STREAMACT: STREAMING OR JUST ANSWERING? DIAGNOSING THE STREAMING INTELLIGENCE ILLUSION IN VIDEO-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Are video-language models truly capable of streaming video understanding, or do they merely answer questions from a given causal context? Many streaming-like evaluations provide a causal prefix or recent window and score the resulting answer at an evaluator-chosen moment. This can make models appear streaming-capable without testing whether they autonomously maintain relevant state, judge evidence sufficiency, or choose appropriate actions as the stream unfolds. We call this mismatch the **Streaming Intelligence Illusion**. To diagnose it, we introduce **StreamAct**, which evaluates streaming video understanding as a strict-causal online trajectory over state, readiness, action, and answer. StreamAct-Bench evaluates 5,436 replayable trajectories under a strict-causal protocol. Target-time answerability alone does not establish this behavior: Qwen3-Omni-30B-A3B achieves 45.02% target-time answer accuracy, yet only 6.68% of its first self-timed responses are timely and correct. To test whether this missing evidence-to-action conversion is learnable, StreamAct-SFT provides explicit supervision over state, readiness, and action, selectively improving these trajectory-level behaviors beyond answer-only and format-only training. These results establish the distinction between streaming and just answering: **Streaming means maintaining state and choosing the right action as evidence unfolds.**

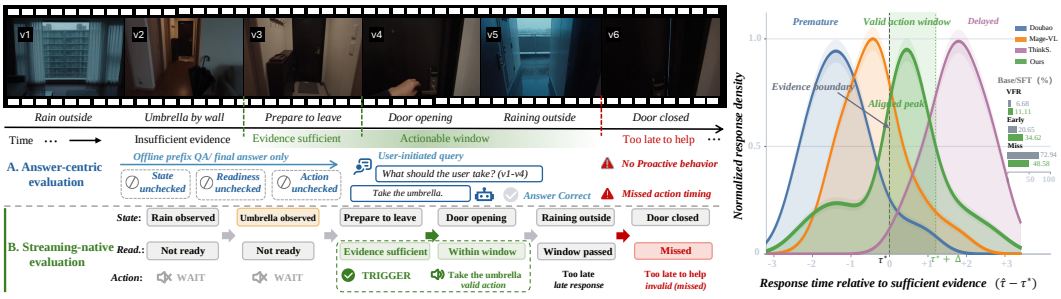


Figure 1: **From Causal Answerability to Streaming Behavior.** Answer-centric evaluation (A) scores responses at evaluator-chosen times, without testing whether models autonomously maintain state, judge evidence sufficiency, and select appropriate actions. Streaming-native evaluation (B) instead assesses the coupled state–readiness–action–answer trajectory as the stream unfolds. Right: premature, delayed, and missing responses expose the Streaming Intelligence Illusion despite target-time answerability; "ThinkS." denotes ThinkStream Liu et al. (2026).

1 INTRODUCTION

Current streaming video models can answer from a causal prefix or recent window, but real-time assistance demands more than reporting what was observed. Consider the AI-glasses assistant in Figure 1: recognizing rain and an umbrella by the door helps only if it reminds the user before they leave. The same answer after that opportunity is correct but no longer useful. A streaming assistant

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.

Protocol	Representative Works	Causal	State	Ready	Action	Ans.
(A) Delayed QA-style protocols						
Offline QA	Video-MME (Fu et al., 2025)	✗	✗	✗	✗	✓
(B) Streaming-like QA protocols						
Windowed QA	StreamingBench (Lin et al., 2026)	✓	✗	✗	✗	✓
Prefix QA	OVO-Bench (Niu et al., 2025)	✓	△	△	△	✓
Evidence-ready	StreamReady (Azad et al., 2026)	✓	✗	✓	△	✓
(C) Streaming-native diagnosis						
Ours	**StreamAct-Bench**	✓	✓	✓	✓	✓

Table 1: Comparison of evaluation protocols. Symbols denote explicit (✓), partial (△), or absent (✗) evaluation; "*Action*" includes waiting, silence, triggering, and answering.

must maintain relevant state and decide whether to wait, remain silent, or act as evidence unfolds. This raises a basic question: *Do video-language models understand an unfolding stream, or merely answer questions from a given causal context?*

These behavioral capabilities are largely absent from current streaming-like evaluation. Many protocols provide a recent window or causal prefix up to time t and score an answer at an evaluator-chosen moment. This prevents direct future leakage but supplies one decision that a streaming model should make for itself: whether current evidence warrants a response. Causal access determines what a model may observe; it does not evaluate what it retains, how it judges evidence sufficiency, or which action it selects. We therefore distinguish *causal answerability*, which measures whether an answer can be recovered from supplied context, from *streaming behavior*, which concerns how a model maintains state, assesses readiness, and controls action as that context evolves.

Streaming evaluation spans temporal access and behavioral requirements. Full-video QA measures answer recovery after observation. Windowed and causal-prefix protocols restrict models to evidence (Lin et al., 2026; Niu et al., 2025; Yang et al., 2026e); evidence-aware protocols probe sufficiency (Azad et al., 2026). Yet these protocols do not necessarily evaluate the same object. Table 1 shows that existing protocols primarily expose answerability or readiness, whereas streaming-native evaluation makes state, readiness, action, and output observable across the full trajectory.

