without restarting the video. This capability is important for interactive storytelling, virtual production, game-like simulation, embodied agents, and open-ended content creation [4, 59, 25]. Recent streaming video diffusion systems have moved toward this goal through causal rollout, prompt switching, and memory-based continuity [70, 41, 44, 71].

However, existing systems are still largely driven by next-scene descriptions: the user provides a new prompt for the next clip, and the model tries to continue rendering it. This interface does not match how human natively interact in their real life, where humans perform actions and operations on their surroundings directly rather than dreaming full details of the next scene. Moreover, a deeper issue arise that world dynamics have not been clearly defined at the operation level. Existing taxonomy often describes the world only by the scene or shot-level information without considering detailed operations at its boundary, which is essential for seamless transition. As a result, models either over-copy the previous scene and produce weak dynamics, or over-follow the new prompt and reset identity, background, or style.

We address this gap by introducing an operation-oriented taxonomy for seamless world dynamics and building an automatic data pipeline. Each segmented scene is represented by a global-local scene schema: global fields describe background, environment, camera, style, and story beat, while

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

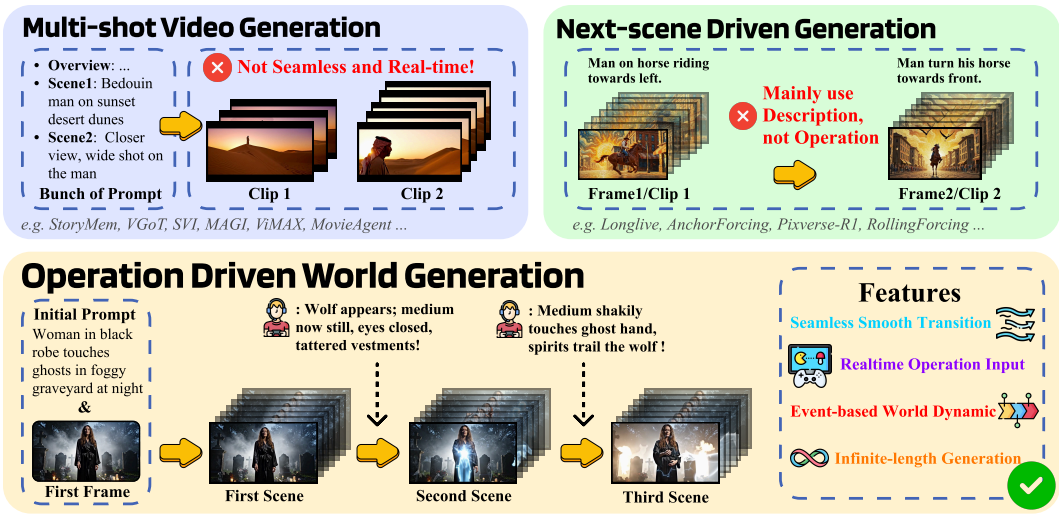


Figure 1: Comparison of multi-shot, next-scene-driven, and operation-driven video generation. Multi-shot methods rely on offline scene prompts and generate separated clips, while next-scene-driven systems switch to a new target-scene description during rollout. Both interfaces leave boundary-level operations and preservation constraints implicit. OOPSWORLD instead receives online operation inputs that executes the requested changes while preserving specified anchors to support seamless transitions, real-time control, event-level dynamics, and long-horizon generation.

local fields describe entities, attributes, actions, and relations. Each boundary is represented as an operator program over this schema, with 8 operation families and 23 atomic operators. This program explicitly specifies what changes, what should be preserved, and how large the update is. We then audit existing multi-scene and long-video corpora under this taxonomy [2, 35, 92, 70], showing that most collapses to small attribute/action changes or over-dense, entangled annotations without clean supervision for any single operation (Figure 2). Based on this finding, we build SEAMLESSWORLD, an operation-oriented synthetic data pipeline that samples valid operator programs, applies them to structured scene states, and renders both explicit target-scene prompts and implicit operation prompts from the same hidden trajectory.

Beyond data, seamless world generation also has a method mismatch. A text-to-video teacher naturally learns from a target-scene prompt, but an interactive student must generate the next scene from a different condition: the user intented interaction. We therefore detach teacher and student conditioning. The teacher remains a strong T2V model supervised by the target scene, while the student receives the operation-oriented instruction and the previous scene information as KV cache. This lets the student learn operation-based dynamics from visual history even when its condition differs from the teacher's. We further introduce Dynamics Curriculum Learning: because high-dynamic switches are compositions of atomic changes, training starts from low-score single-operation updates and gradually moves to higher-score, multi-operator transitions. With only 100K samples, our model OOPSWORLD (**Op**eration-grounded **S**eamless **World** Model) surpasses Anchor-Forcing on VBench Dynamics, improving from 73.61 to 91.67, and raises long-video quality from 75.66 to 80.42 while using only 0.42$\times$ of Anchor-Forcing's reported training budget [30, 71].

Finally, we close the evaluation gap for text-based real-time world generation by introducing two benchmark levels and a standardized evaluation suite grounded in our taxonomy and human-aligned MLLM-as-a-Judge, with probes decomposing generated multi-scene videos into segment state, transition preservation, change, smoothness, and sequence structure. Across 10 representative real-time, long-video, multi-prompt, and multi-shot systems, we find a clear interface gap between explicit next-scene and implicit operation prompting: the average F1 gap is 2.2, with preservation dropping 6.4 points. Several baselines show even larger gaps, indicating that current generators execute changes by weakening preservation rather than editing a persistent world state. In contrast, WORLDOPS is robust to the prompt surface ($\Delta_{\text{avg}} = -1.3$) and qualitatively follows requested changes while composing multiple operations (e.g., jointly changing camera motion and visual style). This suggests seamless generation should be evaluated and optimized as operation-driven world modeling, not next-scene prompt switching.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

2 Operation-Centric Formulation and Data Construction

2.1 Task Formulation: Seamless World Generation

We study **seamless world generation**: a causal process in which a system observes the current visual world, receives an online control, and updates the world without breaking continuity. Unlike standard text-to-video, where a prompt describes one isolated clip, or multi-shot generation, where a sequence of scene captions is given in advance, seamless world generation is driven by *operations*. A user may add an entity, change the weather, move the camera, advance an event, or preserve a character while other parts of the scene evolve. The input is therefore not a full target-scene caption, but a partial world update together with a preservation contract.

Let $W_{<i}$ denote the world generated and displayed before step i. At step i, the system receives a control (u_i, P_i), where u_i is an operation signal and $P_i \subseteq \mathcal{A}$ is a preserve set over a fixed anchor inventory $\mathcal{A}$. The operation signal can be natural language, a symbolic action, a UI event, a policy output, or any other model-consumable control. Future controls $\{(u_j, P_j)\}_{j>i}$ are unknown. The model must extend the displayed world causally,

$$W_{<i} \xrightarrow{(u_i,\, P_i)} W_i, \qquad W_i = f_\theta\big(W_{<i},\, u_i,\, P_i\big),$$

executing u_i while keeping every anchor in P_i valid across the boundary, and committing W_i before the next control is observed. This differs from next-scene prompting, which directly describes W_i and leaves the actual change and preservation constraints implicit. It also differs from video editing, which assumes a single before/after pair; FLF2V, where both endpoints are provided; and long-horizon rollout, which lacks per-boundary user control.

2.2 Operation Taxonomy for World Dynamics

To make operations well-defined, raw pixels are not enough: the model may visually change the world in ways that are hard to attribute, and a preserve contract cannot be checked from the prompt alone. We therefore introduce a structured layer for describing segment states and boundary changes. The design follows the observation that scene understanding is multi-scale, event perception is organized around meaningful state changes, and visual narrative depends on camera, motion, and shot structure [47, 46, 81, 16].

Each segment W_i is represented by a compact schema $\sigma_i = (g_i, \ell_i, d_i)$. The global component g_i records the world backbone, including location, scene type, static anchors, time, weather, atmosphere, camera, lighting, style, palette, and story beat. The local component ℓ_i records salient entities, their attributes, actions, and relations. The dynamics component d_i summarizes within-segment evolution as initial condition, main change, and terminal condition. This schema is not a full simulator state; it is the minimum state needed to define what changed, what stayed, and what should be evaluated.

For two adjacent segments, we compute a realized schema difference: $\Delta_i = \mathrm{Diff}(\sigma_{i-1}, \sigma_i)$, which exposes changed global fields, changed entities, attribute/action updates, relation rewrites, and actually preserved anchors. The preserve set P_i is checked against this realized Δ_i, not against the text: if an instruction claims to keep BACKGROUND but Δ_i shows a location swap, the preservation contract is violated.

An operation is a typed transformation on σ. We use 23 atomic operators in 8 families (Table 9; full inventory in Appendix D.4). Five families act on the global state g (background, environment, camera, style, and story) and three act on the local state ℓ (entity, attribute/action, and relation). At boundary i, the external control u_i is grounded as an operator program π_i that maps σ_{i-1} to σ_i. Valid programs must satisfy three compatibility constraints: anchor conflicts forbid mutating fields declared in P_i; sub-axis redundancy forbids combining operators that determine the same field; and graph validity forbids dangling relation endpoints or removal of preserved identities. Under this formulation, dynamics measure the size and composition of Δ_i, not the length of u_i: a low-dynamic boundary may edit one attribute or action, while a high-dynamic boundary may change camera, add an entity, and advance the story simultaneously.

2.3 SEAMLESSWORLD: Taxonomy-Grounded Syntehtic Dataset

Building on the schema and operator taxonomy, SEAMLESSWORLD materializes seamless world generation as data. The schema σ is the single ground truth; each transition is a sampled operator program π_i applied to σ_{i-1}; and every prompt surface, dynamics score, preserve set, and evaluation

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

probe is derived from the resulting schema artifacts. A seed T2V prompt instantiates σ_0, and subsequent segments are produced by controlled mutations of the previous schema. Dataset and benchmark statistics are shown in Figure 4.

Taxonomy audit of existing data. Before constructing SEAMLESSWORLD, we audit representative multi-scene video-text corpora under our 8-family taxonomy: TALC [2], VISTORY [92], CI-VID [35], and LONGLIVE [70]. Each adjacent caption pair is re-annotated into schema states by a paired-ingest LLM pipeline, then converted into an 8-class binary transition signature. We measure two complementary properties: cross-record diversity $\bar{d}_H = \sum_{k=1}^{8} 2p_k(1-p_k)$, the expected Hamming distance between two random transition signatures; and per-class focal contribution Focal_k, the fraction of transitions where class k is activated inside a focused change with at most two active classes. As shown in Figure 2, prior corpora fail in opposite ways: continuation-style data concentrates on small attribute/action changes, while dense annotation data co-fires many axes without cleanly supervising any one operation. SEAMLESSWORLD is the only audited corpus that scores high on both axes, reaching $\bar{d}_H = 3.41$ out of the theoretical maximum 4.0 and providing non-trivial focal coverage on six of eight operation families.

Synthetic data pipeline. We use VidProM [62] as the seed source, although the pipeline is corpus-agnostic. The construction follows the constraint-driven spirit of recent synthetic data efforts [19, 48], but adds explicit boundary operations and preservation contracts. First, a seed prompt is parsed into σ_0; a second pass fills visually implied fields, normalizes open-text values, validates closed vocabularies, and enforces structural constraints such as one main entity, unique entity identifiers, valid relation endpoints, and bounded list lengths.

For each boundary, we sample a target cell over scope, magnitude, dynamic bin, and switch count, then choose an operator program π_i valid under compatibility rules of §2.2. The boundary score is

$$S_{\text{bdry}}(\pi_i, \Delta_i) = \sum_{c \in C(\pi_i)} w_c + \beta_{\text{mixed}} + \beta_{\text{macro}},$$

with family weights $\text{BG} = 4$, $\text{STYLE} = \text{STORY} = 3$, $\text{ENV} = \text{CAM} = \text{ENTITY} = \text{REL} = 2$, $\text{ATTR} = 1$. β_{mixed} fires when both global and local sides change, and β_{macro} fires when Δ_i alters a backbone anchor such as the global scene backbone or main-character identity. The sequence score $S_{\text{seq}} = \frac{1}{N}\sum_i S_{\text{bdry}} + \frac{1}{2}\max(0, n-2)$ controls construction only and is never shown to the model. Structural operators directly rewrite schema fields and refresh d; semantic-first operators, such as story progress, action progress, entity entry/removal, and main-character swap, first generate a short context-aware continuation before emitting the mutated field and updated dynamics.

Dual prompt surfaces and validation. Each schema trajectory is rendered into two prompt surfaces that share the same hidden σ and visibility mask. The *explicit* surface describes each segment as a standalone cinematic target-scene prompt. The *implicit* surface describes each boundary as an operation from the previous segment: what changes, what should remain, and how the transition should unfold. This lets us compare next-scene prompting and operation-style prompting on identical world trajectories. Before writing a sample, we apply structural validation, round-trip prompt validation, preserve/change validation, and deduplication. These filters reject ill-formed schemas, invalid diffs, prompts that no longer match their source schema, missing keep clauses, empty changes, near-duplicate first scenes, and repeated transition signatures.

Schema-derived evaluation and benchmark. The same schema that builds the data also builds the benchmark. For every surviving trajectory, we compile VQA probes from σ, Δ, and the visibility mask. The probes cover segment state, negative counterfactuals, preserve checks, change checks, transition sameness/difference, smoothness, and sequence structure. They evaluate the semantic contract of seamless world generation: whether the requested operation happened, whether preserved anchors stayed stable, and whether the sequence contains the intended number of transitions. These probes do not replace low-level perceptual metrics; they measure the boundary-level behavior that free-form prompts and generic video metrics cannot localize.

3 OOPSWORLD: The Method

3.1 Preliminary: Causal Operation-Conditioned Rollout

We train a causal video generator in the latent space of a pretrained video diffusion model. Let $\mathbf{x}_i$ denote the i-th generated clip or scene segment, and let $\mathbf{K}_{i-1}$ denote the clean visual KV cache summarized from the already generated prefix. The cache acts as the persistent visual state of the

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

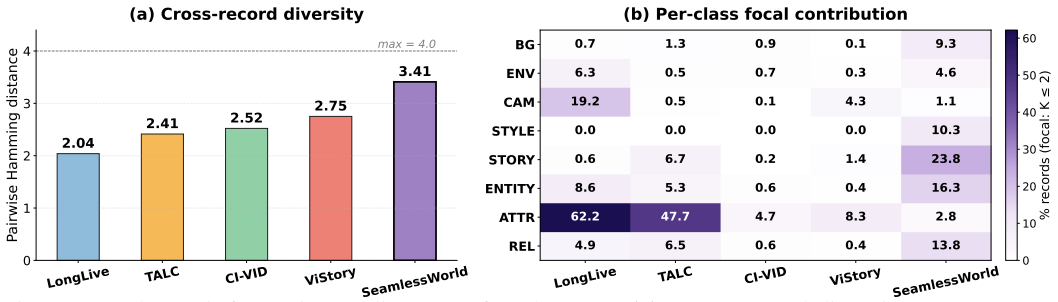


Figure 2: Schema-information audit across five datasets. **(a)** Cross-record diversity: SEAMLESSWORLD reaches 3.41 out of a theoretical maximum of 4.0, while all prior corpora fall between 2.04 and 2.79. **(b)** Per-class focal contribution: SEAMLESSWORLD is the only dataset with substantive focal contribution on every schema axis.

world. At transition i, the student receives the previous clean visual cache and a user operation prompt, and predicts the next clean latent clip:

$$\hat{\mathbf{x}}_{0,i} = G_\theta\left(\mathbf{z}_i, \eta_i; \mathbf{K}_{i-1}, c_i^{\text{op}}\right), \qquad \mathbf{K}_i = \text{sg}\left[\mathcal{C}_\theta\left(\hat{\mathbf{x}}_{0,i}; \mathbf{K}_{i-1}\right)\right]. \tag{1}$$

Here $\mathbf{z}_i$ and η_i are the generator noise and sampling input, c_i^{op} is the operation prompt, $\mathcal{C}_\theta$ extracts the KV states used by later clips, and $\text{sg}[\cdot]$ stops gradients through the history cache. Thus, the current transition is optimized while treating the previous world state $\mathbf{K}_{i-1}$ as fixed. During both training and inference, the cache is computed autoregressively from the student's generated clean prefix, keeping the student condition consistent across optimization and deployment. This follows recent causal video diffusion systems with rollout and KV caching [78, 28, 70], but differs in that our text input describes an operation rather than a full next-scene prompt.

For score distillation, the generated clean latent is re-noised as $\hat{\mathbf{x}}_{\tau,i} = \alpha_\tau \hat{\mathbf{x}}_{0,i} + \sigma_\tau \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and τ is the DMD re-noising level. We build on Distribution Matching Distillation (DMD), which updates a generator using the score difference between a frozen real diffusion model and a trainable fake score model that tracks the student distribution [77, 76]. The fake score model is trained by standard denoising score matching on detached student samples.

3.2 Operation-Conditioned Decoupled Distillation

A direct way to train interactive world generation is to use the same text condition for the student and the teacher. However, this couples two incompatible interfaces. A strong text-to-video teacher is naturally trained to follow an explicit scene description, e.g., "a red car turns into a crowded neon street at night." In contrast, a real-time world student should receive an operation, e.g., "drive forward and enter a busier street," and infer the next scene from the previous visual state. If the student is also given the target-scene prompt, it can solve training as ordinary prompt switching and may ignore the operation program. If the teacher is only given the operation prompt, the real score becomes weak because the teacher is not trained to ground operations in the hidden world state. We therefore decouple the student interface from the score-supervision interface.

