BEYOND PLAUSIBLE MOTION: BENCHMARKING AND IMPROVING ORDERED EVENT-TRANSITION GENERATION IN FEW-STEP AUTOREGRESSIVE VIDEO MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-step autoregressive (AR) video models enable low-latency streaming generation by predicting future frame chunks conditioned on previously generated content, yet this chunk-wise formulation makes temporal transitions in generated videos ambiguous. Existing metrics evaluate perceptual quality and multi-event completion, but do not explicitly assess the internal temporal structure of a single short-horizon event transition. To bridge this gap, we introduce *OETGBench*, an **O**rdered **E**vent-**T**ransition **G**eneration **Bench**mark designed to evaluate whether a short-horizon text-to-video (T2V) generation visibly progresses from an initial state to a target outcome in the intended temporal order. OETGBench comprises 2,500 curated tasks across four broad transition families, systematically decomposing each scenario into structured onset, progression, and outcome phases. Alongside OETGBench, we propose *OETGEval*, a hierarchical framework that couples video-level diagnostics with window-level phase verification to quantify temporal fidelity. Finally, we introduce *Anchor-Guarded Phase-Potential Post-Training* (AG-P3T), which uses OETGEval feedback for reward-guided LoRA optimization with anchor-relative penalties and positive-advantage gating. Experiments show that our methods ensure alignment with human evaluations and reveal persistent event-transition failures in the evaluated few-step AR models despite their visual quality. AG-P3T improves the Event Transition Score while maintaining comparable VBench performance, demonstrating the utility of OETGEval feedback for post-training.

Project website: `https://oetg-bench-demo.pages.dev/`.

1 INTRODUCTION

Few-step autoregressive video generation has emerged as a promising paradigm for low-latency streaming synthesis, offering a practical foundation for interactive simulation, embodied agents, and real-time world models. These models are commonly distilled from pretrained video diffusion models (Huang et al., 2025; Cui et al., 2026; Zhu et al., 2026), often using distribution-matching distillation (DMD) (Yin et al., 2024) to reduce sampling cost while largely preserving generation quality. However, these rollout-based distillation methods primarily optimize low-level denoising objectives or distribution matching with the teacher, rather than explicitly enforcing the correct temporal realization of requested event transitions (Wang et al., 2026).

Recent video-generation benchmarks have progressively expanded beyond perceptual quality and semantic alignment toward finer-grained evaluation of compositionality, physical plausibility, multi-event completion, and temporal instruction following (Meng et al., 2025; Wang et al., 2025b; 2026). Despite this progress, the internal temporal order of a single short-horizon event transition remains insufficiently resolved. In particular, they do not explicitly resolve whether a visually plausible video departs from its initial state, progresses through task-relevant intermediate states, and establishes its target outcome in the intended temporal order. Figure 1 illustrates representative failure cases.

Evaluating such *Ordered Event-Transition Generation* (OETG) in few-step autoregressive T2V models presents two practical challenges. On one hand, there is a lack of benchmarks focused on evaluating short-horizon event transition. Reliable evaluation requires semantically concise and

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: Visual plausibility does not imply ordered event completion. The loaf already appears baked in the first frame, while the requested sequence of scoring, oven insertion, baking, and removal is not fully depicted. Evaluating OETG requires checking observable stages and their temporal order.

visually verifiable events with identifiable initial states, intermediate progression, and target outcomes. On the other hand, there is a lack of corresponding evaluation metrics. Traditional metrics like semantic alignment (Huang et al., 2024) exhibit limitations in detecting implausible motions (Meng et al., 2025). Recent studies have used video-based VLMs for comprehensive video evaluation (He et al., 2024; Sun et al., 2025).

To address these challenges, we propose OETGBench and OETGEval to automate the evaluation of event-transition understanding capability of few-step AR T2V models. We categorize event transitions into four main areas: Action Start, Entity Interaction, Object State Change, Environmental Perturbation and build OETGBench on top of real-world T2V prompts in the VidProM dataset (Wang & Yang, 2024). Beyond prompts that clearly delineate event transitions, OETGBench provides phase-specific descriptions of successful completion and targeted quality-assessment questions to support fine-grained evaluation.

We propose a two-tier hierarchical evaluation strategy comprising Video-level Diagnostics (VLD) and Window-level Phase Verification (WPV). Video-level Diagnostics characterize whole-video quality and event transition-related failure modes. Window-level Phase Verification then evaluates localized temporal windows against the phase-specific descriptions provided by OETGBench, identifying evidence for onset, progression, and outcome and explicitly verifying whether these phases form a valid ordered chain.

Beyond the above, we also investigate whether phase-level evaluation signals can guide improvements in OETG. Inspired by progress-based reward shaping (Ng et al., 1999), we introduce Anchor-Guarded Phase-Potential Post-Training (AG-P3T), a reward-guided LoRA (Hu et al., 2022) method that converts OETGEval scores into anchor-relative rewards. Compared with the pretrained baseline, AG-P3T improves the event score, while keeping VBench Total within 0.65% of its baseline value.

The contributions of our work are three-fold.

i): We introduce OETGBench around the observable onset, progression, and outcome of a goal-directed event transition, a benchmark of 2,500 curated tasks across four event-transition families. Each task includes phase-specific completion descriptions and quality-assessment questions.

ii): Along with the benchmark, we propose a hierarchical evaluation framework - OETGEval, which combines video-level diagnostics of visual quality and degeneracy with window-level phase verification of phase completion and temporal order. It demonstrates consistency with human feedback on OETGBench.

iii): We introduce AG-P3T, a reward-guided LoRA method that evaluates candidate generations against a reference from the frozen base model. Sample-level advantage gating restricts

positive reinforcement to candidates that improve event completion and satisfy quality and shortcut checks.

2 Related Work

Few-Step Autoregressive Video Generation Few-step autoregressive video generation has rapidly evolved toward low-latency streaming synthesis. CausVid (Yin et al., 2025) distills pretrained bidirectional video diffusion models into causal generators through asymmetric distribution matching, while Self-Forcing (Huang et al., 2025) trains on self-generated histories to reduce autoregressive train–test mismatch. Self-Forcing++ (Cui et al., 2026) and LongLive (Yang et al., 2026) further improve long-horizon generation through segment-level supervision and contextual-memory mechanisms. More recent methods such as Causal Forcing (Zhu et al., 2026), Causal Forcing++ (Zhao et al., 2026), and Causal-rCM (Zheng et al., 2026) refine causal initialization and consistency/distillation objectives for efficient one- or few-step autoregressive sampling. These advances substantially improve efficiency and rollout fidelity, motivating finer evaluation of whether efficient generators also faithfully realize requested temporal transitions.

Benchmarks and Evaluation for Video Generation Video-generation evaluation has progressively expanded from perceptual quality and semantic alignment toward finer-grained assessment of physical plausibility, state changes, and event completion. VBench (Huang et al., 2024) provides fine-grained assessment of visual quality, temporal consistency, and condition alignment, while VideoScore2 (He et al., 2025) develops interpretable multi-aspect evaluation aligned with human judgments. T2V-CompBench (Sun et al., 2025) evaluates compositional attributes, actions, motions, and object interactions, whereas PhyGenBench/PhyGenEval (Meng et al., 2025) examine physical commonsense through key-phenomenon detection, order verification, and overall naturalness.

Recent benchmarks increasingly target temporal changes and event completion. TC-Bench (Feng et al., 2025) specifies initial and final scene states and evaluates the completeness of compositional transitions. StoryEval (Wang et al., 2025b) measures whether multiple consecutive events are completed, while OSCBench (Han et al., 2026) evaluates the accuracy and temporal consistency of action-induced object state changes. Concurrent work such as TempAct (Wang et al., 2026) further studies temporal instruction following in autoregressive video generation. These efforts operate at complementary granularities, including compositional transitions, multi-event completion, object-centric state changes, and multi-step temporal instructions. OETGBench instead focuses on the internal realization of an individual short-horizon event transition, explicitly describing event onset, transition progression, and outcome realization. Correspondingly, OETGEval combines video-level diagnostics of quality and transition-specific failure modes with window-level phase verification, explicitly assessing whether the three phases are observable and form a valid temporal order.

Reward-Guided Post-Training for Visual Generation Reward optimization has increasingly been applied to visual generation. DDPO (Black et al., 2024) formulates diffusion fine-tuning as policy optimization, while Flow-GRPO (Liu et al., 2025) and DanceGRPO (Xue et al., 2025) extend group-relative optimization to flow-based image and video generation. DenseGRPO (Deng et al., 2026), related to the broader principle of reward shaping (Ng et al., 1999), introduces step-wise reward gains for finer credit assignment. AR-CoPO (He et al., 2026) and RAVEN (Lu et al., 2026) further adapt reward optimization to autoregressive video. AG-P3T instead converts OETGEval signals into anchor-relative feedback with quality and shortcut checks.

3 Methodology

3.1 OETGBench

We define an *event transition* as a temporally evolving and visually observable process through which a scene, entity, or interaction departs from an identifiable initial state, undergoes task-relevant intermediate changes, and establishes a target outcome. For example, for the event "a person opens an umbrella," the transition is not merely the occurrence of *opening*, but its visible realization from a closed umbrella, through partial unfolding, to a fully opened state. Thus, while an event specifies *what happens*, an event transition characterizes *how the requested change unfolds over time*.

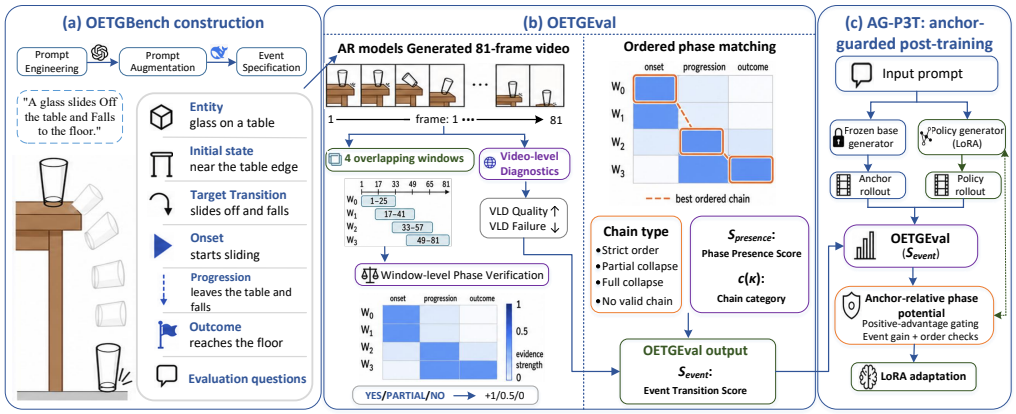


Figure 2: Framework overview: OETGBench defines tasks; OETGEval extracts phase evidence and diagnostics; ordered matching produces the event score; and AG-P3T performs anchor-guarded LoRA post-training.

Overview We organize requested event transitions into four task-oriented categories:

i): Action Start concerns the initiation of a salient action from a quiescent or previously stable behavior, such as beginning to run or initiating a turn. Evidence must show the action starting and developing, rather than only ongoing motion.

ii): Entity Interaction concerns the emergence of a new participating entity or a change in the relationship between entities, such as approaching, making contact, or coordinating movement. The defining evidence is the requested interaction.

iii): Object State Change concerns a visible change in an object's configuration, integrity, or functional state, such as opening or breaking. Evaluation distinguishes the transition process from an object already occupying the target state.

iv): Environmental Perturbation concerns changes in environmental media or surrounding conditions, such as wind-driven foliage or a developing wave. Evidence should reflect the specified disturbance rather than unrelated ambient motion.

These categories describe the dominant requested change, not mutually exclusive object classes or physical laws. The construction pipeline uses keyword-based category proposals; these should not be interpreted as independently verified category annotations.

