DUOSTATE: EVALUATING SELECTIVE STATE CONTROL IN VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Following a video script requires more than keeping characters recognisable: the intended person must change, the other must remain unchanged, and a later instruction may require restoring an earlier state. We introduce DUOSTATE, a framework for evaluating this selective state control. A paired protocol tests four branches of the same two-character scene: no edit, an edit to either character, and an edit followed by restoration after occlusion. A region-grounded evidence-to-ledger judge tracks the characters, records visible attributes from timestamped regions, and checks their state changes against the script. It distinguishes recognition of the requested states from completion of the full process in the prescribed order and time windows. On an independent diagnostic set, regional evidence and ledger-based rules improve success-detection F1 from 0.653 to 0.869 over direct full-frame judging. Strict verification accepts only 5.0% of the benchmark's original reversal clips. Providing enough time for every stage and matching edits across roles improves some results, but state control remains difficult. DUOSTATE connects these failures to specific characters and events, establishing an evaluation target beyond visual consistency.

1 INTRODUCTION

Given a reference image of two people, a video generator is asked to follow a character-specific script: a florist removes her glasses, passes behind a delivery van, and reappears wearing them, while the courier beside her keeps his helmet throughout. Generating this scene requires both stability and change: preserve who each character is, update only the addressed character, and restore the earlier state after occlusion. A video that leaves both characters unchanged can retain identity perfectly while missing the entire script.

This is a problem of *selective state control*: change the intended person's attribute, leave the other person's state alone, and keep both identities recognisable. Video quality, character consistency, and compositional binding each capture part of this problem (Huang et al., 2023b; Zheng et al., 2025; Wu et al., 2025; Sun et al., 2024). But an identity score alone cannot tell whether the requested edit happened, happened to the wrong person, or was followed by the required restoration. Nor does a correct final appearance establish that the intervening change ever occurred.

DUOSTATE organises the evaluation around the same scene and character pair. The null branch asks neither person to change; A_int and B_int request a change to one person at a time; reversal asks A to change and then restore the earlier state after occlusion. This shared context gives each branch a specific diagnostic role: test drift without an edit, execute the requested edit, address the correct person, and restore the state. We require all applicable conditions to hold in the evaluated clip, including execution in reversal itself. A frozen counterpart tests whether an unchanged video can incorrectly receive update credit.

Evaluating these branches also requires resolving small, person-specific changes over time. A whole-frame description can miss an accessory; an isolated close-up can lose the character's identity; a correct final frame can conceal a missing intermediate edit. Our *region-grounded evidence-to-ledger judge* addresses these problems within one pipeline. It binds regions to character references, records visible states with timestamps and full-frame context, and evaluates the resulting trajectories with explicit branch rules (Figure 4). Each decision is therefore linked to a person, an observation, and an event in the script.

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

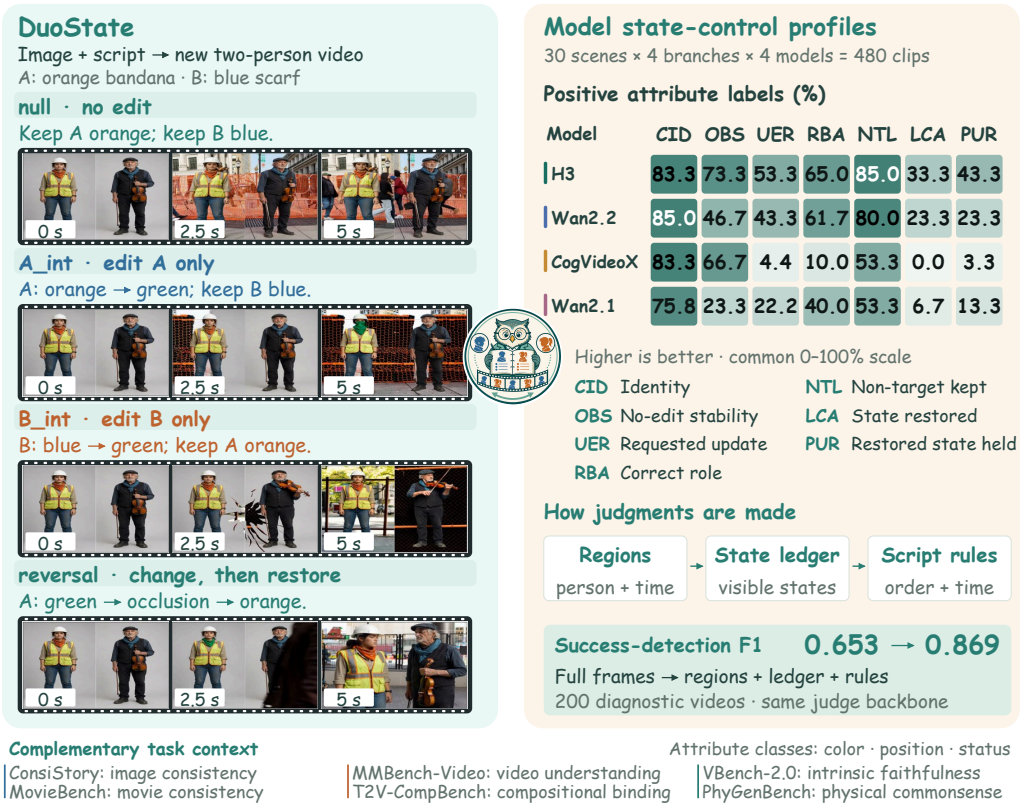


Figure 1: Overview of DUOSTATE. Given a two-person image and a script, a generator must change only the target's state and restore it when requested. Left: q015 requests and Wan2.2 outputs; A_int/B_int update only A/B, null preserves both, and reversal adds post-occlusion restoration (Section 3.1). Right: seven attribute-level positive-label rates; H3 denotes MiniMax-H3. Identity scores remain high while execution and restoration vary substantially across models. CID: Character Identity Consistency; OBS: Off-Branch Stability; UER: Update Execution Rate; RBA: Role-Bound Attribution; NTL: Non-Target Leakage check (1 means preserved); LCA: Legitimate Change Adoption (restoration); PUR: Persistence Under Reappearance. Table 1 defines their conditions and applicable branches. Per-model denominators are $120, 30, 90, 60, 60, 30, 30$ in this order; darker cells indicate higher rates. Local observations feed a state ledger and script checks; F1 compares full-frame and complete-pipeline judging on 200 diagnostic videos (Table 3).

We make three contributions. **(1) A paired benchmark for selective state control** combines no-edit, A-target, B-target, and restoration scripts under shared scene and character references. **(2) An evidence-based evaluator** links local states to character identity and time before checking the script; controlled comparisons show fewer false accepts and missed successes (Section 4). **(3) Paired comparisons and process diagnostics** reveal differences beyond final pass rates: matching edit requirements narrows A/B gaps with little change in overall execution, while equal five-stage pass counts conceal different intermediate progress (Section 6).

2 RELATED WORK

VBench and VBench-2.0 assess video quality and faithfulness, including controllability (Huang et al., 2023b; Zheng et al., 2025); MovieBench emphasises character consistency across scenes (Wu et al., 2025). T2V-CompBench tests dynamic attribute binding, role-associated actions, and contrastive left/right prompts (Sun et al., 2024). DUOSTATE evaluates selective change in a shared-scene, two-character unit: no-edit stability, an edit to either person, and post-occlusion restoration. The paired branches separate drift, execution, attribution, and restoration. ConsiStory (Tewel et al.,

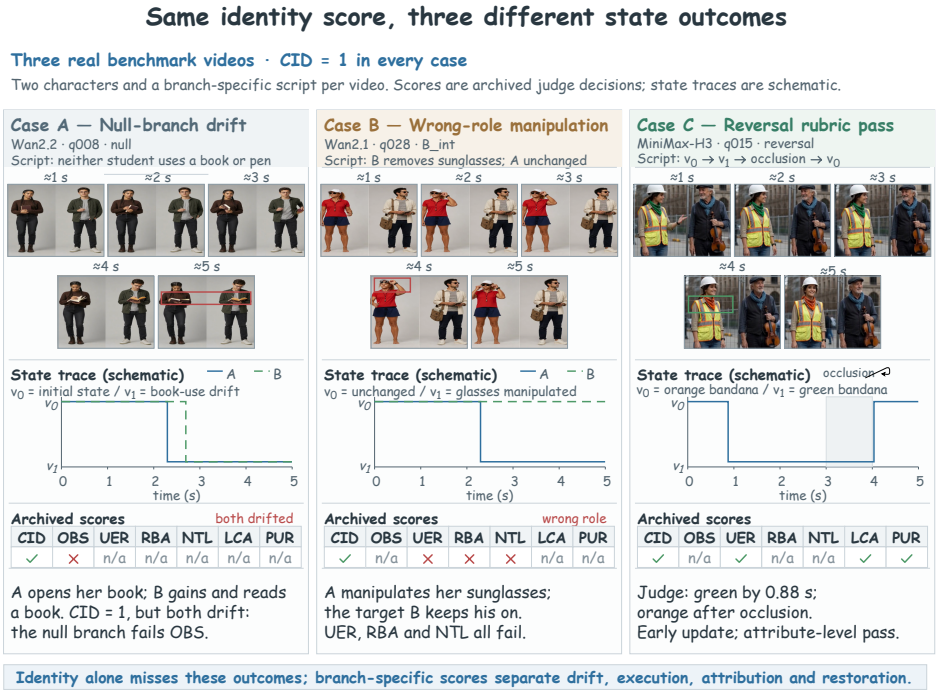


Figure 2: Three benchmark videos receive the same identity score (CID $= 1$), yet exhibit null-branch drift, wrong-role manipulation, and an attribute-level reversal pass. Scores are automatic branch judgments; traces are schematic.

2024), PhyGenBench (Meng et al., 2024), and MMBench-Video (Fang et al., 2024) address complementary subject-consistency, physical-reasoning, and video-understanding problems. Figure 1 and Appendix A expand the protocol and task distinctions.

TIFA (Hu et al., 2023) and DSG (Cho et al., 2024) decompose image faithfulness into checkable questions; VideoArgus (Zeng et al., 2026) uses sample-specific rubrics and visual tools for video evaluation. Our judge connects character-indexed state observations to fixed branch rules for selective edits and ordered restoration.

3 Benchmark and Evaluation Protocol

DuoState contains 30 two-character scenes, ten each for color, position, and status changes. Each provides a reference image and four scripts for generating new videos, yielding 480 outputs across four models. Each scene contains a target character and a non-target character, allowing requested changes and unintended cross-character changes to be evaluated in the same clip. We use two characters to keep this target–non-target contrast explicit under matched scene conditions.

3.1 Matched four-branch design

A quadruple contains distinguishable characters $\mathcal{A}, \mathcal{B}$, scene s, attribute class $a \in \{\text{color}, \text{position}, \text{status}\}$, and character-specific initial and requested states $(v_{0,c}, v_{1,c})$. Its branches share the reference pair, scene, and generation seed. The **null** branch preserves both initial states; **A_int** updates only $\mathcal{A}$; **B_int** applies the paired same-class edit to $\mathcal{B}$; **reversal** adds occlusion and restoration to A_int. Figure 3 gives a concrete instance, and Figure 9 shows all four generated branches.

Matched context holds scene and attribute class fixed while allowing character-specific objects and initial values. Thus glasses removal and helmet removal are paired accessory-status edits, with potentially different difficulty. The standard six-second script requests change at 2–3s, occlusion

Table 1: Binary attribute-level judge labels: 1 denotes a positive assessment, including for NTL. Rates use applicable branches. Timing is requested in the rubric; strict process verification separately checks event evidence against each script. RDR reports paired-update coverage.