For each transition in SEAMLESSWORLD, the hidden trajectory provides a previous state S_{i-1}, an operator program π_i, a target state $S_i = \mathcal{T}(S_{i-1}, \pi_i)$, an operation prompt c_i^{op}, and an explicit target-scene prompt c_i^{scene}. The student only observes the operation prompt and the previous visual cache, while the real and fake score models judge the generated clip under the target-scene prompt:

$$\text{student condition: } c_i^{\text{op}}, \qquad \text{real/fake score condition: } \chi_i^r = \chi_i^f = c_i^{\text{scene}}, \qquad c_i^{\text{scene}} \notin \text{Input}(G_\theta). \tag{2}$$

Thus, the teacher and fake score remain scene-conditioned, while the generator receiving the gradient is operation-conditioned. The target-scene prompt is used only as a supervision interface and is never fed to the student.

This decoupling is supported by the CFG structure of DMD. Let the DMD sample-space direction be $s_{\psi,\gamma}^{\text{real}} - s_\phi^{\text{fake}}$, where ψ is the frozen real teacher, ϕ is the fake score model, and γ is the classifier-free

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

guidance scale. For a CFG teacher, the scene-conditioned DMD direction can be decomposed as

$$\begin{aligned}\Delta_i(\tau) &= s^{\text{real}}_{\psi,\gamma}\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid c_i^{\text{scene}}\right) - s^{\text{fake}}_{\phi}\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid c_i^{\text{scene}}\right) \\ &= \underbrace{s_{\psi}\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid c_i^{\text{scene}}\right) - s^{\text{fake}}_{\phi}\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid c_i^{\text{scene}}\right)}_{\text{distribution-matching regularizer}} \qquad +(\gamma-1)\underbrace{\left[s_{\psi}\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid c_i^{\text{scene}}\right) - s_{\psi}\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid \varnothing\right)\right]}_{\text{CFG augmentation engine}}.\end{aligned} \tag{3}$$

The first term stabilizes distribution matching on student samples, while the second term provides the teacher-side CFG force toward the desired target-scene distribution. Recent analysis of Decoupled DMD suggests that this CFG augmentation term is the dominant engine for few-step distillation, while the distribution-matching term mainly stabilizes training and suppresses artifacts [40]. This view naturally allows the teacher-side condition and the generator-side condition to differ: the teacher provides a scene-level target signal, whereas the student must realize that target from $(\mathbf{K}_{i-1}, c_i^{\text{op}})$. Because c_i^{scene} is hidden from the student, the only learnable path is to interpret the operation program against the persistent visual state.

We optimize the student with the following surrogate opDMD objective:

$$\mathcal{L}^{\text{sur}}_{\text{opDMD}} = \mathbb{E}_{i,\mathbf{z}_i,\eta_i,\tau,\boldsymbol{\epsilon}}\left[-\tilde{w}(\tau)\,\text{sg}\left[\Delta_i(\tau)\right]^{\top}\hat{\mathbf{x}}_{0,i}\right], \tag{4}$$

where $\hat{\mathbf{x}}_{0,i} = G_\theta(\mathbf{z}_i, \eta_i; \mathbf{K}_{i-1}, c_i^{\text{op}})$ and $\tilde{w}(\tau) = w(\tau)\alpha_\tau$ absorbs the re-noising chain-rule factor. The stop-gradient on $\Delta_i(\tau)$ is essential: it prevents backpropagation through the teacher, the fake score model, and score Jacobians. As a result, optimization follows the DMD sample-space direction while only updating the operation-conditioned student.

At inference time, the target-scene prompt c_i^{scene} is absent. The model continues the world autoregressively using only the generated clean visual cache $\mathbf{K}_{i-1}$ and the new user operation c_i^{op}.

3.3 Dynamics Curriculum Learning

High-dynamic transitions are usually compositions of several atomic schema edits (entity insertion, camera motion, layout change, relation update, interaction change), and optimizing them from scratch yields an ambiguous score signal that conflates missed operations, broken visual state, unstable transitions, and confused simultaneous edits. We therefore adopt *Dynamics Curriculum Learning*: each operator program in OOPSWORLD is assigned a dynamics score from our transition taxonomy, where low scores denote simple localized single-operation updates and high scores involve multiple coupled edits or large changes in camera, entities, relations, interactions, or layout. Training begins biased toward low-dynamic transitions, teaching the student the basic alignment among the user operation, the preserved world state in the clean KV cache, and the scene-level teacher score; hard examples are down-weighted rather than removed, so complex transitions remain visible without destabilizing optimization. As training progresses, the bias relaxes and multi-operator high-dynamic transitions become more frequent, forcing the same operation-conditioned mechanism to compose atomic changes into larger world updates—shifting training from local operation grounding to compositional world rollout while preserving identity, background, style, and spatial layout.

4 Experiments

Baselines. We compare against eleven publicly available video-generation systems that collectively span four families of long-video / multi-prompt generation: (i) *streaming causal autoregressive* models trained for indefinite extension (Rolling-Forcing [41], LongLive [70], Anchor-Forcing [71], MAGI-1 [53], YUME 1.5 [44]); (ii) *training-free multi-prompt diffusion control* on a single-shot backbone (DiTCtrl [5]); (iii) *anchor-/memory-conditioned multi-shot* pipelines that chain clips through an explicit visual hand-off (StoryMem [83], Stable-Video-Infinity & Pro [39], VideoGen-of-Thought [88]). Every baseline is driven through a unified adapter (`benchmark_inference.py`) that consumes the same multi-prompt JSONL contract (one prompt per scene, two surfaces `explicit` / `implicit`). A shared post-processing pass re-encodes every output to $n_{\text{scenes}} \times 80$ frames at 16 fps, 832×480 so all systems are evaluated on identical setting. Switch points are placed at evenly-spaced 5 s boundaries so that the prompt change at scene $i \to i+1$ is expected at frame $80i$. We also compare our model with competitive size models and methods [34, 23, 9, 58, 78, 28, 74, 75, 33, 70, 71].

Implementation. Our main experiments are built on Wan-2.1-T2V-1.3B [58], generating 5-second videos at 832×480 resolution. We first train the model on 16k ODE trajectories sampled from the base model, initializing it with causal attention masking following CausVid [78]. We then adopt

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Table 1: Comparison on VBench. *"Our rep."* denotes our reproduction of Anchor-Forcing on their random sampled 100k dataset. Long-Quality reports the average over 6 long-video dimensions.

Model	#Params	Resolution	Dynamic ↑	Evaluation scores ↑			Long-Quality ↑
				Total	Quality	Semantic	
Wan2.1 [58]	1.3B	832 × 480	73.61	84.27	85.24	80.43	-
Streaming AR models							
Self-Forcing [28]	1.3B	832 × 480	63.89	83.83	84.60	80.79	-
Infinity-RoPE [74]	1.3B	832 × 480	58.33	83.55	84.34	80.43	-
DeepForcing [75]	1.3B	832 × 480	63.89	83.85	84.61	80.84	-
MemFlow [33]	1.3B	832 × 480	50.00	81.76	83.74	73.86	-
LongLive [70]	1.3B	832 × 480	37.50	82.81	83.26	**81.01**	-
Anchor-Forcing [71]	1.3B	832 × 480	73.61	83.99	84.84	80.57	75.66
Our Anchor-Forcing reproductions and OOPSWORLD							
AF (our rep., 50k)	1.3B	832 × 480	41.67	83.54	84.19	80.94	73.13
AF (our rep., 100k)	1.3B	832 × 480	65.28	**84.02**	**85.07**	79.83	75.85
OOPSWORLD (50k)	1.3B	832 × 480	41.67	83.20	83.97	80.11	72.55
OOPSWORLD (100k)	1.3B	832 × 480	**91.67**	83.74	84.89	79.18	**80.42**

the switch-prompt training scheme of LongLive [70], where the KV cache employs the tri-RoPE and anchor-frame design from Anchor-Forcing [71]. Since our study focuses on operation-based prompt switching, training is conducted on two-clip switches of 5 seconds per clip, without streaming long-video training. In our implementation, the sink cache and anchor cache store KV states for $N_S = 3$ and $N_J = 3$ latent frames, respectively, while the rolling local cache uses a window size of $W = 9$. Both the actor and critic are optimized with AdamW, using learning rates $lr = 1.0 \times 10^{-5}$ and $lr_{\text{critic}} = 2.0 \times 10^{-6}$. To accelerate training, all text latents are pre-encoded in batch ahead of time. All comparisons are conducted at the 100k-sample scale, and training takes approximately 71 hours on 8 A100 GPUs, amounting to $0.42\times$ the training budget of Anchor-Forcing [71].

Evaluation Protocol. We evaluate generation quality on VBench (plus Long-Quality metric brought from VBench-Long) and scene-transition quality on our proposed two benchmarks (100-sample balanced and 50-sample high-dynamic ones). The sampling is stratified over operator families, dynamic bins, switch counts, prompt surface, and preserve-anchor profiles. This gives a small but controlled benchmark for model comparison, while the full synthetic pool remains available for training and larger-scale analysis. Based on our probe design, the transition evaluation is decomposed into five sub-metrics: *segment_quality*, *switch_preserve*, *switch_change*, *switch_smooth*, and *sequence_structure*. Since all probes are formulated as VideoQA tasks, we adopt the MLLM-as-a-Judge paradigm [8, 49] for evaluation. To select a reliable judge, we collected 5,042 human-annotated samples as ground truth and benchmarked two candidate models: Gemini 3.1 Pro (think_high) [20] and Qwen3.6-35B-A3B (FPS $= 4$, ThinkingBudget $= 4$K) [52]. For each sub-metric, we computed both Pearson correlations against human annotations, with full results reported in Table 2. Guided by these correlations, we use Qwen as the judge for *segment_quality* and *switch_preserve*, and Gemini for *switch_change*, *switch_smooth*, and *sequence_structure*.

Table 2: Validating MLLM-as-a-Judge.

Family	n	Q@OFF	Q@8k	G3.1@1K	Human
Segment	3,995	0.688	<u>0.796</u>	**0.801**	0.787
Preserve	264	0.143	0.765	<u>0.780</u>	**0.790**
Change	638	0.274	0.639	<u>0.745</u>	**0.763**
Smoothness	95	0.061	0.151	<u>0.659</u>	**0.688**
Structure	50	0.586	0.517	**0.704**	<u>0.701</u>

4.1 In-depth Analysis

Operation training improves dynamic responsiveness. Table 1 shows that OOPSWORLD substantially improves dynamic responsiveness while maintaining competitive overall quality. With 100K samples, OOPSWORLD achieves 91.67 VBench Dynamics and 80.42 Long-Quality, outperforming Anchor-Forcing by 18.06 points on Dynamics and 4.76 points on Long-Quality. Compared with our 100K Anchor-Forcing reproduction, the gains are even larger on Dynamics (+26.39) and remain clear on Long-Quality (+4.57). Meanwhile, total quality and semantic quality stay close to the strongest 1.3B baselines, suggesting that operation-conditioned training improves world responsiveness without simply trading off visual fidelity.

Implicit operation prompts expose a preservation bottleneck. Table 3 compares explicit target-scene prompts and implicit operation prompts on the same hidden trajectories. Although the average Δ_{avg} is 2.2, the gap is highly uneven across dimensions: preservation drops by 6.4 points on average,

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.



Figure 3: Qualitative study of OOPSWORLD comparing to AnchorForcing baseline.

while smoothness drops by 5.1 points. In contrast, the change score has a negative gap (-4.8), meaning that implicit prompts often make the desired edit clearer, but existing models tend to realize the edit by breaking anchors or destabilizing the transition. This failure is broad: SVI-Pro, Anchor-Forcing, DiTCtrl, and LongLive all show large Δ_{avg} values, and some models lose more than 15 F1 points on preservation under implicit prompts. OOPSWORLD reduces this surface mismatch, obtaining a small $\Delta_{\text{avg}} = -1.3$ and improving from explicit to implicit on both preserve (87.5 to 90.5) and change (47.0 to 53.1), though smoothness and structure remain open challenges.

Qualitative cases show compositional operation following. Figure 3 further illustrates the difference between prompt switching and operation following. In the first case, OOPSWORLD better follows the requested semantic change from a tropical island shore to a resort entrance while updating the main character composition, whereas Anchor-Forcing tends to preserve the old scene too strongly. In the second case, the requested operation simultaneously changes the visual style from classic western cinematic to low-poly 3D render and changes the camera from a static wide view to pan-right tracking; OOPSWORLD reflects both changes in one coherent transition, while Anchor-Forcing mainly keeps the original look and motion pattern. The remaining cases show similar compositional behavior for palette, entity addition, time/story progression, main-character swap, and relation removal. These examples support the quantitative finding that seamless world generation requires models to execute operations over a persistent state, rather than merely switching to a new descriptive prompt.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Table 3: Baselines on SEAMLESSWORLD benchmark (100 records). E and I denote explicit and implicit settings, respectively. Δ_{avg} reports the average explicit–implicit gap over the four F1 metrics, excluding structural accuracy. Best results are in bold and second-best results are underlined.

Method	Seg. state		Preserve		Change		Smooth		Struct Acc.		Δ_{avg}
	E	I	E	I	E	I	E	I	E	I	E-I
AnchorForcing [71]	80.8	79.9	84.4	76.6	41.7	45.9	60.1	44.1	<u>47.0</u>	45.0	5.1
DiTCtrl [5]	78.7	72.5	64.4	48.3	**65.0**	**71.1**	44.9	42.8	41.0	40.0	4.6
LongLive [70]	<u>81.0</u>	80.2	**91.5**	84.7	31.5	41.1	63.6	45.5	42.0	41.0	4.0
MAGI [53]	<u>81.0</u>	79.7	89.0	**91.3**	33.1	30.3	38.1	<u>58.0</u>	19.0	10.0	-4.5
RollingForcing [41]	80.4	**80.6**	85.9	82.2	35.5	40.6	57.0	50.0	41.0	39.0	1.4
StoryMem [83]	**81.1**	79.6	81.7	70.5	47.2	<u>60.5</u>	31.3	22.2	28.9	**53.1**	2.1
SVI [39]	79.0	76.5	71.9	65.0	48.8	54.5	<u>64.5</u>	**64.3**	43.0	41.0	1.0
SVI-Pro [39]	75.0	72.0	68.9	53.2	<u>53.5</u>	58.2	61.8	53.1	45.0	40.0	5.7
VGOT [88]	68.5	61.7	83.9	76.4	36.0	38.3	44.1	42.8	41.0	45.0	3.3
YuMe 1.5 [44]	80.5	<u>80.3</u>	<u>91.1</u>	<u>91.2</u>	32.0	30.4	**67.7**	57.8	0.0	0.0	2.9
OOPSWORLD	80.1	79.3	87.5	90.5	47.0	53.1	47.8	44.5	**49.0**	<u>46.0</u>	-1.3
Avg. Δ	2.2		6.4		-4.8		5.1		–		2.2

5 Related Works

Video World Models. Recent video generation is moving from short-clip synthesis toward controllable world simulation. Diffusion-transformer backbones [58], chunk-wise autoregressive systems [53, 85, 31], and interactive models for games and embodied agents [25, 65, 27, 91, 12] extend earlier neural game engines and action-controllable world models [59, 4] toward streaming and action-conditioned rollout. A central technical direction is to convert bidirectional diffusion into causal generators through per-token noise schedules, distillation, and long-context teachers [7, 78, 28, 90, 13]. Recent systems further improve the streaming interface with prompt-aware KV re-caching, rolling denoising, adaptive memory, and positional or cache resets [70, 41, 74, 15]. Anchor Forcing [71] is especially relevant to prompt switching because it preserves memory anchors and realigns RoPE across prompt changes. FLEX, Deep Forcing, Relax Forcing, and Knot Forcing [38, 75, 86, 68] instead mainly serve as horizon-extension, cache-compression, or stability modules. In contrast, seamless world generation requires a generator that can accept text-conditioned scene changes while preserving specified anchors across boundaries.

Evaluation for Realtime World Generation. Video generation evaluation has progressed from perceptual clip quality to world-level controllability. General-purpose suites [30, 29, 87, 24, 50, 26] measure visual quality, temporal consistency, prompt alignment, and preference, while temporal and compositional T2V benchmarks [32, 56, 18] test ordered and compositional prompt following. More recent world-model evaluations target physically plausible and controllable environments, including instruction following, 3D or 4D consistency, dynamics, and commonsense physics [37, 17, 42, 14, 64, 3, 45, 22, 21]. WorldMark [69] is closest in standardizing scenes and action interfaces for action-conditioned image-to-video models, but these benchmarks still largely evaluate a single continuous clip. A separate line studies multi-shot, long-video, and transition-aware generation. Time-aligned and interleaved multi-scene corpora [2, 35], start/end-state transition benchmarks [72, 79], narrative and storytelling benchmarks [63, 57, 92, 55], and cinematic or reference-aware generation methods [6, 54, 36, 66, 51, 1, 43, 61] all broaden evaluation beyond isolated shots.

6 Conclusion

We introduced OOPSWORLD, an operation-centric formulation for seamless world generation that shifts the interface from describing the next scene to operating on a persistent visual world. Built on a global-local scene schema, our taxonomy, data pipeline, and SEAMLESSWORLD that explicitly define what should change, what should be preserved, and how each boundary transition should be evaluated. We further proposed an operation-conditioned causal generator with decoupled distillation and dynamics curriculum learning, enabling the model to learn world dynamics from operation inputs rather than target-scene prompts. Experiments show that this formulation improves dynamic responsiveness and long-video quality, suggesting a practical path toward real-time, controllable, and indefinitely extendable video worlds.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

References

[1] Zhaochong An, Menglin Jia, Haonan Qiu, Zijian Zhou, Xiaoke Huang, Zhiheng Liu, Weiming Ren, Kumara Kahatapitiya, Ding Liu, Sen He, et al. Onestory: Coherent multi-shot video generation with adaptive memory. *arXiv preprint arXiv:2512.07802*, 2025.

[2] Hritik Bansal, Yonatan Bitton, Michal Yarom, Idan Szpektor, Aditya Grover, and Kai-Wei Chang. Talc: Time-aligned captions for multi-scene text-to-video generation. *arXiv preprint arXiv:2405.04682*, 2024.

[3] Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025.

[4] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning*, pages 4603–4623, 2024. Also available as arXiv:2402.15391.