Benchmark Construction We structure construction into five stages, adapting source prompts into temporally explicit tasks without introducing unrelated events.

Conceptualization. We restrict tasks to one dominant event intended to be observable within approximately five seconds, matching the 81-frame, 16-FPS generation setting used in our experiments. Eligible tasks require a discernible initial state, event initiation, subsequent progression, and an observable result. Multi-scene narratives, long-horizon dependencies, and viewpoint changes that substitute for an event fall outside this scope. Unlike implicit physical-law prediction, the requested event is explicitly stated: the challenge is its visual realization.

Prompt Engineering. We draw source prompts from the semantic-unique release of VidProM (Wang & Yang, 2024), a real-user text-to-video prompt collection. Rule-based filtering retains prompts with 8–180 English words and observable action or interaction cues, while excluding non-event descriptions, long stories, command-line parameters, and high-risk content. The resulting candidates are ranked by event-related cues and textual conciseness for subsequent structured augmentation.

Prompt Augmentation. DeepSeek-V4-Pro (DeepSeek-AI et al., 2026) judges suitability before rewriting; unsuitable inputs are deferred. Accepted prompts follow a stable-scene, initial-state, and event-transition structure. Rewriting is instructed to preserve source-supported entities, settings, and actions, without adding cinematic modifiers, camera movement, or unrelated content. Separate phase

targets describe the first visible event evidence, its continuation, and the resulting observable state. An outcome cannot merely replace event completion with generic stillness.

Event Specification and Question Construction. For each task, we construct an event specification comprising an initial state, a target transition, and three phase targets. We then formulate task-specific evaluation questions that identify the visual evidence required for each phase. Event specifications and evaluation questions are stored separately from generation prompts and are not provided to the video generator, distinguishing the instructions for generation from the criteria for evaluation.

Quality Control. Quality control focuses on the clarity of generation prompts, consistency of phase targets, and visual assessability of evaluation questions. We apply structured checks for incomplete specifications and degenerate formulations, and prepare random samples for human inspection. Further details of source filtering, prompt augmentation, and quality control are provided in Appendix A.

3.2 OETGEVAL

OETGEval evaluates whether a generated video realizes the event transition specified by a prompt, including its onset, progression, and observable outcome.

We therefore adopt a two-level evaluation framework. Video-level Diagnostics first assesses visual quality and potential failure modes across the video. Window-level Phase Verification then measures event evidence in temporal windows and examines whether that evidence supports the expected phase order. This separation distinguishes visual quality from event completion, and phase presence from temporal order, rather than conflating them in a single global judgment. Following PhyGen-Bench (Meng et al., 2025), we conduct both a human user study and automatic assessment using VLMs.

3.2.1 VIDEO-LEVEL DIAGNOSTICS

We first assess the generated video as a whole to characterize its visual quality and identify failures that may compromise the depiction of the requested event. We use a shared set of natural-language questions to assess four quality criteria and five failure modes. All judgments are based on the supplied frames and must allow changes required by the target event.

Quality Diagnostics. The quality criteria are Motion Plausibility, Transition Continuity, Subject Consistency, and Background Consistency. These criteria characterize the quality of the depiction, not whether the requested event has been completed.

Failure Diagnostics. The failure modes are Motion Collapse, Camera-Motion Dominance, Motion Misalignment, Temporal Artifact, and Identity Drift. Together, these checks distinguish missing event execution from alternative sources of apparent change. Detailed definitions and question templates are provided in Appendix B.

For each question, the judge returns a label $\ell \in \{\mathrm{YES}, \mathrm{PARTIAL}, \mathrm{NO}\}$ and a brief evidence explanation. The mapping $g(\ell)$ assigns scores of 1, 0.5, and 0, respectively. PARTIAL denotes weak, incomplete, or ambiguous evidence. A higher quality score indicates a desirable property, whereas a higher failure score indicates a visible failure. These scores are ordinal judgments, not probabilities. We retain the individual diagnostic scores and report them separately from the event score defined below.

3.2.2 WINDOW-LEVEL PHASE VERIFICATION

Directly asking the Vision-Language Model (VLM) to align the event phase between videos and input prompts is difficult; the prompt specifies what should occur but is not itself visual evidence. To address the challenge, we next evaluate whether the event begins, develops, and produces the specified outcome. For an 81-frame video at 16 FPS, $V = (I_1, \ldots, I_{81})$, we construct four chronologically ordered, overlapping temporal windows $\mathcal{W} = \{W_0, \ldots, W_3\}$. Each window is represented by temporally ordered sampled frames. Overlap accommodates events near window boundaries, while avoiding a fixed assignment of temporal intervals to event phases.

Each task provides a natural-language question q_m for each phase $m \in \mathcal{M} = \{\text{on}, \text{prog}, \text{out}\}$. The window judge answers all phase questions independently using only the current window's visual evidence and the same label scale as the global judge. A window may support multiple phases or none. We therefore retain the complete phase-window score matrix rather than assign a single phase to each window: $S_{m,w} = g(\text{VLM}(W_w, q_m))\,, \quad E_m = \max_{w \in \{0,\ldots,3\}} S_{m,w}$. Here, the VLM term denotes its returned label. The phase score E_m measures the strongest observed evidence for phase m, irrespective of when it appears.

Ordered Phase Matching. Phase presence does not establish temporal order. We therefore search for an onset-progression-outcome chain whose selected evidence is positive and temporally nondecreasing:

$$\mathcal{C} = \{(i,j,k) : 0 \le i \le j \le k < 4,\ S_{\text{on},i} > 0,\ S_{\text{prog},j} > 0,\ S_{\text{out},k} > 0\}\,.$$

When $\mathcal{C}$ is nonempty, we select $(i^*, j^*, k^*) = \arg\max_{(i,j,k)\in\mathcal{C}} (S_{\text{on},i} + S_{\text{prog},j} + S_{\text{out},k})$. The search uses equal phase weights. Ties favor more distinct windows, followed by a wider temporal span and earlier indices. Allowing shared windows accommodates phases occurring close together, while the nondecreasing constraint excludes reversed window assignments.

We classify the selected chain as strict order when $i^* < j^* < k^*$, partial collapse when exactly two adjacent phases share a window, and full collapse when all phases share one window. If $\mathcal{C}$ is empty, the result is no valid chain, indicating a missing phase or evidence that cannot form a compatible sequence. Let κ denote this category. Temporal Order distinguishes these arrangements. Because the windows overlap, it measures coarse temporal order rather than frame-exact ordering; collapse describes limited separation of the selected evidence, not necessarily simultaneous underlying events.

Event Score. We combine phase evidence and constrain its contribution according to the selected chain: $S_{\text{presence}} = \sum_{m\in\mathcal{M}} \alpha_m E_m, \quad S_{\text{event}} = \min(S_{\text{presence}}, c(\kappa))$. The fixed nonnegative weights α_m control the contributions of the phase-presence scores; they do not enter chain search. The category-dependent cap $c(\kappa)$ limits high presence scores when a valid chain is absent or fully collapsed. It is applied after aggregation, not as a penalty within the matching objective. We report S_{event}, the phase scores, Temporal Order, and the global diagnostics separately. Exact weights, categorical mappings, question templates, and sampling settings are provided in Appendix B.

3.3 Anchor-Guarded Post-Training

AG-P3T initializes the policy from CF++ and uses one video generated by the frozen base model per prompt as an *anchor*. This anchor is a task-matched comparator, not a frozen-model regularizer. For rollout v_i and anchor v^a, we compute differences $\Delta f_i = f(v_i) - f(v^a)$. The reward combines training-time episode-reward, phase-score, and Temporal Order differences with diagnostic penalties. Group-standardized rewards yield clipped advantages A_i, whose positive components are gated:

$$\widetilde{A}_i = \min(A_i, 0) + g_i \max(A_i, 0). \tag{1}$$

Here, $g_i \in \{0, 1\}$, and $g_i = 1$ requires episode-reward improvement, non-decreasing outcome or Temporal Order, bounded shortcut deterioration, and no detected severe freezing or premature outcome. Otherwise, positive advantages are zeroed while negative advantages remain unchanged. The gate controls sample contributions, not acceptance of parameter updates. Shared linear anchor offsets cancel under group centering; anchor dependence remains through penalties and gating.

Only LoRA parameters are optimized using clipped, advantage-weighted score-gradient target-shift regression with a current-rollout trace regularizer. Episode advantages are broadcast to selected transitions. Reward construction, diagnostic aggregation, and optimization details are provided in Appendix F.

4 Experiments

4.1 Experimental Setup

We evaluate three representative few-step autoregressive generators: CausVid AR 3-step (Yin et al., 2025), Self-Forcing DMD (Huang et al., 2025), and CF++ frame-wise 2-step (Zhao et al., 2026),

with CF++-initialized AG-P3T reported separately as a post-training reference. Our matched event evaluation comprises 50 human-blind tasks for each model, yielding 150 videos. All 150 videos decode at 832×480 resolution, with 81 frames at 16 FPS. We instantiate OETGEval with Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), and InternVL3.5-8B (Wang et al., 2025a) under the same four-window protocol. External quality is assessed separately using the full VBench suite (Huang et al., 2024). Detailed generation, annotation, and evaluation settings are deferred to Appendices B and C.

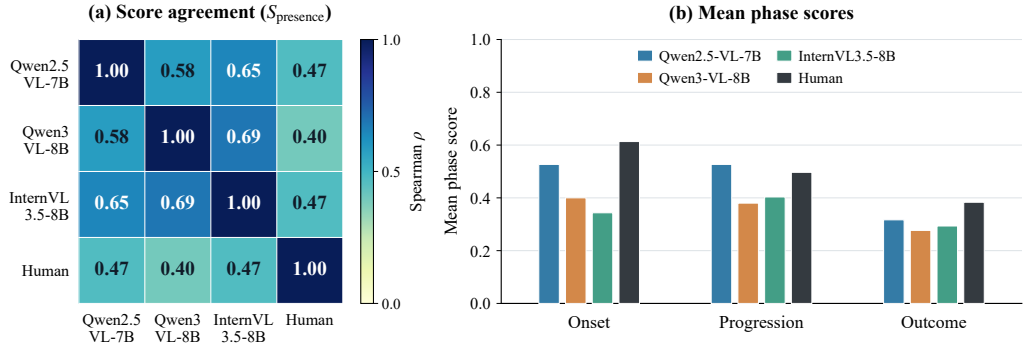


Figure 3: Human agreement and cross-VLM consistency on OETGEval. (a) Pairwise Spearman correlations for the Phase Presence Score, S_{presence}, including the human reference. (b) Mean onset, progression, and outcome scores from three VLMs and the human reference. Colors identify evaluators. Phase-specific correlations appear in Appendix D.

4.2 Evaluator Validation

We assess evaluator reliability using phase ratings from three human reviewers on the same 150 videos. Individual human ratings and automatic phase scores use the scale $\{0, 0.5, 1\}$. For each video and phase, the human reference is the arithmetic mean of the three reviewers' ratings. Table 1 reports video-level Kendall's τ_b and Spearman's ρ for onset, progression, outcome, and the derived Phase Presence Score, S_{presence}. Correlations are computed from matched videos, not from the three generator means; task-cluster bootstrap intervals are provided in Appendix D.

Table 1: Agreement with the human reference on 150 matched videos from 50 tasks. All automatic phase scores use the four-window protocol; the original three-level labels are retained. Kendall's τ_b handles tied ratings.