We call the mismatch between answerability under evaluator-supplied context and autonomous, coherent behavior over an unfolding stream the **Streaming Intelligence Illusion**. Section 2 isolates its signature with a controlled paired diagnosis: video, question, target answer, and causal evidence remain fixed; only who decides when to respond changes. This controlled intervention shows that recovering the right answer from sufficient evidence does not establish coherent behavior as that evidence unfolds.

To expose the behavior, we introduce **StreamAct**, which models streaming video understanding as strict-causal decisions over four coupled variables: *state*, *readiness*, *action*, and *answer*. The trajectory itself becomes the prediction: what the model maintains, whether evidence is sufficient, whether it waits, remains silent, triggers proactively, or answers, and what it emits. **StreamAct-Bench** evaluates these trajectories under strict-causal replay; **StreamAct-SFT** turns the same variables into process supervision to test whether the diagnosed evidence-to-action failure is learnable.

Our experiments progress from separation to diagnosis and intervention. Relative to the similarly sized Qwen3-VL, Qwen3-Omni (Xu et al., 2025) raises target-time answer accuracy from 38.91% to 45.02%, while its four streaming components remain nearly unchanged. Across models evaluated on 5,436 trajectories, native benchmark rankings do not transfer to streaming behavior. Diagnostics then locate the clearest observed gap in converting accumulated evidence into valid autonomous responses, while matched process supervision tests whether this behavior is learnable.

In summary, our contributions are:

- We reveal the **Streaming Intelligence Illusion**: models that answer correctly from evaluator-supplied context can still fail to behave coherently and autonomously as the stream unfolds, despite access to the same causal evidence.

- We introduce **StreamAct** and StreamAct-Bench, a strict-causal framework and benchmark that evaluate state maintenance, evidence readiness, action selection, and answer generation over complete trajectories rather than only at a target moment.
- We show that native benchmark rankings do not reliably predict streaming behavior, localize the largest measured gap to evidence-to-action conversion, and demonstrate that explicit process supervision with StreamAct-SFT partially improves this behavior.

2 The Streaming Intelligence Illusion

2.1 Causal Access Does Not Imply Online Behavior

Causal access is necessary for streaming evaluation, but it is not a behavioral specification. It determines the information available at time t, not whether the model has preserved relevant history, recognized sufficient evidence, or selected an appropriate action. A standard causal-prefix protocol evaluates:

$$Video_{[0:t]} \longrightarrow Answer_t, \qquad t \text{ chosen by the evaluator.} \tag{1}$$

This establishes *causal answerability*: the target answer is recoverable from the observed past. Streaming behavior instead asks whether the model can transform causal observations into coherent state updates and appropriate actions on its own.

Crucially, model-selected response time is not a standalone target, but an outcome of the evolving action trajectory. Models given identical causal evidence may produce the same final answer yet behave differently by forgetting prior evidence, acting without support, or failing to remain silent when no response is warranted.

Final-answer correctness cannot identify this process. The same answer may be guessed before its supporting event, emitted after an actionable opportunity, or produced while the model carries a stale state; conversely, a correct online policy may appropriately defer a difficult answer. A streaming-native protocol must therefore evaluate behavior throughout the trajectory: update state as evidence arrives, withhold unsupported actions, remain silent when no response is warranted, and act appropriately when task conditions are met. The defining shift is from restricting *what the model can see* to evaluating *how it behaves as evidence unfolds* throughout the complete online trajectory.

2.2 A Minimal Signature of the Illusion

We call the discrepancy between answerability under evaluator-selected contexts and behavior under an unfolding stream the **Streaming Intelligence Illusion**. To isolate it without relying on our benchmark, we construct a paired evidence-boundary subset from OVO-Bench (Niu et al., 2025) and StreamingBench (Lin et al., 2026): 1,343 and 3,852 examples, respectively. For each of the 5,195 paired examples, the video, question, target answer, and causal evidence are fixed. The evaluator-timed condition supplies sufficient evidence; self-timed replay makes the model control its response, isolating one necessary but not exhaustive aspect of online behavior. A self-timed response is valid only if its first emission is both supported and semantically correct.

Figure 2 shows a consistent separation between evaluator-timed answerability and self-timed validity. Panel (a) reports a model-level gap of 58.20–70.95 percentage points; its format-matched oracle preserves answerability for Doubao and Qwen3-Omni, ruling out output format as the main cause. Panel (b) reproduces the gap across tasks and benchmarks rather than only in the aggregate. Panel (c) restricts the analysis to originally correct examples and measures how little performance survives under self-directed replay. Together, these views isolate a broad paired-protocol effect while keeping the video, evidence, and answer targets fixed.

This experiment establishes that answerability and online behavior are distinct, but timing outcomes alone do not explain why online behavior fails. Premature, delayed, or missing responses may arise from stale state, misjudged evidence, an inappropriate action, or an incorrect answer. StreamAct makes these sources observable through four coupled variables: *state*, *readiness*, *action*, and *answer*.