Dim.	Name	Branches	Semantics
CID	Character Identity Consistency	all 4	identity $(\mathcal{A}, \mathcal{B})$ preserved across frames
OBS	Off-Branch Stability	null	no unintended drift under null intervention
UER	Update Execution Rate	A_int/B_int/rev	VLM judges whether the intended character updates as requested, including timing
RBA	Role-Bound Attribution	A_int/B_int	change occurs *only* on the target character
NTL	Non-Target Leakage check (↑)	A_int/B_int	1 means no drift of the untargeted character's baseline attribute
LCA	Legitimate Change Adoption	reversal	post-occlusion frames show attribute returned to v_0
PUR	Persistence Under Reappearance	reversal	the reverted value persists after occlusion; no object-physics guarantee
RDR	Reversal Definition Rate	derived	fraction of scenes whose paired A_int clip executes the update

at 3–4s, and restoration at 4–5s. Each item's stored prompt specifies its authoritative schedule, including any item-specific timing.

3.2 Attribute-level joint diagnostic

Table 1 defines seven binary judge labels. Identity failure sets $\mathrm{RBA} = \mathrm{NTL} = \mathrm{LCA} = \mathrm{PUR} = 0$, because role-bound state credit requires a stable character referent. Their attribute-level conjunction is

$$P_{q,b} = \begin{cases} \mathrm{CID}\,\mathrm{OBS}, & b = \mathrm{null}, \\ \mathrm{CID}\,\mathrm{UER}\,\mathrm{RBA}\,\mathrm{NTL}, & b \in \{\mathrm{A_int}, \mathrm{B_int}\}, \\ \mathrm{CID}\,\mathrm{UER}\,\mathrm{LCA}\,\mathrm{PUR}, & b = \mathrm{reversal}. \end{cases} \tag{1}$$

Each factor belongs to the same video (q, b). Equation 1 combines attribute assessments, including the VLM's judgment of requested timing. **Strict reversal success** additionally requires evidence of the complete sequence: correct initial states, a completed target update, scripted occlusion, restoration, and subsequent stability. The intended character must perform the change, the non-target must retain its state, and event order and completion times must satisfy the individual script. Starting an action does not establish completion of the requested state change. Appendix H.2 gives the verification rules.

3.3 Static controls and paired diagnostics

A frozen reversal repeats its first frame, preserving identity and the initial state but making no update. Its $\mathrm{UER} = 0$ suffices for Equation 1 to reject it. All 120 frozen controls receive zero execution and joint passes, despite $\mathrm{CID} = 96.7$–100%. An auxiliary paired analysis reports A_int update coverage and reversal scores on the eligible scenes (Appendices A.4 and G.2). Since A_int and reversal are separate generations, a reversal can pass even when its paired A_int does not.

4 Region-Grounded Evidence-to-Ledger Judge

From regions to state evidence. Our judge turns the protocol into an evidence-based evaluation of character states (Figure 4). Given a clip, character references, and branch script, it binds tracks to A/B and rechecks identity after occlusion. We sample 24 frames approximately uniformly across the

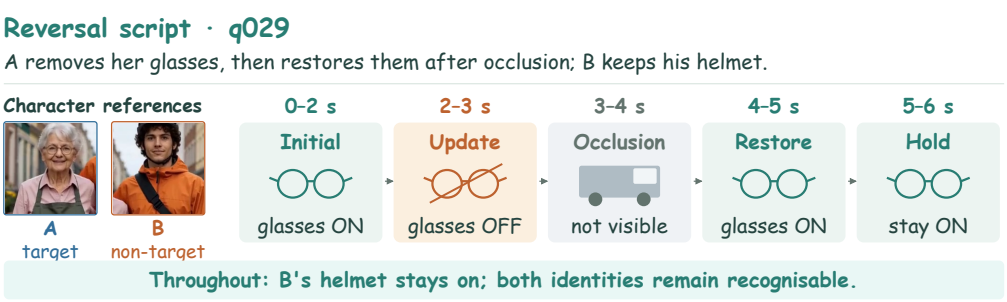


Figure 3: Requested reversal sequence for q029. A removes her glasses, passes behind an occluder, puts them back on, and keeps them on; B's helmet and both identities remain unchanged. Figure 1 shows the four paired branches; Table 1 defines the attribute-level labels.

Table 2: Human agreement on attribute decisions. Judge–majority compares our judge with three-rater majority labels; H1–H2 and Fleiss κ measure human agreement. Each dimension uses its applicable, valid observations.

Dimension	Valid n	Judge–majority	H1–H2	Fleiss κ
Identity (CID)	179	96.1%	83.2%	0.629
Execution (UER)	179	93.3%	89.4%	0.694
Role attribution (RBA)	60	93.3%	80.0%	0.596
Restoration (LCA)	120	80.0%	95.8%	0.876
Persistence (PUR)	120	91.7%	95.8%	0.866

clip, including its first and last frames, and attach their video timestamps. Text-conditioned grounding locates faces, hair, clothing, eyewear, and task-relevant hands; Grounding DINO (Liu et al., 2023a) and SAM 2 (Ravi et al., 2024) provide box and mask-propagation components. Timestamped crops retain their full-frame context. Our vision–language model (VLM), GPT-5.4-Priority, first describes visible states without assigning success, then assesses the full observations against the branch script, producing attribute labels, evidence timestamps, and reasons.

Ledger-conditioned temporal rules. Each observation is represented as $\ell_{c,t,r} = (\hat{s}, v, b, e)$: estimated state, visibility, bounding box, and evidence identifier for character c, time t, and region r. The ledger records observed rather than requested states, with unknown observations retained explicitly. The VLM supplies perceptual evidence; deterministic code applies the identity guard, branch conjunctions, and temporal rules. Strict verification checks event completion and order against each script using the criteria in Section 3; human review uses those same criteria. Appendices B.1 and M detail the observation and rule interfaces.

Human agreement. Three independent blinded raters annotate 120 reversal and 60 intervention clips, with a majority formed separately for each applicable dimension. Against these majorities, agreement is 93.3% for execution and attribution and 80.0% for restoration (Table 2). Appendix B.4 gives the per-rater comparisons and applicable sample counts.

Success detection. Table 3 compares three judges on the same 200 diagnostic videos using a fixed VLM, common human labels, and comparable evaluation budgets. Regional crops reduce false accepts from 16 to 12 and missed successes from 18 to 11; adding the ledger and process rules reduces them further to 6 and 7. The combined ledger–rule layer thus improves both recognition of complete successes and rejection of incomplete sequences. Precision and recall refer to this diagnostic cohort's 50 successes and 150 failures.

Transfer across evaluators. Table 4 compares direct scoring with observation-first assessment and temporal rules on valid, paired attribute labels from the three-rater audit. Each alternative backend receives the same 10 sampled frames in both configurations; structured assessment uses separate observation and assessment calls. Cohen's κ measures agreement with human majorities

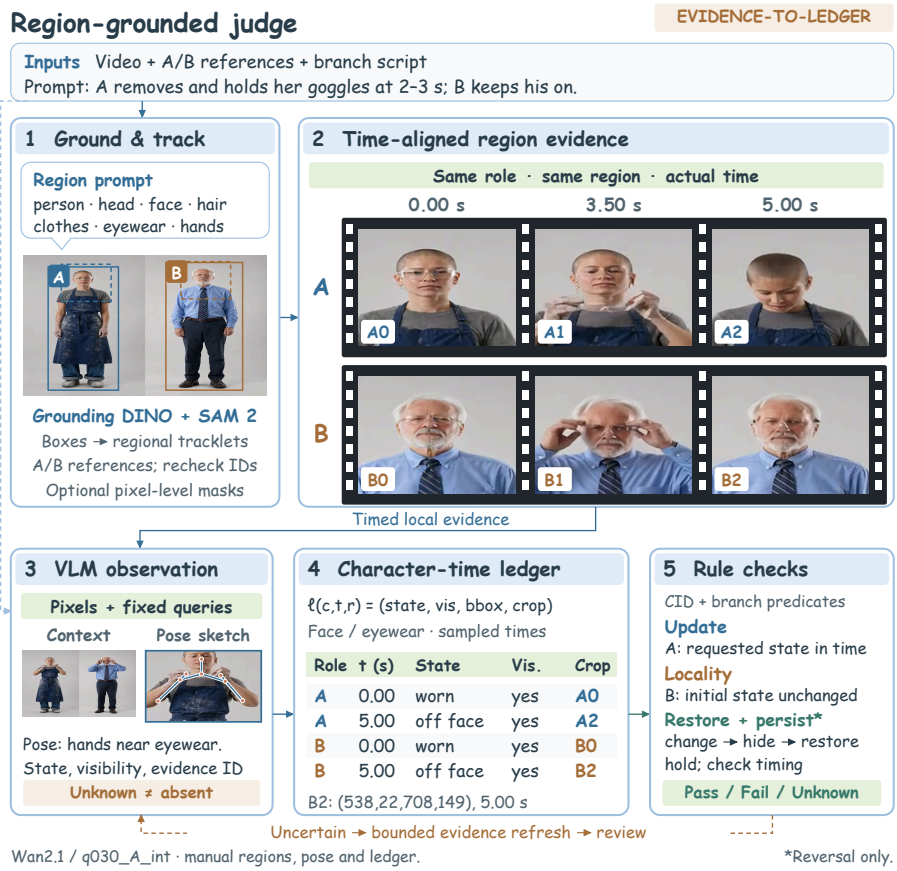


Figure 4: Region-grounded evidence-to-ledger judge. Reference-bound tracks link local observations to role, region, and time; branch rules consume this evidence to check execution, locality, and ordered restoration. Genuine q030 A_int frames illustrate the evidence flow. Boxes, pose cues, and ledger entries are explanatory annotations, not raw detector or VLM outputs; the restoration rule applies only to reversal.

Table 3: Controlled judge comparison on 200 diagnostic videos, with the same VLM and human reference (50 successes, 150 failures) and comparable evaluation budgets. TP: correctly accepted successes; FP: falsely accepted failures; FN: missed successes; TN: correctly rejected failures. Appendix B.5 describes the configurations.

Judge configuration	TP	FP	FN	TN	Precision	Recall	F1
Full frames	32	16	18	134	66.7%	64.0%	0.653
Regional crops	39	12	11	138	76.5%	78.0%	0.772
Crops + ledger + rules	43	6	7	144	87.8%	86.0%	0.869

beyond chance. Qwen3-VL and Claude improve by 0.180 and 0.085, respectively; Gemini changes little, and Qwen3.5-plus decreases on its smaller paired subset. The benefit therefore extends beyond GPT but varies across backends. These within-backend comparisons evaluate attribute agreement; the following studies examine temporal localization and script completion. Appendix B.6 gives configurations and output coverage.

Temporal localization. Median human–judge absolute differences are 0.11s for update onset, 0.08s for occlusion start, 0.10s for occlusion end, and 0.27s for restoration, on 78, 108, 107, and 88 comparable videos. Appendix H reports event coverage and human–human comparisons.

Table 4: Cross-VLM attribute agreement on matched valid labels. Direct and structured configurations use the same labels within each row. n_v counts contributing videos and n_ℓ counts video–dimension labels. Intervals are paired scene-bootstrap 95% intervals for $\Delta\kappa$, computed before rounding. The GPT reference uses the full audit label set.

