WORLDBENCH: SELECTIVE STATE MAINTENANCE IN MULTI-SHOT VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in video generation have enabled increasingly coherent multi-shot video generation with controllable camera viewpoints and recurring subjects. While recent multi-shot video generators have made substantial progress in cross-shot visual consistency, they still frequently exhibit *world-state collapse*, leading to inconsistencies in subject count and spatial assignments across shots. Motivated by this gap, we introduce **WorldBench**, a controlled benchmark for evaluating world-state maintenance in multi-shot video generation. Each episode establishes an initial world state with a wide shot and later re-observes the resulting world from a different viewpoint after controlled narrative changes. Experiments on representative open- and closed-weight models show that even Seedance 2.5 achieves only 56.2% end-to-end success, confirming that world-state maintenance remains a substantial challenge and highlighting the significance of WorldBench for systematically evaluating this capability. Our code is available at https://anonymous.4open.science/r/WorldBench-1.

1 INTRODUCTION

Recent advances in video generation (Wan Team, 2025; ByteDance Seed, 2026) have substantially improved visual fidelity, motion quality, and controllability, enabling increasingly practical applications in filmmaking, advertising, and interactive content creation. As these models move beyond isolated clips toward structured narratives, multi-shot video generation has emerged as an important setting in which a sequence is composed of multiple shots with changing camera viewpoints (Li et al., 2026; Yang et al., 2026). Recent progress (Meng et al., 2026; Yang et al., 2026) has further enabled explicit control over shot composition and camera transitions while preserving the appearance of recurring subjects across cuts. However, rather than merely preserving visual consistency, coherent multi-shot generation also requires models to maintain a persistent world state across shots, while each shot provides only a partial observation of that world.

World-state maintenance can be characterized along two complementary dimensions. When no state-changing event occurs, previously established entities and their spatial relations should persist despite changes in viewpoint and temporary absence from the current observation. When the narrative explicitly changes part of the scene, the generator should incorporate the requested update while preserving all unaffected states. As illustrated in Figure 1, the requested seat exchange should update only the spatial assignments of the two involved participants, while preserving the presence and locations of all unaffected participants. However, the generated final view introduces two additional participants, an unintended corruption of previously established but unaffected state which we refer to as **world-state collapse**. Importantly, this phenomenon persists even among the strongest contemporary video generation models, suggesting that maintaining and selectively updating a persistent world state remains an unresolved challenge. These observations motivate us to evaluate whether multi-shot video generation models today can maintain a persistent world state across changing viewpoints while selectively updating it in response to explicit narrative events.

A central challenge in evaluating world-state maintenance is how to determine whether an observed change reflects an update to the underlying world state or merely a change in how that state is observed by the camera, since an entity may disappear from a shot simply because it falls outside the current view. We therefore introduce **WorldBench**, which uses controlled viewpoint changes to evaluate world-state maintenance across shots. Recent advances in viewpoint controllability make

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



Required and generated final cast

Final state	Count	Participants
Required	4	A, B, C, D
Generated	6	A, B, C, D + two extras

Selective state maintenance

A B C D

Swap **A** and **B**, preserve **C** and **D**.

Two extra people despite no entry.

Figure 1. Example of state maintenance failure in a game-room episode generated by Seedance 2.0 Fast. The opening shot establishes four participants, and the intermediate prompt only changes their spatial arrangement by asking A and B to exchange seats. The expected final state therefore contains the same four participants, but the final wide shot introduces two additional people, highlighted by the outlines, despite no entry event.

such controlled evaluation increasingly available by allowing shot scale and viewing direction to be explicitly constrained. Each episode begins with a wide view that establishes the scene, proceeds through partial observations with or without explicit state updates, and finally re-observes the same world from a different viewpoint. Furthermore, controlled variations in subject count and shot count allow us to analyze how world-state maintenance degrades with scene complexity and sequence length.

Using WorldBench, we conduct extensive experiments on representative open- and closed-weight video generation models. The results show that current models remain limited in their ability to maintain world state across shots, with Seedance 2.5 (Volcengine, 2026), the strongest evaluated model, achieving only 56.2% end-to-end success. We further find that preserving an unchanged state is substantially easier than selectively updating it, with Position Swap emerging as the most challenging transition, while performance consistently declines as the number of subjects or shots increases. Taken together, these results reveal clear limitations in current models' world-state maintenance capabilities and highlight the significance of WorldBench as a diagnostic benchmark beyond conventional cross-shot consistency.

Our work provides a benchmark for multi-shot video generation models with a focus on persistent world-state maintenance. The key contributions are:

- We formulate **world-state maintenance** as a distinct challenge in multi-shot video generation, requiring models to preserve previously established states across partial observations while selectively updating only the components affected by narrative events.
- We introduce **WorldBench**, a controlled benchmark with 112 story-grounded episodes spanning seven scene types and four state transitions, with controlled viewpoint changes to evaluate world-state maintenance.
- We conduct extensive experiments on representative open- and closed-weight video generation models, revealing their limitations in world-state maintenance.

2 Related Work

Multi-Shot Video Generation. Multi-shot video generation seeks to coordinate shot-specific instructions with shared contextual information so that characters, scenes, and ongoing events remain coherent across changes in action and viewpoint. Recent joint generation methods, including ShotAdapter, HoloCine, and MultiShotMaster, combine per-shot conditioning with mechanisms

that model dependencies across shots (Kara et al., 2025; Meng et al., 2026; Wang et al., 2026a). StoryMem and ShotStream instead propagate historical visual context through generated keyframes or cached representations, enabling longer and more interactive sequences (Zhang et al., 2025; Luo et al., 2026), while ShotVerse maps narrative instructions to camera trajectories within a shared coordinate system (Yang et al., 2026). Commercial systems such as Sora 2, Seedance 2.0, and Kling Video 3.0 have also introduced multi-shot generation capabilities (OpenAI, 2025; ByteDance Seed, 2026; Kuaishou Technology, 2026b). Together, these advances make structured generation across changing viewpoints increasingly practical and motivate evaluating whether the underlying world remains stable as the narrative evolves.