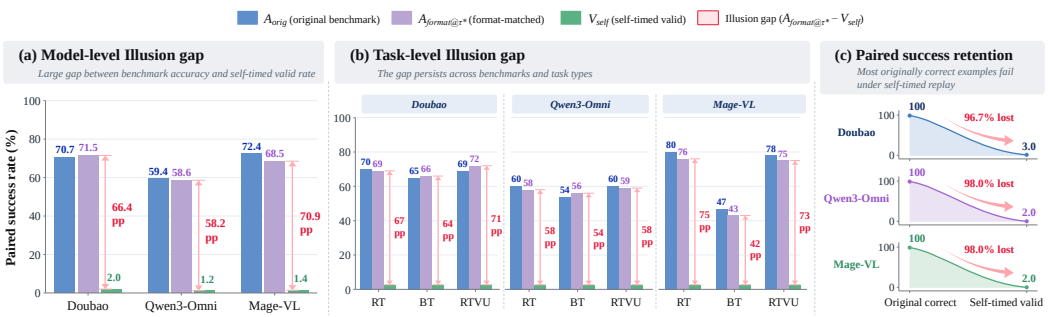


Figure 2: **The Streaming Intelligence Illusion.** (a) contrasts evaluator-timed answer accuracy with valid self-timed responses. (b) breaks this gap down by task. (c) measures how much originally correct performance survives self-directed replay. Oracle timing controls for output-format effects.

3 StreamAct: Streaming as Online Decision-Making

3.1 A Causal Stateful Interface

StreamAct treats streaming intelligence as coherent behavior over an unfolding trajectory, not answer generation at a supplied time. Let $O = (o_1, \ldots, o_T)$ be an ordered video stream and q a query, standing instruction, or task objective. At time t, the model uses only information available so far:

$$\mathcal{F}_t = \sigma(q, o_1, \ldots, o_t), \qquad o_{>t} \notin \mathcal{F}_t. \tag{2}$$

This causal constraint is necessary but does not by itself specify how the model should behave. We model streaming understanding as an online policy that carries task-relevant state forward and converts each new observation into readiness, action, and response:

$$(s_t, r_t, a_t, y_t) = \pi_\theta(o_t, s_{t-1}, q), \qquad \mathcal{A} = \{\textsc{Wait}, \textsc{Silence}, \textsc{Trigger}, \textsc{Answer}\}. \tag{3}$$

Here s_t is the explicit task-relevant state after chunk o_t: what the model carries forward. This inspectable record is not its latent hidden state. Readiness r_t estimates whether that state contains sufficient evidence, a_t selects behavior, and y_t is empty unless that behavior emits. Conceptually, the policy follows:

$$s_{t-1}, o_t \longrightarrow s_t \longrightarrow r_t \longrightarrow a_t \longrightarrow y_t, \tag{4}$$

Without requiring separate architectural heads, the action labels separate distinct behavioral reasons. Wait postpones a pending decision because evidence is still insufficient, whereas Silence is the correct non-emission behavior when no response is warranted. Trigger denotes a self-initiated response, while Answer denotes a response to an explicit query. Readiness alone does not determine action: the task and interaction context distinguish waiting from warranted silence and proactive triggering from answering.

Because the model selects an action at every step, response timing is induced by the action trajectory rather than supplied by the evaluator:

$$e_t = \mathbb{I}\left[a_t \in \{\textsc{Trigger}, \textsc{Answer}\}\right], \qquad \hat{\tau} = \inf\{t : e_t = 1\}. \tag{5}$$

Thus $\hat{\tau}$ is a stopping time with respect to $\{\mathcal{F}_t\}_{t=1}^{T}$, not a prefix chosen retrospectively by the evaluator. The variables decompose streaming behavior: state exposes what is maintained, readiness whether evidence supports action, action how the model behaves, and answer what it communicates when emitting.

3.2 From Valid Responses to Valid Trajectories

Let $E_t \in \{0, 1\}$ indicate whether observations by time t provide sufficient evidence for a task-warranted action, and define the first evidence-sufficient time as

$$\tau^\star = \inf\{t : E_t = 1\}, \qquad \tau^\star = \infty \text{ if no sufficient evidence appears.} \tag{6}$$

Readiness is therefore not confidence in an answer string; it estimates whether the current carried state warrants an action under the task objective (Guo et al., 2017; Geifman & El-Yaniv, 2017):

$$r_t \approx \mathbb{P}(E_t = 1 \mid s_t, q). \tag{7}$$

A model may guess the eventual answer before its supporting event, but that response is not grounded in the stream. For a positive single-event trajectory requiring emission, the first response is streaming-valid only if evidence-aligned and semantically correct:

$$\tau^\star \le \hat{\tau} \le \tau^\star + \Delta, \qquad \mathrm{Correct}(y_{\hat{\tau}}, q, o_{\le \hat{\tau}}) = 1. \tag{8}$$

Here Δ defines the actionable response window, as illustrated by the response-time distributions in the right panel of Figure 1. For a negative trajectory, where no sufficient evidence appears and $\tau^\star = \infty$, valid behavior requires non-emission:

$$e_t = 0, \quad \forall t. \tag{9}$$

These conditions separate premature, delayed, missed, false-triggered, and semantically incorrect responses, but timing and response validity are only projections of the underlying behavior. The complete prediction is the behavioral trajectory:

$$\mathcal{T}_\theta = \left\{(s_t, r_t, a_t, y_t)\right\}_{t=1}^{T}. \tag{10}$$

A valid trajectory should preserve and revise task-relevant state as evidence unfolds, calibrate readiness to that evidence, select contextually valid actions including non-emission, and ground every emitted response in the observations available at that time. StreamAct therefore evaluates not merely whether a model can eventually recover an answer, but whether it can sustain coherent online behavior while the world is unfolding. Operational action semantics are specified in Appendix A.1; multi-event validity and component metrics are defined in Appendix A.2.