[5] Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7763–7772, 2025.

[6] Agneet Chatterjee, Rahim Entezari, Maksym Zhuravinskyi, Maksim Lapin, Reshinth Adithyan, Amit Raj, Chitta Baral, Yezhou Yang, and Varun Jampani. Stable cinemetrics: Structured taxonomy and evaluation for professional video generation. In *Advances in Neural Information Processing Systems*, 2025. Also available as arXiv:2509.26555.

[7] Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *arXiv preprint arXiv:2407.01392*, 2024.

[8] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024.

[9] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

[10] Jiaben Chen, Zixin Wang, Ailing Zeng, Yang Fu, Xueyang Yu, Siyuan Cen, Julian Tanke, Yihang Chen, Koichi Saito, Yuki Mitsufuji, and Chuang Gan. Talkcuts: A large-scale dataset for multi-shot human speech video generation. *arXiv preprint arXiv:2510.07249*, 2025.

[11] Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Daquan Zhou, Huan Ling, Haozhe Liu, Hongwei Yi, Hao Zhang, Muyang Li, Yukang Chen, Han Cai, Sanja Fidler, Ping Luo, Song Han, and Enze Xie. SANA-Video: Efficient video generation with block linear diffusion transformer. *arXiv preprint arXiv:2509.24695*, 2025.

[12] Junyi Chen, Haoyi Zhu, Xianglong He, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Zhoujie Fu, Jiangmiao Pang, et al. Deepverse: 4d autoregressive video generation as a world model. *arXiv preprint arXiv:2506.01103*, 2025.

[13] Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026.

[14] Yubin Chen, Xuyang Guo, Zhenmei Shi, Zhao Song, and Jiahao Zhang. T2VWorldBench: A benchmark for evaluating world knowledge in text-to-video generation. *arXiv preprint arXiv:2507.18107*, 2025.

[15] Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Lol: Longer than longer, scaling video generation to hour. *arXiv preprint arXiv:2601.16914*, 2026.

[16] James E. Cutting, Jordan E. DeLong, and Christine E. Nothelfer. Attention and the evolution of hollywood film. *Psychological Science*, 21(3):432–439, 2010.

[17] Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. *arXiv preprint arXiv:2504.00983*, 2025.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

[18] Weixi Feng, Jiachen Li, Michael Saxon, Tsu-Jui Fu, Wenhu Chen, and William Yang Wang. TC-Bench: Benchmarking temporal compositionality in text-to-video and image-to-video generation. *arXiv preprint arXiv:2406.08656*, 2024.

[19] Ziqi Gao, Weikai Huang, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. Generate any scene: Scene graph driven data synthesis for visual generation training. In *The Fourteenth International Conference on Learning Representations*, 2026. ICLR 2026 poster.

[20] Google DeepMind. Gemini 3.1 pro: A smarter model for your most complex tasks. `https://blog.google/products/gemini/gemini-3-1-pro/`, February 2026. Model card: `https://deepmind.google/models/model-cards/gemini-3-1-pro/`.

[21] Jing Gu, Xian Liu, Yu Zeng, Ashwin Nagarajan, Fangrui Zhu, Daniel Hong, Yue Fan, Qianqi Yan, Kaiwen Zhou, Ming-Yu Liu, and Xin Eric Wang. PhyWorldBench: A comprehensive evaluation of physical realism in text-to-video models. In *International Conference on Learning Representations*, 2026. Also available as arXiv:2507.13428.

[22] Xuyang Guo, Jiayan Huo, Zhenmei Shi, Zhao Song, Jiahao Zhang, and Jiale Zhao. T2VPhysBench: A first-principles benchmark for physical consistency in text-to-video generation. *arXiv preprint arXiv:2505.00337*, 2025.

[23] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

[24] Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, Jie Zhang, et al. Video-bench: Human-aligned video generation benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. Also available as arXiv:2504.04907.

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

[26] Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, et al. Videoscore2: Think before you score in generative video evaluation. *arXiv preprint arXiv:2509.22799*, 2025.

[27] Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, et al. Relic: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

[28] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

[29] Ziqi Huang et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024.

[30] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianyi Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[31] Inferix Team, Tianyu Feng, Yizeng Han, Jiahao He, Yuanyu He, Xi Lin, Teng Liu, Hanfeng Lu, Jiasheng Tang, Wei Wang, et al. Inferix: A block-diffusion based next-generation inference engine for world simulation. *arXiv preprint arXiv:2511.20714*, 2025.

[32] Pengliang Ji, Chuyang Xiao, Huilin Tai, and Mingxiao Huo. T2VBench: Benchmarking temporal dynamics for text-to-video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 5325–5335, 2024. Retracted.

[33] Sihui Ji, Xi Chen, Shuai Yang, Xin Tao, Pengfei Wan, and Hengshuang Zhao. Memflow: Flowing adaptive memory for consistent and efficient long video narratives. *arXiv preprint arXiv:2512.14699*, 2025.

[34] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024.

[35] Yiming Ju, Jijin Hu, Zhengxiong Luo, Haoge Deng, Hanyu Zhao, Li Du, Chengwei Wu, Donglin Hao, Xinlong Wang, and Tengfei Pan. CI-VID: A coherent interleaved text-video dataset. *arXiv preprint arXiv:2507.01938*, 2025.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

[36] Ozgur Kara, Krishna Kumar Singh, Feng Liu, Duygu Ceylan, James M. Rehg, and Tobias Hinz. Shotadapter: Text-to-multi-shot video generation with diffusion models. *arXiv preprint arXiv:2505.07652*, 2025.

[37] Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025.

[38] Jia Li, Xiaomeng Fu, Xurui Peng, Weifeng Chen, Youwei Zheng, Tianyu Zhao, Jiexi Wang, Fangmin Chen, Xing Wang, and Hayden Kwok-Hay So. Train short, inference long: Training-free horizon extension for autoregressive video generation. *arXiv preprint arXiv:2602.14027*, 2026.

[39] Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. *arXiv preprint arXiv:2510.09212*, 2025.

[40] Dongyang Liu, Peng Gao, David Liu, Ruoyi Du, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Hongsheng Li, et al. Decoupled dmd: Cfg augmentation as the spear, distribution matching as the shield. *arXiv preprint arXiv:2511.22677*, 2025.

[41] Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

[42] Yiting Lu, Wei Luo, Peiyan Tu, Haoran Li, Hanxin Zhu, Zihao Yu, Xingrui Wang, Xinyi Chen, Xinge Peng, Xin Li, and Zhibo Chen. 4DWorldBench: A comprehensive evaluation framework for 3d/4d world generation models. *arXiv preprint arXiv:2511.19836*, 2025.

[43] Xiangyang Luo, Qingyu Li, Xiaokun Liu, Wenyu Qin, Miao Yang, Meng Wang, Pengfei Wan, Di Zhang, Kun Gai, and Shao-Lun Huang. Filmweaver: Weaving consistent multi-shot videos with cache-guided autoregressive diffusion. *arXiv preprint arXiv:2512.11274*, 2025.

[44] Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A text-controlled interactive world generation model. *arXiv preprint arXiv:2512.22096*, 2025.

[45] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024.

[46] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[47] Stephen E. Palmer. The effects of contextual scenes on the identification of objects. *Memory & Cognition*, 3(5):519–526, 1975.

[48] Jae Sung Park, Zixian Ma, Linjie Li, Chenhao Zheng, Cheng-Yu Hsieh, Ximing Lu, Khyathi Chandu, Quan Kong, Norimasa Kobori, Ali Farhadi, Yejin Choi, and Ranjay Krishna. Synthetic visual genome. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9073–9086, June 2025.

[49] Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, et al. Judge anything: Mllm as a judge across any modality. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5742–5753, 2025.

[50] Zelu Qi, Ping Shi, Shuqi Wang, Chaoyang Zhang, Fei Zhao, Zefeng Ying, Da Pan, Xi Yang, Zheqi He, and Teng Dai. T2VEval: Benchmark dataset and objective evaluation method for t2v-generated videos. *arXiv preprint arXiv:2501.08545*, 2025.

[51] Lu Qiu, Yizhuo Li, Yuying Ge, Yixiao Ge, Ying Shan, and Xihui Liu. Animeshooter: A multi-shot animation dataset for reference-guided video generation. *arXiv preprint arXiv:2506.03126*, 2025.

[52] Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all. `https://qwen.ai/blog?id=qwen3.6-35b-a3b`, April 2026.

[53] Sand.ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, et al. MAGI-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

[54] Hanwen Shen, Jiajie Lu, Yupeng Cao, and Xiaonan Yang. Enhancing scene transition awareness in video generation via post-training. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 706–721, 2025. Also available as arXiv:2507.18046.

[55] Haoyuan Shi, Yunxin Li, Nanhao Deng, Zhenran Xu, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. MSVBench: Towards human-level evaluation of multi-shot video generation. *arXiv preprint arXiv:2602.23969*, 2026.

[56] 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. *arXiv preprint arXiv:2407.14505*, 2024.

[57] Zhengxu Tang, Zizheng Wang, Luning Wang, Zitao Shuai, Chenhao Zhang, Siyu Qian, Yirui Wu, Bohao Wang, Haosong Rao, Zhenyu Yang, and Chenwei Wu. Seqbench: Benchmarking sequential narrative generation in text-to-video models. *arXiv preprint arXiv:2510.13042*, 2025.

[58] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[59] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations*, 2025. Also available as arXiv:2408.14837.

[60] Dingrui Wang et al. Target-bench: Can video world models achieve mapless path planning with semantic targets? *arXiv preprint arXiv:2511.17792*, 2025.

[61] Qinghe Wang et al. Multishotmaster: A controllable multi-shot video generation framework. *arXiv preprint arXiv:2512.03041*, 2025.

[62] Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. Datasets and Benchmarks Track.

[63] Yiping Wang, Xuehai He, Kuan Wang, Luyao Ma, Jianwei Yang, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Is your world simulator a good story presenter? a consecutive events-based benchmark for future long video generation. *arXiv preprint arXiv:2412.16211*, 2024.

[64] Zeqing Wang, Xinyu Wei, Bairui Li, Zhen Guo, Jinrui Zhang, Hongyang Wei, Keze Wang, and Lei Zhang. Videoverse: Does your t2v generator have world model capability to synthesize videos? *arXiv preprint arXiv:2510.08398*, 2025.

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

[66] Xiaoxue Wu, Xinyuan Chen, Yaohui Wang, and Yu Qiao. Shotdirector: Directorially controllable multi-shot video generation with cinematographic transitions. *arXiv preprint arXiv:2512.10286*, 2025.

[67] Xiaoxue Wu, Bingjie Gao, Yu Qiao, Yaohui Wang, and Xinyuan Chen. Cinetrans: Learning to generate videos with cinematic transitions via masked diffusion models. In *International Conference on Learning Representations*, 2026. Also available as arXiv:2508.11484.

[68] Steven Xiao, Xindi Zhang, Dechao Meng, Qi Wang, Peng Zhang, and Bang Zhang. Knot forcing: Taming autoregressive video diffusion models for real-time infinite interactive portrait animation. *arXiv preprint arXiv:2512.21734*, 2025.

[69] Xiaojie Xu, Zhengyuan Lin, Kang He, Yukang Feng, Xiaofeng Mao, Yuanyang Yin, Kaipeng Zhang, and Yongtao Ge. Worldmark: A unified benchmark suite for interactive video world models. *arXiv preprint arXiv:2604.21686*, 2026.

[70] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

[71] Yang Yang, Tianyi Zhang, Wei Huang, Jinwei Chen, Boxi Wu, Xiaofei He, Deng Cai, Bo Li, and Peng-Tao Jiang. Anchor forcing: Anchor memory and tri-region rope for interactive streaming video diffusion. *arXiv preprint arXiv:2603.13405*, 2026.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

[72] Zuhao Yang, Jiahui Zhang, Yingchen Yu, Shijian Lu, and Song Bai. Versatile transition generation with image-to-video diffusion. *arXiv preprint arXiv:2508.01698*, 2025.

[73] Zhengjian Yao, Yongzhi Li, Xinyuan Gao, Quan Chen, Peng Jiang, and Yanye Lu. Narrative weaver: Towards controllable long-range visual consistency with multi-modal conditioning. *arXiv preprint arXiv:2603.06688*, 2026.

[74] Hidir Yesiltepe, Tuna Han Salih Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. *arXiv preprint arXiv:2511.20649*, 2025.

[75] Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

[76] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024.

[77] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024.

[78] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. Also known as CausVid; arXiv:2412.07772.

[79] Zhiyu Yin, Zhipeng Liu, Kehai Chen, Lemao Liu, Jin Liu, Hong-Dong Li, Yang Xiang, and Min Zhang. VC-Bench: Pioneering the video connecting benchmark with a dataset and evaluation metrics. *arXiv preprint arXiv:2601.19236*, 2026.

[80] Hangjie Yuan, Weihua Chen, Jun Cen, Hu Yu, Jingyun Liang, Shuning Chang, Zhihui Lin, Tao Feng, Pengwei Liu, Jiazheng Xing, Hao Luo, Jiasheng Tang, Fan Wang, and Yi Yang. Lumos-1: On autoregressive video generation with discrete diffusion from a unified model perspective. *arXiv preprint arXiv:2507.08801*, 2025.

[81] Jeffrey M. Zacks, Nicole K. Speer, Khena M. Swallow, Todd S. Braver, and Jeremy R. Reynolds. Event perception: A mind-brain perspective. *Psychological Bulletin*, 133(2):273–293, 2007.

[82] Chenyu Zhang, Daniil Cherniavskii, Antonios Tragoudaras, Antonios Vozikis, Thijmen Nijdam, Derck W. E. Prinzhorn, Mark Bodracska, Nicu Sebe, Andrii Zadaianchuk, and Efstratios Gavves. Morpheus: Benchmarking physical reasoning of video generative models with real physical experiments. *arXiv preprint arXiv:2504.02918*, 2025.

[83] Kaiwen Zhang, Liming Jiang, Angtian Wang, Jacob Zhiyuan Fang, Tiancheng Zhi, Qing Yan, Hao Kang, Xin Lu, and Xingang Pan. Storymem: Multi-shot long video storytelling with memory. *arXiv preprint arXiv:2512.19539*, 2025.

[84] Peixuan Zhang, Zijian Jia, Kaiqi Liu, Shuchen Weng, Si Li, and Boxin Shi. STAGE: Storyboard-anchored generation for cinematic multi-shot narrative. *arXiv preprint arXiv:2512.12372*, 2025.

[85] Zeyu Zhang, Shuning Chang, Yuanyu He, Yizeng Han, Jiasheng Tang, Fan Wang, and Bohan Zhuang. Blockvid: Block diffusion for high-quality and consistent minute-long video generation. *arXiv preprint arXiv:2511.22973*, 2025.

[86] Zengqun Zhao, Yanzuo Lu, Ziquan Liu, Jifei Song, Jiankang Deng, and Ioannis Patras. Relax forcing: Relaxed kv-memory for consistent long video generation. *arXiv preprint arXiv:2603.21366*, 2026.

[87] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, 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.

[88] Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, et al. Videogen-of-thought: Step-by-step generating multi-shot video with minimal manual intervention. *arXiv preprint arXiv:2412.02259*, 2024.

[89] Yang Zhou et al. Drivinggen: A comprehensive benchmark for generative video world models in autonomous driving. *arXiv preprint arXiv:2601.01528*, 2026.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

[90] Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

[91] Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jie Zhou, and Jiwen Lu. Astra: General interactive world model with autoregressive denoising. *arXiv preprint arXiv:2512.08931*, 2025.

[92] Cailin Zhuang, Ailin Huang, Yaoqi Hu, Jingwei Wu, Wei Cheng, Jiaqi Liao, Hongyuan Wang, Xinyao Liao, Weiwei Cai, Hengyuan Xu, et al. Vistorybench: Comprehensive benchmark suite for story visualization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. Also available as arXiv:2505.24862.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the scope and main contributions of the paper, including the operation-centric formulation, SEAMLESSWORLD data/benchmark pipeline, WORLDOPS training method, and the reported improvements in dynamic responsiveness and long-video quality. The experimental claims are supported by the benchmark results and qualitative comparisons in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a Limitations paragraph discussing the compute-constrained 100K-sample training scale, the fact that the full dataset was not used for training, and future ablations on implicit/explicit supervision and dynamics-aware training.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present theorem- or lemma-style theoretical results requiring formal proofs. The equations and appendix material provide task definitions, metric definitions, and method objectives used for data construction, evaluation, and training.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the dataset construction pipeline, operation taxonomy, benchmark construction, evaluation probes, model training setup, and baseline adapters in the main text and appendices. We will release the dataset, code, evaluation suite, and model checkpoint to provide a direct path for reproducing the main results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will publicly release SEAMLESSWORLD, the benchmark and evaluation code, the training/inference code, and the WORLDOPS checkpoint with instructions for reproducing the main experiments. The released assets will be distributed under a CC BY-NC 4.0 license.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 specifies the training and evaluation setup, including backbone, resolution, sample scale, optimizer, learning rates, cache settings, benchmark construction, and evaluation protocol. Appendix D provides detailed baseline configurations and adapter settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Justification: We report aggregate benchmark scores without error bars, confidence intervals, or formal statistical significance tests. Repeating full video-generator training and evaluation across multiple seeds would be computationally expensive, so the results should be interpreted under the fixed benchmark and evaluation protocol described in the paper.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The paper reports the main training budget, including 100K training samples, 8 A100 GPUs, and approximately 71 hours of training, together with model size, resolution, and optimization details. Appendix D further documents the baseline configurations used for evaluation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The authors have reviewed the NeurIPS Code of Ethics and the research conforms to it. The work uses safety-filtered data sources, author/volunteer annotations for evaluation validation, and responsible release practices including non-commercial licensing and documentation.