Backend	n_v	n_ℓ	Direct κ	Structured κ	$\Delta\kappa$	95% interval
GPT-5.4-Priority (reference)	180	719	—	0.486	—	—
Full paired coverage						
Qwen3-VL-235B	180	719	0.289	0.469	+0.180	[0.114, 0.247]
GPT-4o	180	719	0.264	0.429	+0.164	[0.080, 0.248]
GPT-4.1	180	719	0.279	0.371	+0.092	[0.019, 0.154]
Qwen-VL-Plus	180	719	0.146	0.300	+0.155	[0.063, 0.236]
Partial paired coverage						
Claude 4.5 Sonnet	179	715	0.375	0.460	+0.085	[0.024, 0.148]
Gemini 2.5 Flash	174	693	0.413	0.415	+0.002	[−0.076, 0.078]
GLM-5.3-flashx	39	156	0.241	0.408	+0.166	[−0.016, 0.343]
Qwen3.5-plus	33	129	0.062	0.000	−0.062	[−0.154, −0.006]

Joint success and lifecycle validity. Attribute agreement asks whether the requested states are recognised; strict success asks whether the script is completed in order and on time. Under Equation 1, judge and human accept the same eight videos among 119 valid pairs (Table 21). Script-level re-adjudication retains six of these: five H3 and one Wan2.2 clip. No additional original clip passes, giving $6/120$ strict successes (Table 22).

Script completion and physical plausibility are distinct: Appendix C reports physical anomalies in generated clips. A continuity-proxy analysis further distinguishes visual smoothness from script execution (Appendix E.1).

5 Evaluation Setup

The main evaluation comprises ten color, ten position, and ten status scenes, each evaluated under four branches on MiniMax-H3 (MiniMax AI, 2026), Wan2.2-I2V-A14B and Wan2.1-I2V-14B-480p (Wan Team, 2025), and CogVideoX-5B-I2V (Yang et al., 2024): $30 \times 4 \times 4 = 480$ videos. Synthesised portraits form paired character references. We report attribute-level branch conjunctions, paired-update coverage, and strict reversal verification. Appendix F gives generation settings; Appendix G.4 evaluates full-duration and symmetric-edit controls.

6 Results

6.1 Does preserved identity imply selective control?

Figure 5 visualises attribute scores and reversal progress; Table 5 reports joint branch passes. H3 and CogVideoX pass 70.0% and 66.7% of null clips, but 46.7% and 6.7% of A_int clips. Across A_int/B_int, CogVideoX receives $58/60$ identity passes, $4/60$ executed updates, and only $2/60$ joint passes. High identity thus coexists with failure to carry out the addressed edit.

6.2 How do duration and edit requirements affect the results?

Table 6 separates temporal coverage from A/B edit requirements. Strict verification retains $6/120$ original reversals (5.0%); full six-second clips raise this to $11/120$ (9.2%). The full-duration script allocates time to initial states, update, occlusion, restoration, and subsequent stability (Appendix G.4). Duration affects completion, although most reversals still do not complete the script.

At full duration, matching A/B edit requirements reduces the mean absolute branch gap from 13.3 to 5.0 percentage points, while their pooled pass rate changes only from 30.8% to 31.7%. Some branch counts rise while others fall: a smaller A/B gap need not imply better overall execution. The

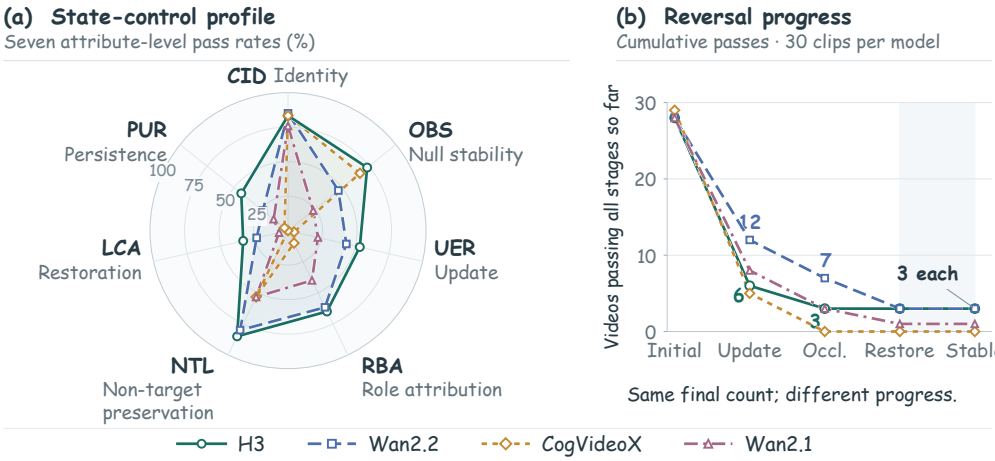


Figure 5: Model profiles and process diagnostics. (a) Positive-label rates for the seven dimensions in Table 1, using applicable branches: per model, $n = 120$ for CID, 90 for UER, 60 for RBA/NTL, and 30 for OBS/LCA/PUR. (b) Auxiliary five-stage cumulative passes (Table 26), requiring every preceding stage. H3 and Wan2.2 reach the same final count with different intermediate progress.

Table 5: Attribute-level joint pass rates (%) under Equation 1, with 95% Wilson intervals and 30 quadruples per cell. Strict reversal rates from script-level re-adjudication appear below and in Table 22.

Model	null	A_int	B_int	reversal
H3[†]	**70.0** [52.1, 83.3]	**46.7** [30.2, 63.9]	**60.0** [42.3, 75.4]	**20.0** [9.5, 37.3]
Wan2.2	43.3 [27.4, 60.8]	43.3 [27.4, 60.8]	33.3 [19.2, 51.2]	6.7 [1.8, 21.3]
CogVideoX	66.7 [48.8, 80.8]	6.7 [1.8, 21.3]	0.0 [0.0, 11.4]	0.0 [0.0, 11.4]
Wan2.1	23.3 [11.8, 40.9]	26.7 [14.2, 44.4]	10.0 [3.5, 25.6]	0.0 [0.0, 11.4]

Strict reversal: H3 5/30 (16.7%), Wan2.2 1/30 (3.3%), others 0/30.
[†] MiniMax-H3 uses a 512×512 canvas; settings are in Section 5.

paired design makes this distinction visible by reporting both the branch contrast and the combined outcome.

6.3 Equal final counts, different progress

In the auxiliary five-stage diagnostic, H3 and Wan2.2 each pass all stages on 3/30 reversals. Their cumulative counts after initial-state and update checks are 6/30 versus 12/30, and after occlusion 3/30 versus 7/30 (Figure 5b; Table 26). More early progress thus need not yield more complete sequences. Across models, 82/120 clips first fail at update; only 13/120 pass the first three stages. The first failure is the earliest unmet criterion in stage order. These diagnostics use the common onset windows in Appendix H, while strict success checks completed states against each video's script. A final pass rate conceals where execution stops.

6.4 Which attribute changes are hardest?

The benchmark has ten quadruples per attribute class. Table 7 pools only the three role-addressed branches (A_int, B_int, reversal), giving 30 videos per model and class. Across models, raw pass is $30/120 = 25.0\%$ for color, $14/120 = 11.7\%$ for position, and $32/120 = 26.7\%$ for status. Position has the lowest pooled score, although it is not the lowest class for every model. These are descriptive class differences; correlated branches and ten scenes per class limit population-level inference.

At scene level, a model is credited when at least two of its three role-addressed branches pass. No model meets this criterion on eight of ten position scenes, versus two of ten in each other class

Table 6: Duration and edit controls. The first two columns report strict reversal passes. The last two compare A/B intervention passes at full duration under original and symmetric edits. Entries are passed/total videos; each A/B pair lists the two branches separately.

	Strict reversal		Full-duration A / B	
Model	Original clips	Full duration	Original edits	Symmetric edits
MiniMax-H3	5/30	5/30	16/30, 21/30	18/30, 19/30
Wan2.2	1/30	3/30	13/30, 9/30	12/30, 10/30
CogVideoX	0/30	1/30	3/30, 1/30	3/30, 2/30
Wan2.1	0/30	2/30	8/30, 3/30	7/30, 5/30

Table 7: Attribute-class breakdown on the same benchmark population as Table 5. Numerators are raw joint passes across A_int, B_int, and reversal; null is excluded. Each denominator contains ten quadruples with three correlated branch observations each.

Model	Color	Position	Status
MiniMax-H3	16/30 (53.3%)	5/30 (16.7%)	17/30 (56.7%)
Wan2.2	12/30 (40.0%)	5/30 (16.7%)	8/30 (26.7%)
CogVideoX	1/30 (3.3%)	0/30 (0.0%)	1/30 (3.3%)
Wan2.1-480p	1/30 (3.3%)	4/30 (13.3%)	6/30 (20.0%)

(Appendix G.7). This groups the correlated branches into a complementary scene-level diagnostic. Overall, 12/30 scenes have no credited model.

7 Discussion and Conclusion

DuoState evaluates whether a generated video follows a character-specific script: the target changes, the other person remains unchanged, and restoration occurs when requested. Its paired design and evidence-to-ledger judge connect these requirements to observable character states and script-level verification.

The paired comparisons and process records reveal differences beyond final scores. Matching A/B edits narrows branch gaps with little change in pooled execution, while the auxiliary five-stage diagnostic distinguishes intermediate progress even when final pass counts coincide. Regional evidence and ledger-based rules reduce both false accepts and missed successes. Full-duration controls improve strict reversal completion, but most scripts remain incomplete under the tested configurations.

Limitations and next steps. The present evaluation covers two-character scenes; extending the reference binding and non-target checks to larger character sets requires additional validation. Appendix K discusses longer sequences, generation settings, and coupled localization components.

AI use statement

Appendix L discloses AI use in data creation.

Ethics statement

Reference portraits are synthetic. Appendix L discusses annotation privacy and responsible release.

Reproducibility statement

Appendix L details seeds, verdicts, controls, annotations, and analysis code. Statistical recomputation uses these records; perceptual re-evaluation requires original video evidence.

REFERENCES

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

Emanuele Bugliarello, H. Hernan Moraldo, Ruben Villegas, Mohammad Babaeizadeh, Mohammad Taghi Saffar, Han Zhang, Dumitru Erhan, Vittorio Ferrari, Pieter-Jan Kindermans, and Paul Voigtlaender. StoryBench: A multifaceted benchmark for continuous story visualization. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. TemporalBench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024.

Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model. In *European Conference on Computer Vision*, 2022.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In *International Conference on Learning Representations*, 2024.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. MMBench-Video: A long-form multi-shot benchmark for holistic video understanding. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024. `https://arxiv.org/abs/2406.14515`.

Alvan R Feinstein and Domenic V Cicchetti. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549, 1990.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. doi: 10.1037/h0031619.

Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008. doi: 10.1348/000711006X126600.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123, 2024. doi: 10.18653/v1/2024.emnlp-main.127.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023a.

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. *arXiv preprint arXiv:2311.17982*, 2023b.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

Mingxiang Liao, Hannan Lu, Xinyu Zhang, Fang Wan, Tianyu Wang, Yuzhong Zhao, Wangmeng Zuo, Qixiang Ye, and Jingdong Wang. Evaluation of text-to-video generation models: A dynamics perspective. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, 2024.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023a.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22139–22149, 2024.

Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. FETV: A benchmark for fine-grained evaluation of open-domain text-to-video generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996.

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.

MiniMax AI. MiniMax-H3: A general-purpose omni-modal video–audio generative system, 2026. Open-source model weights and technical report, `https://huggingface.co/MiniMaxAI/MiniMax-H3`.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition, 2009.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. *arXiv preprint arXiv:2407.14505*, 2024.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via Python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11888–11898, 2023.

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (SIGGRAPH)*, 2024. https://arxiv.org/abs/2402.03286.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

Nahathai Wongpakaran, Tinakon Wongpakaran, Danny Wedding, and Kilem L Gwet. A comparison of Cohen's kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. *BMC Medical Research Methodology*, 13(1):61, 2013.

Weijia Wu, Mingyu Liu, Zeyu Zhu, Xi Xia, Haoen Feng, Wen Wang, Kevin Qinghong Lin, Chunhua Shen, and Mike Zheng Shou. MovieBench: A hierarchical movie level dataset for long video generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. https://arxiv.org/abs/2411.15262.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. Accepted at ICLR 2025.