Multi-Shot Video Evaluation. Video benchmarks assess perceptual quality, prompt adherence, and compositional correctness (Huang et al., 2024; Zheng et al., 2025; Sun et al., 2025), with recent work extending evaluation to narrative, audiovisual, and entity coherence across shots (Wei et al., 2026; Wang et al., 2026b; He et al., 2026). MSVBench and PersonaShot incorporate state persistence into broader continuity assessment (Shi et al., 2026; Wang et al., 2026c), while StEvo-Bench and StateBench study evolution under interrupted observation and continuation from historical states (Ma et al., 2026; Miao et al., 2026). WorldBench focuses on controlled episodes in which a valid generated opening and prescribed discrete events determine the complete expected final cast and assignment of identities to physical places. This enables joint evaluation of requested changes and preservation of unaffected participants, including those omitted from the final content instruction, through exact comparisons of visual state records.

3 World-State Maintenance Across Shots

3.1 Persistent World State

A video can be viewed as a sequence of two-dimensional observations of a dynamic three-dimensional world. In a multi-shot video, the camera observes the same scene from different viewpoints across shots, but these changes affect what is observed rather than the underlying world itself. Coherent generation therefore requires previously established scene information to remain valid unless it is explicitly changed by the narrative. We refer to this evolving yet persistent configuration as the **world state**, which captures scene-level information such as the number of subjects and their locations, and should remain stable unless changed by the narrative.

We consider a multi-shot video episode generated from an ordered sequence of shot prompts:

$$P = (p_1, \ldots, p_T), \qquad p_t = (c_t, v_t), \tag{1}$$

where c_t specifies the narrative content of shot t and v_t specifies its viewpoint. The former governs how the world state evolves, while the latter determines how that state is observed. Let S_t denote the world state at shot t, with the initial state jointly established by the opening content and the model's visual realization, and subsequently updated by the narrative content of later shots,

$$S_1 = \mathrm{Establish}(c_1), \qquad S_t = \mathrm{Update}(S_{t-1}, c_t), \quad t = 2, \ldots, T. \tag{2}$$

Each generated shot is treated as a partial observation of the current world state under the specified viewpoint,

$$O_t = \mathrm{Observe}(S_t, v_t). \tag{3}$$

By separating world-state evolution from camera observation, this formulation defines continuity at the level of the underlying world rather than individual shots. Maintaining such a world state is essential for multi-shot generation, as it allows different views to remain grounded in the same evolving scene while preserving information that is temporarily outside the current observation.

3.2 Selective State Maintenance

A coherent multi-shot video should depict the same underlying world across shots while allowing its state to change only where required by the narrative, a property we refer to as **selective state maintenance**. Let $\mathcal{D}_t$ denote the subset of world-state components affected by the narrative content of shot t. For any state component x, selective state maintenance requires

$$S_t(x) = \begin{cases} \mathrm{Update}(S_{t-1}(x), c_t), & x \in \mathcal{D}_t, \\ S_{t-1}(x), & x \notin \mathcal{D}_t. \end{cases} \tag{4}$$

This formulation naturally distinguishes two complementary cases according to whether a state component is affected by the current narrative. **State persistence** requires previously established states to remain valid when no corresponding change is introduced, even when they are temporarily unobserved or unmentioned, whereas **selective state update** requires explicit events such as entry, exit, or position change to modify the affected states while preserving all unaffected ones. Together, these two cases define the world-state consistency required for successive shots to remain part of the same evolving world.

4 WORLDBENCH CONSTRUCTION

World-state maintenance is difficult to evaluate because changes in observation can be mistaken for changes in the world itself. We therefore construct controlled story episodes that separate viewpoint changes from narrative state changes, and further extend them with harder subject and shot settings to probe the capability boundary of current multi-shot video models.

4.1 CONTROLLED VIEWPOINT PROBING

Evaluating world-state maintenance is challenging because each shot provides only a partial view of the scene, making it difficult to determine whether a missing subject reflects a state change or simply a change in observation. Recent improvements in viewpoint controllability provide a practical way to address this ambiguity by allowing the observed region of a scene to be constrained through shot scale and camera direction. We use this capability to construct episodes that first establish a world state and later expose the scene again from a different viewpoint.

Building on this capability, we therefore construct each episode around three functional stages. **(1) World-state establishment.** An opening wide shot first establishes an observable initial world state by revealing all subjects and their relative spatial arrangement within the scene. **(2) Narrative progression.** Intermediate shots then follow the natural progression of the story, during which subjects may enter, leave, or change positions, while the camera is free to vary with the narrative. **(3) World-state re-observation.** A final wide shot is then explicitly required to re-observe the entire scene, allowing us to probe whether the resulting world state remains consistent with the preceding narrative. To prevent models from achieving apparent continuity by simply reproducing previous frames, we require the final wide shot to adopt a different viewpoint, using either a reverse or overhead view, while preserving sufficient spatial overlap for reliable comparison.

4.2 CONTROLLED STATE TRANSITIONS

To evaluate selective state maintenance under controlled narrative changes, WorldBench defines four state transitions that modify either subject presence or spatial assignment while requiring all unaffected states to remain stable.

1. **Static.** The established world state remains unchanged, requiring previously established subjects and their locations to persist when the scene is re-observed from different viewpoints.
2. **Entry and Exit.** The set of subjects present in the scene changes through the addition or removal of specific participants, while the states of all unaffected subjects remain unchanged.
3. **Position Swap.** The same participants remain in the scene, but selected participants exchange their spatial locations while all other assignments are preserved.

Together, these transitions cover the two primary aspects of world state considered in WorldBench, namely which subjects are present and where they are located. The core benchmark combines these four transition types with seven scene types, yielding 28 configurations and 112 multi-shot prompts.