Guidelines:

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper includes a Broader Impact paragraph discussing both positive AIGC applications and negative societal risks, including misinformation, impersonation, bias, copyright/style concerns, and compute cost. It also discusses responsible release and downstream mitigation practices.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: SEAMLESSWORLD is constructed from a safety-filtered seed dataset and the released dataset, code, and checkpoint will be accompanied by documentation, usage guidance, and a non-commercial CC BY-NC 4.0 license. These safeguards are intended to reduce unsafe or high-risk use of the released AIGC assets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper credits the existing datasets, models, benchmarks, and baseline systems used in the study, including VidProM, Wan-2.1, VBench/VBench-Long, and the compared video-generation systems. We will respect their original licenses and terms of use, and the released SEAMLESSWORLD assets will be provided under CC BY-NC 4.0.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces new assets including SEAMLESSWORLD, the benchmark/evaluation suite, training and inference code, and the WORLDOPS checkpoint. These assets are documented through the dataset statistics, taxonomy, record format, prompt templates, evaluation probes, training details, limitations, and release license.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Justification: The human annotations used to validate the MLLM-as-a-Judge protocol consist of 5,042 samples annotated by the authors and volunteers rather than by a paid crowdsourcing platform. The paper or supplementary material describes the annotation instructions, and no personal, demographic, biometric, or sensitive information was collected.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The work does not conduct a human-subjects study; the author/volunteer annotations are used only as evaluation labels for generated videos, and no personal or sensitive information about annotators is collected or analyzed. Therefore, IRB approval is not applicable under this setting.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs and MLLMs are important components of the data and evaluation pipeline, including schema parsing, semantic-first mutation, prompt rephrasing, roundtrip validation, dataset audit, and MLLM-as-a-Judge evaluation. The paper describes this usage in the data construction, evaluation protocol, and prompt appendices.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Limitations. Due to compute constraints, our main model is trained on a 100K-sample subset of SEAMLESSWORLD rather than the full generated dataset, and several large-scale ablations are left for future work. In particular, we have not fully isolated the effects of implicit versus explicit operation supervision, alternative operation/target-scene conditioning choices, and Dynamics Curriculum Learning under full-data training. Scaling these studies to the complete dataset would help clarify how schema-grounded data, operation-driven prompting, and dynamics-aware training each contribute to seamless world generation, and would provide the community with a more complete understanding of operation-grounded world dynamics.

Broader Impact. This work contributes to AIGC by making video generation more controllable, interactive, and operation-grounded. Potential positive impacts include more accessible tools for storytelling, education, virtual production, game prototyping, simulation, and creative content generation. At the same time, stronger controllable video generation may also increase risks associated with synthetic media, including misinformation, impersonation, deepfakes, biased or stereotyped content, copyright and style-related concerns, and increased compute cost. To support responsible use, we release the dataset, code, and checkpoint under a non-commercial CC BY-NC 4.0 license, build SEAMLESSWORLD from a safety-filtered seed dataset, and recommend that downstream deployments use provenance, watermarking, and usage restrictions for high-risk AIGC applications.

A Additional Related Works

Table 4: Comparison with representative video-generation, world-model, and multi-shot benchmarks/datasets. ✔ denotes explicit support, ✘ denotes absent support, and △ denotes partial or indirect.

Name	Target	Quantity	Taxonomy / labels	Primary focus	Text switch	Global/local GT	Boundary eval.
Benchmarks							
VBench [30]	General T2V quality	∼100 prompts / dim.	16 dimensions; 8 content types	Video quality and alignment	✘	✘	✘
VBench-2.0 [87]	Intrinsic faithfulness	Benchmark suite	5 families; 18 dimensions	Physics, commonsense, control	✘	△	△
T2V-CompBench [56]	Compositional T2V	1.4K prompts	7 composition types	Fine-grained compositionality	△	✘	✘
PhyGenBench [45]	Physical commonsense	160 prompts	27 laws; 4 domains	Physical-law correctness	✘	✘	✘
WorldModelBench [37]	Video world modeling	350 conditions; 67K labels	7 domains; 56 subdomains	Future-state plausibility	✘	△	△
4DWorldBench [42]	3D/4D world generation	Multi-modal tasks	4 evaluation axes	4D alignment and consistency	✘	△	△
WorldMark [69]	Interactive I2V world model	∼500 cases	1st/3rd-person; 3 tiers	Action-conditioned control	✘	✘	△
MSVBench [55]	Multi-shot generation	136 char.; 276 shot refs.	20 metrics; 4 aspects	Multi-shot coherence	△	△	△
Datasets / data-oriented suites							
TALC [2]	Multi-scene T2V data	7,107 videos; 20,177 captions	1–4 scenes	Time-aligned scene captions	△	✘	✘
CI-VID [35]	Interleaved TV2V data	>340K sequences	Clip and transition captions	Coherent clip-to-clip generation	△	✘	△
TransitBench [72]	Transition generation	200 frame pairs	Concept blending; scene transition	First-to-last frame transition	✘	✘	△
VC-Bench [79]	Video connecting	1,579 videos	15 categories; 72 subcategories	Start–end clip connection	✘	✘	△
SEAMLESSWORLD (Ours)	**Text-driven seamless generation**		**Global taxonomy; local scene graph**	**Scene switch with preserve/change probes**	✔	✔	✔

Quantity follows each paper's own reporting unit, e.g., prompts, conditions, videos, reference images; these numbers are therefore not directly comparable.
Text switch indicates whether the benchmark explicitly evaluates natural-language instructions that change one scene/world state into another.
Global/local GT indicates whether the benchmark exposes both global scene attributes and local entity/relation structure as ground truth.
Boundary eval. indicates whether transition behavior is explicitly evaluated. △ includes endpoint-transition, multi-shot, or full-video protocols that only indirectly localize boundary behavior.

A.1 Full Related Works: Video World Models

The recent literature is best understood as a transition from high-fidelity video diffusion to interactive, memory-bearing world simulation. Open video foundation models such as Wan [58], SANA-Video [11], and Lumos [80] improve the backbone side of the problem through scalable diffusion-transformer training, efficient video autoencoding, and data/model scaling, but they are not by

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

themselves sufficient for persistent interactive rollout. The world-model line introduces causal or chunk-wise generation as the organizing principle: MAGI-1 [53] predicts fixed-length video chunks with causal temporal modeling and chunk-wise prompting; BlockVid [85] uses block diffusion, semantic-aware sparse KV cache, and Block Forcing for minute-long videos; and Inferix [31] develops a semi-autoregressive inference engine with interactive streaming and profiling. These systems make long video generation closer to simulator inference, where the model continually updates a latent visual state instead of generating an independent clip.

Interactive world models add an action or control loop to this simulator view. Matrix-Game 2.0 [25] scales real-time streaming interaction with keyboard/mouse control, while Matrix-Game 3.0 [65] adds camera-aware memory retrieval for longer 720p interaction. RELIC [27] emphasizes spatial memory for long-horizon interactive video; Astra [91] introduces action-aware adapters and mixture-of-action experts; and DeepVerse [12] couples autoregressive video with explicit 4D geometry predictions and geometry-aware memory retrieval. Related domains include neural game engines and general-purpose action-controllable virtual worlds [59, 4]. Compared with these action-conditioned systems, our benchmark focuses on text-conditioned scene switching, where the control signal is not a low-level action sequence but a sequence of natural-language scene edits with explicit preservation constraints.

The strongest 2025–2026 progress on real-time text-driven streaming generation comes from forcing- and cache-based causalization. Diffusion Forcing [7] provides a conceptual bridge between next-token prediction and sequence diffusion by assigning independent noise levels to each token. CausVid [78] distills slow bidirectional video diffusion into a fast causal student; Self Forcing [28] trains with self-generated contexts and rolling KV caches; Causal Forcing [90] revisits autoregressive diffusion distillation with an AR teacher; and Context Forcing [13] extends this recipe with long-context teachers and slow-fast memory. Rolling Forcing [41] then reformulates long-video generation as rolling denoising with increasing noise levels, attention sinks, and few-step distillation, making real-time minute-scale diffusion more practical.

Prompt switching and memory control form a more specialized subfield. LongLive [70] supports real-time interactive long video by combining frame-level autoregression, prompt-aware KV re-caching, streaming long tuning, short-window attention, and frame sinks. MemFlow [33] dynamically retrieves relevant historical frames or tokens, reducing the burden of full-history attention and improving scenario switches. ∞-RoPE [74] studies block-relativistic positional encoding, KV flush, and RoPE cut for infinite self-rollout and prompt responsiveness, while LoL [15] addresses hour-scale degradation with multi-head RoPE jitter. Anchor Forcing [71] is particularly aligned with seamless switching because anchor memory, tri-region RoPE, and RoPE realignment are explicitly designed to preserve user-specified anchors during interactive prompt changes. FLEX [38], Deep Forcing [75], Relax Forcing [86], and Knot Forcing [68] complement this line by extending horizons, compressing KV states, stabilizing memory, or adapting real-time rollout to portrait animation. We therefore separate standalone switchable real-time generators from plug-in stability modules: the former determine whether a user can stream and switch prompts, while the latter improve how far and how stably such a system can continue.

A.2 Full Related Works: Evaluation and Datasets for Realtime World Generation

General video-generation evaluation started from disentangled quality and alignment metrics. VBench [30] and VBench++ [29] popularized multi-dimensional evaluation across temporal consistency, subject consistency, motion, aesthetics, imaging quality, and prompt alignment. VBench-2.0 [87] moves this family toward intrinsic faithfulness by adding human fidelity, controllability, creativity, physics, and commonsense dimensions. Video-Bench [24], T2VEval [50], and VideoScore2 [26] further replace single opaque scores with MLLM-based judging, chain-of-query or reasoning traces, and human-aligned multi-dimensional supervision. T2V-Bench [32], T2V-CompBench [56], and TC-Bench [18] focus more directly on temporal dynamics and compositional prompt following. These benchmarks are valuable for calibrating perceptual quality and text-video alignment, but they generally do not encode a ground-truth sequence of segment-level states, boundary deltas, and preservation constraints.

World-level evaluation has expanded rapidly in 2025–2026. WorldModelBench [37] treats video generators as world models and evaluates instruction following and physics adherence; WorldScore [17] defines world generation as next-scene generation under camera and layout conditions; 4DWorld-Bench [42] targets 3D/4D world generation with perceptual quality, condition-4D alignment, physical

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

realism, and 4D consistency; and T2VWorldBench [14] and VideoVerse [64] add broader world-level prompt suites and event-level temporal-causal probing. Physical reasoning benchmarks such as VideoPhy-2 [3], PhyGenBench [45], T2VPhysBench [22], PhyWorldBench [21], and Morpheus [82] curate prompts or real experiments around gravity, collision, conservation laws, human/object motion, and other physical constraints. Domain-specific world-model benchmarks such as DrivingGen [89] and Target-Bench [60] evaluate autonomous-driving or mapless-planning scenarios. WorldMark [69] is the most recent action-conditioned benchmark for interactive image-to-video world models, standardizing WASD-style actions, scenes, trajectories, visual quality, control alignment, and world consistency. SEAMLESSWORLD is orthogonal to these benchmarks: it does not replace physics or action-control tests, but adds a missing text-switching protocol for seamless world generation.

Multi-shot and transition-aware work provides the closest data-level precedent. TALC [2] aligns text descriptions with multiple scenes, while CI-VID [35] collects coherent interleaved text-video data with clip captions and transition captions. VTG/TransitBench [72] and VC-Bench [79] evaluate or train transition generation between visual endpoints, which is useful for video connection but less expressive than open-vocabulary text-conditioned world switching. StoryEval [63], SeqBench [57], ViStoryBench [92], and MSVBench [55] measure whether models can present sequential stories, hierarchical scripts, and multi-shot narratives. Stable Cinemetrics [6], TAV [54], CineTrans [67], ShotAdapter [36], ShotDirector [66], Narrative Weaver [73], AnimeShooter [51], OneStory [1], FilmWeaver [43], TalkCuts [10], MultiShotMaster [61], and STAGE [84] contribute cinematic taxonomies, storyboards, cross-shot memory, speech-shot data, reference-guided consistency, or multi-shot generation pipelines. These works show that the field is moving beyond single clips; however, most of them either provide training data, story-level evaluation, or image-to-video transition tasks rather than a schema-first benchmark where every boundary has derived global/local deltas and explicit preservation probes.

SEAMLESSWORLD is designed to fill this gap. Each sequence is represented by a structured schema whose global fields describe background, camera, look, and story, while its local graph describes entities, attributes, actions, and relations. The benchmark samples multi-boundary delta programs and anchor policies before rendering natural-language prompts, so the ground truth remains available at both semantic and structural levels. During evaluation, visible fields compile into segment-state probes, boundary change probes, preserve probes, smoothness probes, and structure probes, allowing failure analysis by probe family and benchmark cell rather than a single aggregate score. This schema-first design distinguishes SEAMLESSWORLD from prior multi-shot datasets and from action-conditioned world-model benchmarks, and makes it especially suited for evaluating text-conditioned seamless world generation.

B Dataset Audit Details

This appendix provides the full details of the dataset audit summarized in Section 2.3: the comparison corpora, the schema and audit pipeline, the formal definitions of the two schema-information metrics, the rationale behind their design, and the per-dataset breakdown.

B.1 Comparison Datasets and Sample Sizes

We compare SEAMLESSWORLD against four representative multi-shot video-text corpora that together span the dominant annotation paradigms in the field:

- **LONGLIVE** [70] (248,221 transitions, 50,000 sampled): five-second continuation pairs within ongoing scenes.
- **TALC** [2] (13,070 transitions, all used): consecutive-caption pairs from multi-shot videos.
- **CI-VID** [35] (552,784 transitions, 50,000 sampled): consecutive video-clip captions with header-level dense annotations.
- **VISTORY** [92] (1,237 transitions, all used): consecutive-shot pairs with structured per-shot metadata (Setting, Static-Shot, Plot, Perspective).

SEAMLESSWORLD reports its full $259{,}455$ boundaries. After dropping LLM/JSON failures, parse-rates are between 91.6% and 97.3% across datasets (Table 5). A held-out 50/50 dev/val split was used to lock the paired-ingest prompt before any of the headline numbers were computed.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Table 5: Dataset audit summary. N_{ok} is the count of records that survived JSON validation; $\bar{K}$ is the average number of activated classes per record; σ_p is the standard deviation of the per-class marginal fire rates across the 8 classes.

Dataset	N_{ok}	parse	$\bar{K}$	σ_p
LONGLIVE	48,645	97.3%	2.20	26.8
TALC	12,360	94.6%	2.77	27.5
CI-VID	45,797	91.6%	5.93	18.5
VISTORY	1,174	94.9%	5.52	20.5
SEAMLESSWORLD	259,455	N/A	2.97	14.2

B.2 Audit Protocol

Schema. Every transition (boundary between two consecutive scenes) is represented as a delta over an 8-class taxonomy: BG (background / location), ENV (environment: time, weather, atmosphere, lighting), CAM (camera: shot scale, angle, motion, composition), STYLE (style, palette), STORY (story beat), ENTITY (entity add / remove / main-character swap), ATTR (appearance / state / action), and REL (relation add / remove / change). The 8 classes are partitioned into 23 fine-grained ops.

Pipeline. For every dataset D, we materialise its transitions as ordered pairs (BEFORE, AFTER) of single-shot textual descriptions following each corpus's native structure (consecutive-caption pairs in TALC; consecutive-shot pairs concatenating Setting/Static-Shot/Plot/Perspective fields in VISTORY; consecutive video-clip captions for CI-VID; the original five-second continuation pairs in LONGLIVE). Each pair is fed to a single LLM call (Qwen3.6-35B-A3B [52], served through sglang, temperature 0.1, max_tokens $= 6144$) under the locked v5 paired-ingest prompt, which returns two Pydantic-validated `Segment` JSON objects σ_a, σ_b subject to the verbatim-carry-forward rule: any field not positively asserted as different by AFTER must equal σ_a's value character-for-character. We then run the same structural diff routine `derive_boundary`(σ_a, σ_b) used by the synthesis pipeline, and a deterministic lookup `diff_to_ops` produces the per-transition op list. This yields, for every transition r, an 8-dimensional binary class-signature $\mathbf{s}^{(r)} \in \{0,1\}^8$ where $s_i^{(r)} = 1$ iff at least one op of class i fires at r. SEAMLESSWORLD's transitions are read directly from the synthesis pipeline's own `ops` field, which uses the same vocabulary.

B.3 Metric A: Pairwise Hamming Distance

Definition. Let $p_i = \frac{1}{N}\sum_{r=1}^{N} s_i^{(r)}$ be the marginal probability that class i fires at a transition. The expected Hamming distance between two i.i.d.-sampled transitions' class-signatures is

$$\bar{d}_H \;=\; \mathbb{E}\Big[\, d_H(\mathbf{s}^{(a)}, \mathbf{s}^{(b)}) \,\Big] \;=\; \sum_{i=1}^{8} 2\, p_i\,(1 - p_i). \tag{5}$$

$\bar{d}_H \in [0, C/2]$ where $C = 8$, so $\bar{d}_H \in [0, 4.0]$. The maximum 4.0 is attained only when every $p_i = 0.5$.

What it measures. The expected number of schema axes (out of 8) on which two random transitions of the dataset disagree in fire pattern. It captures *across-record heterogeneity of transition types*.

Why this design. The metric is robust against both common failure modes of schema coverage. A high $\bar{d}_H$ (close to 4.0) requires that different boundaries activate different class subsets, so no single fire-pattern dominates and no class is always-on or always-off. A low $\bar{d}_H$ (close to 0) arises in one of two dual ways: either a single class fires most of the time and the rest are silent (a single-axis corpus, e.g. LONGLIVE's ATTR-only continuations), or the same large subset of classes always co-activates (an everything-fires corpus, e.g. CI-VID's dense header annotations). Both pathologies push every p_i toward 0 or 1, which collapses $p_i(1 - p_i)$ toward 0. Crucially, the metric cannot be inflated by simply firing every class at every record.