Ziyun Zeng, Zixuan Wang, Yongsheng Yu, Hang Hua, and Jiebo Luo. VideoArgus: Agentic rubric-grounded unified evaluation for video generation and editing. *arXiv preprint arXiv:2608.05485*, 2026.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. StoryDiffusion: Consistent self-attention for long-range image and video generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

A PROTOCOL DETAILS AND EXTENDED POSITIONING

Meaning of intervention. A *role-addressed intervention* changes the branch instruction while retaining the scene and character references. This is an intervention on the evaluation prompt: it tests the generator's response to a specified edit under matched context. The interpretation concerns observable behavior, rather than identification of internal causal variables or counterfactual mechanisms (Pearl, 2009).

A.1 COMPREHENSIVE VIDEO-QUALITY BENCHMARKS

Recent large-scale evaluation suites operate at the granularity of the whole video: VBench (Huang et al., 2023b) dissects video-generation quality into 16 hierarchical dimensions, including subject consistency and motion smoothness, and provides human-preference calibration per dimension. Its successor VBench-2.0 (Zheng et al., 2025) extends coverage to intrinsic faithfulness dimensions such as physics, commonsense, and controllability. Both are complementary to our contribution: they aggregate over whole videos with prompt-driven evaluation, while our unit of analysis is a matched *quadruple* of parallel branches on the *same* scene, per named character.

EvalCrafter (Liu et al., 2024) combines visual, content, motion, and alignment metrics with human-opinion calibration. FETV (Liu et al., 2023b) organises prompts by content, controllable attributes, and complexity, including temporal categories. DEVIL (Liao et al., 2024) examines motion dynamics and their relation to quality scores, while VideoScore (He et al., 2024) learns multi-aspect video assessment from human feedback. These works motivate reporting several dimensions and validating automatic judgments against people; our branch profiles ask how those judgments change under matched character-specific instructions.

A.2 ADJACENT BENCHMARKS AND THEIR POSITIONING

The footer of Figure 1 locates six adjacent works. **MovieBench** (Wu et al., 2025) provides movie-level, multi-scene data with character identities and shot descriptions. **T2V-CompBench** (Sun et al., 2024) tests dynamic attribute changes, multi-object action binding, and contrastive spatial prompts. Building on this distinction between consistency and change, DUOSTATE organises no-edit, A-target, B-target, and restoration branches around a shared character pair and scene. This joint structure supplies a common context for examining whether an edit occurs, who receives it, and whether the earlier state is restored. **ConsiStory** (Tewel et al., 2024) studies consistent subjects in image generation; **PhyGenBench** (Meng et al., 2024) evaluates physical commonsense; **MMBench-Video** (Fang et al., 2024) evaluates video understanding. Together, these works cover complementary questions of consistency, compositionality, physical plausibility, and video understanding.

T2I-CompBench (Huang et al., 2023a) separates attribute binding from object relationships in generated images. StoryBench (Bugliarello et al., 2023) evaluates action execution, story continuation, and story generation; StoryDiffusion (Zhou et al., 2024) develops consistent image sequences and video transitions. Our paired protocol concentrates on selective changes within a shared two-character scene. VideoPhy (Bansal et al., 2024) tests physical commonsense in material interactions, complementing our character-state criteria.

On the evaluator side, MVBench (Li et al., 2024) tests video understanding that requires more than a single frame, and TemporalBench (Cai et al., 2024) examines fine-grained motion and event order. These tasks motivate checking temporal evidence explicitly. In DUOSTATE, that evidence must establish the requested sequence for the specified character.

A.3 VIDEO GENERATION MODELS USED AS BASELINES

We evaluate publicly released I2V baselines: CogVideoX-5B (Yang et al., 2024), HunyuanVideo-I2V (Kong et al., 2024), the Wan family (Wan Team, 2025), and MiniMax-H3 (MiniMax AI, 2026). The main evaluation includes four models; HunyuanVideo and Wan2.1-720p are restricted to supplementary exploratory configurations. The comparison measures performance under the specified generation settings, which vary alongside model architecture.

A.4 PAIRED-UPDATE COVERAGE AND CONDITIONAL SCORING

Raw attribute-level reversal averages $P_{q,\text{reversal}}$ over all N scored quadruples. An auxiliary analysis selects scenes where the separate A_int clip receives update credit. We report this conditional attribute score together with its eligible fraction, RDR:

$$G_q = \mathbf{1}\{\text{UER}(q, \text{A_int}) = 1\}, \qquad N_G = \sum_q G_q,$$
$$\text{Pass}_{\text{cond}} = \frac{\sum_q G_q P_{q,\text{reversal}}}{N_G}, \qquad \text{RDR} = \frac{N_G}{N}. \tag{2}$$

When $N_G = 0$, the conditional rate is N/A. Ineligible reversals are excluded from this denominator; eligible clips still require their own four reversal labels. The gate uses another clip, so it selects scenes rather than establishing update and restoration within the reversal itself. Equation 1 already rejects frozen reversals through their zero execution label. Appendix G.2 relates the conditional and full-population counts.

A.5 EVIDENCE-BASED EVALUATION AND HUMAN VALIDATION

The evidence-to-ledger judge (Section 4) associates local observations with character and time before applying branch-specific rules. Grounding and visual observation establish the evidence; ledger queries determine whether it satisfies the script. We examine this evaluator through human agreement on attributes, localization of event times, and recognition of complete lifecycles. Appendices B.4, H.

TIFA (Hu et al., 2023) assesses image faithfulness through visual questions, and DSG (Cho et al., 2024) adds dependencies to keep their answers consistent. VQAScore (Lin et al., 2024) uses a visual question-answering model to score text–visual alignment. ViperGPT (Surís et al., 2023) separates visual processing from reasoning through executable programs. For generated and edited videos, VideoArgus (Zeng et al., 2026) gathers tool-based evidence under sample-specific rubrics. Our protocol fixes the branch criteria and uses a character–time ledger to check their temporal conditions, making the unit of judgment an observed state trajectory under a matched instruction.

A.6 BRANCH SCRIPTS AND CONSTRUCTION

Each quadruple consists of two visually distinguishable characters $\mathcal{A}$ and $\mathcal{B}$, a shared scene s, an attribute class $a \in \{\text{color}, \text{position}, \text{status}\}$, and character-specific initial and requested states $(v_{0,c}, v_{1,c})$. A side-by-side reference image specifies the character pair. We abbreviate the target's states as v_0, v_1 when its identity is clear. The four branch scripts are:

- **null:** no intervention; both characters retain their own initial states.
- **A_int:** $\mathcal{A}$ changes from $v_{0,\mathcal{A}}$ to $v_{1,\mathcal{A}}$; $\mathcal{B}$ is unchanged.
- **B_int:** the paired same-class intervention on $\mathcal{B}$; $\mathcal{A}$ is unchanged.
- **reversal:** the A_int change is followed by an occlusion and restoration of $\mathcal{A}$'s initial state.

The standard six-second schedule requests a change at 2–3s, occlusion at 3–4s, and restoration at 4–5s. Some items use different schedules; the stored branch prompt is authoritative. The branches share a scene, reference pair, and generation seed. *Matched* denotes this shared context and attribute class: the accessory, initial colour, or concrete operation may differ between roles. Accordingly, the contrast diagnoses role addressing while retaining differences in concrete edit difficulty. RBA is scored per intervention video; the paired branch supplies an additional control.

A.7 DATA CONSTRUCTION PIPELINE

The main seed set contains 10 color, 10 position, and 10 status quadruples. Character reference portraits are generated by an image model and concatenated into a two-character reference image. Each branch script is run under the model-specific configuration in Appendix F. Human validation comprises an outcome-stratified audit, a three-rater attribute study, and a process study of event times and complete lifecycles.

Table 8: Evidence-to-ledger judge agreement with attribute annotations. Each row reports the applicable valid observations; agreement is the binary-label match rate.

Dimension	n	Agreement
Character identity (CID)	179	96.1%
Update execution (UER)	179	93.3%
Role attribution (RBA)	60	93.3%
Non-target leakage (NTL)	60	95.0%
Restoration (LCA)	120	80.0%
Persistence (PUR)	120	91.7%

B Human agreement and annotation sensitivity

B.1 Visual observation and attribute assessment

The attribute assessment uses GPT-5.4-Priority with 24 timestamped frames spanning each clip. For a clip of T frames, the sampled indices are $\min(\lfloor Ti/23 \rfloor, T-1)$ for $i = 0, \ldots, 23$; timestamps are computed from the frame rate. A describe–then–judge procedure separates visual description from rubric assessment: the observation step describes visible characters and state changes; the assessment step receives the full description and branch script and produces applicable labels, evidence timestamps, and reasons. Each call has a 3,000-token output cap. Deterministic code applies the identity guard and branch conjunction. Regional observations support character-specific inspection, while temporal queries examine event order and timing. The human studies below evaluate attribute decisions.

B.2 Character-state records and evidence links

The judge records per-dimension labels, confidence, evidence timestamps, and reasons, and links decisions to character–time observations. A semantic view of the character-state ledger is

$$\ell(q, b, t) = \{c : \{\text{attr}, \text{pos}, \text{action}\}\}_{c \in \{\mathcal{A}, \mathcal{B}\}} \cup \{\text{occlusion}\}.$$

Appendix M specifies the region, visibility, bounding-box, and crop fields that make these observations inspectable. The stored main-benchmark observation field retains only the first 800 characters of each description, although attribute assessment used the full description. Recomputing aggregate scores is possible from the saved verdicts; repeating perceptual or temporal assessment requires the original video evidence.

B.3 Full-frame assessment: outcome-stratified audit

The full-frame baseline supplies timestamped frames directly to the VLM, without bounding-box crops or SAM masks. Its 100-video audit is stratified by model, branch, and automatic pass label, with additional sampling of reversal cases. The sampling supports inspection across outcomes rather than uniform estimation over all 480 videos. The annotation guide asks for CID, UER, RBA on A_int/B_int or LCA on reversal, and a separate physical-anomaly flag. It accepts some partial changes for UER, whereas the judge is instructed to enforce the script; disagreement can therefore reflect the acceptance threshold as well as perception. We retain a *primary* annotation sheet and an *alternative* sheet for sensitivity analysis. Agreement between these sheets measures annotation sensitivity; independent-rater agreement is evaluated in the three-rater study below.

Table 9 reports full-frame baseline agreement with the primary annotations. We report raw agreement, Cohen's κ (Cohen, 1960), and Gwet's AC_1 (Gwet, 2008), since chance correction is sensitive to the strongly skewed identity marginals (Feinstein & Cicchetti, 1990; Wongpakaran et al., 2013). Pooled AC_1 is 0.797, 0.366, and 0.215 for CID, UER, and LCA; corresponding raw agreements are 84.0%, 68.0%, and 60.7%. These results characterize full-frame assessment under this audit's acceptance thresholds.

Error structure. Against the primary annotations, the full-frame baseline has 22 false negatives and 2 false positives on UER, and 11 false negatives and 0 false positives on LCA. Its execution and

Table 9: Full-frame baseline agreement on the 100-video outcome-stratified audit using primary annotations. Only branch-applicable binary labels are included. κ is undefined for B_int CID because all paired labels equal 1.

Branch	Dim.	n	Agree	Cohen's κ	Gwet's AC_1
null	CID	25	84.0%	0.405	0.781
A_int	CID	24	87.5%	−0.06	0.858
A_int	UER	24	75.0%	0.482	0.529
B_int	CID	23	100.0%	n/a	1.000
B_int	UER	23	69.6%	0.439	0.396
reversal	CID	28	67.9%	0.113	0.500
reversal	UER	28	60.7%	0.287	0.215
reversal	LCA	28	60.7%	0.319	0.215
Pooled CID		100	84.0%	0.245	0.797
Pooled UER		75	68.0%	0.398	0.366
Pooled LCA		28	60.7%	0.319	0.215

restoration decisions are conservative relative to this audit's acceptance thresholds. On RBA, the 47 applicable rows yield 59.6% agreement, $\kappa = 0.154$, and $AC_1 = 0.295$. These dimension-level results use the outcome-stratified sample; the annotation sheet does not cover OBS, NTL, or PUR.