	Onset		Progression		Outcome		S_{presence}	
Evaluator	τ_b	ρ	τ_b	ρ	τ_b	ρ	τ_b	ρ
Qwen2.5-VL-7B	0.319	0.350	0.262	0.288	0.322	0.354	**0.363**	0.465
Qwen3-VL-8B	**0.414**	**0.454**	0.147	0.167	0.475	0.512	0.301	0.395
InternVL3.5-8B	0.289	0.303	**0.303**	**0.336**	**0.478**	**0.518**	0.358	**0.466**

All three VLMs correlate positively with the human ratings, although no evaluator is strongest across all phases. Qwen3-VL has the highest onset correlation ($\rho = 0.454$), Qwen2.5-VL and InternVL3.5 show similar agreement on S_{presence} (0.465 and 0.466), and InternVL3.5 has the highest outcome correlation (0.518). Progression is less consistently judged. Despite differences in absolute scores, both evaluation methods exhibit highly consistent trends across models.

Temporal Order Assessment. Table 2 compares OETGEval against direct full-video evaluations from three VLMs on 150 videos. OETGEval achieves the highest correlations and AUROC, demonstrating stronger alignment with human temporal order judgments on this set. The lower direct-VLM scores reveal a persistent gap in capturing ordered event transitions. Coupled with the incomplete event realization reported in Section 4.3, these findings highlight limitations of both the evaluated generators and VLM judges, while validating OETGEval as an effective tool for diagnosing OETG failures.

Table 2: Temporal Order assessment.

Evaluator	Spearman ρ	Kendall τ_b	AUROC
Qwen2.5-VL	0.127	0.127	0.525
Qwen3-VL	0.296	0.296	0.630
InternVL3.5-8B	0.193	0.193	0.603
OETGEval	**0.406**	**0.381**	**0.708**

Figure 3(a) displays cross-VLM and human–VLM correlations, while Figure 3(b) compares average scoring levels. Relative to the human reference, Qwen2.5-VL has signed mean differences of -0.087, $+0.030$, and -0.067 for onset, progression, and outcome, respectively. Thus, its average scoring level is relatively close to the human reference, while the other two VLMs are generally more conservative. Nevertheless, similar means can conceal different individual decisions. In particular, progression exhibits higher cross-VLM correlations than human–VLM correlations, indicating closer agreement among evaluators than with the human reference. These results highlight a limitation of the evaluated VLMs in OETG assessment. This gap suggests that cross-VLM consistency alone does not establish human-aligned assessment of event transition.

4.3 Quantitative Evaluation

Table 3: Event-transition and external quality results. Higher is better except VLD-F.

	OETGEval Score							**VBench (%)**		
Method	S_{event}	E_{on}	E_{prog}	E_{out}	Temporal Order	VLD-Q	VLD-F↓	Total	Quality	Semantic
CausVid AR 3-step	0.3435	0.530	**0.570**	**0.340**	0.310	0.7220	0.0285	80.79	83.93	68.23
Self-Forcing DMD	**0.3540**	**0.550**	0.560	0.330	**0.350**	**0.8205**	**0.0180**	**82.13**	**84.96**	**70.82**
Causal Forcing++ 2-step	0.2905	0.500	0.450	0.280	0.280	0.7090	0.0375	81.85	84.88	69.73
Post-training reference										
AG-P3T	0.4038	0.400	0.650	0.478	0.509	0.6463	0.0738	81.32	83.96	70.76

Event realization remains incomplete. Among the three few-step AR generators, Self-Forcing obtains the highest mean Event Transition Score. However, Self-Forcing's advantage over CausVid is not resolved by the paired confidence interval. Outcome remains the weakest phase for all three models, compared with progression scores. Mean Temporal Order further indicates limited ordered phase evidence. These results suggest that producing visible activity remains easier than completing the specified transition.

Visual consistency does not imply event completion. To compare these properties without changing the evaluation videos, we additionally compute four VBench consistency-related dimensions. Table 4 summarizes subject consistency, background consistency, temporal flickering, and motion smoothness using their equal-weight mean. CF++ ranks highest on all four dimensions but lowest on S_{event}, whereas Self-Forcing exhibits the reverse ranking. Across individual videos, the temporal consistency summary correlates negatively with S_{event}. These observations support evaluating event completion separately from visual consistency, without establishing a causal trade-off between the two.

Table 4: Event and temporal consistency scores. TC is the equal-weight mean of four VBench dimensions.

Generator	S_{event} ↑	TC↑
CausVid 3-step	0.3435	0.9526
Self-Forcing	**0.3540**	0.9396
CF++ 2-step	0.2905	**0.9623**

Figure 4 shows representative event-transition cases; detailed qualitative analysis is provided in Appendix G.



Figure 4: Qualitative examples across four event-transition families: Entity Interaction (top left), Environmental Perturbation (top right), Object State Change (bottom left), and Action Start (bottom right). Frames follow temporal order within each model row.

4.4 AG-P3T Post-Training

Reward and gating ablation. Table 5 compares four cumulative configurations: event-only optimization using S_{event} (A1), additional phase/order reward weighting (A2), anchor-relative quality and shortcut penalties (A3), and positive-advantage gating (A4). All configurations use the same fixed event scorer, four policy rollouts per prompt, and identical group normalization and advantage clipping.

Table 5: Cumulative ablation of AG-P3T.

ID	S_{event} ↑	E_{out} ↑	VLD-Q↑	VLD-F↓
A1	0.3361	0.200	0.6023	0.0638
A2	0.3492	0.200	0.6010	0.0775
A3	0.3746	0.350	0.4725	**0.0047**
A4	**0.4038**	**0.478**	**0.6463**	0.0738

Phase/order weighting yields a modest event-score gain but leaves outcome realization unchanged and VLD Quality nearly unchanged, suggesting that additional reward emphasis alone is insufficient in this setting. Anchor-relative penalties improve outcome realization and reduce VLD Failure, but also lower VLD Quality; suppressing measured failures therefore does not necessarily preserve overall quality. Adding the positive-advantage gate achieves the strongest event, outcome, and quality scores, although the ungated penalized variant retains the lowest failure score. This comparison supports gating as a complement to reward penalties for balancing event realization and quality, rather than uniformly minimizing failures. The gate filters positive sample advantages without subsequent renormalization; it does not certify post-update model quality. Detailed configurations are provided in Appendix F.

5 Conclusion

We introduced OETGBench and OETGEval to assess ordered event-transition generation. Our experiments show that the evaluated few-step AR models still struggle to realize requested transitions despite high visual quality, while direct VLM judgments have limited alignment with human temporal assessments. OETGEval helps expose these failures through phase verification and video-level diagnostics. We also explore using OETG feedback to improve generation through AG-P3T.

AI USE STATEMENT

Generative AI tools were used for prompt augmentation and automated video assessment, as described in the methodology. In addition, AI assistance was also used to refine the manuscript.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training Diffusion Models with Reinforcement Learning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.13301.

Jiaxing Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-Forcing++: Towards Minute-Scale High-Quality Video Generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8a30aba6514b56d02976f49797f6338a-Abstract-Conference.html.

DeepSeek-AI et al. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence. *arXiv preprint arXiv:2606.19348*, 2026. URL https://arxiv.org/abs/2606.19348.

Haoyou Deng, Keyu Yan, Chaojie Mao, Xiang Wang, Yu Liu, Changxin Gao, and Nong Sang. DenseGRPO: From Sparse to Dense Reward for Flow Matching Model Alignment. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2601.20218.

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

Xianjing Han, Bin Zhu, Shiqi Hu, Franklin Mingzhe Li, Patrick Carrington, Roger Zimmermann, and Jingjing Chen. OSCBench: Benchmarking Object State Change in Text-to-Video Generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30867–30884, 2026. doi: 10.18653/v1/2026.acl-long.1425. URL https://aclanthology.org/2026.acl-long.1425/.

Dailan He, Guanlin Feng, Xingtong Ge, Yi Zhang, Bingqi Ma, Guanglu Song, Yu Liu, and Hongsheng Li. AR-CoPO: Align Autoregressive Video Generation with Contrastive Policy Optimization. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2603.17461.

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. URL https://aclanthology.org/2024.emnlp-main.127/.

Xuan He, Dongfu Jiang, Ping Nie, Minghao Liu, Zhengxuan Jiang, Mingyi Su, Wentao Ma, Junru Lin, Chun Ye, Yi Lu, Keming Wu, Benjamin Schneider, Quy Duc Do, Zhuofeng Li, Yiming Jia, Yuxuan Zhang, Guo Cheng, Haozhe Wang, Wangchunshu Zhou, Qunshu Lin, Yuanxing Zhang, Ge Zhang, Wenhao Huang, and Wenhu Chen. VideoScore2: Think before You Score in Generative Video Evaluation. *arXiv preprint arXiv:2509.22799*, 2025. URL https://arxiv.org/abs/2509.22799.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2506.08009.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024. URL https://arxiv.org/abs/2311.17982.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training Flow Matching Models via Online RL. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3a10c46572628d58cb44fb705f25cbbf-Abstract-Conference.html.

Yanzuo Lu, Ronglai Zuo, and Jiankang Deng. RAVEN: Real-time Autoregressive Video Extrapolation with Consistency-model GRPO. *arXiv preprint arXiv:2605.15190*, 2026. URL https://arxiv.org/abs/2605.15190.

Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards World Simulator: Crafting Physical Commonsense-Based Benchmark for Video Generation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43781–43806. PMLR, 2025. URL https://proceedings.mlr.press/v267/meng25c.html.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy Invariance under Reward Transformations: Theory and Application to Reward Shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 278–287, 1999. URL https://people.eecs.berkeley.edu/~russell/papers/icml99-shaping.pdf.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A Comprehensive Benchmark for Compositional Text-to-video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8406–8416, 2025. URL https://arxiv.org/abs/2407.14505.

Jing Wang, Xiangxin Zhou, Jiajun Liang, Kaiqi Liu, Wanyuan Pang, Zhenyu Xie, Tianyu Pang, and Xiaodan Liang. TempAct: Advancing Temporal Plausibility in Autoregressive Video Generation via Planner-Executor RL. *arXiv preprint arXiv:2606.28016*, 2026. URL https://arxiv.org/abs/2606.28016.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025a. URL https://arxiv.org/abs/2508.18265.

Wenhao Wang and Yi Yang. VidProM: A Million-Scale Real Prompt-Gallery Dataset for Text-to-Video Diffusion Models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 65618–65642, 2024. URL https://arxiv.org/abs/2403.06098.

Yiping Wang, Xuehai He, Kuan Wang, Luyao Ma, Jianwei Yang, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. Is Your World Simulator a Good Story Presenter? A Consecutive Events-Based Benchmark for Future Long Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13629–13638, 2025b. URL https://arxiv.org/abs/2412.16211.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. DanceGRPO: Unleashing GRPO on Visual Generation. *arXiv preprint arXiv:2505.07818*, 2025. URL https://arxiv.org/abs/2505.07818.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time Interactive Long Video Generation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/91a1610c6ed9e02d33f826b46f472b92-Abstract-Conference.html.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-Step Diffusion with Distribution Matching Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024. URL https://arxiv.org/abs/2311.18828.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From Slow Bidirectional to Fast Autoregressive Video Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22963–22974, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Yin_From_Slow_Bidirectional_to_Fast_Autoregressive_Video_Diffusion_Models_CVPR_2025_paper.pdf.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal Forcing++: Scalable Few-Step Autoregressive Diffusion Distillation for Real-Time Interactive Video Generation. *arXiv preprint arXiv:2605.15141*, 2026. URL https://arxiv.org/abs/2605.15141.

Kaiwen Zheng, Guande He, Min Zhao, Jintao Zhang, Huayu Chen, Jianfei Chen, Chen-Hsuan Lin, Ming-Yu Liu, Jun Zhu, and Qianli Ma. Causal-rCM: A Unified Teacher-Forcing and Self-Forcing Open Recipe for Autoregressive Diffusion Distillation in Streaming Video Generation and Interactive World Models. *arXiv preprint arXiv:2606.25473*, 2026. URL https://arxiv.org/abs/2606.25473.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal Forcing: Autoregressive Diffusion Distillation Done Right for High-Quality Real-Time Interactive Video Generation. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.02214.