4 Operationalizing Streaming-Native Behavior

Section 3 defines streaming understanding through the joint evolution of state, readiness, action, and answer. We now turn this definition into a data object that is both evaluable and trainable. The same replayable trajectory schema underlies **StreamAct-Bench**, comprising 5,436 evaluation trajectories and 91,839 causal decision nodes, and **StreamAct-SFT**, comprising 60K source-video-disjoint training trajectories. The former evaluates streaming-native behavior, while the latter supervises the same behavior through structured causal turns.

4.1 From Formulation to Replayable Trajectories

The shared unit is a replayable trajectory rather than a static video–question–answer tuple. Each trajectory contains a temporally ordered stream, a query or standing objective, and chunk-level targets aligned with Section 3:

$$\mathcal{T} = \{(o_t, s_t^\star, r_t^\star, a_t^\star, y_t^\star)\}_{t=1}^{T}, \qquad \{\mathcal{W}_k^+\}_{k=1}^{K},\ \mathcal{W}^-. \tag{11}$$

Here $(s_t^\star, r_t^\star, a_t^\star, y_t^\star)$ specifies the reference state, evidence sufficiency, valid action, and response at step t. The trajectory is accompanied by positive response windows $\{\mathcal{W}_k^+\}_{k=1}^{K}$ and negative intervals $\mathcal{W}^-$, denoting when emission is warranted or must be suppressed. During primary replay, the evaluator reveals only o_t, the objective, and the model's own previously predicted state s_{t-1}. Reference states are used for scoring, not supplied as input. This makes state maintenance, evidence judgment, action selection, and response generation jointly observable under a common causal interface.

4.2 StreamAct-Bench

StreamAct-Bench instantiates trajectories requiring causal answering, evidence-aware waiting, proactive triggering, state revision, audio-visual grounding, and silence under negative or distracting evidence. Together, these task families exercise the coupled variables in Section 3: state revision

tests what is carried forward, evidence-aware waiting tests appropriate deferral, triggering and silence test action control, and grounded answering tests emitted content.

The benchmark is built through a staged media-grounded pipeline. Candidate videos are first decoded into chunk timelines, and each retained chunk is annotated only from evidence available by that point. Event and state tracks are then mined from these chunk records, evidence boundaries and valid/negative response windows are calibrated, and executable task schemas are written from the grounded records. An independent causal verifier checks answer support, evidence sufficiency, timing, action validity, state continuity, and future-leakage risk. Ambiguous records are revised or rejected rather than silently retained. The 5,436 trajectories expand into 91,839 temporally ordered decision nodes under replay, making the evaluation unit denser than one final answer per video.

The primary protocol replays each stream strictly causally and scores the complete prediction trajectory. Table 2 reports target-time answer accuracy (Ans.), phase macro-F1 (State), evidence-sufficiency AUROC (Ready), and four-action macro-F1 (Action). StreamScore equally averages State, Ready, Action, and valid-first-response rate (VFR), which requires the first response to be timely and correct; Ans. remains separate. Calibration, timing, and strict trajectory diagnostics remain auditable, with full definitions in Appendices B and A.2.

4.3 StreamAct-SFT as Behavioral Intervention

StreamAct-SFT uses the same trajectory schema but is constructed from an audited source pool. Videos overlapping StreamAct-Bench evaluation media are removed before splitting. After source-video deduplication and causal validation, the resulting dataset contains 60K replayable training trajectories, each serialized into structured causal turns. A turn represents one observable decision point rather than an independent video or static question–answer pair. Training targets are written only after chunk-level observation, trajectory construction, and causal verification, so the model is supervised on process variables rather than benchmark answers or post-hoc summaries. Instead of supervising only the final answer, StreamAct-SFT provides stepwise targets:

$$\{(s_t^\star, r_t^\star, a_t^\star, y_t^\star)\}_{t=1}^{T}. \tag{12}$$

At each step, the input contains the current chunk, objective, and preceding trajectory state; the target updates state, marks evidence status, chooses an action, and emits a response when warranted. The response field is populated only for a supported TRIGGER or ANSWER; otherwise, the turn supervises WAIT or SILENCE. These fields are serialized into a structured response and optimized with a standard autoregressive objective, without dedicated controller heads. Thus, a trajectory yields multiple causally grounded learning decisions rather than a single final-answer target.

StreamAct-SFT serves as a behavioral intervention, testing whether diagnosed state–readiness–action behaviors can be learned through explicit process supervision rather than assumed to emerge from answer training. We compare process supervision with answer-only, format-only, and component-removed controls. Matched comparisons determine whether the diagnosed trajectory behaviors are learnable targets rather than merely post-hoc metrics. Construction, review, split policy, and optimization details are given in Appendix C.

5 Experiments

5.1 Experimental Setup

Section 2 established the Streaming Intelligence Illusion through a controlled paired experiment. We investigate three questions: (1) Do benchmark-native rankings transfer to complete streaming trajectories? (2) Where does the mismatch manifest, and what can controlled interventions localize? (3) Can explicit process supervision improve these behaviors?