B.4 Human agreement on attribute decisions

This study comprises 120 reversal and 60 intervention clips annotated by three independent blinded raters. We compare the judge's attribute decisions with each rater and with per-dimension majority labels. CID and UER have 179 valid observations; RBA and NTL apply to 60 intervention clips, and LCA and PUR to 120 reversals. OBS is not covered. The study measures agreement on individual dimensions. Its cohort and three-rater reference differ from the outcome-stratified audit, so differences between the two studies reflect different evaluation samples and annotation procedures.

Table 10: Human agreement on attribute decisions. J denotes our judge; H1–H3 are independent raters. Percentages compare J with each rater and their majority; H1–H2 and Fleiss κ measure human agreement.

Dim.	n	J–H1	J–H2	J–H3	Majority	H1–H2	Fleiss κ
CID	179	87.7%	91.1%	88.8%	96.1%	83.2%	0.629
UER	179	89.4%	89.9%	87.2%	93.3%	89.4%	0.694
RBA	60	81.7%	91.7%	86.7%	93.3%	80.0%	0.596
LCA	120	80.8%	78.3%	80.8%	80.0%	95.8%	0.876
PUR	120	91.7%	89.2%	90.8%	91.7%	95.8%	0.866

Reference construction and consistency checks. Majority agreement compares judge decisions with the majority label on each item, rather than averaging the three judge–rater agreements. Fleiss κ (Fleiss, 1971) characterizes agreement among the human raters. Attribute-majority labels and complete-process annotations evaluate different targets, so their agreement statistics are reported separately.

B.5 Contribution of regional evidence and process checks

We compare three judges on an independent set of 200 diagnostic videos, with 50 complete successes and 150 failures under the human reference. All three use the same VLM, videos, human labels, and comparable evaluation budgets. Full-frame assessment scores the visible sequence directly; regional assessment adds character and part crops; the full judge additionally records states in a ledger and applies process rules. A positive verdict requires the script-level conditions in Appendix H.2.

Table 3 reports all three configurations. Crops provide local evidence; the last comparison measures the combined contribution of the ledger and rules. The diagnostic cohort's success prevalence is independent of the main benchmark's prevalence.

B.6 Transfer of attribute assessment across VLMs

We compare direct scoring with observation-first assessment and temporal rules on the 180-video, three-rater audit. Both configurations use the same 10 timestamped frames per clip; each call has a 3,500-token output cap. Direct scoring produces attribute labels in one call. Structured assessment first records timestamped observations, then judges the requested states in a second call and applies onset gates for update ($[1.3, 3.2]$s) and restoration ($[3.5, 5.0]$s). The output cap is per call, rather than a matched total token budget. This comparison evaluates the observation-and-rule stages on branch-applicable attribute labels. The GPT-5.4-Priority reference configuration yields pooled Cohen's $\kappa = 0.486$ on 719 valid video–dimension labels.

For each alternative backend, we pair the two configurations on identical video–dimension labels with valid human-majority references. Cohen's κ is computed over these pooled binary decisions, rather than averaged over dimensions. We resample scenes jointly across configurations for 10,000 bootstrap replicates, retaining all videos and dimensions within each sampled scene. Table 4 reports percentile intervals for the paired difference; Table 11 records the available output coverage and API model identifiers.

The paired subsets determine the scope of each within-backend comparison in Table 4. Doubao and Kimi have direct-scoring outputs but no paired structured outputs in the evaluated records.

Table 11: Available cross-VLM outputs before pairing. Each cell gives valid videos / valid labels among the 180 audit videos; a video contributes if at least one applicable label has a valid human reference and binary prediction. Zero coverage denotes no valid structured output, not a failed benchmark video. Model names in Table 4 correspond to these recorded API identifiers.

API model identifier	Direct	Structured
`qwen3-vl-235b-a22b-instruct`	180 / 719	180 / 719
`gpt-4o`	180 / 719	180 / 719
`gpt-4.1`	180 / 719	180 / 719
`qwen-vl-plus`	180 / 719	180 / 719
`claude-4-5-sonnet-20250929`	180 / 719	179 / 715
`gemini-2.5-flash`	180 / 719	174 / 693
`glm-5.3-flashx`	172 / 687	42 / 168
`qwen3.5-plus`	157 / 620	38 / 152
`doubao-seed-1-6-vision-250815`	179 / 715	0 / 0
`kimi-k2.8-preview`	135 / 540	0 / 0

B.7 Sensitivity to the annotation reference

We compare the primary and alternative attribute annotations on their common branch-applicable binary entries. Null-branch execution and address entries are excluded even if filled in a sheet; missing, non-binary, and N/A entries are not converted to failures. Table 12 measures the sensitivity of the attribute reference labels, rather than independent annotator agreement. Physical-anomaly flags are reported separately in Section C.

Against the alternative annotations, judge agreement is 85.0% on CID, 73.3% on UER, and 64.3% on LCA ($AC_1 = 0.808, 0.468, 0.286$). Execution and restoration decisions remain conservative relative to this reference. The alternative sheet contains 16 anomaly positives among 92 binary-labelled rows, including three H3 reversal clips not flagged in the primary sheet; eight anomaly labels are missing. The main anomaly analysis uses the complete primary annotations, with this comparison measuring reference sensitivity.

Scope of prevalence correction. The Rogan–Gladen expression $\hat{\pi}_{\mathrm{RG}} = (\hat{\pi}_{\mathrm{obs}} + \mathrm{Spec} - 1)/(\mathrm{Sens} + \mathrm{Spec} - 1)$ recovers the human prevalence algebraically when all inputs come from

Table 12: Agreement between primary (P) and alternative (A) attribute annotations. Marginals are positive/negative counts. $236/242 = 97.5\%$ of applicable paired attribute labels agree; entries across dimensions are not independent observations.

Dimension	n	Agree	κ	AC_1	Marginals P $\rightarrow$ A
CID	100	99.0%	0.947	0.988	90/10 $\rightarrow$ 89/11
UER	75	94.7%	0.883	0.903	51/24 $\rightarrow$ 47/28
LCA/RBA	67	98.5%	0.954	0.978	54/13 $\rightarrow$ 53/14

Table 13: Judge versus alternative annotations on branch-applicable binary labels. LCA uses reversal rows only; it is not pooled with RBA. Sensitivity/specificity CIs are Wilson intervals. The final column compares automatic and human prevalence within the same audit subset.

Dim.	n	TP/FP/TN/FN	Sens. (95% CI)	Spec. (95% CI)	Auto $\rightarrow$ human
CID	100	80/6/5/9	0.899 [0.82, 0.95]	0.455 [0.21, 0.72]	0.860 $\rightarrow$ 0.890
UER	75	29/2/26/18	0.617 [0.47, 0.74]	0.929 [0.77, 0.98]	0.413 $\rightarrow$ 0.627
LCA	28	9/0/9/10	0.474 [0.27, 0.68]	1.000 [0.70, 1.00]	0.321 $\rightarrow$ 0.679

the same confusion matrix. That identity is not independent evidence that calibration transfers to new videos or individual models. The audit's outcome-stratified sampling, incomplete dimension coverage, and small per-model cells do not support correcting composite benchmark passes with pooled dimension-level sensitivity and specificity. Attribute-level rates remain uncorrected label conjunctions; strict reversal results come from script-level verification, not this prevalence transform (Appendix H.2).

C PHYSICAL-ANOMALY TAXONOMY

Primary annotations flag a physical anomaly in $12/100$ audited videos: $8/25$ (32.0%) for Wan2.1-480p, $3/21$ (14.3%) for CogVideoX, $1/25$ (4.0%) for Wan2.2, and $0/29$ for MiniMax-H3. These are rates within the outcome-stratified audit sample. The alternative annotations differ on some cases and omit eight flags (Section B.7); the absence of primary H3 flags therefore does not establish anomaly-free generation.

Annotator notes and Figure 6 illustrate seven qualitative categories: object duplication, hand–object clipping, scene replacement or character disappearance, morphological deformation, object levitation, implausible occluder appearance, and discontinuous scene transitions. These categories can overlap and are not exhaustive.

The taxonomy also identifies two measurable predicates that are outside the current metric bank: an inventory-consistency predicate ($\Delta N_{\text{obj}} = 0$ modulo occlusion) and a scene-continuity predicate. Both can be expressed over the same structured records without altering the current metric semantics; categories (a), (c), and (e) would then become explicit metric failures.

D CONTROLLED DIAGNOSTIC VALIDATION

D.1 STATE DECISIONS ON PAPER-DOLL CLIPS

The controlled paper-doll set (MVS-A) tests recognizable state events and deliberately invalid scripts. It separates 140 positive-scene clips from 60 adversarial negatives. Each dimension is evaluated only on its applicable branches.

On positive-scene clips, F1 is at least 0.85 on six of seven dimensions. Controlled clips isolate structural events, while the human studies assess judgments on generated videos.

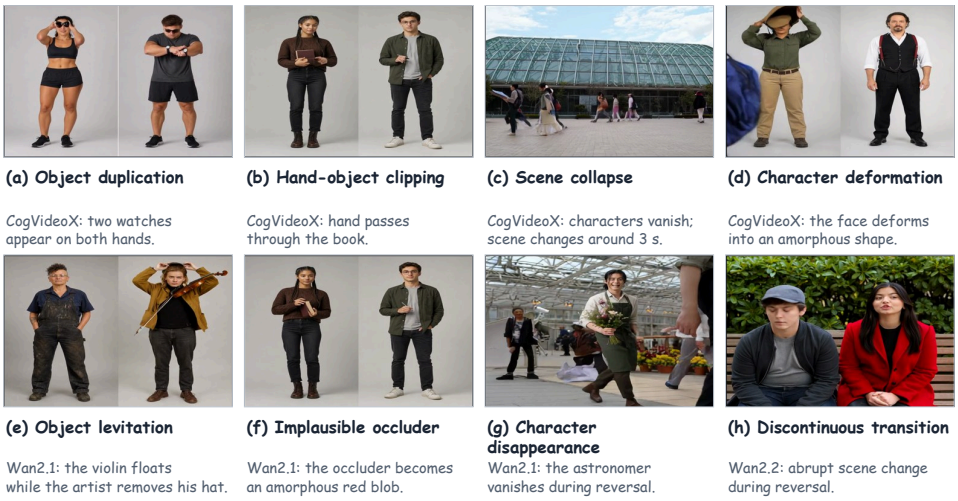


Figure 6: Physical anomalies in the outcome-stratified audit. Primary annotations flag $12/100$ clips. Panels illustrate failure categories rather than an exhaustive or mutually exclusive taxonomy. Model-wise counts and annotation sensitivity appear in Appendix C.

Table 14: Controlled diagnostic clips. The table reports branch-applicable labels on 140 positive-scene clips; a positive scene need not satisfy every dimension. Precision is 1.000 on six dimensions and 0.923 on PUR. The separate 60 adversarial negatives test invalid event sequences.