A OETGBENCH: CONSTRUCTION, STATISTICS, AND TASK EXAMPLES

A.1 FROM USER PROMPTS TO OBSERVABLE TRANSITIONS

We construct tasks from VidProM's semantic-unique user prompts. A suitable source describes one event that can visibly unfold within a short video, rather than merely specifying its entities or final state.

Candidate selection. We retain English prompts containing 8–180 words and observable action or interaction cues, with priority given to descriptions of 25–100 words. We exclude static descriptions, camera-only reveals, extended narratives, command-line parameters, and high-risk content. For example, a camera approaching an already open umbrella does not demonstrate the umbrella opening. Available toxicity-related scores are screened at 0.3, with a 0.2 threshold for sexual explicitness; missing scores do not automatically imply rejection. These filters provide candidates, not certificates of semantic suitability.

Structured augmentation. DeepSeek-V4-Pro converts each candidate into a generation prompt and an evaluation specification. The instructions require a stable setting, an identifiable initial state, one dominant transition, and a visible result. Onset, progression, and outcome must describe successive aspects of the same event. Rewriting may clarify source-supported details but must not introduce unrelated entities, actions, camera movements, or cinematic effects. Unsuitable sources are deferred instead of being forced into the schema. The generation prompt follows *First, . . . Then, . . .*; 45–100 words are recommended. Subsequent question augmentation makes evaluation criteria task-specific while preserving the generation prompt, phase targets, identifier, and split assignment.

Quality control. Automatic validation checks the schema, nonempty specifications, complete phase questions, and the five shortcut criteria. It rejects forbidden camera/style additions and flags tasks requiring long memory, shot changes, or multiple unrelated events. An outcome described only as *still* or *settled* is insufficient unless the task-specific result remains observable. Prompt length outside 45–110 words produces a warning rather than automatic rejection.

Table 6: Construction statistics. The final selection is drawn from validated, deduplicated records. API failures and deferred sources are not treated as accepted tasks.

Stage	Count	Stage	Count
Source prompts	1,038,805	Accepted rewrites	3,501
Rule-selected candidates	12,000	Deferred / failed	2,076 / 413
Rewrite requests	5,990	Validator accepted / rejected	2,935 / 566
After exact deduplication	2,935	Final tasks	2,500

A.2 EVENT FAMILIES, DIVERSITY, AND SPLITS

Action Start concerns initiation of an action; *Entity Interaction* emphasizes contact or coordinated behavior; *Object State Change* concerns an object's form, configuration, or material state; and *Environmental Perturbation* concerns change in the surrounding medium or scene. These boundaries can overlap. The statistics below use rule-based category proposals, not independently adjudicated semantic labels.

Diversity derives primarily from distinct source situations. Source-level semantic deduplication is complemented by exact checks on prompts, entity–transition descriptions, and source identifiers. A bounded text-similarity screen flags potentially redundant tasks at Jaccard similarity 0.86; it is not exhaustive semantic clustering. Thus, construction preserves source variation without treating arbitrary object substitutions as additional independent events.

Figure 5 summarizes the category composition and prompt vocabulary of the full corpus. Action Start is the largest proposed category, whereas Entity Interaction is the smallest. The vocabulary visualization provides a complementary view of recurring entities, actions, and settings; word frequency alone does not establish semantic diversity.

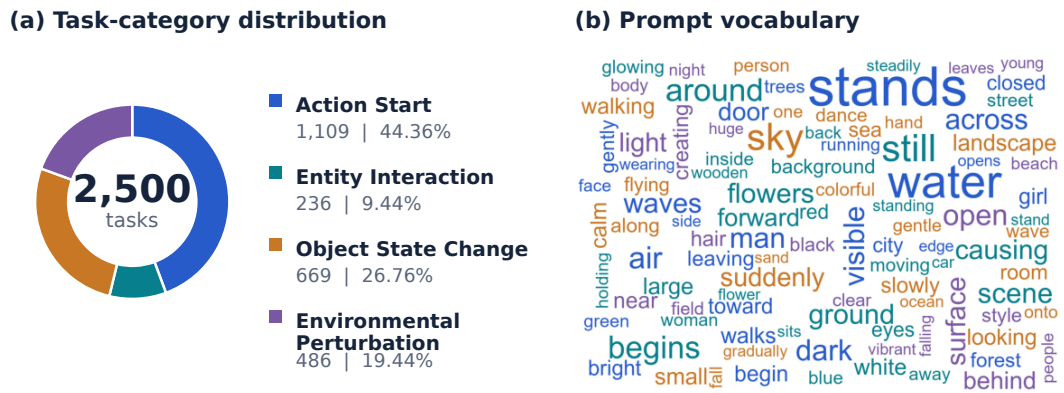


Figure 5: **Category composition and prompt vocabulary of OETGBench.** (a) Counts cover all 2,500 tasks, including the training pool, and reproduce keyword-based category proposals rather than independently adjudicated labels. (b) Word size reflects frequency among the 100 most frequent retained words in the generation prompts, after lowercasing and fixed-stopword removal. Phase targets and evaluation questions are excluded; no stemming or lemmatization is applied. These descriptive summaries are not estimates of evaluation-set failure rates or independently validated semantic coverage.

A.3 Task Representation and Examples

Each task specifies a generation prompt, initial state, target transition, three phase targets, three phase questions, and task-conditioned diagnostic questions. Only the generation prompt is supplied to the generator. Phase targets and questions are evaluator-side annotations. The stored field `development` corresponds to *progression* in the paper. Table 7 gives examples from all four families; the companion JSON contains their complete specifications and unabridged questions.

Table 7: Four generation prompts, reproduced from the task records. Complete phase targets and questions for all four examples accompany the source.

Action Start (TDV3-0856915). **Generation prompt.** A girl with long, flowing gray hair and a black jacket stands in an anime-style landscape with a wide, open field and a distant horizon. First, she is standing still, facing the horizon. Then, she starts walking forward, her hair flowing gently, and she continues walking until she has visibly moved closer to the horizon, with the background perspective shifting slightly.
Entity Interaction (TDV3-0657846). **Generation prompt.** In a simple, colorful illustration style like a first-grade book, Lamu and Alfio are in a sunny park. First, Lamu is standing a short distance away from Alfio, looking at him with a smile. Then, she runs towards him with open arms and finally hugs him tightly, both smiling happily.
Object State Change (TDV3-0367425). **Generation prompt.** The word 'yandex' is sculpted from clear ice, standing on a snowy surface in a winter landscape with gently falling snowflakes. First, the ice text is solid and fully intact, with sharp edges and a frozen appearance. Then, the ice begins to melt slowly, causing the letters to soften, drip, and gradually lose their shape until only a puddle of water remains on the snow.
Environmental Perturbation (TDV3-0105167). **Generation prompt.** A wooden boat with a single sailor sails on a calm sea under a cloudy sky. First, the sea is calm and the boat moves smoothly. Then, the sea rapidly becomes stormy, with large waves crashing against the boat, causing the sailor's face to show worry and concern.

B OETGEVAL: EVALUATION PROTOCOL AND SCORING DETAILS

B.1 VISUAL INPUTS AND QUESTION CONSTRUCTION

OETGEval first characterizes video-level quality and failures, then verifies phase evidence within temporal windows. These operations answer different questions. A coherent-looking video may never complete the event, whereas locally recognizable phase evidence may coexist with identity changes or discontinuities elsewhere. VLD therefore does not replace WPV, and the two outputs are reported separately rather than conflated into a single measure of correctness.

For an 81-frame video at 16 FPS, VLD uses frames $0, 10, \ldots, 80$ arranged in temporal order in a 3×3 contact sheet. WPV uses the four overlapping intervals in Table 8. Each window is represented by its first, middle, and last frame.

Table 8: WPV inputs. Indices are zero-based and endpoints are inclusive. All three phase questions are asked independently for every window.

Window	Start	End	Sampled frame indices
w_0	0	24	0, 12, 24
w_1	16	40	16, 28, 40
w_2	32	56	32, 44, 56
w_3	48	80	48, 64, 80

The generation prompt is provided to VLD as task context, explicitly not as visual evidence. WPV receives the natural-language questions from the task's phase specification, without field-name prefixes. A window can support several phases or none; it is never assigned a phase by taking the largest of its three scores. This retains the complete 4×3 evidence matrix.

B.2 VIDEO-LEVEL DIAGNOSTICS

Table 9 defines the nine criteria. Quality questions are positive: YES supports the requested property. Failure questions have the opposite direction: YES indicates an observed failure. Identity and scene consistency allow changes required by the target event, such as ice losing its original shape during melting. Such a change is not, by itself, identity drift.

Table 9: VLD criteria. The experiments use task-conditioned questions implementing these criteria, not retrospectively substituted generic questions.

Criterion	Evidence assessed
Motion plausibility	Motion is natural and causally compatible with the transition.
Transition continuity	Change unfolds without an abrupt or impossible jump.
Subject consistency	The main entity retains its identity, allowing intended changes.
Background consistency	The scene remains coherent without unrelated changes.
Motion collapse	The video freezes instead of performing the event.
Camera-motion dominance	Viewpoint motion or a reveal substitutes for the event.
Motion misalignment	Visible motion is generic or unrelated to the target transition.
Temporal artifacts	Flicker, noise, or distortions create apparent motion.
Identity drift	Unintended deformation, disappearance, or replacement occurs.

All answers use $\{\text{YES}, \text{PARTIAL}, \text{NO}\} \mapsto \{1, 0.5, 0\}$. Let $q_{\text{motion}}, q_{\text{continuity}}, q_{\text{entity}}, q_{\text{scene}}$ denote the four quality scores, and p_f, p_c, p_g, p_a, p_i the five failure scores in table order. We directly aggregate these scores into two video-level diagnostics:

$$\begin{aligned} \text{VLD Quality} &= 0.30q_{\text{motion}} + 0.20q_{\text{continuity}} + 0.25q_{\text{entity}} + 0.25q_{\text{scene}}, \\ \text{VLD Failure} &= 0.25p_f + 0.20p_c + 0.20p_g + 0.15p_a + 0.20p_i. \end{aligned} \tag{2}$$

Both diagnostics lie in $[0, 1]$. Higher VLD Quality indicates better video-level quality, whereas higher VLD Failure indicates stronger evidence of the specified failure modes. We report both separately from S_{event}; neither contributes to its computation. Motion-based measurements and DINO persistence are evaluated separately and do not enter the aggregates in Equation (2).

B.3 Judge Instructions and Response Validation

The following templates specify the role and decision rules. Bracketed slots are replaced with task text. Exact instantiated prompts and construction instructions accompany the source.

> **VLD instruction.** You are an evaluator for a model-generated video. You will be given the original generation prompt and the full generated video. Your task is to answer the following questions based only on the visible content of the video. The original prompt is provided only as task context. It is not visual evidence.
>
> **Original generation prompt:** [generation prompt]
>
> **Questions:** [four quality questions and five failure questions, as natural-language text]
>
> Answer YES only if the video clearly satisfies the question based on visible evidence. Answer PARTIAL if the evidence is weak, incomplete, ambiguous, or only partly satisfies the question. Answer NO if the evidence is absent, contradicts the question, or is only implied by the prompt. Do not infer events that are not visibly shown. Do not rely only on the text prompt; judge the actual video. For failure-case questions, YES means the failure is present; NO means the failure is not visible. Evaluate the full video globally, not a single frame or a short segment.

> **WPV instruction.** You are an evaluator for a model-generated video segment. You will be given one temporal video segment from the generated video. Your task is to answer the following questions based only on the visible content of this video segment.
>
> **Questions:** [onset question] [progression question] [outcome question]
>
> Answer YES only if this video segment clearly shows visible evidence that satisfies the question. Answer PARTIAL if the evidence is weak, incomplete, ambiguous, or only partly visible in this segment. Answer NO if this segment does not show visible evidence for the question. Judge only what is visible in this segment. The same segment may satisfy more than one question. It is also valid for a segment to satisfy none of the questions. Do not force the segment to belong to exactly one event phase.
>
> **Shared output requirement.** Return valid JSON only. Use the original question text as the key. Each entry contains a label (YES, PARTIAL, or NO) and a brief explanation based on visible evidence.