4.3 STORY-GROUNDED EPISODE CONSTRUCTION

Story Source Selection. To combine experimental control with natural narrative structure, we construct each episode from a predefined scene setting and a traceable story source. We consider seven representative scene types covering common multi-person interactions in a café, meeting room, living room, dining room, seminar room, game room, and kitchen. For each scene type, we draw on

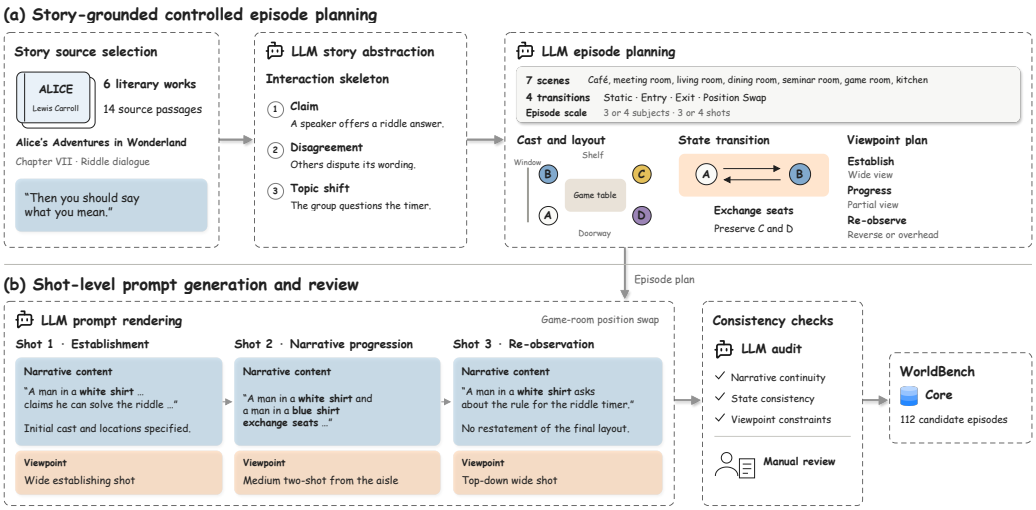


Figure 2. Construction pipeline of WorldBench. Story passages are abstracted into interaction skeletons and instantiated under controlled cast, state-transition, and viewpoint constraints, after which the episode plans are rendered into shot-level prompts and checked for consistency. The game-room example illustrates how a requested seat swap is combined with a final overhead observation.

real story sources to obtain natural interaction patterns and adapt them into benchmark episodes under controlled state-transition and viewpoint constraints. Story passages that match the required scene and interaction settings are directly selected from six literary works and adapted into benchmark episodes under predefined state-transition and viewpoint constraints.

Prompt Generation. Each selected passage is first abstracted into a concise story skeleton that preserves the core interaction while removing visual details unnecessary for controlled generation. Based on this skeleton, the cast and initial layout are instantiated, the target world-state transition is specified, and a camera plan is constructed according to the controlled observation protocol described above. The resulting episode plan is then rendered into shot-level prompts, with **narrative content** specifying the local event and **viewpoint instructions** determining how the scene is observed. To systematically carry out these steps, a staged language-model pipeline is designed for story abstraction, episode planning, prompt generation, and consistency checking, with GPT-5.6 Sol used throughout the pipeline, followed by manual review to remove cases with ambiguous state transitions or unreliable camera requirements.

5 Evaluation Protocol

WorldBench evaluates whether the final world state remains consistent with the generated opening and the narrative changes introduced across the episode. As shown in Figure 3, the evaluation derives an expected final state from the observed opening and the requested updates, then compares it with the final re-observation.

5.1 Opening World-State Reference

The opening wide shot defines the world established by the model and serves as the reference for subsequent evaluation. We first detect hard cuts using FFmpeg scene-change detection, from which shot boundaries and aligned video segments are obtained. For each episode e, two frames are then sampled from the opening shot to identify the generated participants and their coarse physical locations. For each relevant state component x, the evaluator records its observed value and represents the opening state as

$$S_{\text{open}} = \{(x, S_{\text{open}}(x))\}, \tag{5}$$

where $S_{\text{open}}(x)$ denotes the observed value of each evaluated world-state component x, primarily subject presence and coarse physical location, and S_{open} is fixed as the reference for subsequent state

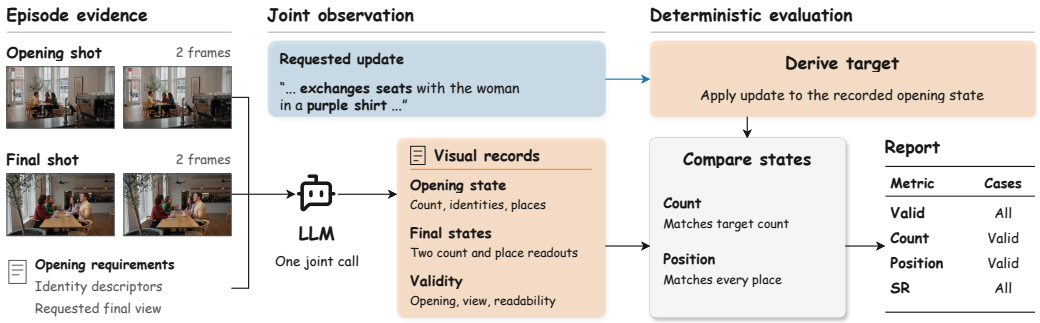


Figure 3. Evaluation protocol of WorldBench. The generated opening provides a reference world state, to which the requested narrative update is applied to derive the expected final state. The final shot is then re-observed from the prescribed viewpoint and compared with this target. Opening and final frames are sampled at relative times 0.25 and 0.75 within their aligned shots.

derivation and cross-view comparison. An opening is accepted as a valid reference only when all intended participants can be reliably identified and their physical locations clearly distinguished. The observed appearance descriptions are retained as identity references for later re-observation, with entrants anchored at their first visible shot.