Reading Figure 2 (a). Each bar shows $\bar{d}_H$ for one dataset (Eq. 5); the dashed reference line is the theoretical maximum 4.0. SEAMLESSWORLD's value of 3.41 (i.e., a random pair of its boundaries differs on 3.4 of the 8 schema axes) is uniquely above 3.0 and approximately $1.5\times$ LONGLIVE's 2.04. The four prior corpora cluster between 2.0 and 2.8, each failing the metric for opposite reasons:

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

continuation corpora because their transitions all activate ATTR alone, dense-annotation corpora because their transitions all activate ATTR, BG, ENV, STORY, ENTITY, and REL together.

B.4 Metric B: Per-Class Focal Contribution

Definition. Let $K^{(r)} = \sum_i s_i^{(r)}$ be the number of activated classes at transition r. We call a transition *focal* if $K^{(r)} \leq 2$, i.e. at most two schema axes change. The focal contribution of class i to dataset D is

$$\mathrm{Focal}_i(D) \;=\; \frac{1}{N}\,\big|\{r : s_i^{(r)} = 1 \;\wedge\; K^{(r)} \leq 2\}\big|. \tag{6}$$

$\mathrm{Focal}_i \in [0, p_i]$, achieving its upper bound p_i only when class i never co-fires with ≥ 2 other classes.

What it measures. The fraction of the dataset's transitions in which class i is genuinely the focus of a small, targeted change rather than passively riding along inside a five-to-eight-class blanket activation. It directly answers "does any meaningful number of transitions in this dataset use class i as their primary axis of variation?"

Why this design. This metric is, by construction, the strict counterpart of cross-record marginals: a dataset that achieves high marginal coverage by always co-activating every class will fail this metric uniformly, because none of those activations is focal. Three regimes are possible. *High* Focal_i *on most or all classes*: every schema axis is exercised as a clean focal change in some non-trivial subset of transitions; the dataset truly covers the schema, not by piling everything together but by isolating each axis at different times. *Concentrated* (high on one class, near zero elsewhere): the corpus is a single-axis show. LONGLIVE puts 62.2% of its records into the ATTR row (action progression) and $\leq 0.7\%$ on BG, STYLE, and STORY; TALC is similar with ATTR at 47.7%. *Uniformly low* (near zero everywhere): the corpus has no focal transitions at all because $K^{(r)} \leq 2$ effectively never happens; the dataset's annotations mention many schema axes simultaneously, so every class is buried in co-firing noise. CI-VID (max 8.3%) and VISTORY (max 8.3%) are in this regime: broad coverage, but no class is ever *the* change.

Reading Figure 2 (b). Cell (i, j) is Focal_i in dataset j (Eq. 6). SEAMLESSWORLD is the only dataset with non-trivial focal contribution ($\geq 4\%$) on six of eight classes (BG 9.3, STYLE 10.3, STORY 23.8, ENTITY 16.3, REL 13.8, ENV 4.6) while no class exceeds 25%. LONGLIVE and TALC are visibly concentrated on the ATTR row alone; CI-VID and VISTORY are mostly empty (dense co-firing offers no focal transitions). STYLE is the most striking column: 0.00% to 0.02% for the four prior corpora, 10.3% for SEAMLESSWORLD.

B.5 Joint Reading

The two metrics are designed as complementary checks, not redundant ones.

Metric A asks *"do different boundaries change different things?"*. A dataset that always changes the same large subset (CI-VID, VISTORY) and a dataset that always changes a single small subset (LONGLIVE, TALC) both fail. Only a dataset whose boundaries span varied class subsets passes.

Metric B asks *"does each individual schema axis ever serve as the primary axis of change?"*. A dataset that never isolates any class as the primary change (everything-fires corpora) fails on every class. A dataset that only isolates one class (continuation corpora) fails on the other seven.

A corpus can score high on Metric A by combining "focused per record" with "varied across records", *or* by combining "everything per record" with "varied marginals"; only the former configuration also scores high on Metric B. SEAMLESSWORLD's schema-driven generation is the only audited corpus that exhibits both properties simultaneously.

C Detailed Formulation of Operation-Conditioned Decoupled DMD

This appendix provides the complete mathematical form of the operation-conditioned decoupled distillation objective used in Sec. 3. The main paper keeps only the essential equations; here we make the re-noising process, score parameterization, fake-score training, CFG expansion, and stop-gradient implementation explicit.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

C.1 Causal Rollout and Re-noising

Let $\mathbf{x}_i$ denote the i-th clip or scene segment. The student predicts a clean latent clip from generator noise $\mathbf{z}_i$, generator sampling input η_i, the previous clean visual cache $\mathbf{K}_{i-1}$, and a text condition c_i:

$$\hat{\mathbf{x}}_{0,i} = G_\theta\left(\mathbf{z}_i, \eta_i; \mathbf{K}_{i-1}, c_i\right). \tag{7}$$

The cache for future clips is updated as

$$\mathbf{K}_i = \mathrm{sg}\left[\mathcal{C}_\theta\left(\hat{\mathbf{x}}_{0,i}; \mathbf{K}_{i-1}\right)\right], \tag{8}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient. Thus, losses on later clips do not backpropagate through the cache into earlier generated clips. During both training and inference, we use the same autoregressive generated-prefix cache, so the student observes the same type of conditioning state in optimization and deployment.

For DMD, the generated clean latent is re-noised at diffusion level τ:

$$\hat{\mathbf{x}}_{\tau,i} = \alpha_\tau \hat{\mathbf{x}}_{0,i} + \sigma_\tau \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{9}$$

Since $\partial \hat{\mathbf{x}}_{\tau,i} / \partial \hat{\mathbf{x}}_{0,i} = \alpha_\tau \mathbf{I}$, the re-noising chain rule contributes a factor α_τ to the generator gradient. In the main paper, this factor is absorbed into $\tilde{w}(\tau) = w(\tau)\alpha_\tau$.

C.2 Score Parameterization

When the real teacher and the fake score model are parameterized as noise predictors, their scores are obtained from predicted noise by

$$s_\psi\left(\mathbf{x}_\tau, \tau \mid c\right) := -\frac{1}{\sigma_\tau} \boldsymbol{\epsilon}_\psi\left(\mathbf{x}_\tau, \tau, c\right), \qquad s_\phi^{\mathrm{fake}}\left(\mathbf{x}_\tau, \tau \mid c\right) := -\frac{1}{\sigma_\tau} \boldsymbol{\epsilon}_\phi\left(\mathbf{x}_\tau, \tau, c\right). \tag{10}$$

The frozen teacher is denoted by ψ, and the trainable fake score model is denoted by ϕ.

The CFG-guided real score is

$$s_{\psi,\gamma}^{\mathrm{real}}\left(\mathbf{x}_\tau, \tau \mid c\right) = s_\psi\left(\mathbf{x}_\tau, \tau \mid \varnothing\right) + \gamma\left[s_\psi\left(\mathbf{x}_\tau, \tau \mid c\right) - s_\psi\left(\mathbf{x}_\tau, \tau \mid \varnothing\right)\right], \tag{11}$$

where γ is the CFG scale and $\varnothing$ denotes the unconditional prompt.

C.3 General DMD Gradient

Distribution Matching Distillation compares the score of a generated sample under a frozen real diffusion model and a fake score model that tracks the current generator distribution. For general real and fake conditions χ_i^r and χ_i^f, the generator gradient is

$$\nabla_\theta \mathcal{L}_{\mathrm{DMD}} \doteq \mathbb{E}_{i, \mathbf{z}_i, \eta_i, \tau, \boldsymbol{\epsilon}}\left[w(\tau)\alpha_\tau \left(s_\phi^{\mathrm{fake}}\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid \chi_i^f\right) - s_{\psi,\gamma}^{\mathrm{real}}\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid \chi_i^r\right)\right)^\top \frac{\partial \hat{\mathbf{x}}_{0,i}}{\partial \theta}\right]. \tag{12}$$

Equivalently, under gradient descent, the sample is moved in the direction

$$s_{\psi,\gamma}^{\mathrm{real}}\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid \chi_i^r\right) - s_\phi^{\mathrm{fake}}\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid \chi_i^f\right). \tag{13}$$

C.4 Decoupled Operation and Scene Conditions

In our setting, each transition provides a previous state S_{i-1}, an operator program π_i, a target state $S_i = \mathcal{T}(S_{i-1}, \pi_i)$, an operation prompt c_i^{op}, and a target-scene prompt c_i^{scene}. The student is conditioned only on the previous visual cache and the operation prompt:

$$\hat{\mathbf{x}}_{0,i} = G_\theta\left(\mathbf{z}_i, \eta_i; \mathbf{K}_{i-1}, c_i^{\mathrm{op}}\right). \tag{14}$$

The real teacher and the fake score model judge the generated clip under the target-scene prompt:

$$\chi_i^r = \chi_i^f = c_i^{\mathrm{scene}}. \tag{15}$$

Thus, the student-side condition and score-side condition are intentionally decoupled. The target-scene prompt is never provided to the student; it is only used to query the teacher and to train the fake score model.

The fake score model can be viewed as estimating the scene-conditioned marginal distribution induced by operation-conditioned student samples:

$$s_\phi^{\mathrm{fake}}\left(\mathbf{x}_\tau, \tau \mid c_i^{\mathrm{scene}}\right) \approx \nabla_{\mathbf{x}_\tau} \log p_\theta\left(\mathbf{x}_\tau \mid c_i^{\mathrm{scene}}\right), \tag{16}$$

where this marginal is formed by pairing samples from $G_\theta(\mathbf{z}_i, \eta_i; \mathbf{K}_{i-1}, c_i^{\mathrm{op}})$ with their corresponding target-scene prompts c_i^{scene}.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

C.5 Fake Score Training

The fake score model is trained on detached student samples. For fake-score training, we use

$$\hat{\mathbf{x}}_{\tau,i}^{\mathrm{sg}} = \alpha_\tau \mathrm{sg}\left[\hat{\mathbf{x}}_{0,i}\right] + \sigma_\tau \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{17}$$

and optimize the standard denoising score matching objective

$$\mathcal{L}_{\mathrm{fake}}(\phi) = \mathbb{E}_{i,\mathbf{z}_i,\eta_i,\tau,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\phi\left(\hat{\mathbf{x}}_{\tau,i}^{\mathrm{sg}}, \tau, c_i^{\mathrm{scene}}\right)\right\|_2^2\right]. \tag{18}$$

Detaching $\hat{\mathbf{x}}_{0,i}$ here ensures that this loss updates only the fake score model ϕ, not the student generator θ.

C.6 CFG Decomposition of the Decoupled DMD Direction

For the operation-conditioned student, we define the DMD sample-space direction

$$\Delta_i(\tau) = s_{\psi,\gamma}^{\mathrm{real}}\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid c_i^{\mathrm{scene}}\right) - s_\phi^{\mathrm{fake}}\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid c_i^{\mathrm{scene}}\right). \tag{19}$$

Substituting the CFG score in Eq. (11) gives

$$\begin{aligned}\Delta_i(\tau) &= s_\psi\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid \varnothing\right) + \gamma\left[s_\psi\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid c_i^{\mathrm{scene}}\right) - s_\psi\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid \varnothing\right)\right] - s_\phi^{\mathrm{fake}}\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid c_i^{\mathrm{scene}}\right) \\ &= \left[s_\psi\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid c_i^{\mathrm{scene}}\right) - s_\phi^{\mathrm{fake}}\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid c_i^{\mathrm{scene}}\right)\right] \\ &\quad + (\gamma - 1)\left[s_\psi\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid c_i^{\mathrm{scene}}\right) - s_\psi\left(\hat{\mathbf{x}}_{\tau,i}, \tau \mid \varnothing\right)\right].\end{aligned} \tag{20}$$

The first bracket is the distribution-matching regularizer, and the second bracket is the CFG augmentation term. In our decoupled setting, the CFG augmentation term points toward the target-scene distribution specified by c_i^{scene}, while the student must realize this transition using only $(\mathbf{K}_{i-1}, c_i^{\mathrm{op}})$.

C.7 Generator Objective and Stop-Gradient Implementation

Using $\Delta_i(\tau)$, the operation-conditioned DMD gradient is

$$\nabla_\theta \mathcal{L}_{\mathrm{opDMD}} \doteq \mathbb{E}_{i,\mathbf{z}_i,\eta_i,\tau,\boldsymbol{\epsilon}}\left[-w(\tau)\alpha_\tau \,\mathrm{sg}\left[\Delta_i(\tau)\right]^\top \frac{\partial G_\theta\left(\mathbf{z}_i, \eta_i; \mathbf{K}_{i-1}, c_i^{\mathrm{op}}\right)}{\partial \theta}\right]. \tag{21}$$

Equivalently, this gradient can be implemented with the scalar surrogate loss

$$\mathcal{L}_{\mathrm{opDMD}}^{\mathrm{sur}} = \mathbb{E}_{i,\mathbf{z}_i,\eta_i,\tau,\boldsymbol{\epsilon}}\left[-w(\tau)\alpha_\tau \,\mathrm{sg}\left[\Delta_i(\tau)\right]^\top G_\theta\left(\mathbf{z}_i, \eta_i; \mathbf{K}_{i-1}, c_i^{\mathrm{op}}\right)\right]. \tag{22}$$

Or, using $\tilde{w}(\tau) = w(\tau)\alpha_\tau$,

$$\mathcal{L}_{\mathrm{opDMD}}^{\mathrm{sur}} = \mathbb{E}_{i,\mathbf{z}_i,\eta_i,\tau,\boldsymbol{\epsilon}}\left[-\tilde{w}(\tau) \,\mathrm{sg}\left[\Delta_i(\tau)\right]^\top \hat{\mathbf{x}}_{0,i}\right]. \tag{23}$$

The stop-gradient on $\Delta_i(\tau)$ is necessary. Without it, automatic differentiation would introduce extra terms through the teacher score, the fake score model, and their score Jacobians. With the stop-gradient, the loss implements the intended DMD update: the sample is moved along the score-difference direction while gradients flow only through the operation-conditioned generator output.

C.8 Dynamics Curriculum Details

We use Dynamics Curriculum Learning to stabilize optimization on interactive transitions. The key difficulty is that high-dynamic transitions often combine several schema edits: camera movement, entity insertion or deletion, attribute change, spatial relation update, interaction change, and layout transformation. When such examples dominate from the start, the DMD signal can become hard to interpret. A failed sample may be wrong because the operation was not followed, because the previous visual state was not preserved, because the camera motion was unstable, or because multiple edits interfered with one another.

To avoid this, every operator program is assigned a dynamics score according to our transition taxonomy. Simple local edits receive low scores. Examples include changing a minor attribute, moving a single entity, or applying a small camera update while preserving the rest of the scene.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Complex transitions receive high scores. These include large camera motions, scene layout changes, insertion of multiple entities, relation updates between objects, and multi-step interaction changes.

At the beginning of training, the sampler favors low-dynamic transitions. This stage teaches the student the basic operation-grounding mechanism: the clean KV cache specifies what should be preserved, the operation prompt specifies what should change, and the scene-conditioned teacher score provides the target distribution. Hard examples are not removed entirely; they are only sampled less frequently. This prevents the model from overfitting to trivial edits while still keeping early optimization stable.

As training progresses, the sampler gradually increases the frequency of high-dynamic examples. The model is then required to compose multiple atomic operations into larger world transitions while maintaining identity, layout consistency, background continuity, and style. In this way, the curriculum moves from local operation grounding to compositional world rollout. The same operation-conditioned mechanism learned on easy examples is reused for increasingly complex transitions, rather than training a separate model or objective for different dynamic ranges.

D SeamlessWorld Details: Definition, Taxonomy, and Pipeline

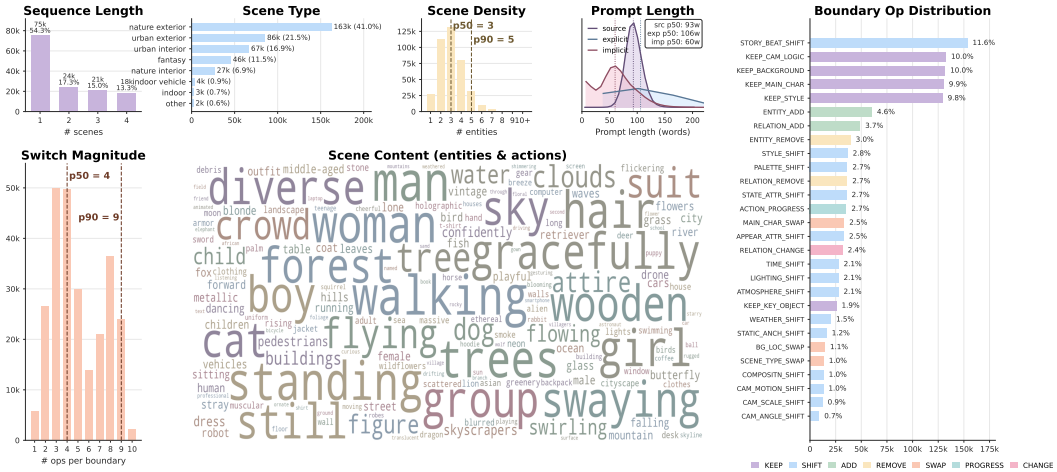


Figure 4: Statistics of SeamlessWorld Dataset.

D.1 Formal Task Definition

A SeamlessWorld example is a sequence

$$x = (\sigma_0, \sigma_1, \ldots, \sigma_K),$$

where each σ_i is a segment schema and each boundary $b_i = (\sigma_i, \sigma_{i+1})$ is associated with an operator program δ_i and a realized preserve set P_i. The desired generated video should satisfy four conditions:

1. **Segment fidelity:** each segment depicts the schema-visible state of σ_i.
2. **Change fidelity:** each boundary executes the schema deltas induced by δ_i.
3. **Anchor preservation:** each anchor in P_i remains visually consistent across the boundary.
4. **Seamlessness:** the transition avoids uncontrolled flicker, identity reset, or abrupt artifacts not implied by the requested event.

We instantiate each segment as a 5-second state. The duration is not part of the task definition; it is a practical unit that makes segment-level and boundary-level VQA evaluation stable.