We validate every requested question, label, and evidence field. Missing questions, unknown labels, and empty explanations are invalid outputs, not NO answers. Format retries retain the first valid response rather than the most favorable score. A video-level result is complete only when all required window questions are available. This separates evaluator coverage from generation success.

B.4 Ordered Matching and Score Aggregation

Let $s_{w,m} \in \{0, 0.5, 1\}$ be the answer for phase $m \in \{\text{on}, \text{prog}, \text{out}\}$ in window w. We compute

$$E_m = \max_w s_{w,m}, \qquad S_{\text{presence}} = 0.25E_{\text{on}} + 0.25E_{\text{prog}} + 0.30E_{\text{out}}. \tag{3}$$

An admissible chain (i, j, k) satisfies $0 \le i \le j \le k < 4$ and has positive evidence at all three selected positions. Among the 20 nondecreasing triples, we maximize

$$s_{i,\text{on}} + s_{j,\text{prog}} + s_{k,\text{out}}. \tag{4}$$

Thus, the implemented search uses equal phase weights and no explicit collapse penalty. Ties prefer more distinct windows, then a larger span $k - i$, and finally the lexicographically earliest triple. The selected chain determines the order score and a categorical cap:

Chain type κ	Condition	Temporal Order	$c(\kappa)$
No valid chain	No admissible triple	0	0.25
Full collapse	$i = j = k$	0.50	0.60
Partial collapse	$i = j < k$ or $i < j = k$	0.75	0.80
Strict ordered chain	$i < j < k$	1.00	1.00

The final score is $S_{\text{event}} = \min(S_{\text{presence}}, c(\kappa))$. Its maximum is 0.80, not 1.00, because the presence weights sum to 0.80. The partial-collapse and strict-chain caps are therefore nonbinding. No valid chain may indicate either a missing phase or positive evidence occurring only in an incompatible order.

Worked examples. For window rows $(0, 0, 0)$, $(1, 0.5, 0)$, $(0.5, 1, 0.5)$, and $(0, 0.5, 1)$, the best chain is $w_1 \to w_2 \to w_3$. All phase maxima are 1, giving Temporal Order 1 and $S_{\text{event}} = 0.8$. If all three phases instead have evidence only at w_1, their maxima remain 1 but full collapse caps the score at 0.6. If onset is visible only at w_3, progression only at w_1, and outcome only at w_0, no admissible chain exists and the cap is 0.25.

These rules verify compatibility of coarse evidence positions, not fine-grained causal order. An all-0.5 matrix selects $(0, 1, 3)$ through tie-breaking and receives Temporal Order 1 despite ambiguous localization; its event score is only 0.4. This counterexample motivates reporting the phase matrix and chain type alongside the scalar order score.

C Experimental Setup and Human Annotation Protocol

C.1 Generation and Evaluation Cohorts

The principal cross-generator cohort contains 50 human-blind tasks and one generation per task from CausVid AR 3-step, Self-Forcing DMD, and CF++ frame-wise 2-step. All 150 videos are decodable, with 81 frames, 16 FPS, and resolution 832×480. Generation prompts and task identifiers are matched across generators. The evaluators are Qwen2.5-VL-7B, Qwen3-VL-8B, and InternVL3.5-8B. Changing the VLM does not change the phase questions or deterministic aggregation rules. Separate cohorts support full VBench aggregates. Completed Qwen2.5 WPV results cover 600 windows and 1,800 answers.

C.2 Human Judgments

The 150-video human reference contains 1,350 phase ratings from three reviewers. Generator identities were anonymized. Each reviewer used the full clip, prompt, and phase descriptions: 1 denotes clear realization, 0.5 partial realization, and 0 absence. Notes identify missing actions, premature outcomes, or ambiguity. Three malformed entries were corrected using the reviewer's explicit clarifications.

For each video and phase, we compute the human reference as the arithmetic mean of the three reviewers' ratings, without rounding or discretization. The human Phase Presence Score is then computed by applying the weighted aggregation in Equation (3) to the three reviewer-averaged phase scores.

Human order reference and AUROC. The temporal-order reference is retrospectively constructed from the three reviewers' phase ratings and written notes. For each video and phase, we use the arithmetic mean of the three reviewers' ratings. A video is labeled positive if all three averaged phase scores are greater than zero and none of the reviewers explicitly reports a violation of the required temporal sequence, such as a skipped necessary process, reversed order, or premature outcome. Otherwise, it is labeled negative. The absence of a reported temporal violation is treated as no veto, not as independent confirmation of correct order. AUROC uses these binary reference labels.

OETGEval evaluator. The OETGEval row uses Qwen2.5-VL-7B under the four-window protocol. Ordered phase matching determines the chain category and its corresponding Temporal Order value, which is used for this comparison rather than the Event Transition Score, S_{event}.

Direct full-video question. Each VLM receives all 81 frames in chronological order, the generation prompt, and the task-specific phase questions. It answers the following shared question:

> Does the video visibly show the specified onset, progression, and outcome unfolding in the required temporal order, without missing a phase, skipping a required process, or presenting the outcome prematurely?

The instructions permit partial realization, provided that all three phases have visible evidence and unfold in the required order. Responses are `YES`, `NO`, or `UNDETERMINABLE`. The first two map to 1 and 0, respectively; undetermined responses remain missing rather than being assigned zero.

Direct-VLM AUROC is computed from these binary predictions, whereas OETGEval uses its graded Temporal Order values.

C.3 Statistical Analysis and Computational Settings

We report Kendall's τ_b to accommodate ties and Spearman's ρ for rank association. Confidence intervals use 5,000 task-cluster bootstrap samples, retaining all three generator videos for each sampled task. Paired differences use the same resampled tasks for both conditions. Intervals are marginal 95% percentile intervals, not multiplicity-adjusted superiority tests. Undefined correlations from constant resamples are not assigned zero.

The controlled perturbation experiment uses Qwen2.5-VL with BF16 inference, SDPA attention, PyTorch 2.6/CUDA 12.4, and Transformers 4.51.3 on an A800 80 GB GPU.

D Evaluator Validation and Robustness

D.1 Human Agreement and Cross-VLM Variation

Table 10 reports phase-wise agreement and its uncertainty. Agreement varies by phase and evaluator: Qwen3-VL and InternVL3.5 have the strongest outcome correlations, while progression remains less reliably aligned. Human presence correlations are moderate ($\rho = 0.395$–0.466), and the overlapping intervals do not support a uniform evaluator ranking. Figure 6 further distinguishes human alignment from cross-VLM agreement.

Table 10: Human agreement with 5,000 task-cluster bootstrap replicates. Each resampled task retains the three generator videos.

Evaluator	Metric	τ_b [95% CI]	ρ [95% CI]
Qwen2.5-VL-7B	Onset	0.319 [0.177, 0.453]	0.350 [0.195, 0.492]
Qwen3-VL-8B	Onset	0.414 [0.283, 0.531]	0.454 [0.311, 0.579]
InternVL3.5-8B	Onset	0.289 [0.145, 0.426]	0.303 [0.152, 0.445]
Qwen2.5-VL-7B	Progression	0.262 [0.075, 0.431]	0.288 [0.080, 0.471]
Qwen3-VL-8B	Progression	0.147 [-0.004, 0.296]	0.167 [-0.004, 0.333]
InternVL3.5-8B	Progression	0.303 [0.133, 0.461]	0.336 [0.148, 0.507]
Qwen2.5-VL-7B	Outcome	0.322 [0.176, 0.460]	0.354 [0.193, 0.505]
Qwen3-VL-8B	Outcome	0.475 [0.353, 0.586]	0.512 [0.381, 0.630]
InternVL3.5-8B	Outcome	0.478 [0.337, 0.599]	0.518 [0.364, 0.649]
Qwen2.5-VL-7B	S_{presence}	0.363 [0.220, 0.497]	0.465 [0.287, 0.623]
Qwen3-VL-8B	S_{presence}	0.301 [0.157, 0.433]	0.395 [0.209, 0.555]
InternVL3.5-8B	S_{presence}	0.358 [0.238, 0.471]	0.466 [0.313, 0.599]

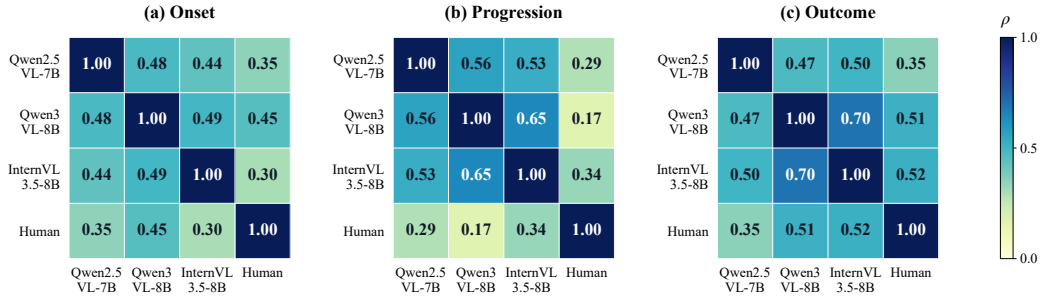


Figure 6: Phase-specific Spearman correlations on the matched 150-video cohort. The human reference is the phase-rating reference described in Appendix C. Agreement among VLMs and agreement with humans are distinct comparisons.

D.2 Controlled Temporal Perturbations

Design. We apply eight conditions to each of the 150 source videos: the original, two technical controls, and five temporal perturbations. Using half-open frame intervals, define $A = [0, 16)$, $B = [16, 40)$, $C = [40, 64)$, and $D = [64, 81)$. Re-encoding preserves $A + B + C + D$; the split–rejoin control partitions and rejoins the same sequence before a single final encode. The perturbations are full reversal, $A + C + B + D$ (middle swap), $A + \text{repeat}(f_{15}, 48) + D$ (deletion and hold), $A + B + B + D$ (repetition), and $A + B + \text{reverse}(C) + D$ (backtracking). All outputs retain 81 frames, 16 FPS, and the source resolution, using H.264 CRF 18 and the medium preset.

The boundaries are fixed temporal partitions, not human-verified phase boundaries. Consequently, an edited clip is not automatically a known semantic violation, especially when the original already fails. The experiment measures score sensitivity to controlled edits rather than an accuracy against synthetic ground-truth order labels.

Coverage and controls. The 1,200 videos require 4,800 window judgments; 4,799 pass validation. One original-video window remains invalid, yielding 1,199 complete scores and 149 sources complete in every condition. All re-encoded and perturbed conditions are complete on 150 sources. On the common 149, original/re-encoded/split–rejoin mean Temporal Order is $0.3121/0.3054/0.2903$, and mean S_{event} is $0.3304/0.3240/0.3168$. Re-encoding changes order by -0.0067 (95% CI $[-0.0422, 0.0267]$), providing no resolved systematic shift from encoding alone.