We reproduce all results on our infrastructure. OVO-Bench (Niu et al., 2025) and Streaming-Bench (Lin et al., 2026) follow their native protocols, whereas StreamAct-Bench uses strict-causal full-trajectory replay. The comparison spans general-purpose, video-language, and streaming-oriented systems (Bai et al., 2025; Xu et al., 2025; Yang et al., 2026e; Liu et al., 2026; Wang et al., 2026; Lu et al., 2026; Yao et al., 2026; Li et al., 2026). All StreamAct scores use the same canonical evaluator.

		OVO-Bench			StreamingBench				StreamAct-Bench				
Model	**Size**	**RT**	**BT**	**All**	**RTV**	**OS**	**Ctx.**	**All**	**Ans.**	**State**	**Ready**	**Action**	**Stream.**
Closed general-purpose models (reference only)													
Doubao-Seed-2.0-Lite	–	59.55	63.45	61.50	74.29	62.60	38.40	65.21	44.34	40.82	60.89	37.19	36.12
Open-source video-language models													
Qwen3-VL-30B-A3B	30B-A3B	59.62	51.66	55.64	71.88	44.96	32.13	58.53	38.91	**39.71**	59.86	36.61	35.53
Qwen3-VL-8B	8B	55.08	38.67	46.88	67.69	34.99	34.40	54.12	26.77	33.81	57.93	30.19	31.51
Qwen3-VL-4B	4B	54.96	41.36	48.16	66.98	31.63	29.87	52.11	7.99	26.32	51.18	19.30	24.37
Online and streaming-oriented models													
JoyAI-VL-Interaction	9B	61.71	56.38	59.05	76.35	41.50	41.73	62.04	33.52	21.25	49.66	22.65	25.03
VideoChat3-4B	4B	58.90	55.26	57.08	83.01	43.00	36.97	65.47	13.18	10.01	46.55	27.08	21.38
ThinkStream-3B	3B	55.56	49.60	52.58	70.23	39.16	41.07	57.77	56.32	28.36	52.28	22.69	26.38
Mage-VL-4B	4B	**82.68**	49.29	65.99	79.92	46.50	**59.73**	68.49	30.74	17.17	53.85	18.32	23.43
AURA	9B	79.80	60.40	65.30	**83.20**	45.20	59.00	69.99	55.39	22.03	50.34	21.23	23.40
MOSS-VL-Realtime	11B	75.90	**72.60**	**70.20**	82.90	**48.80**	56.40	**70.20**	32.60	21.64	48.06	27.85	24.94
Matched-backbone comparison													
Qwen3-Omni-30B-A3B	30B-A3B	63.68	51.03	57.36	73.32	45.50	39.47	60.80	45.02	39.59	59.28	36.58	35.53
StreamAct-SFT (Ours)	30B-A3B	62.13	55.94	59.04	73.91	44.80	38.80	60.87	**67.90**	39.49	**60.55**	**53.93**	**41.27**
Δ *vs. Base (pp)*	–	-1.55	+4.91	+1.68	+0.59	-0.70	-0.67	+0.07	**+22.88**	-0.10	+1.27	**+17.35**	**+5.74**

Table 2: Benchmark-native performance and strict trajectory-level behavior. Darker shading marks larger gaps to column-best non-reference State/Ready/Action/Stream scores. Bold denotes column best; Δ reports SFT−Base changes (pp).

Table 2 compares benchmark-native performance with strict trajectory-level behavior for the same models. OVO-Bench uses its benchmark aggregate, while StreamingBench aggregates RTV, OS, and Ctx. StreamAct instead evaluates state, readiness, action, and grounded responses over complete trajectories. Implementation details are provided in Appendices A and B.

5.2 Does Benchmark Answerability Transfer to Streaming Behavior?

Answerability gains do not necessarily transfer to trajectory behavior. Qwen3-Omni raises Ans. from 38.91% to 45.02% over the similarly sized Qwen3-VL, yet their State, Ready, and Action scores differ by at most 0.58 points, and both obtain a StreamScore of 35.53. The separation also appears from another direction: MOSS-VL-Realtime and Mage-VL-4B perform strongly under native protocols but reach StreamScores of only 24.94 and 23.43, whereas ThinkStream attains 26.38 despite lower native aggregates. Together, these comparisons expose a clear ranking mismatch: native-to-StreamScore rank correlations are $\rho = -0.24$ for OVO-Bench and $\rho = -0.42$ for StreamingBench. Native protocols therefore favor models differently from complete trajectory evaluation. Target-time answerability remains positively associated with StreamScore ($\rho = 0.54$), showing that content understanding matters but does not determine online control. Sections 5.3–5.5 next locate this gap through component diagnostics and controlled interventions.

5.3 Where Does Streaming Behavior Fail?

Streaming behavior fails when readiness and action lose temporal alignment with evidence. Table 3 relates component quality to valid, premature, and missed responses; Figure 3 visualizes this timing and action–readiness mismatch. Doubao, Qwen3-Omni, and Qwen3-VL-30B obtain similar Ready scores near 60 and Action scores near 37, yet only 5.58–6.68% of trajectories produce a valid first response and 73–87% miss the valid window. Thus, moderate component scores do not ensure that the corresponding decisions occur when supported by evidence. StreamAct-SFT raises Action by 17.35 points and VFR by 4.43 points over its base while reducing Miss by 24.36 points. Early also rises by 13.97 points, reflecting a shift from conservative non-response toward more proactive emission: more trajectories receive an action, although some responses precede the valid window.