D.2 Schema

The schema is the only ground-truth artifact. Prompts, transition descriptions, dynamic scores, preserve sets, and VQA probes are derived from it.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Table 6: Static op-pair exclusions used by the redundancy filter. Some pairs are also covered by the one-op-per-family cap; we keep them here because they document the semantic reason for exclusion and protect non-greedy builders.

Pair	Type	Reason
Any two non-KEEP ops from the same family	Class cap	A boundary program activates at most one operator per family, so the dynamic score remains interpretable at the family level.
STYLE_SHIFT + PALETTE_SHIFT	Sub-axis redundancy	Palette is a sub-axis of visual style; evaluating both separately over-specifies one rendering axis.
TIME_SHIFT + LIGHTING_SHIFT	Causal redundancy	Lighting often follows time of day; pairing them makes it unclear which operation caused the visual change.
BG_LOCATION_SWAP + SCENE_TYPE_SWAP	Backbone redundancy	Location swap already jointly samples scene type and static anchors.
CAMERA_ANGLE_SHIFT + COMPOSITION_SHIFT	Frame-geometry redundancy	Both constrain the same camera-layout axis.
MAIN_CHARACTER_SWAP + ENTITY_REMOVE	Graph ambiguity	After swapping the main character, removing another entity can make the post-boundary local graph ambiguous.
RELATION_ADD + ENTITY_REMOVE	Graph invalidity	Entity removal may orphan the newly added relation or make relation endpoints ill-defined.
BG_LOCATION_SWAP + STATIC_ANCHOR_SHIFT	Backbone redundancy	A location swap already refreshes static scene anchors.
ACTION_PROGRESS + STATE_ATTR_SHIFT	Event-state redundancy	Action progression usually implies a new state on the same subject.
WEATHER_SHIFT + ATMOSPHERE_SHIFT	Causal redundancy	Atmosphere often follows weather; pairing them makes the target transition less disentangled.

Sequence. A sequence contains a unique `seq_id`, a `switch_count` K, a list of $K+1$ segments, one boundary policy per transition, and metadata including the visibility mask, canonical signature, and applied delta programs.

Segment. Each segment contains:

- `general`: background, story, camera, and look fields.
- `local_graph`: entities and relations.
- `dynamics`: first state, last state, and progression.
- `_meta`: field-source tags, e.g., explicit, inferred, or not applicable.

Table 7: Global segment fields.

Group	Fields
Background	`location`, `scene_type`, `static_anchors`, `time_of_day`, `weather`, `atmosphere`
Story	`beat`
Camera	`shot_scale`, `angle`, `motion`, `composition`
Look	`style`, `lighting`, `palette`

Local graph. An entity is represented as

$$e = (\texttt{entity_id}, \texttt{role}, \texttt{class}, \texttt{salience}, \texttt{appearance_attrs}, \texttt{state_attrs}, \texttt{action}).$$

Allowed roles are `main`, `support`, `prop`, `bg_anchor`, and `key_object`. Exactly one entity must have role `main`. A relation is a triple $(\texttt{subj}, \texttt{predicate}, \texttt{obj})$. The predicate vocabulary is:

spatial: `left_of`, `right_of`, `near`, `inside`, `on`, `behind`, `in_front_of`;

interaction: `looking_at`, `touching`, `following`, `pushing`, `pulling`;

possession: `holding`, `carrying`, `wearing`.

Dynamics fields. The dynamics block stores:

`first_state`, `last_state`, `progression_nl`.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

These are kept as short natural-language fields rather than fully structured trajectories, because many 5-second micro-events are easier to express linguistically while still being grounded by the schema.

D.3 Boundary Derivation

For consecutive segments σ_i and σ_{i+1}, we derive a boundary descriptor

$$D_i = (\Delta_g, \Delta_\ell, m_g, m_\ell, \textit{scope}, \textit{magnitude}, P_i).$$

Global delta. We compare 14 general paths: `background.location`, `background.scene_type`, `background.static_anchors`, `background.time_of_day`, `background.weather`, `background.atmosphere`, `story.beat`, `camera.shot_scale`, `camera.angle`, `camera.motion`, `camera.composition`, `look.style`, `look.lighting`, `look.palette`. List fields are compared as normalized multisets.

Local delta. The local delta contains entity additions/removals, class changes, appearance/state attribute changes, action changes, relation additions/removals, and a `main_swap` flag.

Mass weights.

Table 8: Mass weights for scope and magnitude derivation.

Global path	Weight	Local key	Weight
`background.location`	3	`entity_added`	1 per item
`background.scene_type`	3	`entity_removed`	1 per item
`background.static_anchors`	1	`attr_changed`	1 per entity
`background.time_of_day`	1	`action_changed`	2 per entity
`background.weather`	1	`relation_added`	1 per item
`background.atmosphere`	1	`relation_removed`	1 per item
`story.beat`	2	`main_swap`	3 boolean
`camera.shot_scale`	1		
`camera.angle`	1		
`camera.motion`	1		
`camera.composition`	1		
`look.style`	2		
`look.lighting`	1		
`look.palette`	1		

A boundary is *micro* if the scene backbone is unchanged and $m_g \leq 3$; otherwise it is *macro*. With ratio threshold $\rho = 1.5$, the boundary scope is: $\begin{cases} \textit{local}, & m_\ell > \rho m_g, \\ \textit{global}, & m_g > \rho m_\ell, \\ \textit{mixed}, & \text{otherwise.} \end{cases}$

Realized preserve set. The realized preserve set is re-derived from the two schemas:

- MAIN_CHARACTER: same main entity id and class.
- KEY_OBJECT: same key-object id and class when a key object exists in both segments.
- STYLE: same normalized `look.style`.
- BACKGROUND: same location and scene type.
- CAMERA_LOGIC: same shot scale and angle.

Downstream rendering and VQA construction use this realized set rather than the originally sampled anchor policy.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Table 9: Operator taxonomy. Operators are typed transformations on the schema σ. Five families act on the global state g; three on the local state ℓ. Full 23-operator inventory in Appendix D.4.

Family	Representative operators	Purpose
	Global Operation	
Background (BG)	location swap, scene-type swap, static-anchor shift	Changes the world backbone while testing whether declared identities remain continuous.
Environment (ENV)	time shift, weather shift, atmosphere shift, lighting shift	Changes global conditions without necessarily changing place or identity.
Camera (CAM)	shot-scale shift, angle shift, motion shift, composition shift	Changes viewpoint and framing under content-preservation constraints.
Style (STYLE)	style shift, palette shift	Changes rendering distribution while preserving semantic content.
Story (STORY)	story-beat shift	Advances the event without forcing a new place or subject.
	Local Operation	
Entity (ENTITY)	entity add, entity remove, main-character swap	Tests object birth/death and identity-level changes.
Attribute / Action (ATTR)	appearance shift, state shift, action progress	Tests fine-grained changes on existing entities.
Relation (REL)	relation add, relation remove, relation change	Tests compositional interaction changes among entities.

D.4 Full Operator Inventory

D.5 Operation Compatibility Rules

Anchor/operator mutual exclusion. The following pairs are hard conflicts:

$(\text{KEEP_MAIN_CHARACTER}, \text{MAIN_CHARACTER_SWAP})$,
$(\text{KEEP_STYLE}, \text{STYLE_SHIFT})$, $(\text{KEEP_STYLE}, \text{PALETTE_SHIFT})$,
$(\text{KEEP_BACKGROUND}, \text{BG_LOCATION_SWAP})$, $(\text{KEEP_BACKGROUND}, \text{SCENE_TYPE_SWAP})$,
$(\text{KEEP_CAMERA_LOGIC}, \text{CAMERA_SCALE_SHIFT})$, $(\text{KEEP_CAMERA_LOGIC}, \text{CAMERA_ANGLE_SHIFT})$.

Redundancy and graph-validity pairs. When two sampled operators match one of the following unordered pairs, the second one is dropped:

$(\text{STYLE_SHIFT}, \text{PALETTE_SHIFT})$,
$(\text{TIME_SHIFT}, \text{LIGHTING_SHIFT})$,
$(\text{BG_LOCATION_SWAP}, \text{SCENE_TYPE_SWAP})$,
$(\text{BG_LOCATION_SWAP}, \text{STATIC_ANCHOR_SHIFT})$,
$(\text{CAMERA_ANGLE_SHIFT}, \text{COMPOSITION_SHIFT})$,
$(\text{ACTION_PROGRESS}, \text{STATE_ATTR_SHIFT})$,
$(\text{WEATHER_SHIFT}, \text{ATMOSPHERE_SHIFT})$,
$(\text{MAIN_CHARACTER_SWAP}, \text{ENTITY_REMOVE})$,
$(\text{RELATION_ADD}, \text{ENTITY_REMOVE})$.

Precondition checks. RELATION_ADD requires at least two entities. RELATION_REMOVE and RELATION_CHANGE require at least one existing relation. ENTITY_REMOVE requires at least one non-main entity, excluding preserved key objects. WEATHER_SHIFT is skipped for indoor scenes. Micro boundaries cannot contain BG_LOCATION_SWAP or SCENE_TYPE_SWAP.

D.6 Sampling and Balancing

Target grid. The benchmark grid is

$$(\textit{scope}, \textit{magnitude}, \textit{dynamic_bin}, \textit{switch_count_category}),$$

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Table 10: Full operator inventory. S denotes structural mutation; M denotes semantic-first mutation with refreshed segment dynamics.

Family	Operator	Type	Mutated field(s)
BG	BG_LOCATION_SWAP	S	`location`, `scene_type`, `static_anchors`
	SCENE_TYPE_SWAP	S	`scene_type`, `location`, `static_anchors`
	STATIC_ANCHOR_SHIFT	M	add/remove one `static_anchors` item
ENV	TIME_SHIFT	M	`background.time_of_day`
	WEATHER_SHIFT	S	`background.weather`; skipped for indoor scenes
	ATMOSPHERE_SHIFT	M	`background.atmosphere`
	LIGHTING_SHIFT	M	`look.lighting`
CAM	CAMERA_SCALE_SHIFT	S	`camera.shot_scale`
	CAMERA_ANGLE_SHIFT	S	`camera.angle`
	CAMERA_MOTION_SHIFT	S	`camera.motion`
	COMPOSITION_SHIFT	S	`camera.composition`
STYLE	STYLE_SHIFT	S	`look.style`
	PALETTE_SHIFT	S	`look.palette`
STORY	STORY_BEAT_SHIFT	M	`story.beat`
ENTITY	ENTITY_ADD	M	append one support or prop entity
	ENTITY_REMOVE	M	remove one non-main entity
	MAIN_CHARACTER_SWAP	M	replace main entity and remove invalid relations
ATTR	APPEARANCE_ATTR_SHIFT	S	`appearance_attrs` of main entity
	STATE_ATTR_SHIFT	M	`state_attrs` of main entity
	ACTION_PROGRESS	M	`action` of main entity
REL	RELATION_ADD	S	add one relation triple
	RELATION_REMOVE	S	remove one relation triple
	RELATION_CHANGE	M	change one relation predicate or object

where:

$$scope \in \{local, global, mixed\}, \quad magnitude \in \{micro, macro\},$$
$$dynamic_bin \in \{low, mid, high, very\text{-}high\}, \quad switch_count_category \in \{single, multi\}.$$

The *single* category contains one transition; *multi* samples two to four transitions.

Excluded cells. Very-high local cells are excluded because local-only operators cannot reliably reach the very-high score range under the family weights:

$$(local, *, very\text{-}high, *).$$

Coverage tracker. A coverage tracker maintains a target quota per valid cell. At each synthesis step, we sample a deficit cell in proportion to its remaining gap, then attempt sequence construction, rendering, filtering, validation, and deduplication. Surface mode is not a grid axis because every canonical sequence is rendered into both explicit and implicit surfaces.

Default hyperparameters.

D.7 Rendering

Visibility mask. A visibility mask is sampled once per sequence and shared by explicit and implicit prompt surfaces. Core paths are always visible:

`location`, `scene_type`, `shot_scale`, `angle`, `style`, `main.class`, `main.action`.

Optional global fields, main attributes, non-main entities, and relations are independently dropped according to Table 11. This creates prompts that are realistic rather than exhaustively schema-like, while preserving exact knowledge of which fields are evaluable.

Explicit segment prompt. The explicit prompt is rendered from ordered clauses: camera → location → time/weather → atmosphere → look → main entity → other entities → relations → story → dynamics.

A lightweight LLM rephrase makes the prompt fluent while being instructed to copy schema-visible literal values exactly.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Table 11: Default synthesis and filtering hyperparameters.

Parameter	Default
Target score distribution	$\mathrm{TruncNormal}(5, 2, [1, 12])$
Anchor sampling	$\mathrm{Bernoulli}(0.5)$ per anchor type
Boundary retry budget	12
Optional-field visibility drop	0.3
Relation visibility drop	0.3
Non-main entity visibility drop	0.4
Direct implicit prompt probability	0.3
Explicit roundtrip threshold	0.75
Explicit roundtrip retries	1
Lexical deduplication threshold	0.75 word-3-gram Jaccard
Negative VQA probe rate	1.0

Implicit transition prompt. The implicit prompt is rendered from the realized boundary descriptor. It contains:

- keep clauses derived from P_i;
- concrete global changes, e.g., "time becomes night";
- concrete local changes, e.g., entity add/remove, attribute change, action change, relation add/remove;
- a short transition tail connecting the previous last state and the next first state.

A direct template-only mode is used with probability 0.3 to mimic terse user edit prompts.

D.8 Roundtrip Filtering

Explicit roundtrip. An LLM parses each explicit prompt back into schema-aligned fields. We compare only visible fields. String matching proceeds through loose normalization, substring/stem overlap, token Jaccard, embedding similarity, and an optional LLM yes/no fallback for borderline cases. List-valued fields use overlap against the smaller list. A record passes when every segment has visible-field pass rate at least 0.75. One re-render retry is allowed.

Implicit roundtrip. A lightweight parser extracts `keeps` and `changes` from each implicit prompt. Missing keep claims relative to the realized preserve set are hard failures. Extra keep claims are warnings. Empty change content is a failure.

Deduplication. The canonical signature concatenates normalized location, scene type, main class, main action, look style, and boundary preserve sets, then hashes the string. We additionally reject lexical near-duplicates using word-3-gram Jaccard similarity on the first explicit segment prompt.

D.9 VQA Probe Compilation

For each sequence, probes are compiled from the schema and visibility mask. Each probe stores `probe_id`, `type`, `question`, `expected`, and indexing metadata.

Probe families.

1. **Segment state:** questions for visible fields and counterfactual questions sampled from closed vocabularies or cross-record field pools.
2. **Preserve:** questions for realized preserved anchors, including whether changed fields stayed the same.
3. **Change:** questions for global and local deltas, including whether preserved anchors changed.
4. **Smoothness:** yes questions about flicker, reset, and transition artifacts.
5. **Structure:** integer question asking for the number of scene transitions.

Aggregation. Positive and negative probes are aggregated into parent families. We compute yes/no confusion tables and report F_1 for segment state, preserve, change, and smoothness. The structure family is reported as integer accuracy. We report metrics overall and by benchmark cell.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

E Detailed Experiment Setups

- **Anchor-Forcing** [71]. A streaming, causal autoregressive long-video model that fine-tunes a WAN2.1-T2V-1.3B backbone with a LoRA adapter and a denoising-distillation schedule, and adds a *junction KV cache* plus *global attention sink* so that prompt switches are conditioned on *anchor* tokens that persist across blocks. We use upstream's `interactive_inference` entry, which materialises switch points through a YAML field `switch_frame_indices`. *Hyperparameters:* backbone WAN2.1-T2V-1.3B; distilled denoising step list $[1000, 750, 500, 250]$ with `warp_denoising_step`; `num_frame_per_block`$=3$, `local_attn_size`$=12$, `timestep_shift`$=5.0$, `sink_size`$=3$, `global_sink`, `rope_enhance`, `use_prompt_cache` (junction KV cache); LoRA rank $=$ alpha $=256$, dropout 0.0, `bfloat16`; seed 6979, 1 sample per case; `num_output_frames` chosen per case as the smallest multiple of `num_frame_per_block` that is $\geq n_{\text{scenes}} \times 20 + 1$ (Wan VAE temporally decompresses $4\times$); switch indices spaced evenly across the latent axis.
- **LongLive** [70]. A T2V variant of the same Wan-1.3B + LoRA recipe used by Anchor-Forcing but trained without a first-frame anchor and without the junction prompt cache, and intended for arbitrarily long generation with mid-stream prompt updates. We use upstream's `interactive_inference` loop (`MultiTextDataset + switch_frame_indices`). *Hyperparameters:* backbone WAN2.1-T2V-1.3B; denoising step list $[1000, 750, 500, 250]$ with `warp_denoising_step`; `num_frame_per_block`$=3$, `local_attn_size`$=12$, `timestep_shift`$=5.0$, `sink_size`$=3$, `global_sink`; LoRA rank $=$ alpha $=256$, dropout 0.0, `bfloat16`; seed 1; `num_output_frames` derived per case as in Anchor-Forcing, switch indices spaced evenly.
- **Rolling-Forcing** [41]. A streaming long-video diffusion model trained with a Distribution Matching Distillation (DMD) loss and a *rolling* KV-cache schedule that recurrently slides the conditioning window forward. It is built on the larger WAN2.1-T2V-14B backbone and produces a single continuous shot rather than reconditioning per segment, so for our benchmark the adapter composes the per-scene prompts into one ordered "Scene 1: ...; Scene 2: ..." header and issues one upstream call per case (this caveat is documented in the released codebase, which encodes a single prompt batch before generation). *Hyperparameters:* backbone WAN2.1-T2V-14B; distilled checkpoint `rolling_forcing_dmd.pt` (`-use-ema`); denoising step list $[1000, 800, 600, 400, 200]$ with `warp_denoising_step`; `num_frame_per_block`$=3$, `timestep_shift`$=5.0$, guidance scale 3.0, mixed precision, hybrid-FSDP; `image_or_video_shape`$=[1, 21, 16, 60, 104]$ (native 832×480 at 5 s/segment latent); per-case latent T derived from n_{scenes}.
- **MAGI-1** [53]. An autoregressive *chunk-wise* diffusion video model (`videodit_ardf`) that generates fixed-width chunks left-to-right, using a multi-stage classifier-free guidance schedule that decreases text guidance and increases prior-chunk guidance over the rollout. The upstream entry script accepts a single `-prompt`, so the adapter collapses each case's prompts into one ordered header ("`compose`" strategy; one MAGI invocation per case); `v2v-chain` is also supported but was not used in our reported runs. *Hyperparameters:* 4.5B base configuration (`example/4.5B/4.5B_base_config.json`); 34 transformer layers, hidden 3072, FFN 12288, 24 heads; `num_steps`$=64$, `chunk_width`$=6$, `window_size`$=4$; guidance: `cfg_number`$=3$, `cfg_t_range`$=[0, 0.0217, 0.1, 0.3, 0.999]$, `prev_chunk_scales`$=[1.5, 1.5, 1.5, 1.0, 1.0]$, `text_scales`$=[7.5, 7.5, 7.5, 0, 0]$; `noise2clean_kvrange`$=[5, 4, 3, 2]$, `clean_chunk_kvrange`$=1$; native output 96 frames at 720×720, 24 fps, `bfloat16`; seed 1234. Outputs are then padded/letterboxed/re-fps'd to the shared spec by the post-processing step.
- **DiTCtrl** [5]. A *training-free* multi-prompt long-video method that adapts a fixed text-to-video DiT backbone (here COGVIDEOX-2B) by inserting a KV-sharing AdaLN mixin between adjacent prompt segments and performing latent blending across overlapping noisy latents. *Hyperparameters:* backbone COGVIDEOX-2B (sat checkpoint), `fp16`; T5-XXL text encoder; ZeroSNR DDPM discretisation, shift_scale 3.0; multi-prompt sampler `VPSDEDPMPP2MSampler_MultiPrompt` with `num_steps`$=50$ and DynamicCFG (scale 6, exp 5); sampling resolution 720×480 native, 13 latent frames at 16 fps; KV-sharing AdaLN mixin active over steps $[2, 25]$ and layers $[25, 30]$; latent blending with `overlap_size`$=9$, `num_transition_blocks`$=2$, `longer_mid_segment`$=0$ (total segments $= N + 2(N-1)$ for N prompts); seed 42. Outputs are letterboxed and length-enforced to the shared 832×480 / $n_{\text{scenes}}\times80$f spec. We note that for $n_{\text{scenes}}=3$ cases the native length (241f) overshoots the target by a single frame and the upstream output arrives at 720×480; the post-processing pass corrects both.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