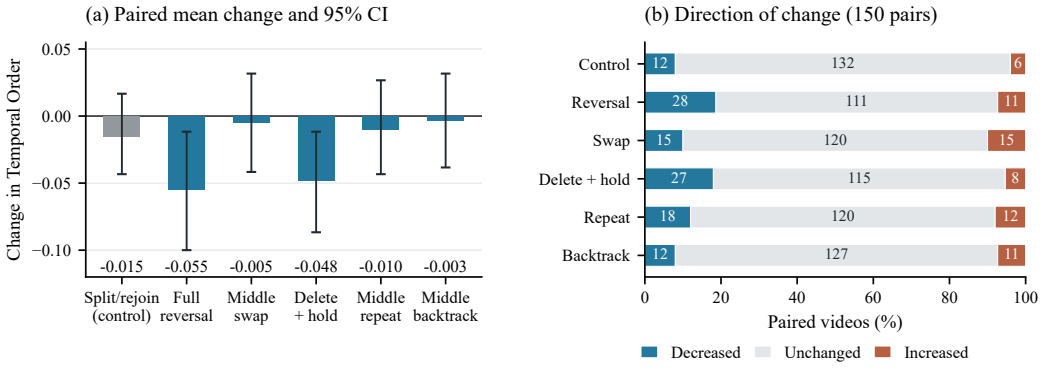


Figure 7: Temporal perturbation sensitivity relative to re-encoding on 150 paired sources. Left: mean changes with task-cluster 95% intervals. Right: numbers of decreased, unchanged, and increased order scores. A large tied fraction limits average sensitivity.

Table 11: Temporal Order under perturbation: 150 matched sources in every row. The re-encoded reference mean is 0.303. Differences include 95% task-cluster intervals.

Condition	Mean TO	Δ vs. re-encode	Δ vs. split–rejoin
Split–rejoin	0.288	-0.015 [-0.043, +0.017]	Reference
Full reversal	0.248	-0.055 [-0.100, -0.012]	-0.040 [-0.092, +0.008]
Middle swap	0.298	-0.005 [-0.042, +0.032]	+0.010 [-0.018, +0.038]
Deletion + hold	0.255	-0.048 [-0.087, -0.012]	-0.033 [-0.073, +0.005]
Repetition	0.293	-0.010 [-0.043, +0.027]	+0.005 [-0.030, +0.043]
Backtracking	0.300	-0.003 [-0.038, +0.032]	+0.012 [-0.023, +0.048]

Results. Full reversal and process deletion reduce Temporal Order relative to re-encoding, whereas swapping, repetition, and backtracking have small mean effects.

E Additional Generator Results and Visual-Consistency Analysis

E.1 Phase Realization Across Event Families

Table 12 pairs automatic and human phase judgments within each family. The small, unequal groups (8–18 prompts) support descriptive comparisons, not definitive domain rankings.

The Action Start results illustrate the distinction between beginning an action and realizing its outcome: human outcome scores remain below onset scores for all three generators. Human ratings also distinguish Self-Forcing from CF++ more strongly on progression in this family. Interaction tasks reveal a different evaluator limitation: visible entities and contact can receive substantial automatic progression credit even when the human regards the intervening action as incomplete. Category-wise agreement should therefore be inspected phase by phase rather than inferred from a single overall average.

Table 12: Category-wise phase means on matched videos. Each entry is automatic / human (Qwen2.5 / the phase-rating reference). Category sizes count prompts per generator.

Family	Generator	Tasks	Onset	Progression	Outcome
Action Start	CausVid	18	0.417/0.694	0.528/0.611	0.250/0.250
	Self-Forcing	18	0.500/0.667	0.556/0.639	0.278/0.444
	CF++	18	0.472/0.472	0.361/0.222	0.111/0.222
Entity Interaction	CausVid	8	0.562/0.688	0.562/0.188	0.625/0.688
	Self-Forcing	8	0.562/0.625	0.500/0.562	0.438/0.625
	CF++	8	0.500/0.188	0.375/0.250	0.375/0.250
Object State Change	CausVid	14	0.571/0.714	0.536/0.500	0.179/0.321
	Self-Forcing	14	0.500/0.857	0.500/0.607	0.143/0.393
	CF++	14	0.500/0.714	0.500/0.500	0.214/0.179
Environmental Perturbation	CausVid	10	0.650/0.450	0.700/0.500	0.500/0.600
	Self-Forcing	10	0.700/0.700	0.700/0.700	0.600/0.550
	CF++	10	0.550/0.350	0.600/0.550	0.600/0.450

Table 13: Chain classifications on the same 150 videos. No valid chain includes absent phase evidence as well as incompatible evidence ordering.

Evaluator	No valid chain	Full collapse	Partial collapse	Strict chain
Qwen2.5-VL	92	13	18	27
Qwen3-VL	121	7	10	12
InternVL3.5	141	0	1	8

The chain counts in Table 13 additionally reveal evaluator-dependent thresholds for recognizing complete phase evidence. A large no-valid-chain fraction should not be interpreted as an equally large frequency of pure order reversals: it includes videos for which at least one phase is missing. This distinction is essential when diagnosing which generation capability remains unresolved.

E.2 Event Realization Versus Visual Consistency

We compute four VBench dimensions on the same 150 videos, thereby holding generation content fixed. Their equal-weight mean, denoted TC here, is a descriptive temporal-consistency aggregate rather than the official VBench Total score:

$$\mathrm{TC} = \tfrac{1}{4}(\mathrm{Subject} + \mathrm{Background} + \mathrm{Flicker} + \mathrm{Smoothness}). \tag{5}$$

CF++ has the highest consistency scores but the lowest event score. Relative to CF++, Self-Forcing improves S_{event} by 0.0635 (95% CI $[0.0165, 0.1150]$) while reducing TC by 0.0227 ($[-0.0310, -0.0146]$). CausVid's event difference is 0.0530 ($[0.0065, 0.1060]$). The event difference between Self-Forcing and CausVid is not resolved by the paired interval.

Table 14: Same-video event and visual-consistency results. All 50 prompts and 150 videos are shared across the measurements. Larger values are preferred in every column.

Generator	S_{event}	Subject	Background	Flicker	Smoothness	TC
CausVid AR 3-step	.3435	.9161	.9388	.9730	.9826	.9526
Self-Forcing DMD	.3540	.8940	.9172	.9665	.9806	.9396
CF++ frame-wise 2-step	.2905	.9409	.9452	.9776	.9856	.9623

Across individual videos, TC and S_{event} have Spearman correlation -0.360 ($[-0.555, -0.142]$). Removing temporal flickering yields -0.369 ($[-0.563, -0.154]$) and preserves the generator ranking. Thus, the observation does not depend solely on flicker being favorable to nearly static videos. It supports evaluating event realization separately from visual consistency, but does not establish a causal trade-off.

Full VBench Total, Quality, and Semantic aggregates use a different prompt suite. They characterize broader generation quality and are not paired measurements of event performance on the human-blind tasks. In particular, a correlation between a full-suite model mean and a human-blind event mean would conflate both content and metric changes.

F AG-P3T: Implementation and Additional Results

F.1 Anchor-relative Feedback and Positive-Advantage Gating

AG-P3T starts from CF++ frame-wise 2-step with zero-output LoRA initialization. A frozen-base video provides a cached reward anchor for each prompt. The policy produces four rollouts with seeds 0–3; the anchor uses seed 0. Rollouts, rewards, and advantages are refreshed for each committed update.

Let $\Delta x_i = x_i - x^a$ denote the difference between policy sample i and its anchor. Let $q_{\text{motion}}, q_{\text{continuity}}, q_{\text{entity}}, q_{\text{scene}}$ denote the four quality scores underlying VLD Quality, and let F_{VLD} denote VLD Failure. Specifically,

$$Q_{\text{VLD}} = 0.30q_{\text{motion}} + 0.20q_{\text{continuity}} + 0.25q_{\text{entity}} + 0.25q_{\text{scene}}. \tag{6}$$

The training score uses the Event Transition Score S_{event}:

$$\begin{aligned} r_i = {} & \Delta S_{\text{event},i} + 0.1\Delta E_i^{\text{on}} + 0.2\Delta E_i^{\text{prog}} + 0.4\Delta E_i^{\text{out}} + 0.3\Delta O_i \\ & - 0.5\big[-0.6\Delta q_{\text{motion},i} - 0.4\Delta q_{\text{continuity},i} - 0.02\big]_+ \\ & - 0.4\big[-0.5\Delta q_{\text{entity},i} - 0.5\Delta q_{\text{scene},i} - 0.02\big]_+ \\ & - 0.8\big[\Delta F_{\text{VLD},i} - 0.01\big]_+, \end{aligned} \tag{7}$$

where $[z]_+ = \max(z, 0)$. The quality penalties separately constrain motion and continuity, and entity and scene preservation; they are not a single penalty on the aggregate VLD Quality. Scores are centered within the four-sample group, divided by the population standard deviation with a numerical floor, and clipped to $[-1.5, 1.5]$. A positive advantage is retained only if $\Delta S_{\text{event},i} > 0$, either $\Delta E_i^{\text{out}} \geq 0$ or $\Delta O_i \geq 0$, $F_{\text{VLD},i} \leq F_{\text{VLD}}^a + 0.03$, and neither severe freezing nor a premature outcome is detected. Failed positive advantages become zero; negative advantages are retained. There is no subsequent recentering.

For the linear component L of the training score, a shared anchor offset cancels under group centering:

$$(L_i - L^a) - \overline{(L - L^a)} = L_i - \bar{L}. \tag{8}$$

The anchor therefore affects this implementation through the nonlinear penalties and eligibility conditions, not through the centered linear offset alone. Gating restricts individual samples' positive gradient contribution; it does not inspect and accept or reject the resulting model update. Across 20 committed groups, 34 of 80 advantages are initially positive; 29 remain positive after gating, with 46 negative and 5 zero.

F.2 Optimization and Interpretation

Only LoRA parameters are optimized. The objective is score-gradient target-shift regression, not an explicit policy-ratio GRPO objective. For a captured transition with next state a, noise scale σ, and predicted mean μ_θ, its form is

$$g = \widetilde{A}\,\frac{a - \mathrm{sg}(\mu_\theta)}{\sigma^2 + \epsilon}, \qquad y = \mathrm{sg}(\mu_\theta) + \eta\ \mathrm{clip}(g),$$
$$\mathcal{L} = \mathrm{MSE}(\mu_\theta, y) + \beta_{\mathrm{tr}}\,\mathrm{MSE}(\mu_\theta, \mu_{\mathrm{tr}}). \tag{9}$$

Here sg denotes stop-gradient and μ_{tr} is the stored current-policy rollout mean. It is not the frozen-base prediction or a base-model KL constraint. Episode advantages are broadcast to selected transitions, rather than separately assigned to onset, progression, and outcome.

Table 15: Settings of the post-training run used in the matched re-evaluation. The transition selection retains the earliest five captured rows per rollout, not five video frames.

Setting	Value	Setting	Value
Policy group size	4	Optimizer steps	400
LoRA rank / alpha	8 / 16	Learning rate	10^{-6}
Target shift η	10^{-4}	Trace weight β_{tr}	0.02
Gradient clip	1.0	Score-gradient clip setting	0.25
Transition rows per step	20	Checkpoint contents	LoRA only

The differentiable forward does not reconstruct the complete autoregressive history and positional state of the rollout. This limits its interpretation as an exact gradient of the original conditional transition kernel. The run is a limited post-training study, not evidence of realized phase-local credit assignment or exact likelihood-ratio optimization.

G Qualitative Examples and Evaluator Failure Cases

Qualitative examples are useful for locating disagreements that averages conceal, but selected clips cannot estimate failure frequency. Table 16 complements the videos in Figure 4 with their phase ratings. We use the original three-level labels rather than selecting a binary recoding for each example.

Table 16: Scores for the twelve videos corresponding to Figure 4, using original three-level judgments. Selected-case evidence does not estimate category-wide performance.