Dimension	n (applicable)	Precision	Recall	F1
CID	140	1.000	1.000	**1.000**
UER	105	1.000	0.952	**0.976**
RBA	70	1.000	0.976	**0.988**
NTL	70	1.000	0.743	**0.852**
LCA	35	1.000	0.857	**0.923**
PUR	35	0.923	0.571	**0.706**
OBS	35	1.000	1.000	**1.000**
unweighted mean		**0.989**	**0.871**	**0.921**

E Continuity-proxy implementation and pairing

E.1 Continuity-proxy comparison

We compare rubric passes with a lightweight continuity proxy: the geometric mean of centre-crop CLIP similarity, border-crop CLIP similarity (Radford et al., 2021), and optical-flow smoothness. This is a custom VBench-inspired proxy, not the official VBench evaluation. Proxy measurements are paired with the same main-benchmark verdicts used in Table 5, using video-path identity rather than scene ID alone. On 839 paired videos (480 generated and 359 static/pan/loop controls), video-level Spearman correlation with binary rubric pass is 0.182 overall and 0.351 for generations. Controls have mean proxy score 0.801 and raw pass rate 0.170, versus 0.558 and 0.285 for generated videos. Figure 7 shows that high visual continuity is insufficient for role-addressed execution; the two measures nevertheless have a nonzero association.

F Generation configurations and evaluation matrix

F.1 Baselines

Table 15 distinguishes the four primary baselines from supplementary configurations. The main evaluation uses the stated model-specific canvas and frame settings. The supplementary comparison uses the configurations listed in Table 19.

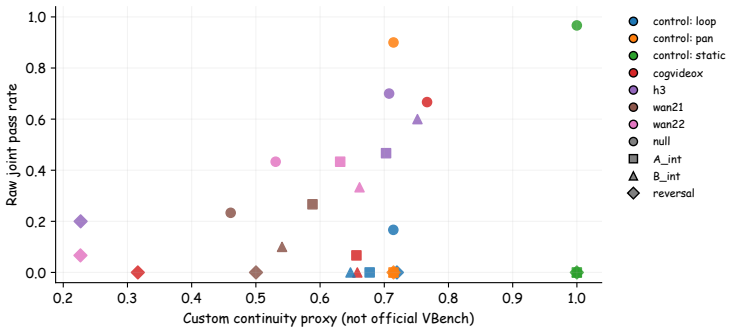


Figure 7: Custom continuity proxy versus raw rubric pass, aggregated by model and branch for display. Correlations quoted in the text are computed on individual videos, not these cell means. The proxy is not an official VBench score. Static, pan, and loop controls can pass null stability, so their aggregate pass rate is not expected to be identically zero.

Table 15: Baseline scope. The main evaluation contains 120 scored videos per model. H3 uses a 512×512 canvas; CogVideoX uses 720×480. The CogVideoX q029 clip in Figure 8 contains 49 frames at 16 fps. Generation settings and actual frame timestamps accompany the evaluation records.

Abbr.	Model	Evaluation scope
Wan21	Wan2.1-I2V-14B-480p	Main evaluation; CFG analysis
Wan22	Wan2.2-I2V-A14B	Main evaluation; CFG analysis
Cog	CogVideoX-5B-I2V	Main evaluation
H3	MiniMax-H3	Main evaluation, 512×512 on one A800-80G
HYV	HunyuanVideo-I2V	Supplementary exploratory configurations
Wan21-720p	Wan2.1-I2V-14B-720p	Supplementary exploratory configurations

F.2 Data scale

The main comparison comprises 30 quadruples $\times$ 4 models $\times$ 4 branches $=$ 480 videos. Attribute-class and difficulty analyses use this same set. Supplementary studies either reuse specified subsets or evaluate additional configurations; their observation counts are not summed as independent videos.

F.3 Guidance sensitivity under matched observations

We examine CFG values $\{5.0, 7.5, 12.0\}$ on Wan2.1-480p and Wan2.2. For each model/quadruple/branch/CFG key, repeated judge records are resolved by retaining the last record, independently of its score. We retain only quadruple–branch keys available at all three settings within each model: 63 keys for Wan2.1 (189 observations) and 53 for Wan2.2 (159 observations). This holds item composition fixed across CFG. The source data contain 311 Wan2.1 records (215 unique keys) and 207 unique Wan2.2 records before complete-case matching. The Wan2.2 sweep and a separate 135-video run both use 40 inference steps; the comparison varies guidance, not the inference-step budget.

F.4 Evaluation coverage

G Detailed result analyses

G.1 Raw comparisons and qualitative adjudication

Table 5 reports attribute-level joint passes on a common denominator of 30 quadruples per model and branch. H3 and CogVideoX pass 70.0% and 66.7% of null videos, respectively, but their A_int pass rates are 46.7% and 6.7%. Attribute-level reversal passes are $6/30$ for H3, $2/30$ for Wan2.2,

Table 16: Evaluation coverage. Counts refer to each study's retained observations, not disjoint sets of videos. Attribute agreement, event localization, and lifecycle validity answer different validation questions.

Evaluation	Count	Scope
Main benchmark, four baselines	480	Branch scores and conditional coverage
Frozen reversal controls	120	All raw joint passes zero
Wan2.1 matched CFG observations	189	63 keys × three settings
Wan2.2 matched CFG observations	159	53 keys × three settings
Additional Wan2.2 40-step run	135	null, A_int, reversal
Supplementary short-clip comparison	240	Unequal durations
Available 512 × 512 canvas records	55	Incomplete; 23 Wan2.1, 32 Cog
Outcome-stratified human audit	100	Primary and alternative annotations
Three-rater attribute agreement	180	Valid counts depend on dimension
Event-time analysis	180	Intervention and reversal clips
Human process validation	120	Event times and complete lifecycles
Script-level reversal verification	120	Same original clips; 6 strict passes
Controlled judge comparison	200	50 human successes, 150 failures
Paper-doll diagnostic clips	200	140 positive scenes, 60 negatives
Continuity-proxy paired records	839	480 real, 359 controls
Post-hoc lifecycle records	120	Excluded: incomplete descriptions



Figure 8: Reversal frames for q029. The florist is asked to remove her glasses and restore them after a van occlusion. H3's branch labels are $(\mathrm{CID}, \mathrm{UER}, \mathrm{LCA}, \mathrm{PUR}) = (1, 1, 0, 0)$: removal is accepted, restoration is not.

and $0/30$ for the other models. Two-sided exact paired McNemar tests (McNemar, 1947) on these attribute-level labels give $p = 0.219$ for H3 versus Wan2.2, $p = 0.031$ for H3 versus either zero-pass model, and $p = 0.500$ for Wan2.2 versus either zero-pass model. These are exploratory, unadjusted comparisons. Wan2.2 versus CogVideoX on A_int gives $p = 0.0034$. Wilson intervals (Wilson, 1927) quantify per-cell binomial uncertainty, including at zero observed successes. Script-level verification gives $5/30$ strict reversal passes for H3 and $1/30$ for Wan2.2 (Table 22); the duration and edit controls below examine configuration effects.

Sampling and qualitative adjudication. The q029 reversal script asks the florist to remove her glasses and restore them after a delivery van occlusion. Rows show MiniMax-H3, Wan2.2, Wan2.1,



Figure 9: Four Wan2.2-I2V-A14B branches for q015. A is the engineer (orange bandana); B is the busker (blue scarf). Both intervention branches request green on their respective targets; reversal restores A's orange bandana after occlusion. Left labels specify the scripts and columns share sampling times. Temporal validity is assessed separately in Section 4.

and CogVideoX. H3's branch labels are $(\mathrm{CID}, \mathrm{UER}, \mathrm{LCA}, \mathrm{PUR}) = (1, 1, 0, 0)$: the judge recognises removal but not restoration. The still images do not resolve whether a faint face-level feature is a restored or duplicated pair, so this example is not counted as a confirmed anomaly or a successful lifecycle. All samples are $[0, 1.25, 2.5, 3.75, 5.0]$s. Timestamps refer to each row's own clip, not a shared column time.

The aggregate table uses these q029 branch labels. The audited cases in Section C demonstrate that physical violations can coexist with an automatic pass.

Matched-branch illustration (Figure 9). Four branches of q015 generated by Wan2.2-I2V-A14B. A is an engineer with an orange bandana; B is a busker with a blue scarf. A_int requests orange $\rightarrow$ green on A's bandana, whereas B_int requests blue $\rightarrow$ green on B's scarf. Thus the edit class and destination colour are shared, but the initial colours and objects differ. The reversal script asks A's bandana to return to orange after occlusion. Frames are sampled at $[0, 1.25, 2.5, 3.75, 5.0]$s. The strips illustrate the branch design; sparse frames alone do not certify timing or a complete lifecycle.

G.2 Interpreting conditional numerators and denominators

Table 17 separates unconditional attribute-level performance from performance on quadruples whose paired A_int receives execution credit. H3 has 19 eligible A_int videos, of which four paired reversals pass the attribute conjunction; Wan2.2 has 17 eligible videos, of which two pass. CogVideoX and Wan2.1 have 3 and 12 eligible videos, respectively, but no passing reversals. A_int joint passes $(14, 13, 2, 8)$ are not the eligibility counts. The gate changes both numerator and denominator, so a conditional rate need not be lower than a raw rate. Reporting coverage alongside the conditional score identifies the model-dependent subset being evaluated.

For $N_G > 0$, splitting the raw numerator gives

$$\mathrm{Pass}_{\mathrm{raw,rev}} = \mathrm{RDR} \cdot \mathrm{Pass}_{\mathrm{cond}} + \frac{1}{N} \sum_{q} (1 - G_q) P_{q,\mathrm{reversal}}. \tag{3}$$

The second term counts passing reversals whose paired A_int update fails. H3's six attribute-level reversal passes include two such cases, yielding $4/19$ conditionally. Raw rates compare the common population; conditional rates and RDR describe model-specific eligible scenes.

Table 17: Attribute-level reversal scores and paired-update coverage. Raw uses all 30 quadruples; conditional uses paired A_int UER $= 1$. Intervals are 95% Wilson intervals. These scores use Equation 1, not strict process verdicts.

Model	Raw ($k/30$)	Conditional (k/N_G; %, CI)	RDR
MiniMax-H3	$6/30 = 20.0\%$	$4/19 = 21.1$ [8.5, 43.3]	$19/30 = 63.3\%$
Wan2.2	$2/30 = 6.7\%$	$2/17 = 11.8$ [3.3, 34.3]	$17/30 = 56.7\%$
CogVideoX	$0/30 = 0.0\%$	$0/3 = 0.0$ [0.0, 56.1]	$3/30 = 10.0\%$
Wan2.1-480p	$0/30 = 0.0\%$	$0/12 = 0.0$ [0.0, 24.2]	$12/30 = 40.0\%$

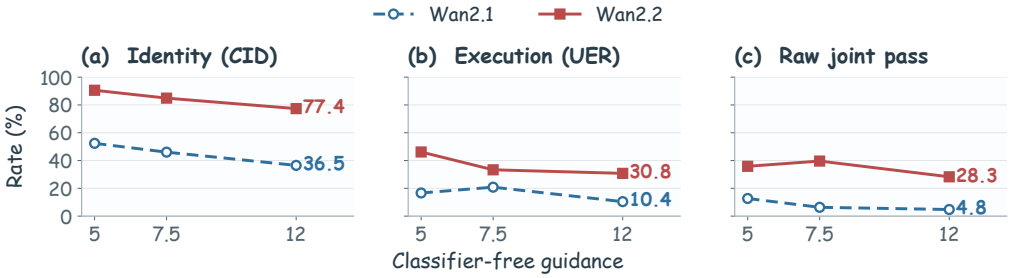


Figure 10: Guidance sensitivity on matched observations (Table 18). Raising CFG from 5 to 12 lowers identity and execution in both sweeps; intermediate execution and joint-pass trends differ by model. These comparisons concern the tested configurations.

The product $\mathrm{RDR} \cdot \mathrm{Pass}_{\mathrm{cond}}$ puts the paired event back on the common scene denominator: $4/30$ for H3 and $2/30$ for Wan2.2. It measures how often both the paired update and reversal pass, with each component evaluated in its own clip. Coverage also determines statistical resolution: CogVideoX's $0/3$ conditional result has a Wilson upper endpoint of 56.1%, versus 24.2% for Wan2.1's $0/12$. Reporting these denominators distinguishes a sparsely observed conditional result from a failure observed over a larger eligible set.