5.2 Target State Derivation

Once the opening reference is established, the intermediate shots specify how this world should evolve before it is re-observed in the final shot. Let U denote the ordered sequence of narrative state changes introduced between the opening shot and the final world-state re-observation. Starting from S_{open}, these changes are applied in narrative order to derive the expected final state,

$$S_{\text{target}} = \text{Apply}(S_{\text{open}}, U). \tag{6}$$

Following the selective state maintenance principle in Section 3, each update changes only the state components involved in the narrative while preserving all others. Under this formulation, Static keeps the opening state unchanged, Entry and Exit modify subject presence, and Position Swap changes the locations of the specified participants. The resulting S_{target} therefore represents how the world established in the opening should have evolved after all requested narrative events.

5.3 Final Re-observation and State Evaluation

Re-observation Validity. Given a valid opening reference and its derived target state S_{target}, the final shot is first checked to ensure that it re-observes enough of the scene to evaluate the resulting world state. We sample two frames from the final shot and evaluate each for viewpoint compliance, coverage of the relevant interaction area, and sufficient visual clarity to identify the subjects and their physical locations. A final observation is considered invalid only when the required scene region is not sufficiently visible or when key physical locations cannot be reliably identified. Observable discrepancies, such as an unexpected participant, a missing occupant, or an incorrect spatial assignment, are instead treated as world-state errors rather than observation failures. For each episode e, we define the validity indicator as

$$V_e = \mathbb{I}[\text{valid opening reference} \wedge \text{valid final re-observation}], \tag{7}$$

where $V_e = 1$ indicates that both the opening and final observations provide sufficient evidence for world-state evaluation, and $V_e = 0$ otherwise.

Final-State Recovery. For each valid episode, the evaluator recovers the final world state from the two sampled frames using the same state components x defined in Section 3. Subject identities are matched to the references established in the opening, while their physical locations are recorded under the new viewpoint. The resulting observation is represented as

$$S_{\text{final}} = \{(x, S_{\text{final}}(x))\}, \tag{8}$$

where both sampled frames are required to support the recovered state to avoid relying on a transiently correct observation. The recovered S_{final} is then compared with S_{target} to determine whether the world state has evolved consistently with the requested narrative.

State Correctness and Metrics. With S_{final} recovered from the final re-observation, we quantify both whether an episode provides sufficient evidence for evaluation and whether the observed world state matches S_{target}. We first report **Valid**, which measures the fraction of episodes with both a valid opening reference and a valid final re-observation,

$$\text{Valid} = \frac{1}{N} \sum_{e \in \mathcal{D}} V_e, \tag{9}$$

where $\mathcal{D}$ contains N benchmark episodes. For valid episodes, we compare the recovered final state S_{final} with S_{target} along two aspects of world-state correctness. **Count** evaluates whether the final number of subjects matches the target, capturing unexpected additions or omissions, while **Position** evaluates whether subjects occupy the expected physical locations. Let C_e and P_e indicate whether both sampled final frames satisfy the corresponding count and position criteria for episode e. We aggregate these judgments over valid episodes as

$$\text{Count} = \frac{\sum_{e \in \mathcal{D}} V_e C_e}{\sum_{e \in \mathcal{D}} V_e}, \qquad \text{Position} = \frac{\sum_{e \in \mathcal{D}} V_e P_e}{\sum_{e \in \mathcal{D}} V_e}. \tag{10}$$

Finally, an episode is considered successful only when it is valid and both its count and position are correct. The overall success rate is therefore

$$\text{SR} = \frac{1}{N} \sum_{e \in \mathcal{D}} V_e C_e P_e. \tag{11}$$

While **Count** and **Position** measure state correctness over evaluable episodes, **SR** additionally provides an end-to-end measure of world-state maintenance over the full benchmark.

6 Experiments

Experimental Setup. We evaluate a diverse set of open- and closed-weight video generation models on WorldBench, which contains 112 multi-shot episodes spanning seven scene types and four state transitions: Persistence, Entry, Exit, and Position Swap. Controlled subject-count and shot-count subsets are further used to analyze the effects of scene complexity and sequence length. We use GPT-5.6 Sol (OpenAI, 2026) as the multimodal evaluator to extract the structured visual records required by our evaluation protocol.

Baselines. Our baselines cover representative open-weight and closed-weight video generation models, following the grouping in Table 1. **(1) Open-weight models** include LTX-2.5 (Lightricks, 2026), HoloCine 14B (Meng et al., 2026) and CineTrans-DiT (Wu et al., 2026). **(2) Closed-weight models** include Seedance 2.0 (Team Seedance et al., 2026), Seedance 2.5 (Volcengine, 2026), Kling 2.6 (Kuaishou Technology, 2026a), Kling 3.0 (Kuaishou Technology, 2026b), Wan 3.0 (Alibaba Cloud, 2026), and MiniMax H3 Max (MiniMax, 2026), representing commercial video generation systems.

6.1 Benchmark Performance

Table 1 reports the performance of representative open- and closed-weight video generation models on WorldBench. The results provide both an overall assessment of world-state maintenance and a breakdown across different state transitions.

Main Results. Table 1 shows substantial variation in world-state maintenance across current multi-shot video generators. Seedance 2.5 achieves the highest end-to-end success rate (SR) of 56.2%, followed by Seedance 2.0 at 39.3% and Wan 3.0 at 35.7%. Other closed-weight models obtain SRs between 20.5% and 30.4%, while all evaluated open-weight baselines remain below 18%. More broadly, stronger recent models tend to perform better on WorldBench, with Seedance 2.5 substantially improving over Seedance 2.0. Nevertheless, even the best-performing model succeeds on only 56.2% of the episodes, indicating that reliable world-state maintenance remains an open challenge.

Table 1. Comparison of open weight and closed weight video generation models on our WorldBench benchmark. Valid measures the fraction of evaluable episodes. Count and Position measure final state correctness over valid episodes, while SR measures end to end success over all cases. The last four columns report SR for Persistence, Entry, Exit, and Position Swap. All values are percentages and higher is better. The best result in each column is **bolded**.