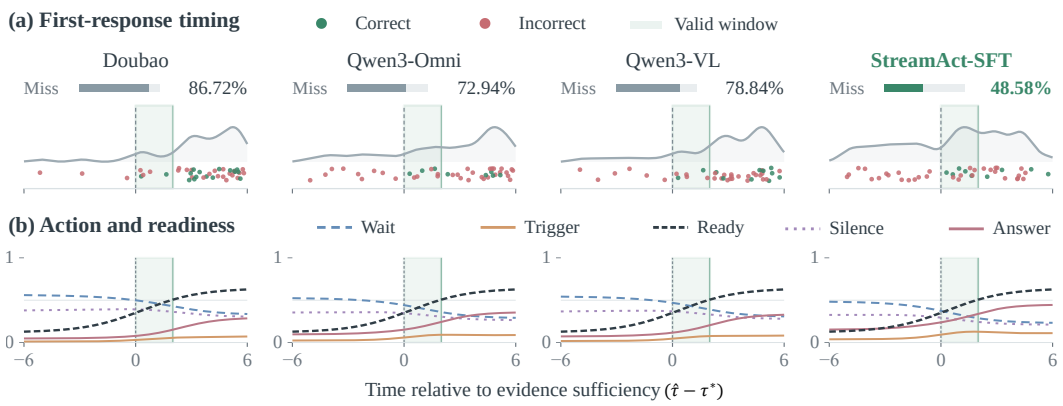


Figure 3: **Evidence-aligned behavior.** (a) schematically contrasts premature, valid, delayed, and missed first responses; annotated Miss rates are measured. (b) illustrates how readiness and action may diverge over time. Shading denotes valid response windows; curves and points are conceptual.

Model	State↑	Ready↑	Action↑	VFR↑	Early↓	Miss↓
Doubao	40.82	60.89	37.19	5.58	8.12	86.72
Qwen3-Omni	39.59	59.28	36.58	6.68	20.65	72.94
Qwen3-VL-30B	39.71	59.86	36.61	5.92	15.08	78.84
StreamAct-SFT	39.49	**60.55**	**53.93**	**11.11**	34.62	**48.58**

Table 3: Component quality and response failures under strict trajectory replay.

Intervention	VFR↑	Early↓	Miss↓	ΔVFR [95% CI]
Full SFT	8.84	34.56	53.30	—
State oracle	8.18	24.01	63.32	−0.66 [−3.30, 1.98]
Readiness oracle	9.63	30.47	57.92	+0.79 [−0.79, 2.37]
Reference-timed	80.61	0.00	18.07	+71.77 [68.47, 74.93]

Table 4: Inference controls and response validity on 758 shared trajectories.

Reference scheduling reveals substantial headroom in autonomous response control. On the common subset in Table 4, replacing State or Readiness alone yields no reliable VFR gain because both paired confidence intervals cross zero; neither local intervention reliably improves the remaining model-predicted trajectory. By contrast, supplying the reference response schedule raises VFR from 8.84% to 80.61% while leaving response content model-generated. The model can often answer when told when to act, but cannot yet convert its evolving evidence into that decision autonomously. Together with the component results, this shows that the largest measured gap lies in coordinating evolving evidence with autonomous response decisions.

5.4 Process Supervision as a Behavioral Intervention

The diagnosis predicts that answer or format supervision alone cannot recover coherent online behavior. We test this on a fixed Qwen3-Omni backbone with matched budgets and the same source-video-disjoint pool, under a shared strict-causal evaluation protocol for every variant, separating answer learning, interface compliance, process control, and joint optimization.

Supervision	State	Ready	Action	VFR	Stream.
Base	**39.59**	59.28	36.58	6.68	35.53
Answer-only	30.20	58.00	37.40	8.20	33.45
Format-only	30.90	58.80	42.20	7.50	34.85
$s{+}r{+}a$	37.55	59.95	50.45	8.95	39.23
Full SFT	39.49	**60.55**	**53.93**	**11.11**	**41.27**

Table 5: Process-supervision strategies under matched evaluation.

Answer-only raises VFR from 6.68 to 8.20 but lowers StreamScore to 33.45 as State falls to 30.20; format-only improves Action to 42.20 but not the aggregate. In contrast, $s{+}r{+}a$ reaches 50.45 Action and 39.23 StreamScore without an answer target, showing that intermediate behavior is learnable. Full SFT performs best at 53.93 Action, 11.11 VFR, and 41.27 StreamScore, while Table 2 shows its answerability gain. Answer and process targets are therefore complementary: one supplies semantic competence, the other teaches when and how to use it. Their joint optimization improves both emitted content and the decisions governing whether and when it is produced throughout the evolving trajectory under strict causality.

5.5 Which Process Signals Support Streaming Behavior?

Table 6 asks whether the full gain comes from one target or their coordination. Removing any process target degrades not only its corresponding component but also VFR, showing that state maintenance, evidence judgment, and action selection are learned as an interdependent trajectory rather than isolated objectives.

Supervision	State	Ready	Action	VFR	Stream.
w/o State	35.20	59.35	51.45	9.30	38.83
w/o Readiness	37.10	55.85	48.95	7.55	37.36
w/o Action	37.35	59.60	44.65	8.05	37.41
w/o Answer	37.55	59.95	50.45	8.95	39.23
Full SFT	**39.49**	**60.55**	**53.93**	**11.11**	**41.27**

Table 6: Component-target ablations under matched conditions.