G.3 Complete guidance results

We match the same quadruple–branch keys at classifier-free guidance (CFG) scales (Ho & Salimans, 2022) $\{5, 7.5, 12\}$ within each Wan variant: 63 keys for Wan2.1 and 53 for Wan2.2, yielding 189 and 159 observations. Repeated judge records are resolved by retaining the last record independently of its score; only keys observed at all three settings are included. This fixes item composition within each sweep, while the two models retain different complete-case subsets.

Figure 10 and Table 18 show identity declining at each guidance step in both variants. Execution is non-monotone for Wan2.1 ($16.7 \to 20.8 \to 10.4\%$) and declines for Wan2.2 ($46.2 \to 33.3 \to 30.8\%$). Both have lower identity and execution at CFG 12 than at CFG 5, while Wan2.2's null stability and joint pass peak at 7.5. Thus the high-guidance endpoint degrades performance in these settings, with intermediate behaviour varying by model.

Additional 40-step Wan2.2 run. A separate run contains 15 quadruples $\times$ three branches (null, A_int, reversal) $\times$ three CFG values, totalling 135 videos. Raw pass rates are 0.467, 0.422, and 0.289 at CFG 5, 7.5, and 12. Null pass is $0.933 \to 0.867 \to 0.533$, A_int is $0.467 \to 0.333 \to 0.267$, and reversal is $0.000 \to 0.067 \to 0.067$. This supports the lower-pass high-CFG endpoint under another setting.

G.4 Duration and edit controls

We test whether temporal coverage and unequal edit demands account for the observed branch profiles. The duration control uses full six-second clips: correct initial states at 0–2s, a completed target update at 2–3s, occlusion and reappearance at 3–4s, restoration at 4–5s, and a stable restored state at

Table 18: CFG analysis on matched observations. Wan2.1: 63 keys per setting (189 observations); Wan2.2: 53 per setting (159). Each dimension averages only applicable branches. Raw pass is the conjunction rate over all retained keys, not a conditional reversal rate.

Model	CFG	CID	OBS	UER	LCA	Raw pass
Wan2.1	5.0	52.4%	26.7%	16.7%	0.0%	12.7%
Wan2.1	7.5	46.0%	20.0%	20.8%	7.1%	6.3%
Wan2.1	12.0	36.5%	6.7%	10.4%	0.0%	4.8%
Wan2.2	5.0	90.6%	42.9%	46.2%	33.3%	35.8%
Wan2.2	7.5	84.9%	64.3%	33.3%	16.7%	39.6%
Wan2.2	12.0	77.4%	35.7%	30.8%	8.3%	28.3%

Table 19: Supplementary 15-scene comparison (15 videos per branch, 60 per model; 240 total), under the unequal-duration settings in Appendix F. Overall averages the four equal-sized branch rates. Wilson intervals are descriptive per-video intervals; branches within a quadruple may be dependent.

Model	null	A_int	B_int	reversal	Overall (CI)
MiniMax-H3	0.333	**0.533**	**0.533**	**0.400**	**0.450** [0.331, 0.575]
Wan2.2-I2V-A14B	**0.867**	0.467	0.333	0.000	0.417 [0.301, 0.543]
CogVideoX-5B	**0.867**	0.000	0.000	0.000	0.217 [0.131, 0.336]
Wan2.1-480p	0.400	0.133	0.000	0.000	0.133 [0.069, 0.242]

5–6s. The non-target character must preserve its state throughout. Original clips are verified against their own scripts; the common six-second schedule applies to the full-duration control.

At full duration, we also compare the original character-specific edits with symmetric edits that match the requested change across A and B. Table 6 reports strict reversal passes and the intervention outcomes. All cells contain 30 videos; the original reversal column is the same re-adjudicated set as Table 22.

Full duration increases strict reversal passes for Wan2.2, CogVideoX, and Wan2.1, while H3 remains at $5/30$; the controlled rates range from $1/30$ to $5/30$. Symmetric edits reduce the absolute A/B count gaps from $5, 4, 2, 5$ to $1, 2, 1, 2$, respectively, and preserve the ordering of models by combined A/B passes. The pooled outcome changes little even as the branch gaps narrow.

G.5 Supplementary short-clip comparison

A supplementary comparison evaluates 15 quadruples $\times$ four branches per model. CogVideoX, Wan2.1, and Wan2.2 use 49-frame clips, while H3 uses 124 frames. Table 19 reports performance under the settings documented in Appendix F, with unequal temporal coverage.

H3 scores higher on the intervention branches, whereas Wan2.2 and CogVideoX attain higher null stability. This branch-dependent pattern is more informative than the overall order alone. Overlapping H3/Wan2.2 intervals and the configuration differences in Appendix F limit architectural attribution.

G.6 Attribute-class comparison

Table 7 reports the per-model class scores in Section 6.4. Position has the lowest pooled pass rate, while individual model profiles differ. The scene-level analysis below groups the three role-addressed branches before assessing difficulty.

G.7 Quad-level difficulty distribution

A model passes a quadruple for this diagnostic if at least two of its three role-addressed branches (A_int, B_int, reversal) receive a raw joint pass. We define easy as at least three models passing, medium as one or two, and hard as none. The benchmark contains 1 easy quadruple, 17 medium,

Table 20: Scene-level difficulty across four baselines ($n = 30$ quadruples). A quadruple is credited to a model if it passes ≥ 2 of the three role-addressed branches (A_int, B_int, reversal). Position concentrates in hard ($8/10$); colour spans medium; few scenes meet the easy-tier criterion.

Class	Easy (≥ 3 pass)	Medium (1–2 pass)	Hard (0 pass)
color	0	8	2
position	0	2	8
status	1	7	2
Total	1 (3%)	17 (57%)	12 (40%)

and 12 hard (Table 20). The null branch cannot by itself earn credit. These tiers describe difficulty for the evaluated models and settings.

Eight of the ten position quadruples are hard under this rule, compared with two of ten in each other class. This agrees with position's lower pooled score in Section 6.4. Eight color quadruples fall in the medium tier, providing variation between the evaluated models. Temporal coverage and judging errors can also contribute to a hard-tier assignment.

H TEMPORAL LOCALIZATION AND LIFECYCLE VALIDITY

We evaluate human agreement on the attribute conjunction, verify complete execution against each video's script, and measure event localization. Script-level verification defines strict reversal success; the attribute audit and auxiliary timing analyses show which observations support that decision.

H.1 HUMAN VALIDATION OF ATTRIBUTE-LEVEL JOINT DECISIONS

Of 120 reversal videos, 119 have valid paired attribute-level judge and human verdicts. Both accept exactly the same eight videos: six from MiniMax-H3 and two from Wan2.2. The remaining 111 valid pairs are joint rejections (Table 21). This establishes agreement on individual videos under Equation 1. The script-level verification below additionally checks completed events, non-target preservation, and their timing.

Table 21: Paired human validation of attribute-level joint reversal decisions under Equation 1. TP, FP, FN, and TN use the human attribute verdict as reference. One Wan2.1 video lacks a valid paired attribute label and is excluded. N/A denotes undefined precision, recall, and F1 when neither judge nor human labels any video positive.

Model	Valid n	TP	FP	FN	TN	Precision	Recall	F1
MiniMax-H3	30	6	0	0	24	100%	100%	1.000
Wan2.2	30	2	0	0	28	100%	100%	1.000
CogVideoX	30	0	0	0	30	N/A	N/A	N/A
Wan2.1	29	0	0	0	29	N/A	N/A	N/A
Total	119	8	0	0	111	100%	100%	1.000

H.2 SCRIPT-LEVEL VERIFICATION

Human and machine verification use the same success conditions for each reversal. Both characters must begin in the prescribed states. The designated character must complete the requested update, undergo the scripted occlusion, return to the initial state, and retain that state afterwards. The other character's relevant state must remain unchanged. Each event is checked for role attribution, order, and compliance with the applicable window or deadline in that video's script. Completion requires observing the requested state, not merely the onset of a movement. A verdict therefore draws on the initial state, the changed state before occlusion, the restored state, and the subsequent hold interval.

Table 22 reports re-adjudication of all 120 original reversal videos. Five of H3's six attribute-positive clips and one of Wan2.2's two satisfy the full criteria; none of the 112 attribute-negative clips passes. Strict verification therefore yields $6/120$ successes. The attribute agreement in Table 21 and this verification have distinct denominators: the former contains 119 valid attribute-label pairs, while the latter checks the full original reversal set against its scripts.

Table 22: Strict re-adjudication of the original reversal videos. Retained counts strict successes among attribute-positive clips; additional counts strict successes among attribute-negative clips. Each model has 30 videos. A dash means there are no attribute-positive clips to recheck.

Model	Attribute passes	Retained	Additional	Strict passes
MiniMax-H3	$6/30$	$5/6$	$0/24$	$5/30$ (16.7%)
Wan2.2	$2/30$	$1/2$	$0/28$	$1/30$ (3.3%)
CogVideoX	$0/30$	—	$0/30$	$0/30$ (0.0%)
Wan2.1	$0/30$	—	$0/30$	$0/30$ (0.0%)
Total	$8/120$	$6/8$	$0/112$	$6/120$ (5.0%)

H.3 AUXILIARY TIMING AND STAGE DIAGNOSTICS

Across 180 intervention/reversal clips, event-time assessment reports 55 update onsets within the study's $[1.3, 3.2]$s analysis window, 41 earlier, 6 later, and 78 undetected. Thus $47/180$ onsets fall

Table 23: Attribute conjunctions and auxiliary process readouts. Temporal uses the event-time assessment's common analysis windows; Lifecycle uses its five-stage labels. Their $2/120$ and $7/120$ counts are not script-level strict pass rates. Strict verification with per-script windows and non-target preservation is reported in Table 22.

Model	Attribute	Temporal	Lifecycle	Human attribute majority
H3	$6/30$	$1/30$	$3/30$	$6/30$
Wan2.2	$2/30$	$1/30$	$3/30$	$2/30$
CogVideoX	$0/30$	$0/30$	$0/30$	$0/30$
Wan2.1	$0/30$	$0/30$	$1/30$	$0/29$
Total	$8/120$	$2/120$	$7/120$	$8/119$

outside this window. This distribution describes the judge's temporal output; its analysis window is distinct from the item-specific script schedules in Appendix A. Human localization accuracy is evaluated below on comparable timestamps.

Table 24: Agreement between attribute and event-time decisions. Counts show whether the two assessments accept each dimension. Disagreement occurs in both directions, so these readouts are not nested acceptance tests; the table measures decision differences rather than human-verified corrections.

Dimension	n	Both pass	Attribute only	Temporal only	Neither
UER	180	22	21	16	121
LCA	120	10	9	10	91
PUR	120	9	16	11	84

Table 25: Lifecycle decomposition on 30 reversals per model. Each stage column gives its marginal pass count over the full sample, independently of other stages. Joint is the intersection of all five criteria.

Model	Initial	Update	Occlusion	Restoration	Stable	Joint
H3	28	7	19	19	22	3
Wan2.2	28	14	14	10	13	3
CogVideoX	29	5	0	0	2	0
Wan2.1	28	9	6	1	5	1
Total ($/120$)	113	35	39	30	42	7

Cumulative progress and first failures. Using the same 120 auxiliary lifecycle records, Table 26 counts videos that satisfy every stage up to a given point. The first failure is the earliest failed criterion in the prescribed stage order; videos passing all five criteria have no failed stage. Post-restoration stability participates in both calculations. These counts use the auxiliary lifecycle criteria of Table 23, with update and restoration onset windows of $[1.3, 3.2]$s and $[3.5, 5.0]$s; script-specific strict verdicts are reported in Table 22.