	All Cases				Persistence	Entry	Exit	Position Swap
Model	**Valid↑**	**Count↑**	**Position↑**	**SR↑**	**SR↑**	**SR↑**	**SR↑**	**SR↑**
Open-weight								
LTX-2.5 (Lightricks, 2026)	47.3	60.4	34.0	15.2	46.4	0.0	7.1	7.1
HoloCine 14B (Meng et al., 2026)	28.6	34.4	21.9	1.1	4.5	0.0	0.0	0.0
CineTrans-DiT (Wu et al., 2026)	22.3	20.0	4.0	0.9	3.6	0.0	0.0	0.0
Closed-weight								
Kling 2.6 (Kuaishou Technology, 2026a)	25.9	**100.0**	**79.3**	20.5	50.0	0.0	32.1	0.0
Kling 3.0 (Kuaishou Technology, 2026b)	64.3	72.2	50.0	30.4	42.9	32.1	42.9	3.6
Wan 3.0 (Alibaba Cloud, 2026)	66.1	83.8	62.2	35.7	71.4	25.0	39.3	7.1
MiniMax H3 Max (MiniMax, 2026)	62.5	77.1	57.1	29.5	60.7	28.6	25.0	3.6
Seedance 2.0 (Team Seedance et al., 2026)	67.9	85.5	57.9	39.3	60.7	35.7	46.4	14.3
Seedance 2.5 (Volcengine, 2026)	**81.2**	90.1	71.4	**56.2**	**92.9**	**53.6**	**57.1**	**21.4**

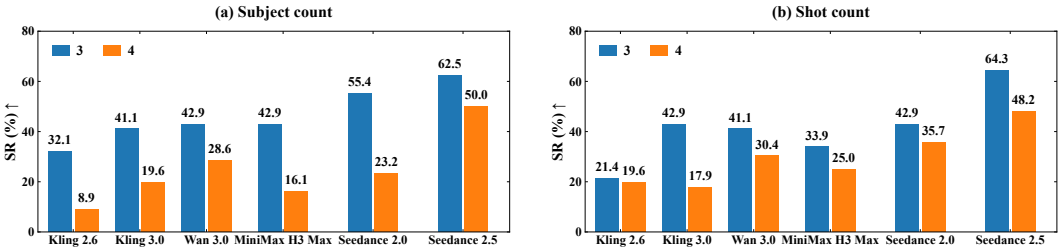


Figure 4. Effect of subject count and shot count on world state maintenance. (a) SR on episodes with three and four subjects. (b) SR on episodes with three and four shots. These comparisons evaluate the effects of scene complexity and sequence length on world state maintenance. All values are percentages and higher is better.

Validity versus State Correctness. Strong performance on evaluable episodes does not necessarily imply reliable end-to-end world-state maintenance. Kling 2.6 achieves 100.0% Count and 79.3% Position accuracy on valid episodes, but its Valid rate is only 25.9%, resulting in an SR of 20.5%. In contrast, Seedance 2.5 combines a much higher Valid rate of 81.2% with 90.1% Count and 71.4% Position accuracy, yielding an SR of 56.2%. This contrast validates our evaluation design of separating observation validity from state correctness, preventing strong conditional performance on a small evaluable subset from overstating end-to-end world-state maintenance.

Performance across State Transitions. Performance varies substantially across state-transition types. Persistence is consistently the easiest setting, while transitions requiring explicit state updates are considerably more challenging. For Seedance 2.5, SR decreases from 92.9% on Persistence to 53.6% on Entry and 57.1% on Exit, and further to 21.4% on Position Swap. Position Swap is particularly difficult across models, with all evaluated methods achieving at most 21.4% SR. These results indicate that current models are substantially better at preserving an established world state than selectively modifying it while keeping unaffected states unchanged.

6.2 Effect of subject and shot count.

Figure 4 reports SR under controlled subject-count and shot-count settings. These comparisons isolate the effects of scene complexity and sequence length on world-state maintenance by varying the number of subjects and shots, respectively.

Increasing either the number of subjects or the number of shots consistently reduces SR across all evaluated models. Raising the subject count from three to four leads to an average drop of 21.8 percentage points, while increasing the shot count from three to four results in a smaller average drop



Figure 5. Representative failure cases in WorldBench. (a,b) Invalid episodes caused by an incorrect opening cast or an unobservable final state. (c,d) Count errors caused by extra or missing participants. (e,f) Position errors caused by an unexecuted swap or inconsistent physical layout across shots.

of 11.6 points. These results show that both scene complexity and sequence length make world-state maintenance more difficult, with subject count exerting a stronger effect in the current setting. This suggests that tracking multiple identities and their spatial assignments remains a major challenge for current multi-shot video generators.

6.3 Failure Mode Analysis

Figure 5 presents representative failure modes captured by WorldBench, covering validity failures, count errors, and position errors.

- **Validity failures.** Figure 5(a) shows a case where the opening shot contains an incorrect cast, preventing a valid reference state from being established. Figure 5(b) instead fails at re-observation, where the final view does not expose enough of the scene for reliable state recovery.
- **Count errors.** Figs. 5(c,d) show that subject presence can still collapse even when the episode is valid. Models may introduce unexpected participants or remove subjects that should remain, revealing difficulty in preserving the correct subject set throughout state transitions.
- **Position errors.** Figure 5(e) shows a requested swap that is not executed, while Figure 5(f) exhibits a change in the underlying physical layout. These failures indicate that models struggle to preserve stable identity-to-place correspondences across shots.

Additional examples and visualizations of these failure modes are provided in Appendix D.