- **StoryMem** [83]. A multi-shot story-video pipeline built on WAN2.2: shot 1 is generated with a T2V model, and shots $2{:}N$ are produced by an M2V (memory-to-video) loop where a LoRA-adapted I2V model conditions on a bounded memory of keyframes from previous shots. The adapter packs each benchmark case into a single scene with per-shot `cut` flags, runs `pipeline.py` via `torchrun`, and renames upstream's concatenated output to `{seq_id}.mp4`. *Hyperparameters:* T2V backbone WAN2.2-T2V-A14B, I2V backbone WAN2.2-I2V-A14B, StoryMem MI2V LoRA (`Kevin-thu/StoryMem`, `Wan2.2-MI2V-A14B`); size 832×480; `max_memory_size`$=10$; LoRA rank 128; `sample_solver`=`unipc`; `sample_guide_scale`$=3.5$; mode `mi2v` (start each shot from the last frame of the previous shot under the memory LoRA); 8 GPUs; `frame_num` $=4n+1$, `fix`$=3$ keyframes (StoryMem default).
- **VideoGen-of-Thought (VGoT)** [88]. A shot-by-shot story-video pipeline that authors a screenplay, synthesises a per-character avatar, and renders each shot independently with DYNAMICRAFTER (image-to-video) using the avatar as an identity anchor that is reused across shots. We bypass the GPT-4 script stage with `-use_exist_prompt`: the adapter materialises `short_shot_description.txt`, a one-entry `avatar_prompt.json`, and a per-shot `image_prompt_pairs.json` (each shot points at the same avatar image so cross-shot identity propagation has a stable reference); each prompt is wrapped in VGoT's five-dimension format (*Character / Background / Relation / Camera Pose / HDR Description*). *Hyperparameters:* backbone DYNAMICRAFTER (`weights/DynamiCrafter/model.ckpt`), config `configs/inference_module3.yaml`; height 576, width 1024, video_length 40 frames per shot (native 8 fps; the post-pass fps-norms to 16 fps for 80f/shot); `ddim_steps`$=50$, `ddim_eta`$=1.0$; `unconditional_guidance_scale`$=7.5$; `frame_stride`$=10$; seed 3407, 1 sample per shot; identity propagation through Kolors text encoder + Kolors-IP-Adapter-Plus. *Caveat (audited):* VGoT's per-shot native window is $\sim$2 s of motion, so for our 5 s/scene spec the post-pass clone-pads ~ 75–80% of every output (`tail/head` motion ratio ≈ 0.06 in our audit). Numbers from VGoT should therefore be read as a stress-test rather than a fair operating-point comparison.
- **Stable-Video-Infinity (SVI 2.0)** [39]. A streaming, image-to-video multi-clip pipeline that fine-tunes a WAN2.2-I2V-A14B backbone with high- and low-noise LoRAs and generates a sequence of 81-frame clips, where each clip is conditioned on the previous clip's final frame plus an optional anchor frame and the next prompt. The adapter writes a Python-literal prompt list to `prompt.txt` and a per-case starting frame (`frame.jpg`; falls back to a shared starter when the case has no `first_frame`); the upstream script emits per-clip `mp4`s plus a `*_streaming_final.mp4` which we rename to `{seq_id}.mp4`. *Hyperparameters:* backbone WAN2.2-I2V-A14B; LoRAs `SVI_Wan2.2-I2V-A14B_high_noise_lora_v2.0` and `..._low_noise_lora_v2.0`, $\alpha = 1$; `height`$=480$, `width`$=832$; `frames_per_clip`$=81$, native `fps`$=15$; `num_clips` $= n_{\text{scenes}}$ (one clip per prompt); seed `clip_idx`$\times 42$ per clip; `use_anchor`=`True` (the case's starting frame is reused as a global anchor for every clip); `torch.bfloat16`; T5 `umt5-xxl-enc-bf16`; Wan2.1 VAE.
- **Stable-Video-Infinity Pro (SVI 2.0 Pro)** [39]. The Pro variant of SVI 2.0 with the same backbone and infrastructure but a different pair of LoRAs (`..._high/low_noise_lora_v2.0_pro`) that the authors release for higher-quality streaming. Adapter, prompt-list contract, anchor frame, resolution, clip length, and seed schedule are identical to SVI 2.0; only the LoRA pair changes. *Audit caveat:* our *explicit*-surface SVI Pro outputs landed at 15 fps / 158 frames before normalisation; the post-processing re-encode to 16 fps is what produces the spec-compliant `mp4`s used in our metrics.
- **YUME 1.5** [44]. A long-form text-conditioned video model whose "prompt switching" API is implemented as a multi-line caption file: each line of the caption is the text condition for a $\sim$ 2 s segment, while `-prompt` is a global context appended to every line (see `fastvideo/sample/sample_5b.py:824-831`). The adapter writes one caption file per case (one line per benchmark prompt, wrapped in YUME's "Actual distance moved:" sentinel that upstream uses to derive the output filename) and a synthesised global summary; the largest `mp4` emitted under `_yume_runs/{seq_id}/` is renamed to `{seq_id}.mp4`. *Hyperparameters:* backbone YUME-5B-720P (`stdstu123/Yume-5B-720P`), script `fastvideo/sample/sample_5b.py`, mode `t2v`; `num_euler_timesteps`$=4$ (distilled); `cfg`$=0.1$, `distill_cfg`$=3.0$; native height 704, width 1280, `num_frames`$=163$; 1 GPU per case; flash-attn 2.7.0.post2 backend. Outputs are letterboxed/length-enforced to the shared 832×480 / $n_{\text{scenes}}\times 80$f spec.
- **Veo 3.1 Extended (reference)**. A closed-source reference baseline using Google's Veo 3.1 native video model with the `extend` operation. For a case with N prompts we issue N Veo calls — one base call (prompt 0, 8 s) and $N-1$ extend calls (one per remaining prompt, $+7$ s each, conditioning

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

on the final second of the prior clip). From the final cumulative video ($8+7(N-1)$ s) we trim per-prompt 5 s windows (prompt 0: $[0, 5]$; prompt $i \geq 1$: $[8+7(i-1),\ 8+7(i-1)+5]$) and concatenate them into the `{seq_id}.mp4`, yielding $N \times 5$ s of output that matches the seamless-bench eval contract. *Hyperparameters:* model `veo-3.1-generate-preview`; `per_clip_duration`$= 5.0$ s (≤ 7 s, the Veo extend hop); `aspect_ratio`$=$ 16:9, `resolution`$=$ 720p; sync `google-genai` client (the async path uses `aiohttp`, which does not honour `ALL_PROXY` on our cluster); cross-case parallelism via a `ThreadPoolExecutor` with `concurrency`$= 4$; poll interval 10 s, `max_wait_per_clip`$= 1800$ s. Because the extend operation conditions on the final second of the prior clip, the first $\sim$2–3 s of each 5 s post-switch window is a *gradual* blend rather than a hard cut — this is a property of Veo, not of the adapter, and is reflected in our metrics. Due to API cost (Veo 3.1 preview pricing $\sim\$0.40$–$0.75$ per call, $\sim\$2$–$\$4$ per 5-prompt case), we report this baseline only on a 10-case smoke subset of the full 100-case benchmark.

F Prompts

Prompt for Schema Parser

System:
You are a scene-schema parser for a video-generation benchmark. Read a natural-language single-shot video prompt and output strict JSON only. Do not include markdown or commentary.

User:
Convert the video prompt into the schema below.

Rules:

- Use short phrases for atomic fields such as location, style, action, appearance attributes, and state attributes.
- Use `null` for fields that are not stated or visually implied.
- Exactly one entity must have role `"main"`.
- Relation predicates must come from the allowed vocabulary: `left_of, right_of, near, inside, on, behind, in_front_of, looking_at, touching, following, pushing, pulling, holding, carrying, wearing`.
- Closed vocabularies: `shot_scale` $\in$ {extreme_close, close, medium, wide, extreme_wide}; `angle` $\in$ {low, eye_level, high, overhead}; `salience` $\in$ {high, medium, low}.
- Dynamics should describe a natural 5-second progression. If the source prompt is static, infer a small plausible progression.

Return JSON with: `general.background`, `general.story`, `general.camera`, `general.look`, `local_graph.entities`, `local_graph.relations`, `dynamics`.

Input prompt:
`<RAW_PROMPT>`

Figure 5: Prompt for schema parser.

G Case Study

Probe-set anatomy for benchmark case 1 (`b3_bench_040325`; cf. Fig. 13)

seq_id=`b3_bench_040325` **cell_id**=(general,micro,high,single) **switch_count**=1 **#probes**=59
Type breakdown: `segment_state`=28, `segment_state_negative`=25, `preserve`=1, `transition_differ`=1, `change`=1, `transition_same`=1, `smoothness`=1, `structure`=1.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Prompt for Semantic-First Action Progress

System:
You generate natural 5-second scene continuations for a structured video dataset. Output strict JSON only.

User:
Current scene:
<COMPACT_SCHEMA_SUMMARY>

Task:
The main subject is currently "<OLD_ACTION>". Generate a plausible next action 5 seconds later in the same scene. The new action must be different but continuous with the current event.

Return JSON:

```
{
  "brief_rationale": "one short sentence explaining why the
   new action is plausible",
  "main_action": "2-5 word present-continuous phrase",
  "dynamics": {
    "first_state": "<=15 words",
    "last_state": "<=15 words",
    "progression_nl": "<=25 words"
  }
}
```

Figure 6: Prompt for semantic-first action progress.

Prompt for Explicit Rephrase

System:
You rewrite a structured video prompt into fluent cinematic prose. Preserve all literal field values exactly. Do not substitute synonyms for location, main subject class, main action, style, shot scale, or camera angle. Write 3 to 6 short sentences. Output text only.

User:
Rewrite the following draft while preserving every concrete value:

<DRAFT_PROMPT>

Figure 7: Prompt for explicit rephrase.

Prompt for Implicit Transition Rephrase

System:
You rewrite a transition edit instruction for a video generator. Keep every listed change and every concrete new value. Do not describe the previous shot from scratch. Do not invent extra changes. Output one concise header.

User:
Rewrite this transition instruction:

<DRAFT_TRANSITION>

Figure 8: Prompt for implicit transition rephrase.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Prompt for Explicit Roundtrip Parse

System:
You extract schema fields from a rendered video prompt. Output strict JSON only. Extract only what is explicitly present in the prompt.

User:
Parse the prompt into: `location`, `scene_type`, `static_anchors`, `time_of_day`, `weather`, `atmosphere`, `story_beat`, `shot_scale`, `angle`, `motion`, `composition`, `style`, `lighting`, `palette`, `main_class`, `main_action`, `main_appearance_attrs`, `main_state_attrs`.

Prompt:
`<EXPLICIT_PROMPT>`

Figure 9: Prompt for explicit roundtrip parse.

Output Record Format

```
{
  "seq_id": "<unique id>",
  "switch_count": 1,
  "cell_id": "(scope,magnitude,dynamic_bin,switch_count_category)",
  "scope": "local|global|mixed",
  "magnitude": "micro|macro",
  "dynamic_bin": "low|mid|high|very_high",
  "switch_count_category": "single|multi",
  "scenes": [
    {
      "scene_idx": 0,
      "seg_id": "s0",
      "prompts": {"explicit": "..."}
    },
    {
      "scene_idx": 1,
      "seg_id": "s1",
      "prompts": {"explicit": "...", "implicit": "..."}
    }
  ],
  "visibility_mask": {"s0.general.background.location": true},
  "canonical_signature": "<sha1>",
  "source_prompt": "<original seed prompt>",
  "sequence": "<full schema dump>",
  "delta_programs": [["ACTION_PROGRESS", "RELATION_ADD"]],
  "probes": [
    {
      "probe_id": "...",
      "type": "segment_state",
      "segment_idx": 0,
      "field": "general.background.location",
      "question": "In segment 0, is the scene set in ...?",
      "expected": "yes"
    }
  ]
}
```

Figure 10: Output record format.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Held-out high-dynamic case 1: local-scope multi-entity edit (`ENTITY_ADD+RELATION_ADD+STATE_ATTR_SHIFT`)

scope=local **magnitude**=macro **dynamic_bin**=high (score 7)
ops: `RELATION_ADD`, `ENTITY_ADD`, `STATE_ATTR_SHIFT`
preserved: `style`, `background`

Source prompt (VidProM)
A young girl and a young boy, both around 6 years old, stand hand-in-hand at the edge of a dense, lush jungle. They are dressed in casual play clothes suitable for an adventurous outing. The girl has long brown hair tied in a ponytail and wears a floral sundress, while the boy has short, curly hair and is wearing shorts and a t-shirt. Both children look curiously towards the jungle, their faces expressing a mix of excitement and wonder. Behind them, the jungle is a vibrant sea of green, with tall trees and thick undergrowth creating a mysterious backdrop. The scene is captured from a medium shot, focusing on the children and the jungle edge.

Scene 1 – explicit prompt (`from_explicit`)
A centered, eye-level medium shot with static camera motion frames the edge of a dense lush jungle during the day under clear and sunny skies. The scene is set in nature exterior with tall trees, thick undergrowth, and lush greenery, rendered in cinematic realistic style with natural daylight and a vibrant green palette. The atmosphere feels serene as a girl with brown ponytail in floral standing still, wearing brown ponytail and floral sundress, appears curious and excited. A boy with curly hair in shorts stands still nearby, and together two children gaze curiously into the jungle. The shot opens as children hold hands, staring intently into the green foliage, they remain rooted but shift their weight forward in awe of the jungle, and closes as children slowly lean forward slightly, eyes wide with wonder.

Scene 2 – explicit prompt (`to_explicit`)
A medium shot at eye level with a static centered camera frames the edge of a dense lush jungle during the day. The scene is rendered in cinematic realistic style with natural daylight and a vibrant green palette. A girl with brown ponytail in floral standing still, wearing brown ponytail and floral sundress, is accompanied by a boy with curly hair in shorts. A bright blue morpho butterfly fluttering gently near the children's field of view draws their attention. Two children gaze curiously into the jungle, holding hands and eyes fixed on foliage, then tilt their heads upward to follow the butterfly's ascent into the sky.

Scene 2 – implicit (continuation) prompt (`to_implicit`)
Continue from the previous shot, but keep the same visual style and the same setting and backdrop. Change: a bright blue morpho butterfly enters the scene; the main subject's state becomes awestruck, entranced; the boy with curly hair in shorts is now near the main subject. Transition: the previous shot ends with children slowly lean forward slightly, eyes wide with wonder, and the new shot opens with Children hold hands, eyes fixed on the dense jungle foliage.

Figure 11: Held-out high-dynamic case 1: local-scope multi-entity edit (`ENTITY_ADD+RELATION_ADD+STATE_ATTR_SHIFT`).

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Held-out high-dynamic case 2: mixed-scope scene-type swap (`SCENE_TYPE_SWAP`+5 co-ops)

scope=mixed **magnitude**=macro **dynamic_bin**=very_high (score 17)
ops: `SCENE_TYPE_SWAP`, `STORY_BEAT_SHIFT`, `LIGHTING_SHIFT`, `ENTITY_ADD`, `RELATION_ADD`, `STATE_ATTR_SHIFT`
preserved: `style`, `camera_logic`

Source prompt (VidProM)
A close-up medium shot of a smiling woman in a red swimsuit, standing confidently on a sandy beach. She has a warm, inviting smile and her hair cascades down her shoulders. The sun sparkles off the ocean behind her, creating a bright, sunny day atmosphere. The water gently laps at the shore, and palm trees sway in the background. The scene captures her youthful energy and allure in a natural, lively setting.