H.4 Temporal localization

Three independent blinded raters annotate 120 reversals, 30 per model, using four binary fields (A/B initial correctness, post-restoration stability, and non-target drift), four event-time fields, and notes. Table 27 reports absolute timing differences only when both sources supply comparable timestamps. Each video can contribute up to three human–human pairs; those pairs are not independent videos.

The first three median differences are 0.08–0.11s; restoration differs more and is comparable on only 73.3% of videos. Unmatched times are not automatically treated as either correct or incorrect: absent events, missed detections, uncertain observations, and missing annotations require separate accounting. A difference below 1s does not establish agreement on a timing-window boundary. For example, 0.88s and 1.40s differ by 0.52s but lie on opposite sides of a 1.3s threshold.

Table 26: Cumulative progress and first failures in the auxiliary five-stage diagnostic. Each model contributes 30 reversals. Panel (a) requires every preceding criterion to pass; panel (b) assigns each unsuccessful video to exactly one first failed stage. The 7 all-stage passes in (a), together with the 113 first failures in (b), account for all 120 videos.

Criterion	H3	Wan2.2	CogVideoX	Wan2.1	Total
(a) Cumulative passes through each stage					
Initial state	28	28	29	28	113
Update	6	12	5	8	31
Occlusion	3	7	0	3	13
Restoration	3	3	0	1	7
Post-restoration stability	3	3	0	1	7
(b) First failed stage					
Initial state	2	2	1	2	7
Update	22	16	24	20	82
Occlusion	3	5	5	5	18
Restoration	0	4	0	2	6
Post-restoration stability	0	0	0	0	0

Table 27: Temporal localization agreement. HH denotes human–human pairs; HM denotes comparable human–judge videos. All reported HH pairs differ by at most 1s. HM coverage is reported out of 120 videos; median absolute differences (MedAD) and counts within 1s use only comparable observations.

Event	HH pairs	HH MedAD (s)	HM $n/120$	HM MedAD (s)	HM $\leq$ 1s
Update onset	169	0.18	$78/120$	0.11	78
Occlusion start	218	0.18	$108/120$	0.08	108
Occlusion end	217	0.18	$107/120$	0.10	107
Restoration onset	231	0.30	$88/120$	0.27	88

For the four binary fields, all three raters agree on 86, 79, 86, and 85 of 120 clips, with Fleiss κ values 0.354, 0.191, 0.596, and 0.057, respectively. The weakest chance-corrected agreement concerns non-target drift. We report both unanimity and κ to characterize this field-level uncertainty; the mere existence of a majority with three binary votes does not measure reliability.

I LIMITS OF PER-MODEL PREVALENCE CORRECTION

Dimension-level sensitivity and specificity in Table 13 describe agreement with the alternative annotation reference. A Rogan–Gladen transform fitted and evaluated on the same confusion matrix returns its human prevalence by construction. Applying the transform to model-specific populations or a conjunction of dimensions additionally requires calibration transfer. The transform is affine, not generally a fixed additive offset, and pooled dimension-level errors need not preserve composite-pass rankings. We therefore report uncorrected attribute conjunctions and, separately, script-verified strict reversal outcomes. Per-model calibrated estimates would require representative calibration samples, complete joint-dimension annotations, and uncertainty propagation.

J PER-MODEL FAILURE SIGNATURES

- **H3.** Highest observed main-benchmark branch-pass rates, with $6/30$ attribute-level and $5/30$ strict reversal passes.
- **Wan2.2.** Competitive A_int execution and joint scores, with $2/30$ attribute-level and $1/30$ strict reversal passes on the original clips; full duration raises strict passes to $3/30$.
- **CogVideoX.** Identity passes on $58/60$ intervention videos, but only $4/60$ executed updates and $2/60$ joint passes.
- **Wan2.1.** Null stability and reversal pass are low; nevertheless $12/30$ A_int videos receive UER $= 1$, so reversal eligibility is not zero.
- **HunyuanVideo.** Evaluated only in supplementary exploratory configurations and excluded from the main benchmark ranking.

K EXTENDED DISCUSSION AND LIMITATIONS

K.1 THREE EMPIRICAL OBSERVATIONS

Branch profiles contain information that an overall rank hides. CogVideoX's high null score coexists with very low intervention passes, whereas H3's intervention rates are higher but its reversal score remains low. These patterns concern the evaluated settings; release recency and architecture were not independently manipulated.

Identity alone is insufficient. On the main A_int/B_int evaluation, CogVideoX retains identity on $58/60$ clips, but executes the requested update on only $4/60$ (3 A_int, 1 B_int). The frozen controls sharpen this distinction: identity is usually preserved while no reversal joint pass is awarded. High identity is therefore compatible with failure to carry out a state-changing script.

Paired-update coverage as an auxiliary diagnostic. RDR reports the fraction of scenes on which the paired A_int clip executes the requested change. Conditional reversal describes the reversal outputs for that subset. It does not establish update success inside those reversal videos, and a low RDR does not make their raw reversal scores undefined. The useful distinction is between full-population results and this explicitly selected, model-dependent scene subset.

K.2 TASK-BASED POSITIONING

We position DUOSTATE by *evaluation stance* and *intervention target*. VBench and VBench-2.0 emphasise video quality and faithfulness using prompt-level evaluation, whereas our protocol uses paired role-addressed interventions on a fixed scene. Physics-oriented dimensions concern the physical scene, while our focus is the atomic state of a named character. Figure 1 groups these complementary tasks and highlights our joint test of role-specific edits, attribution, and post-occlusion restoration.

K.3 LIMITATIONS

Scale and generation settings. The main evaluation contains 30 scenes and short I2V clips under the generation settings and temporal coverage documented in Appendix F. Extending to long multi-shot narratives requires broader scene collections and longer sequences. Matched generation budgets would further separate model capability from configuration effects.

Judge and annotation limits. The controlled success-detection study yields 6 false accepts and 7 missed successes among 200 diagnostic videos (Table 3). Regional evidence and process checks improve both precision and recall, but their perceptual inputs remain VLM-dependent. The outcome-stratified audit also shows sensitivity to the annotation reference. These studies motivate reporting complete-success errors alongside agreement on individual attributes.

Temporal validation. The paper-doll negative audit exposes accepted out-of-window edits and partial occlusion. Script-level verification explicitly checks event completion, order, and each requested window; it retains six of the eight attribute-positive original reversals. Localization accuracy is measured only where human and judge timestamps are comparable, so event coverage remains relevant. The controlled comparison establishes gains from regional evidence and the combined ledger–rule layer; it does not isolate masks, tracking, or evidence refresh within the full judge.

Scope of controls. The protocol shares scenes, references, and attribute classes. Full-duration and symmetric-edit controls reduce temporal-coverage and edit-difficulty differences, and improve some outcomes, while strict restoration remains uncommon (Table 6). The available canvas audit is small and incomplete; it does not isolate H3's architecture from resolution. Closed-source commercial and memory-augmented baselines are absent from the present comparison.

Synthetic references and guidance sweeps. Model-generated portraits may introduce biases that frozen controls do not address. Mirrored-role and real-footage evidence is limited to the reported audit coverage. The guidance study covers two Wan variants and uses last-record deduplication with complete-case matching. Testing unrelated architectures and independently manipulating identity constraints would be necessary to establish a general retention–plasticity mechanism.

L AI USE, ETHICS, AND REPRODUCIBILITY DETAILS

AI USE STATEMENT

AI models assisted seed verbalisation. The authors validated the data, analyses.

ETHICS STATEMENT

The benchmark uses synthetic character portraits rather than intentionally reproducing named real people. Generated media can nevertheless exhibit unwanted associations or physical and identity artefacts. Evaluation code, annotation instructions, and aggregate results are research artifacts with stated intended use and limitations; release of generated media requires attention to the applicable model and asset terms.

Human annotation. Annotation covers the specified characters, branch script, and physical anomalies; the outcome-stratified audit form omits automatic verdicts. Primary and alternative annotations support sensitivity analysis. The attribute-agreement and process studies use three independent blinded raters, with dimension-majority and complete-process references kept separate. Aggregate results contain no identifying annotator information.

REPRODUCIBILITY STATEMENT

The supplementary analysis includes the 30-quadruple seed definitions, main-benchmark verdicts, frozen-control verdicts, primary and alternative annotations, auxiliary result records, and code for recomputing branch scores, conditional rates, and validation statistics. VLM labels are measured inputs to deterministic aggregation, rather than ground truth. The stored main-benchmark descriptions

are truncated, so reproducing perceptual or temporal judgments requires the original videos and additional inference. Sample membership, acceptance criteria, and human reference construction must accompany the reported studies for full replication.

M EVIDENCE-TO-LEDGER IMPLEMENTATION DETAILS

M.1 ROLE-BOUND REGIONAL EVIDENCE

The evidence-to-ledger judge in Figure 4 associates observations with characters before evaluating their states. Text-conditioned grounding provides candidate person and part boxes; SAM 2 supports prompted mask propagation. Reference binding assigns A/B independently of screen position and rechecks identity after occlusion. Face, hair, clothing, eyewear, and task-relevant hand regions are retained with the full frame, preserving the relationship between a local observation and its scene context. Boxes and masks locate evidence; the VLM interprets visible attributes, and temporal rules assess their consistency with the script.

Promptable segmentation in SAM (Kirillov et al., 2023) and memory-based video segmentation in XMem (Cheng & Schwing, 2022) provide related approaches to selecting regions and maintaining them over time. Here, Grounding DINO (Liu et al., 2023a) and SAM 2 (Ravi et al., 2024) supply the localization components; the ledger records the visible states of those regions under their assigned character identities.

M.2 OBSERVATION INTERFACE AND RULE EVALUATION

For role c, time t, and region r, the ledger records $\ell_{c,t,r} = (\hat{s}, v, b, e)$, where $\hat{s}$ is an observed attribute state, v its visibility status, b a bounding box, and e the linked crop identifier. The VLM produces perceptual observations from the full frame and local evidence. The rule layer consumes these records with the branch script to check execution, attribution, non-target stability, and the temporal sequence of restoration. This separation makes an observation error distinguishable from a predicate or ordering error.

If a region is not visible, the ledger records an unknown state rather than inferring that no change occurred. Unresolved observations remain in coverage accounting; the conditional denominator in Equation 2 is defined by the paired A_int update. Strict reversal verification checks the initial state, completed update, occlusion, restoration, and subsequent stability, together with role attribution, non-target preservation, and the individual script's timing (Appendix H.2). Attribute conjunctions retain the definition in Equation 1.

Diagram provenance. Figure 4 uses genuine Wan2.1 q030 A_int frames at 0.00, 3.50, and 5.00s. The overlaid boxes, pose sketch, and ledger rows are explanatory annotations of the implemented method, not raw detector, pose-estimator, SAM, or VLM outputs. The depicted clip illustrates target and non-target inspection; the restoration predicate describes reversal branches separately.

M.3 DIAGNOSTIC OUTPUTS AND VALIDATION TARGETS

Human agreement evaluates attribute decisions against majority labels; temporal localization compares observed event times; lifecycle validity verifies the complete sequence against the script. The original reversal set has $8/120$ attribute-level passes and $6/120$ strict successes. Paired-update coverage describes the separate A_int branch rather than replacing within-video verification (Appendices B.4 and H).

The controlled comparison fixes the videos, VLM, and human reference at comparable budgets. Regional crops raise success-detection F1 from 0.653 to 0.772; adding the ledger and process rules raises it to 0.869 (Table 3). These results quantify the contributions of local evidence and explicit process checking. Individual grounding, mask-propagation, and evidence-refresh effects remain coupled within the evaluated configurations.