7 Conclusion

We introduce WorldBench, a controlled benchmark for evaluating selective world-state maintenance in multi-shot video generation. By disentangling narrative state changes from viewpoint changes, WorldBench assesses whether models can preserve established states while selectively updating only those affected by the narrative. Experiments on representative open- and closed-weight models reveal substantial limitations in this capability, with even the best-performing model achieving only 56.2% end-to-end success. These results establish selective world-state maintenance as a critical yet unresolved challenge for coherent multi-shot video generation.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and improving the clarity of the manuscript. In addition, an LLM assistant was used in benchmark construction and as a multimodal evaluator to extract structured visual records from generated videos. All AI-assisted outputs were reviewed and verified by the authors.

REFERENCES

Alibaba Cloud. Wan3.0 video generation. Alibaba Cloud Model Studio documentation, 2026. URL https://www.alibabacloud.com/help/en/model-studio/wan3-video-generation-guide. Accessed September 10, 2026.

ByteDance Seed. Seedance 2.0 Official Launch. Official model announcement, 2026. URL https://seed.bytedance.com/en/blog/seedance-2-0-official-launch. Accessed September 8, 2026.

Ruozhen He, Meng Wei, Ziyan Yang, and Vicente Ordonez. EntityBench: Towards Entity-Consistent Long-Range Multi-Shot Video Generation, 2026. URL https://arxiv.org/abs/2605.15199. arXiv preprint.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28405–28415, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Kara_ShotAdapter_Text-to-Multi-Shot_Video_Generation_with_Diffusion_Models_CVPR_2025_paper.html.

Kuaishou Technology. Kling 2.6: Text-to-Video API. Official API documentation, 2026a. URL https://kling.ai/document-api/api/video/2-6/text-to-video. Accessed September 20, 2026.

Kuaishou Technology. Kling AI Launches 3.0 Model, Ushering in an Era Where Everyone Can Be a Director. Official model announcement, 2026b. URL https://ir.kuaishou.com/news-releases/news-release-details/kling-ai-launches-30-model-ushering-era-where-everyone-can-be. Accessed September 8, 2026.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2858f8c8683aaa8c12d487354cf328dc-Abstract-Conference.html.

Lightricks. LTX-2.5: Open source documentation. Official model documentation, 2026. URL https://docs.ltx.io/open-source-model/getting-started/overview. Accessed September 10, 2026.

Yawen Luo, Xiaoyu Shi, Junhao Zhuang, Yutian Chen, Quande Liu, Xintao Wang, Pengfei Wan, and Tianfan Xue. ShotStream: Streaming multi-shot video generation for interactive storytelling. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2603.25746.

Ziqi Ma, Mengzhan Liufu, and Georgia Gkioxari. Out of sight, out of mind? evaluating state evolution in video world models, 2026. URL https://arxiv.org/abs/2603.13215. arXiv preprint.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Shuailei Ma, Yixuan Li, Cheng Chen, Yanhong Zeng, Yujun Shen, and Huamin Qu. HoloCine: Holistic generation of cinematic multi-shot long video narratives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://holo-cine.github.io/.

Yingmao Miao, Pengfei Zhang, Chaoran Xu, Meng Yu, Jing Tang, Xiangxiang Chu, Chao Shen, and Chenhao Lin. Do video generators track the world across segments? a benchmark and method for world-state reasoning in video continuation, 2026. URL https://arxiv.org/abs/2609.03673. arXiv preprint.

MiniMax. MiniMax H3 Max: Video Generation. Official API documentation, 2026. URL https://platform.minimax.cn/docs/guides/video-generation. Accessed September 20, 2026.

OpenAI. Sora 2 is here. Official model announcement, 2025. URL https://openai.com/index/sora-2/. Accessed September 8, 2026.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, July 2026. Accessed: 2026-09-26.

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. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. URL https://aclanthology.org/2026.findings-acl.1203/.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Sun_T2V-CompBench_A_Comprehensive_Benchmark_for_Compositional_Text-to-video_Generation_CVPR_2025_paper.html.

Team Seedance et al. Seedance 2.0: Advancing video generation for world complexity, 2026. URL https://arxiv.org/abs/2604.14148. Technical report.

Volcengine. Seedance 2.5: Model Catalogue. Official API documentation, 2026. URL https://docs.volcengine.com/docs/82379/1330310. Accessed September 20, 2026.

Wan Team. Wan2.2: Open and advanced large-scale video generative models. Official model repository, 2025. URL https://github.com/Wan-Video/Wan2.2. Accessed September 10, 2026.

Qinghe Wang, Xiaoyu Shi, Baolu Li, Weikang Bian, Quande Liu, Huchuan Lu, Xintao Wang, Pengfei Wan, Kun Gai, and Xu Jia. MultiShotMaster: A controllable multi-shot video generation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a. URL https://arxiv.org/abs/2512.03041.

Shengyi Wang, Niantong Li, Guangzheng Hu, Hong Qi, Fei Ding, Weixu Qiao, Jinlin Wang, Xiaotong Lv, Peng Han, Zimeng Li, Fanshu Ding, Yushu Wang, Han Wu, Jingjing Chen, Chongxiao Wang, Yanhao Wu, Chenglong Huang, Xiaoqian Zhu, Jie Tian, Hua Li, Jingjing Fan, Mingshuang Tang, Zhong Li, Hengxia Qiang, Weibin Chen, Jinyang Zhen, Bing Zhao, Lin Qu, Jing Li, and Hu Wei. FilmBench: A Film-Grade Benchmark for Cinematic Video Generation, 2026b. URL https://arxiv.org/abs/2607.24241. arXiv preprint.

Yuji Wang, Yuheng Chen, Teng Hu, Ran Yi, Yijia Hong, Han Feng, Weijian Cao, Chengjie Wang, Lizhuang Ma, and Jiangning Zhang. PersonaShot: Benchmarking Person-Centric Narrative Continuity in Multi-Shot Video Generation, 2026c. URL https://arxiv.org/abs/2608.16717. arXiv preprint.