Case / generator	Human on/prog/out	Qwen2.5 on/prog/out
Molten iron / CausVid AR 3-step	0/0.5/0.5	1/1/0.5
Molten iron / Self-Forcing DMD	1/1/1	0.5/1/0.5
Molten iron / CF++ frame-wise 2-step	0.5/0.5/0.5	0.5/1/0.5
Coastal wave / CausVid AR 3-step	1/1/1	1/1/1
Coastal wave / Self-Forcing DMD	1/1/0	1/1/1
Coastal wave / CF++ frame-wise 2-step	1/1/0.5	1/1/1
Glowing stone / CausVid AR 3-step	0/0/0	0/0/0
Glowing stone / Self-Forcing DMD	0/0/1	0/0/0
Glowing stone / CF++ frame-wise 2-step	0/0/0	0/0/0
Girl rising / CausVid AR 3-step	1/0.5/0.5	0/0/0
Girl rising / Self-Forcing DMD	1/1/0.5	0.5/0.5/0
Girl rising / CF++ frame-wise 2-step	0.5/0.5/0	0/0/0

Incomplete transitions despite visible activity. For the coastal-wave case, Qwen2.5 gives full credit to all phases for every generator. The human distinguishes CausVid's complete outcome from Self-Forcing's absent outcome and CF++'s partial outcome. This contrast illustrates how conspicuous motion can support an overly generous outcome judgment. The aggregate score should not obscure which requested result was actually verified.

Outcome appearance is not a complete process. In the glowing-stone case, the human gives Self-Forcing an outcome score of 1 but zero onset and progression. This is precisely why a visible final condition cannot substitute for transition completion. The remaining two generators receive zero throughout from both the human and Qwen2.5. These examples motivate retaining phase-specific evidence and examining temporal order, while the controlled tests in Appendix D establish how much confidence the current matching procedure warrants.

G.1 Window-Level Views of Annotated Failures

Figures 8–12 show five selected clips with their task-specific phase questions, sampled frames, and archived evaluator summaries. Each row represents one overlapping window, not a preassigned event phase. The frames reproduce the first, middle, and last sampling positions in Table 8; they are reconstructed from the source videos rather than copied from the archived JPEG inputs.

The cases distinguish failure to depict a prerequisite from failure to establish an outcome. In the pool and molten-iron examples, a salient visual effect appears without the requested action being fully demonstrated. In the flight example, the human attributes apparent progress to camera movement. The remaining examples expose evaluator limitations: a strict chain can be retrieved despite reported identity discontinuity or an unfinished action.

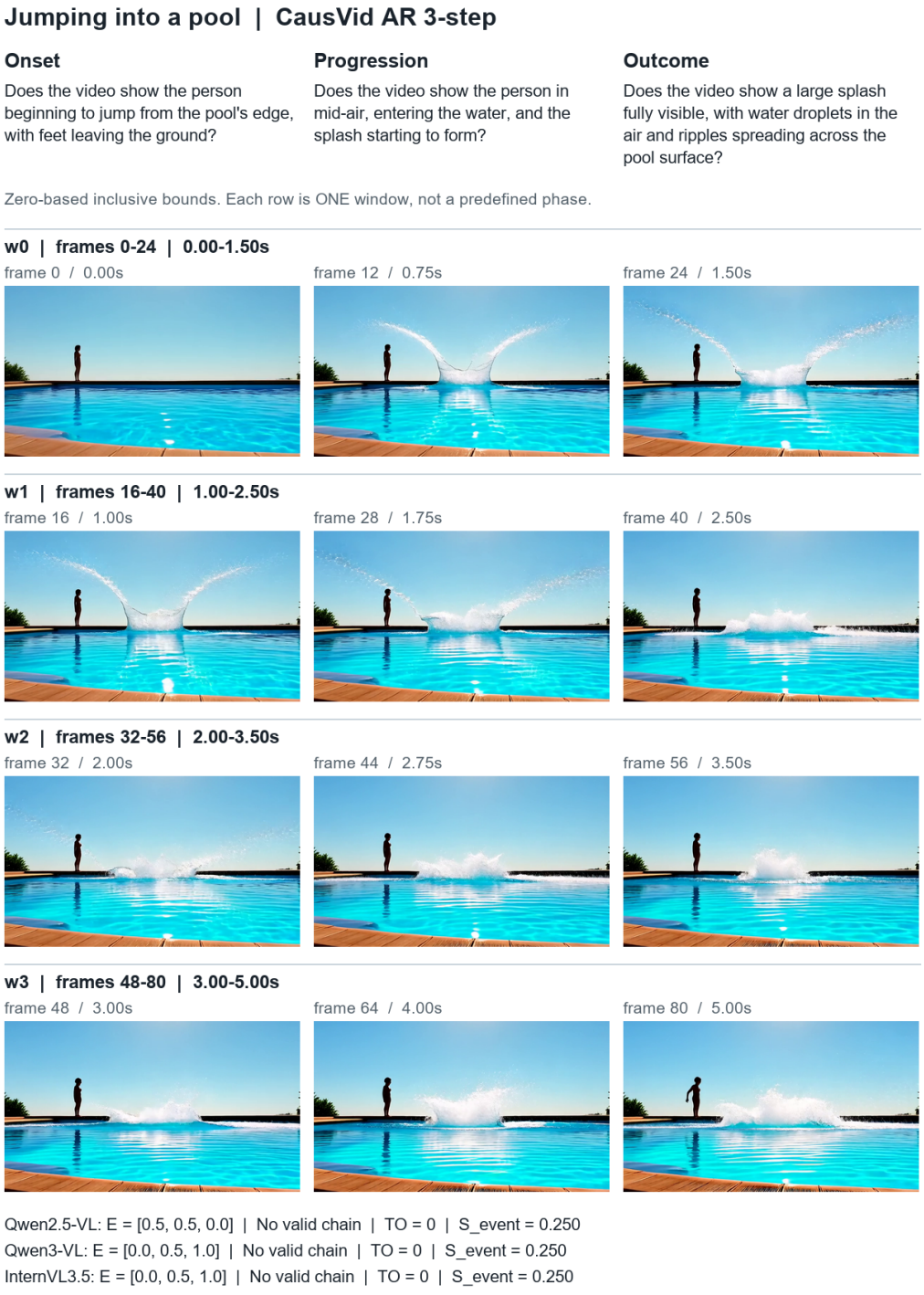


Figure 8: **A visible consequence without the requested initiating action.** CausVid's pool-jump video is annotated as $(0, 0, 1)$: a large splash appears while the required jump is not established. All three evaluators return no valid chain and $S_{\text{event}} = 0.25$, although their phase judgments differ. The example illustrates why a visible outcome cannot alone establish transition completion.

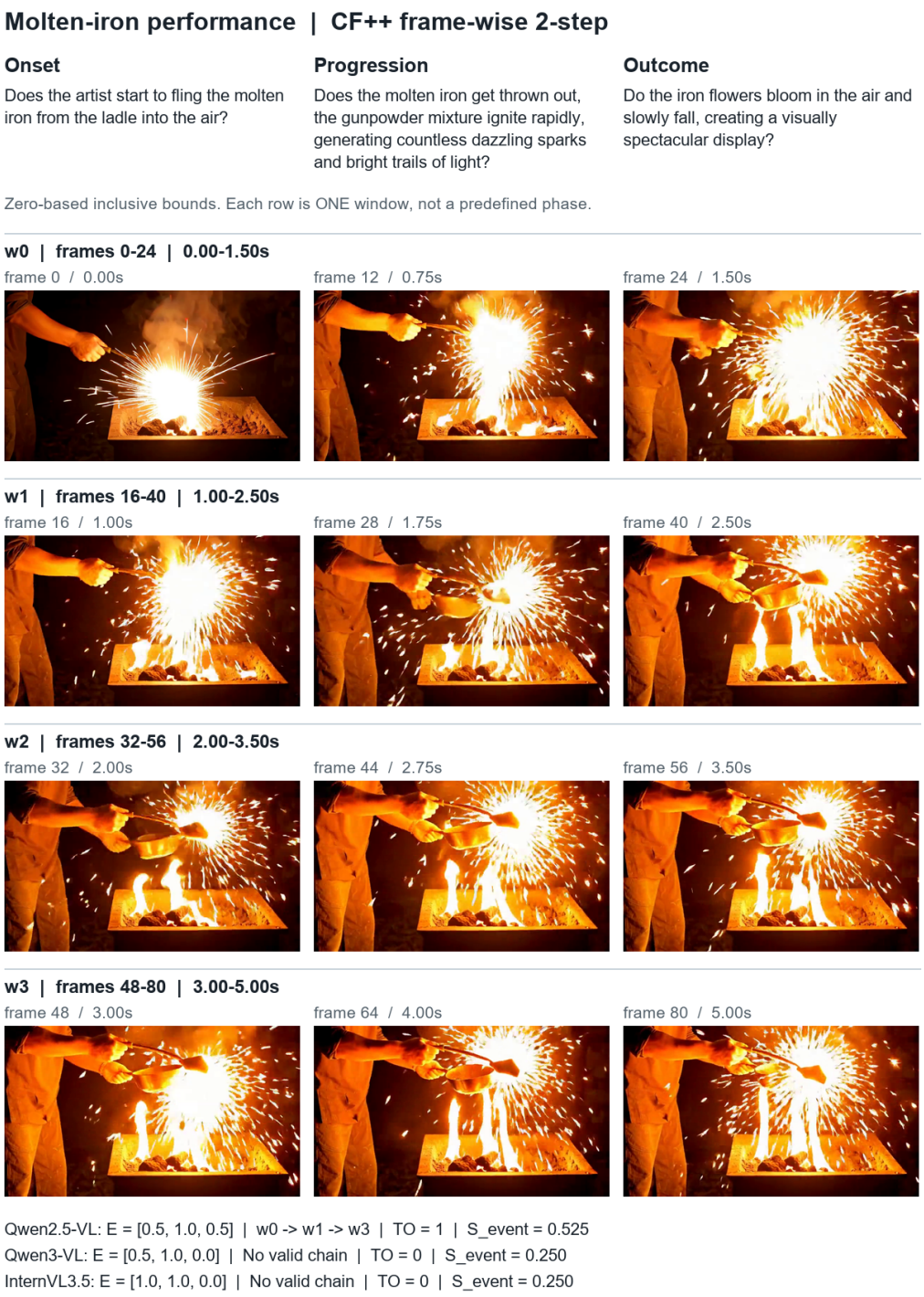


Figure 9: **Persistent effects can obscure a missing transition.** In CF++'s molten-iron video, the human assigns $(0.5, 0.5, 0.5)$ and notes that sparks are already present without a clear throwing action. Qwen2.5 retrieves a strict chain, whereas Qwen3 and InternVL3.5 find no valid chain. Repeated evidence across overlapping windows can therefore support an ordered match without demonstrating the intended process.