Readiness and Action supervision exert the strongest cross-component effects: removing Readiness lowers StreamScore by 3.91 points, while removing Action lowers Action by 9.28 points. Readiness links accumulated evidence to whether action is warranted, and Action supervision turns that judgment into valid emission or non-emission. Answer supervision also improves behavioral metrics, showing that semantic response learning supports rather than substitutes for process control. The full gains therefore arise from coordinating what the model knows, when that knowledge is sufficient, and how it should act.

6 Related Work

6.1 Video-Language Models and Video Understanding Benchmarks

Video-language models extend vision-language pretraining and instruction tuning to temporal inputs and long contexts (Alayrac et al., 2022; Li et al., 2023a; Liu et al., 2023; Maaz et al., 2024; Lin et al., 2024; Wang et al., 2024b; Ren et al., 2024; Wang et al., 2024a; Bai et al., 2025; Xu et al., 2025; Wang et al., 2026). Their benchmarks span localized QA, long-form understanding, and embodied video (Xu et al., 2016; Krishna et al., 2017; Lei et al., 2018; Grauman et al., 2022; Fu et al., 2025; Wu et al., 2024; Li et al., 2024), but usually score an answer after a sampled clip or supplied context. StreamAct changes the evaluation object, not merely the video access pattern: it measures whether state, evidence sufficiency, action, and emission remain jointly valid while the stream unfolds.

6.2 Streaming Evaluation, Online Agents, and Process Supervision

Streaming work introduces partial observation, online context, response timing, and efficient streaming architectures (Lin et al., 2026; Niu et al., 2025; Azad et al., 2026; Wang et al., 2025; Chen et al., 2025; Zhang et al., 2026; Shen et al., 2026; Yang et al., 2026e; Lu et al., 2026; Yao et al., 2026). These advances address different layers: efficient ingestion concerns how video arrives, while StreamReady (Azad et al., 2026) models when accumulated evidence supports an answer. StreamAct targets the broader behavioral trajectory, additionally auditing state revision and the choice among waiting, warranted silence, proactive triggering, and answering. It connects streaming VLMs to partially observable decision-making (Kaelbling et al., 1998) and process supervision (Wei et al., 2022; Lightman et al., 2024), while grounding every target in observable media evidence. Thus, it complements rather than replaces answerability and efficiency benchmarks by testing whether their capabilities compose into coherent online control. Appendix F provides an expanded comparison.

7 Conclusion

StreamAct exposes the Streaming Intelligence Illusion: answerability does not ensure coherent online behavior. We define streaming intelligence as an auditable state–readiness–action–answer trajectory. Under strict-causal replay, answer gains need not improve trajectory behavior. Diagnostics locate the clearest gap in autonomous evidence-to-response control; process supervision shows it is learnable but unsolved. By aligning evaluation and supervision under the same causal interface, StreamAct connects behavioral diagnosis directly to targeted improvement across complete trajectories rather than isolated answers. StreamAct thus makes this failure diagnosable and shifts evaluation from what systems answer to how they behave as the world unfolds.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing and LaTeX formatting. We did not use generative AI tools to generate research ideas, experimental results, or benchmark annotations. All AI-assisted text and code were reviewed and verified by the authors, and we take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work introduces an evaluation benchmark and a supervised-tuning dataset built from video sources. All source videos were collected and used in compliance with their licenses, and benchmark records are grounded only in visible or native-audio evidence. We release trajectory-level annotations rather than re-distributing raw media where licensing restricts it, and human verification was performed by trained annotators under fair working conditions. The study does not involve human-subject experimentation. We are not aware of ethical concerns beyond those inherent to video-understanding research.

REPRODUCIBILITY STATEMENT

We describe the benchmark construction, canonical evaluation protocol, and metric definitions in Appendices B and A, and the StreamAct-SFT data construction, splitting, and training setup in Appendix C. The frozen 5,436-trajectory evaluation release and the canonical evaluator will be released to enable exact reproduction of all reported results.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

Shehreen Azad, Vibhav Vineet, and Yogesh S Rawat. Streamready: Learning what to answer and when in long streaming videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40494–40504, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. Livecc: Learning video llm with streaming speech transcription at scale. In *CVPR*, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, 2025.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 706–715, 2017.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. Tvqa: Localized, compositional video question answering. In *Conference on Empirical Methods in Natural Language Processing*, 2018. URL https://api.semanticscholar.org/CorpusID:52171684.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023a.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. VideoChat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023b. URL https://arxiv.org/abs/2305.06355.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 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.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, et al. Videochat3: Fully open video mllm for efficient and generalist video understanding. *arXiv preprint arXiv:2607.14935*, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12147–12151. IEEE, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Zikang Liu, Longteng Guo, Handong Li, Ru Zhen, Xingjian He, Ruyi Ji, Xiaoming Ren, Yanhao Zhang, Haonan Lu, and Jing Liu. Thinking in streaming video, 2026. URL https://arxiv.org/abs/2603.12938.

Xudong Lu, Yang Bo, Jinpeng Chen, Shuhan Li, Xintong Guo, Huankang Guan, Fang Liu, Dunyuan Xu, Peiwen Sun, Heyang Sun, Rui Liu, and Hongsheng Li. Aura: Always-on understanding and real-time assistance via video streams. *arXiv preprint arXiv:2604.04184*, 2026.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18902–18913. IEEE, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323. IEEE, 2024.