Yujie Wei, Yujin Han, Zhekai Chen, Yongming Li, Kaixun Jiang, Zhihang Liu, Quanhao Li, Zhiwu Qing, Xiang Wang, Zhen Xing, Ruihang Chu, Lingyi Hong, Yefei He, Junjie Zhou, Junqiu Yu, Yang Shi, Difan Zou, Kai Zhu, Shiwei Zhang, Yingya Zhang, Yu Liu, Xihui Liu, and Hongming Shan. MSAVBench: Towards Comprehensive and Reliable Evaluation of Multi-Shot Audio-Video Generation, 2026. URL https://arxiv.org/abs/2605.20183. arXiv preprint.

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. URL https://openreview.net/forum?id=955hVLJdfP.

Songlin Yang, Zhe Wang, Xuyi Yang, Songchun Zhang, Xianghao Kong, Taiyi Wu, Xiaotong Zhao, Ran Zhang, Alan Zhao, and Anyi Rao. ShotVerse: Advancing cinematic camera control for text-driven multi-shot video creation, 2026. URL https://arxiv.org/abs/2603.11421. arXiv preprint.

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, 2025. URL https://arxiv.org/abs/2512.19539. arXiv preprint.

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

A Benchmark Construction Details

A.1 Episode Construction Pipeline

Episode construction follows a staged pipeline that separates narrative abstraction, controlled state specification, and prompt rendering. Table 2 summarizes the five stages.

Table 2. Stages used to construct WorldBench episodes. Each stage operates under the state and viewpoint constraints defined by the benchmark.

Stage	Input	Output
Source abstraction	Selected story passage	Interaction skeleton preserving the core narrative structure.
Subject casting	Interaction skeleton and attribute constraints	Cast with distinguishable appearance descriptors.
Story planning	Cast and target state transition	Narrative beats satisfying the required event and mention constraints.
Prompt rendering	Episode plan and viewpoint specification	Shot-level narrative and camera instructions.
Prompt audit	Episode specification and rendered prompts	Consistency checks for narrative, state transitions, and viewpoints.

The episode plan determines the initial cast, physical layout, requested state transition, subject mentions, and viewpoint sequence before the shot-level prompts are rendered. This separation prevents incidental wording choices from altering the underlying state-transition specification.

B World-State and Transition Specification

B.1 Worked Example of Target Derivation

Consider an episode in which four subjects A, B, C, and D are observed in the opening at physical places q_A, q_B, q_C, and q_D. The intermediate instruction requests that A and B exchange places. The final content does not need to restate the complete cast or layout.

The expected final state is obtained by applying the requested swap to the observed opening. Subjects A and B exchange their assignments, while the states of C and D remain unchanged.

Table 3. Example target state for a requested exchange between A and B. The target is derived from the physical places observed in the opening.

Identity	Opening place	Target place	Requirement
A	q_A	$\mathbf{q_B}$	Requested update
B	q_B	$\mathbf{q_A}$	Requested update
C	q_C	q_C	Preserved state
D	q_D	q_D	Preserved state

This construction allows the evaluation to test both selective updating and state preservation. The changed subjects must reach their requested target places, while subjects that are not involved in the event must retain their previously established states.

C Limitations

Evaluation Scope. WorldBench focuses on subject presence and coarse physical-place assignments as two fundamental components of world state. This controlled setting does not cover more general forms of state evolution such as continuous motion, fine-grained spatial relations, object states, or long-range narrative dependencies. The benchmark also considers a limited set of indoor scenes and

structured multi-subject interactions, leaving broader environments and more diverse state transitions for future study.

D FURTHER VISUALIZATIONS

Figures 6–11 expand the six failure modes in Figure 5. Each figure presents four episodes from different models, with two opening frames, two final readouts, and the recorded scores. The final line states the expected behavior and the observed failure.

1. Seedance 2.5 | Person Exit

Valid: FAIL Count: N/A Position: N/A SR: FAIL

Opening 1 | 0.76 s Opening 2 | 2.28 s Final 1 | 7.04 s Final 2 | 8.38 s



Failure: opening has 5 people; 4 required.

2. Wan 3.0 | Static Viewpoint Change

Valid: FAIL Count: N/A Position: N/A SR: FAIL

Opening 1 | 0.68 s Opening 2 | 2.02 s Final 1 | 6.50 s Final 2 | 8.17 s



Failure: opening has 4 people; 3 required.

3. MiniMax H3 Max | Static Viewpoint Change

Valid: FAIL Count: N/A Position: N/A SR: FAIL

Opening 1 | 0.58 s Opening 2 | 1.75 s Final 1 | 6.42 s Final 2 | 8.43 s



Failure: opening has 7 people; 4 required.

4. Kling 2.6 | Static Viewpoint Change

Valid: FAIL Count: N/A Position: N/A SR: FAIL

Opening 1 | 1.69 s Opening 2 | 5.06 s Final 1 | 7.95 s Final 2 | 9.34 s

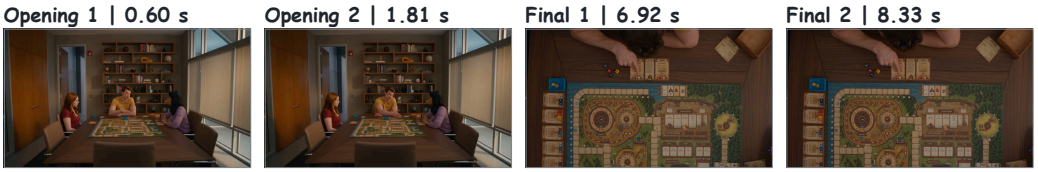


Failure: opening has 3 people; 4 required.

Figure 6. Invalid opening references across four models. Extra or missing participants violate the required opening cast, preventing reference-based evaluation. Count and Position are N/A, and SR fails.

1. Seedance 2.5 | Static Viewpoint Change

Valid: FAIL Count: N/A Position: N/A SR: FAIL



Failure: all 3 people should remain in place; the final crop hides two participants.

2. Wan 3.0 | Static Viewpoint Change

Valid: FAIL Count: N/A Position: N/A SR: FAIL



Failure: all 3 people should remain; the final crop excludes the blue participant.