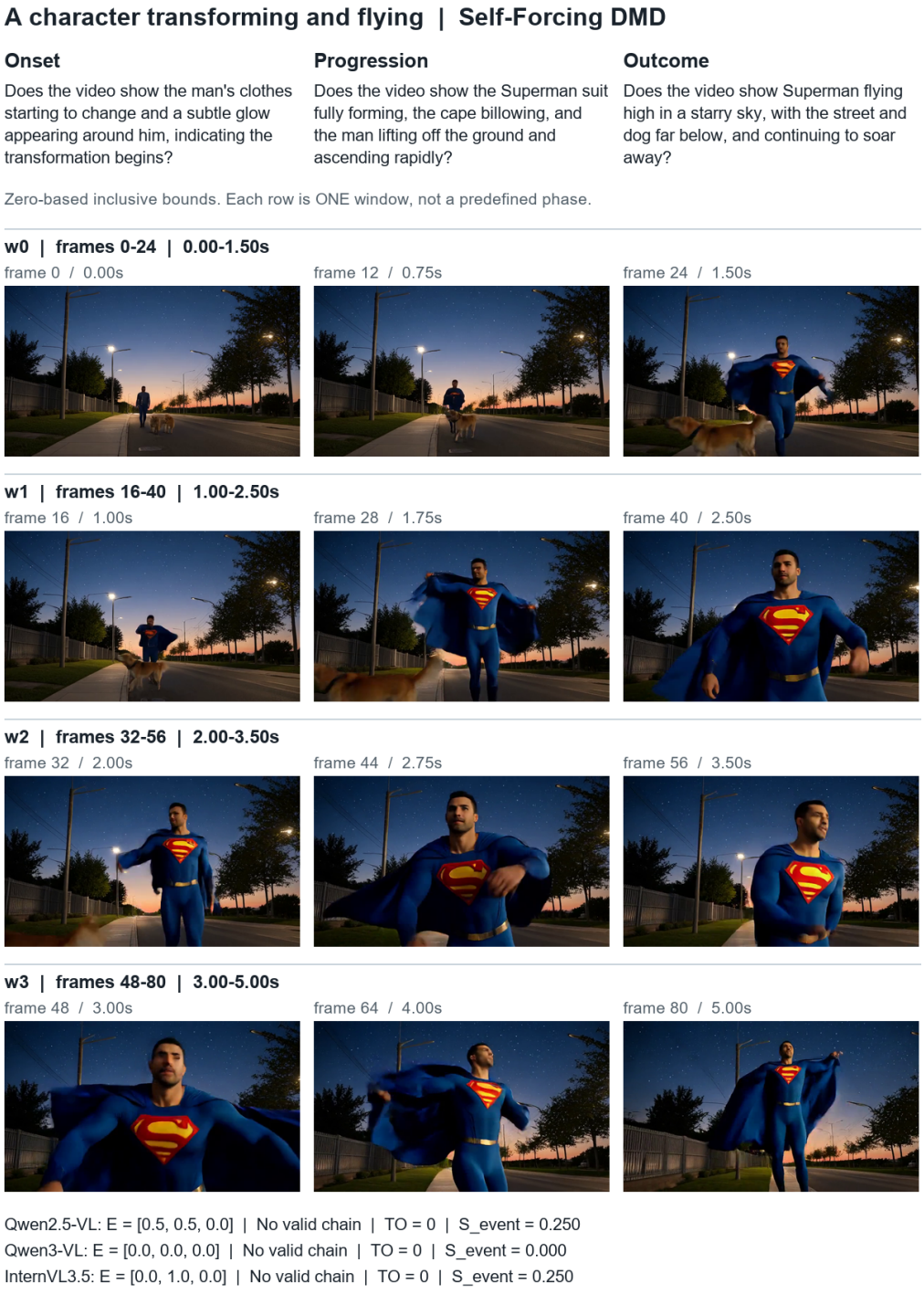


Figure 10: **Apparent progress attributed to camera movement.** Self-Forcing's transformation-and-flight video receives human scores $(1, 0.5, 0)$; the annotation attributes apparent flight to camera movement rather than the requested ascent. All three evaluators find no valid chain, while disagreeing on onset and progression. This case motivates distinguishing task-relevant action from other sources of visible change.

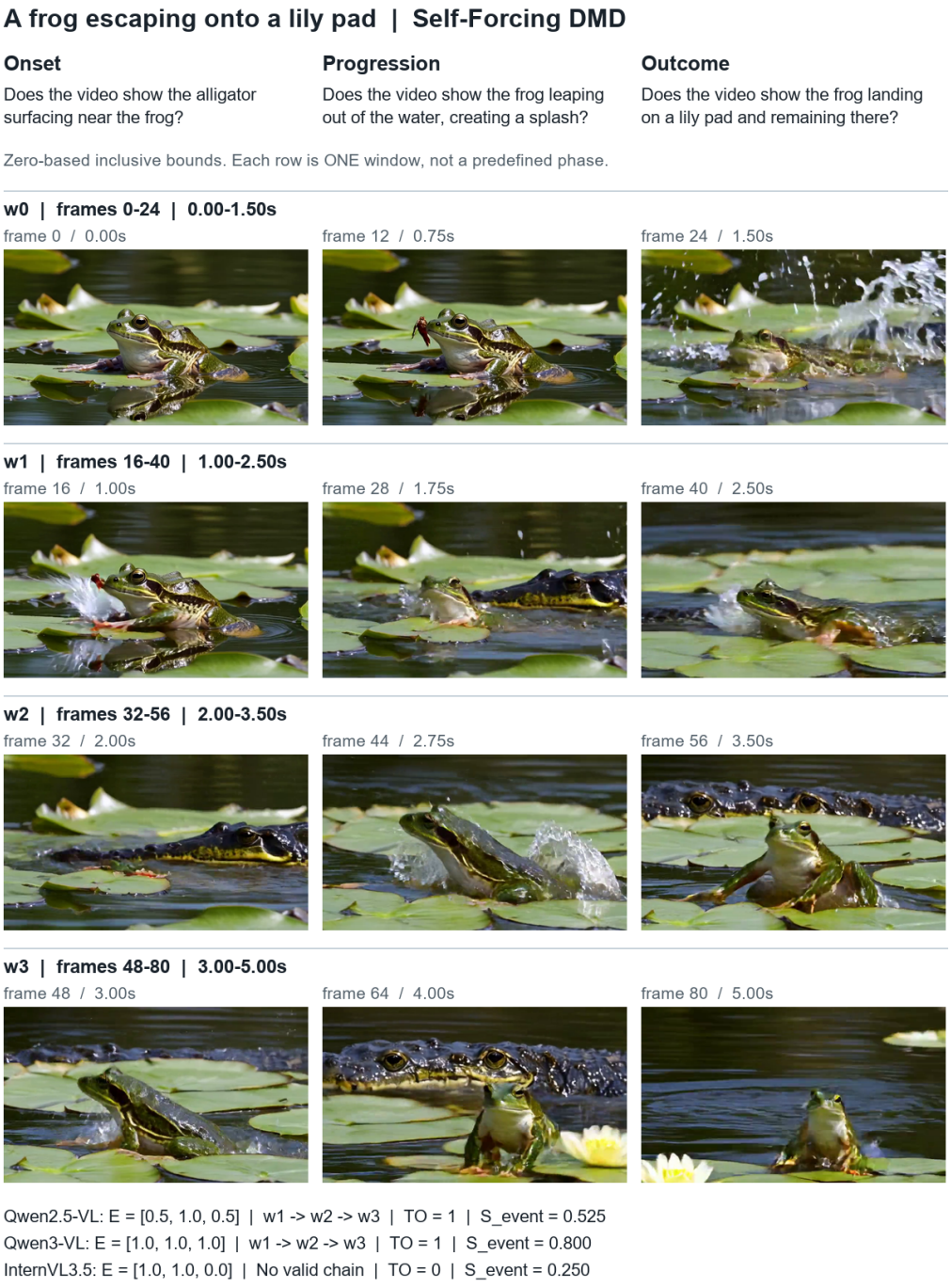


Figure 11: **An ordered match does not ensure identity continuity.** The human scores Self-Forcing's escape video as $(1, 0.5, 0)$ and reports identity drift, teleportation, and discontinuous motion. Qwen2.5 and Qwen3 nevertheless retrieve strict chains, with event scores of 0.525 and 0.800; InternVL3.5 finds no outcome evidence. The discrepancy shows why phase matching and video-level diagnostics should be interpreted jointly.

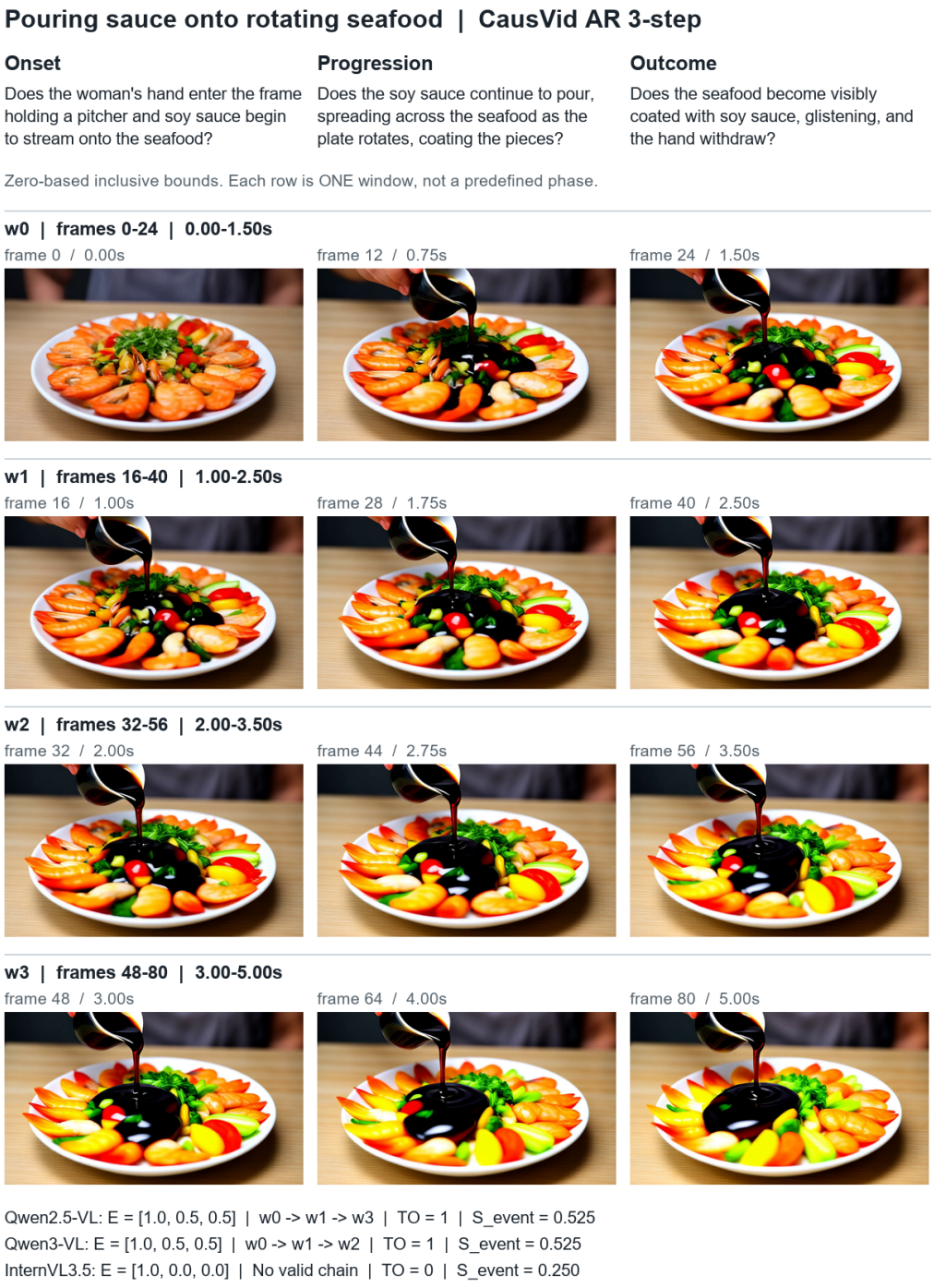


Figure 12: **Continued action without a completed outcome.** CausVid's sauce-pouring video is annotated as $(1, 1, 0)$ because pouring continues without the requested conclusion. Qwen2.5 and Qwen3 assign partial outcome credit and retrieve strict chains; InternVL3.5 finds no valid chain. Visible activity across successive windows is thus insufficient to establish that the target outcome has been reached.

H LIMITATIONS, DATA RELEASE, AND REPRODUCIBILITY

Scope and annotation. OETGBench focuses on short, visible transitions, not long-term planning or proof of physical causality. Automatic curation and overlapping rule-based categories limit claims of balanced coverage.

Temporal resolution and evaluator dependence. Sparse overlapping windows can miss brief changes or spread the same evidence across several positions. A strict chain need not establish clearly separated phases. The controlled tests and cross-VLM variation therefore motivate reporting evidence matrices and uncertainty, not treating the scalar order score as ground truth.

Companion material. The source package provides task examples, construction/evaluation prompts, split statistics, result tables, perturbation summaries, and ordered-matching code with worked-example tests. Figures use standard plotting software.