Scene 1 – explicit prompt (`from_explicit`)
A centered medium shot at eye level captures a jennifer love hewitt in red bikini standing confidently on a sandy beach with ocean. It is day and sunny with a lively atmosphere, set against palm trees swaying in the wind. The scene is rendered in a cinematic realistic style with a vibrant warm palette. The camera remains static as the jennifer love hewitt in red bikini smiles at the camera while maintaining her pose. The shot closes as wind shifts her hair slightly while the smile remains.

Scene 2 – explicit prompt (`to_explicit`)
A static medium shot at eye level captures a lively scene at a cave entrance by the sea. The setting is a sandy beach with ocean lapping at the shore and palm trees swaying under a sunny day. Bright sunlight slightly diffused by passing clouds illuminates the vibrant warm palette in a cinematic realistic style. Jennifer Love Hewitt in red bikini stands confidently, with red bikini and long hair, smiling directly at the camera. A terrier trots into frame, causing Jennifer to look down and smile wider at the interruption. She laughs warmly, looking down at the playful dog.

Scene 2 – implicit (continuation) prompt (`to_implicit`)
change the location to cave entrance by sea; change the lighting to bright sunlight slightly diffused by passing clouds; change the scene type to nature interior; the palm trees is now near the main subject; the main subject is now playful, charmed.

Figure 12: Held-out high-dynamic case 2: mixed-scope scene-type swap (`SCENE_TYPE_SWAP`+5 co-ops).

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Curated 100-record benchmark case 1: cell (general,micro,high,single), single switch

cell_id=(general,micro,high,single) **switch_count**=1 **surface_mode**=explicit **#probes**=59

Source prompt (VidProM)
A quick sketch in bold monochrome imagery showcasing a flag with three horizontal stripes waving dramatically against a backdrop of a bustling grandstand filled with excited spectators. The flag appears to be fluttering in the wind, creating a sense of motion and excitement. The grandstand is packed with people standing and cheering, their faces partially visible, adding to the vibrant atmosphere. The scene is rendered in stark black and white, emphasizing the dynamic movement and energy of the moment. Medium shot focusing on the flag and the crowd behind it.

Scene 1 – explicit
A medium shot from an eye level angle, with static camera motion, composed off center. The setting is a bustling grandstand during windy overcast weather, creating a tense atmosphere. Rendered in a bold monochrome sketch style, the scene features high contrast black and white tones. A three striped flag with three horizontal stripes and waving fabric, appearing fluttering, is the main subject, accompanied by excited spectators cheering loudly and a grandstand structure. The flag waves dramatically in front of the cheering crowd, opening as the flag holds steady while the crowd begins to cheer, continuing as wind gusts catch the flag causing it to snap while the crowd energy builds, and closing as the flag flaps vigorously and the crowd jumps in excitement.

Scene 2 – explicit
A static eye-level medium shot captures the bustling grandstand with its tiered seating and spectators during the day under windy overcast skies. Rendered in a bold monochrome sketch style with a monochrome palette, the image focuses on a three striped flag waving dramatically with three horizontal stripes and waving fabric alongside excited spectators cheering loudly. The shot opens as wind catches the flag while the crowd murmurs with anticipation, then shows a strong gust snapping the flag to trigger an immediate roar from the spectators. The sequence closes as the flag whips hard and the crowd erupts loudly.

Scene 2 – implicit (continuation)
Continue from the previous shot, but keep the same setting and backdrop, the same camera framing, the same main subject, and the same visual style. The transition begins with the flag flapping vigorously as the crowd jumps in excitement. Then the wind catches the flag and the crowd murmurs in anticipation.

Sample probes (3 / 59)

segment_state: In segment 1, is the scene set in bustling grandstand? expected: `yes`

preserve: Is the camera framing consistent across segments 1 and 2? expected: `yes`
change: Between segments 1 and 2, does the story beat progress from 'flag waves dramatically in front of cheering crowd' to 'flag flutters wildly as crowd roars in approval'? expected: `yes`

Figure 13: Curated 100-record benchmark case 1: cell (general,micro,high,single), single switch.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Curated 100-record benchmark case 2: cell `(mixed,macro,very_high,single)`

cell_id=(mixed,macro,very_high,single) **switch_count**=1 **surface_mode**=`explicit` **#probes**=72

Source prompt (VidProM)
A cute, little pink cat illustration with expressive eyes and a friendly smile, saying "hello." The cat has small, round ears and a fluffy pink tail. It is standing on its hind legs, paws clasped together as if waving.

The background is a soft pastel color with subtle, whimsical patterns. The cat's face shows warmth and happiness. Close-up view, static shot.

Scene 1 – explicit
A close shot at eye level with static camera motion is composed centered against a soft pastel background with whimsical patterns. The scene takes place during bright daylight with a tender atmosphere. It is rendered in a cute illustration style with soft even lighting and a pastel pink palette. A little pink cat

stands on hind legs, appearing happy and warm. The shot opens with the cat standing still with paws clasped, the cat lifts clasped paws gently to wave hello, and closes as the cat raises paws in a waving gesture.

Scene 2 – explicit
A close, eye-level, centered shot of a little pink cat with round ears, a fluffy tail, and expressive eyes. The scene is set in a soft pastel background with whimsical patterns during late afternoon, creating a

tender atmosphere. Rendered in a cute illustration style with a pastel pink palette, the cat sits gently, appearing happy and warm. A small pastel butterfly hovers near the cat's raised paw as it smiles broadly and waves. The shot opens with the pink cat standing on its hind legs, waving near the butterfly, then smoothly lowers into a seated position, maintaining a tender gaze as shadows lengthen. It closes with the cat sitting gently, eyes warm, maintaining focus on the hovering pastel butterfly.
Scene 2 – implicit (continuation)
Continue from the previous shot. Only the time of day, the story beat, a new small pastel butterfly, and the main subject's action change; everything else stays the same (the same setting and backdrop, the

same camera framing, the same main subject, and the same visual style). The time of day becomes late afternoon, the story beat becomes cat smiles broadly while continuing to wave, a small pastel butterfly enters the scene, and the main subject is now sitting down gently. The previous shot ends with cat raises paws in a waving gesture, and the new shot opens with Pink cat stands on hind legs, smiling broadly, waving paw near butterfly.

Sample probes (3 / 72)

segment_state: In segment 1, is the scene set in soft pastel background with whimsical patterns? expected: `yes`
preserve: Is the same main subject present in both segments 1 and 2? expected: `yes`
change: Between segments 1 and 2, does the time of day change from bright daylight to late afternoon? expected: `yes`

Figure 14: Curated 100-record benchmark case 2: cell `(mixed,macro,very_high,single)`.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Training dataset case 1: `mid` dynamic, local-scope macro (`RELATION_ADD+STATE_ATTR_SHIFT`)

scope=local **magnitude**=macro **dynamic_bin**=mid (score 5)
ops: `RELATION_ADD`, `STATE_ATTR_SHIFT`
preserved: `style`, `background`

Source prompt (VidProM)
A Bollywood actress and actor smiling at each other in a vibrant Indian setting. They are standing close together, both wearing traditional colorful Bollywood attire. The actress is in a bright red sari with intricate gold embroidery, while the actor is in a matching red sherwani. Both have joyful expressions, with the actress playfully tilting her head to one side. The background showcases a festive Indian street with ornate decorations, vibrant lights, and bustling crowds. The scene is filled with warmth and happiness. Medium shot, static camera.

Scene 1 – explicit prompt (`from_explicit`)
A medium shot, from a eye level angle, with static camera motion frames a festive indian street (urban exterior) during the day. The atmosphere feels tender, rendered in a cinematic realistic style with warm natural daylight and a vivid saturated reds palette. A bollywood actress in red sari wearing a bright red sari and gold embroidery stands near a bollywood actor in red sherwani smiling, surrounded by ornate decorations and vibrant lights. Couple shares joyful moment in festive street. A bollywood actress in red sari tilting head playfully while the bollywood actor in red sherwani is near the bollywood actress in red sari. The shot opens as couple stands close, maintaining eye contact, they share a joyful gaze while she playfully tilts her head slightly, and closes as actress tilts head playfully, smiles widen.

Scene 2 – explicit prompt (`to_explicit`)
A medium shot from a eye level angle with static camera motion, composed centered. Set in festive indian street (urban exterior) with ornate decorations, vibrant lights, bustling crowds during clear bright day. Rendered in cinematic realistic style with a tender atmosphere. A bollywood actress in red sari tilting head playfully, appearing affectionate and softened gaze, also present: a vibrant lights, the bollywood actor in red sherwani is near the bollywood actress in red sari, the bollywood actress in red sari is looking at the bollywood actor in red sherwani, and the bollywood actor in red sherwani is looking at the bollywood actress in red sari. The couple shares joyful moment in festive street as the shot opens as playful head tilt, wide smiles under vibrant lights, the couple's playful expressions shift into a quiet, loving stare as the festive buzz fades into their shared intimacy, and closes as softened gazes, tender affection between the couple.

Scene 2 – implicit (continuation) prompt (`to_implicit`)
The next shot changes only the main subject's state and the relation between the ornate decorations and the bollywood actor in red sherwani; everything else stays the same (the same setting and backdrop, the same camera framing, the same main subject, and the same visual style). The main subject's state becomes affectionate, softened gaze and the ornate decorations is now on the bollywood actor in red sherwani. Transition: the previous shot ends with actress tilts head playfully, smiles widen, and the new shot opens with Playful head tilt, wide smiles under vibrant lights.

Figure 15: Training dataset case 1: `mid` dynamic, local-scope macro (`RELATION_ADD+STATE_ATTR_SHIFT`).

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Training dataset case 2: `very_high` dynamic, mixed-scope macro, 5 co-ops

scope=mixed **magnitude**=macro **dynamic_bin**=very_high (score 13)
ops: `STORY_BEAT_SHIFT`, `ATMOSPHERE_SHIFT`, `ENTITY_REMOVE`, `CAMERA_ANGLE_SHIFT`, `ACTION_PROGRESS`
preserved: `style`, `background`

Source prompt (VidProM)
A joyful scene featuring LEGO brick dolls dancing happily together in a vibrant 4K animation. The dolls come in various colors and sizes, each displaying unique expressions of joy and excitement. They are arranged in a circle, performing simple yet synchronized dance moves. The background is a colorful, playful environment with LEGO blocks forming a stage and decorative elements. The camera captures the lively scene from a medium shot, maintaining a static view to highlight the detailed movements and interactions of the LEGO dolls.

Scene 1 – explicit prompt (`from_explicit`)
A centered medium shot from eye level captures the scene with static camera motion. Set in a colorful lego stage environment with lego block stage and decorative lego elements, the scene takes place during day. Rendered in vibrant 4k animation style, the atmosphere feels tender. A group of lego brick dolls dancing happily, appearing joyful and excited, perform simple synchronized dance moves in a circle.

Lego dolls dance happily in a circle as the shot opens with them starting movement and closes after completing the synchronized dance step.

Scene 2 – explicit prompt (`to_explicit`)
A medium shot from a high angle captures a colorful lego stage environment. The scene takes place during day under an atmosphere of affection. Rendered in vibrant 4k animation style with bright even light, the visuals feature a vivid saturated palette. A group of lego brick dolls sways in empty void,

appearing joyful and excited. Lego dolls break circle and pair up dancing as the shot opens with them dancing in a circle on colorful bricks. Stage bricks vanish and circle breaks as dolls pair up and float freely. The sequence closes as pairs float joyfully in open space as bricks fade.

Scene 2 – implicit (continuation) prompt (`to_implicit`)
Cutting to the next shot, still keeping the same visual style and the same setting and backdrop. The atmosphere becomes affectionate and the story beat shifts as lego dolls break circle and pair up dancing.

The camera angle becomes high and the lego block stage leaves, leaving the main subject swaying in empty void. Transition: the previous shot ends with lego dolls complete synchronized dance step, and the new shot opens with Lego dolls dance in a circle on colorful bricks.

Figure 16: Training dataset case 2: `very_high` dynamic, mixed-scope macro, 5 co-ops.

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

(1) Per-segment state probes (`segment_state` + `segment_state_negative`)
One YES probe and one matched NO probe are emitted per (segment, field) pair so the judge cannot pass by always answering YES.

• field=`general.background.location`, segment 1
Q: In segment 1, is the scene set in bustling grandstand?
Expected: `yes`
Q: In segment 1, is the scene set in vibrant underwater ocean?
Expected: `no`

• field=`general.background.scene_type`, segment 1
Q: In segment 1, is this a urban exterior scene?
Expected: `yes`
Q: In segment 1, is this a nature interior scene?
Expected: `no`

• field=`general.camera.shot_scale`, segment 1
Q: In segment 1, is this framed as a medium shot?
Expected: `yes`
Q: In segment 1, is this framed as a wide shot?
Expected: `no`

---

(2) Preserved-anchor probes (`preserve` + `transition_differ`)
For each anchor declared kept across the boundary, ask both whether it is consistent and whether it differs — two phrasings testing the same fact from opposite sides.

• anchor=`camera_logic`, boundary 1
Q: Is the camera framing consistent across segments 1 and 2?
Expected: `yes`
Q: Does the camera framing differ between segments 1 and 2?
Expected: `no`

---

(3) Changed-field probes (`change` + `transition_same`)
For each field that the synthesized op-program mutates, emit a positive "does it progress?" probe and a matched negative "is it the same?" probe.

• field=`general.story.beat`, boundary 1
Q: Between segments 1 and 2, does the story beat progress from 'flag waves dramatically in front of cheering crowd' to 'flag flutters wildly as crowd roars in approval'?
Expected: `yes`
Q: Is the story beat the same across segments 1 and 2?
Expected: `no`

---

(4) Boundary smoothness probe (`smoothness`)
One per boundary; tests perceptual continuity rather than schema content.

Q: Is the transition from segment 1 to segment 2 smooth, without abrupt flicker, identity reset, or obvious glitches?
Expected: `yes`

---

(5) Sequence-level structure probe (`structure`)
Asks the judge to count observable scene transitions; answer is the integer ground-truth switch_count.

Q: The video should contain 1 distinct scene transitions, where a scene transition is defined as a noticeable change in setting, background, or main subject. How many scene transitions do you observe? Answer with a single integer.
Expected: `1`

Figure 17: Probe-set anatomy for benchmark case 1 (`b3_bench_040325`; cf. Fig. 13).

Probe-set anatomy for benchmark case 2 (`b3_bench_024021`; **cf. Fig. 14**)

seq_id=`b3_bench_024021` **cell_id**=(mixed,macro,very_high,single) **switch_count**=1
#probes=72

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Type breakdown: `segment_state`=29, `segment_state_negative`=27, `preserve`=4, `transition_differ`=4, `change`=4, `transition_same`=2, `smoothness`=1, `structure`=1.

---

(1) Per-segment state probes (`segment_state` + `segment_state_negative`)
One YES probe and one matched NO probe are emitted per (segment, field) pair so the judge cannot pass by always answering YES.

• field=`general.background.location`, segment 1
Q: In segment 1, is the scene set in soft pastel background with whimsical patterns?
Expected: `yes`
Q: In segment 1, is the scene set in blurred cozy indoor setting?
Expected: `no`

• field=`general.background.scene_type`, segment 1
Q: In segment 1, is this a nature interior scene?
Expected: `yes`
Q: In segment 1, is this a urban exterior scene?
Expected: `no`

• field=`general.camera.shot_scale`, segment 1
Q: In segment 1, is this framed as a close shot?
Expected: `yes`
Q: In segment 1, is this framed as a extreme wide shot?
Expected: `no`

---

(2) Preserved-anchor probes (`preserve` + `transition_differ`)
For each anchor declared kept across the boundary, ask both whether it is consistent and whether it differs — two phrasings testing the same fact from opposite sides.

• anchor=`main_character`, boundary 1
Q: Is the same main subject present in both segments 1 and 2?
Expected: `yes`
Q: Does the main subject differ between segments 1 and 2?
Expected: `no`

• anchor=`style`, boundary 1
Q: Do both segments 1 and 2 share the same visual style?
Expected: `yes`
Q: Does the visual style differ between segments 1 and 2?
Expected: `no`

---

(3) Changed-field probes (`change` + `transition_same`)
For each field that the synthesized op-program mutates, emit a positive "does it progress?" probe and a matched negative "is it the same?" probe.

• field=`general.background.time_of_day`, boundary 1
Q: Between segments 1 and 2, does the time of day change from bright daylight to late afternoon?
Expected: `yes`
Q: Is the time of day the same across segments 1 and 2?
Expected: `no`

• field=`general.story.beat`, boundary 1
Q: Between segments 1 and 2, does the story beat progress from 'cute pink cat greets viewer with a wave' to 'cat smiles broadly while continuing to wave'?
Expected: `yes`
Q: Is the story beat the same across segments 1 and 2?
Expected: `no`

---

(4) Boundary smoothness probe (`smoothness`)
One per boundary; tests perceptual continuity rather than schema content.

Q: Is the transition from segment 1 to segment 2 smooth, without abrupt flicker, identity reset, or obvious glitches?
Expected: `yes`

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

(5) Sequence-level structure probe (`structure`)
Asks the judge to count observable scene transitions; answer is the integer ground-truth switch_count.

Q: The video should contain 1 distinct scene transitions, where a scene transition is defined as a noticeable change in setting, background, or main subject. How many scene transitions do you observe? Answer with a single integer.
Expected: `1`

Figure 18: Probe-set anatomy for benchmark case 2 (`b3_bench_024021`; cf. Fig. 14).

Confidential reviewer copy. This manuscript is submitted to the 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Unauthorized sharing, redistribution, or disclosure is strictly prohibited.