3. Kling 3.0 | Static Viewpoint Change

Valid: FAIL Count: N/A Position: N/A SR: FAIL



Failure: all 3 people should remain; the final camera shows only the man in white.

4. Seedance 2.0 | Person Exit

Valid: FAIL Count: N/A Position: N/A SR: FAIL



Failure: white should exit, leaving 2 people; the final crop prevents counting.

Figure 7. Unobservable final states. The openings are valid, but final crops exclude participants or their places. Insufficient evidence makes Count and Position N/A; participants outside the image are not assumed to be absent.

1. Kling 3.0 | Static Viewpoint Change

Valid: PASS Count: FAIL Position: FAIL SR: FAIL



Failure: the same 4 people should remain; an extra blue participant makes 5.

2. Wan 3.0 | Person Exit

Valid: PASS Count: FAIL Position: FAIL SR: FAIL



Failure: yellow should exit, leaving 3 people; all 4 remain.

3. MiniMax H3 Max | Person Exit

Valid: PASS Count: FAIL Position: FAIL SR: FAIL



Failure: green should exit, leaving 3 people; all 4 remain.

4. Seedance 2.0 | Position Swap

Valid: PASS Count: FAIL Position: FAIL SR: FAIL



Failure: blue and yellow should swap, keeping 3 people; an extra purple makes 4.

Figure 8. Too many final participants. The examples include duplicated participants and people who remain despite an instructed exit. Each final readout exceeds the expected count after the requested update.

1. Kling 3.0 | Person Exit

Valid: PASS Count: FAIL Position: FAIL SR: FAIL



Failure: only blue should exit, leaving 3 people; yellow also disappears, leaving 2.

2. Seedance 2.5 | Person Entry

Valid: PASS Count: FAIL Position: FAIL SR: FAIL



Failure: purple should enter, making 3 people; only the original 2 remain.

3. Seedance 2.0 | Position Swap

Valid: PASS Count: FAIL Position: FAIL SR: FAIL



Failure: blue and yellow should swap, keeping 4 people; purple disappears, leaving 3.

4. Wan 3.0 | Person Entry

Valid: PASS Count: FAIL Position: FAIL SR: FAIL

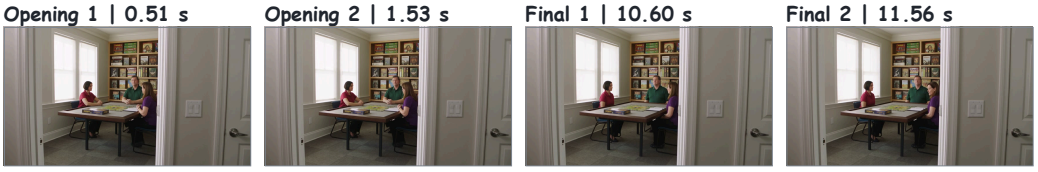


Failure: white should enter, making 4 people; only the original 3 remain.

Figure 9. Too few final participants. Unaffected people disappear after an exit or swap, or a required entrant is absent from the final state. The two final readouts contain fewer people than the requested update requires.

1. Seedance 2.5 | Position Swap

Valid: PASS Count: PASS Position: FAIL SR: FAIL



Failure: purple and red should exchange places; both stay in their opening places.

2. Wan 3.0 | Position Swap

Valid: PASS Count: PASS Position: FAIL SR: FAIL



Failure: green and orange should exchange places; both stay in their opening places.

3. Seedance 2.0 | Position Swap

Valid: PASS Count: PASS Position: FAIL SR: FAIL



Failure: blue and yellow should exchange places; both stay in their opening places.

4. Kling 2.6 | Position Swap

Valid: PASS Count: PASS Position: FAIL SR: FAIL



Failure: green and orange should exchange places; both stay in their opening places.

Figure 10. Requested swaps not executed. The final participant counts are correct, but the two selected identities retain their opening places. Both final readouts therefore fail Position and SR.

1. Kling 3.0 | Position Swap

Valid: PASS Count: PASS Position: FAIL SR: FAIL

Opening 1 | 0.89 s | Opening 2 | 2.66 s | Final 1 | 10.07 s | Final 2 | 11.39 s

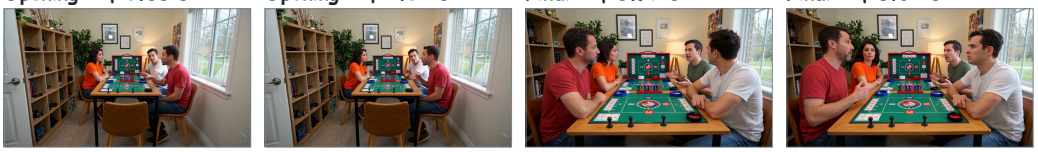


Failure: red and green should swap; one opening sofa is replaced by two armchairs.

2. Wan 3.0 | Person Entry

Valid: PASS Count: PASS Position: FAIL SR: FAIL

Opening 1 | 0.68 s | Opening 2 | 2.02 s | Final 1 | 6.97 s | Final 2 | 8.32 s



Failure: green should fill the empty end seat; that seat is removed and seating changes.

3. Seedance 2.0 | Static Viewpoint Change

Valid: PASS Count: PASS Position: FAIL SR: FAIL

Opening 1 | 0.50 s | Opening 2 | 1.50 s | Final 1 | 10.29 s | Final 2 | 11.46 s



Failure: all 3 people and the layout should stay fixed; the table shape changes.

4. Seedance 2.5 | Person Exit

Valid: PASS Count: PASS Position: FAIL SR: FAIL

Opening 1 | 0.72 s | Opening 2 | 2.16 s | Final 1 | 6.89 s | Final 2 | 8.32 s



Failure: white should exit and leave her chair fixed; the empty chair moves.

Figure 11. Physical-layout failures despite correct counts. Furniture replacement, removed or moved seats, and a changed table structure break correspondence with the opening places. Requested changes to participants should preserve this physical layout.