Yujiao Shen, Shulin Tian, Jingkang Yang, and Ziwei Liu. A simple baseline for streaming video understanding. *arXiv preprint arXiv:2604.02317*, 2026.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Pengyu Wang, Chenkun Tan, Shaojun Zhou, Qirui Zhou, Yanxin Chen, Xingyang He, Huazheng Zeng, Jijun Cheng, Chenghao Wang, Xiaomeng Qian, et al. Moss-vl technical report. *arXiv preprint arXiv:2608.15045*, 2026.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pp. 396–416. Springer, 2024b.

Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. Omnimmi: A comprehensive multi-modal interaction benchmark in streaming video contexts. *arxiv*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5288–5296, 2016. URL https://api.semanticscholar.org/CorpusID:206594535.

Panqi Yang, Haodong Jing, Jiahao Chao, Tingyan Xiang, Li Lin, Yao Hu, Yang Luo, and Yongqiang Ma. MUSE: Resolving manifold misalignment in visual tokenization via topological orthogonality. *arXiv preprint arXiv:2605.05646*, 2026a.

Panqi Yang, Haodong Jing, Nanning Zheng, and Yongqiang Ma. InstrucRobo: Object-centric multi-instruction decoupling model for explainable robotic manipulation. *Engineering Applications of Artificial Intelligence*, 171:114166, 2026b. ISSN 0952-1976.

Panqi Yang, Haodong Jing, Nanning Zheng, and Yongqiang Ma. UniBVR: Balancing visual and reasoning abilities in unified 3D scene understanding. *Neurocomputing*, 671:132599, 2026c. doi: 10.1016/j.neucom.2025.132599.

Panqi Yang, Haodong Jing, Nanning Zheng, and Yongqiang Ma. UniHOI: Unified human-object interaction understanding via unified token space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11640–11648, 2026d. doi: 10.1609/aaai.v40i14.38148.

Senqiao Yang, Kaichen Zhang, Zhaoyang Jia, Jinghao Guo, Yifei Shen, et al. Mage-vl: An efficient codec-native streaming multimodal foundation model. *arXiv preprint arXiv:2607.24904*, 2026e.

Dingyu Yao, Junhao Zhou, Chenxu Yang, Chuanyu Qin, Haowen Hou, Zheming Liang, Congcong Wang, Yuhang Cao, Shenglong Ye, Shuai Xie, Shuhuan Gu, Haoyang Huang, Qingyi Si, Nan Duan, and Jiaqi Wang. JoyAI-VL-Interaction: Real-time vision-language interaction intelligence. Technical report, JD.com, June 2026. URL https://echovideo.jd.cn/JoyAI-VL-Interaction/JoyAI-VL-Interaction-Reportv1.pdf.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. URL https://arxiv.org/abs/2306.02858.

Xichen Zhang, Guankai Li, Yinghao Zhu, Shijian Wang, Sitong Wu, Shaozuo Yu, Meng Chu, Yuan Lu, and Jiaya Jia. Streamarena: Toward continuous, interactive, and long-horizon agentic streaming video understanding, 2026. URL https://arxiv.org/abs/2608.05703.

APPENDIX CONTENTS

Appendix	Contents	Page
A Evaluation protocol	Action semantics, video protocol, metrics, and replay	14
B StreamAct-Bench	Release, task scope, construction, and provenance	17
C StreamAct-SFT	Sources, causal supervision, validation, and isolation	19
D Additional results	Matched supervision results	21
E Runtime interface	Prompt contract, option mapping, and reproduction	22
F Extended related work	Models, benchmarks, online control, and supervision	23

The appendices follow the evaluation chain: define valid behavior (A), construct grounded trajectories (B–C), inspect controlled results (D), and examine recorded behavior in video (E). Runtime contracts and related work follow in F–G.

A EVALUATION PROTOCOL AND METRICS

The protocol separates the information available to a model, its node-level behavior, and the validity of its responses. This distinction makes state, readiness, action, and answer jointly inspectable without treating a correct final answer as evidence that the preceding behavior was valid.

A.1 OPERATIONAL ACTION SEMANTICS

The four actions encode why a model emits or refrains from emitting. They are semantic labels implemented through a common autoregressive interface, not separate architectural heads.

Action	Emit	Valid behavioral condition
Non-emission actions		
WAIT	No	The task remains unresolved and currently available evidence is insufficient.
SILENCE	No	No response is warranted, including under negative or distractor evidence.
Emission actions		
TRIGGER	**Yes**	A standing instruction or task condition warrants a self-initiated response.
ANSWER	**Yes**	An explicit query is active and the available evidence supports a response.

Table 7: Operational semantics of the StreamAct action space. Actions are grouped by whether they emit a response.

Wait versus silence. A pending question about where a cup will be placed warrants WAIT while it is still being carried. Under "alert me if someone falls," normal walking warrants SILENCE. The distinction concerns an unresolved decision versus the absence of a warranted response; it is not determined by whether a query has been issued.

A.2 GENERAL TRAJECTORY VALIDITY

For response opportunity k, an emission must satisfy

$$t \in \mathcal{W}_k^+, \qquad \mathrm{Correct}(y_t, y_k^\star, o_{\le t}) = 1, \tag{13}$$

and $e_t = 0$ throughout negative intervals $\mathcal{W}^-$. The release contains one scored response opportunity per trajectory; replay nevertheless continues after emissions to expose later state changes and invalid actions.

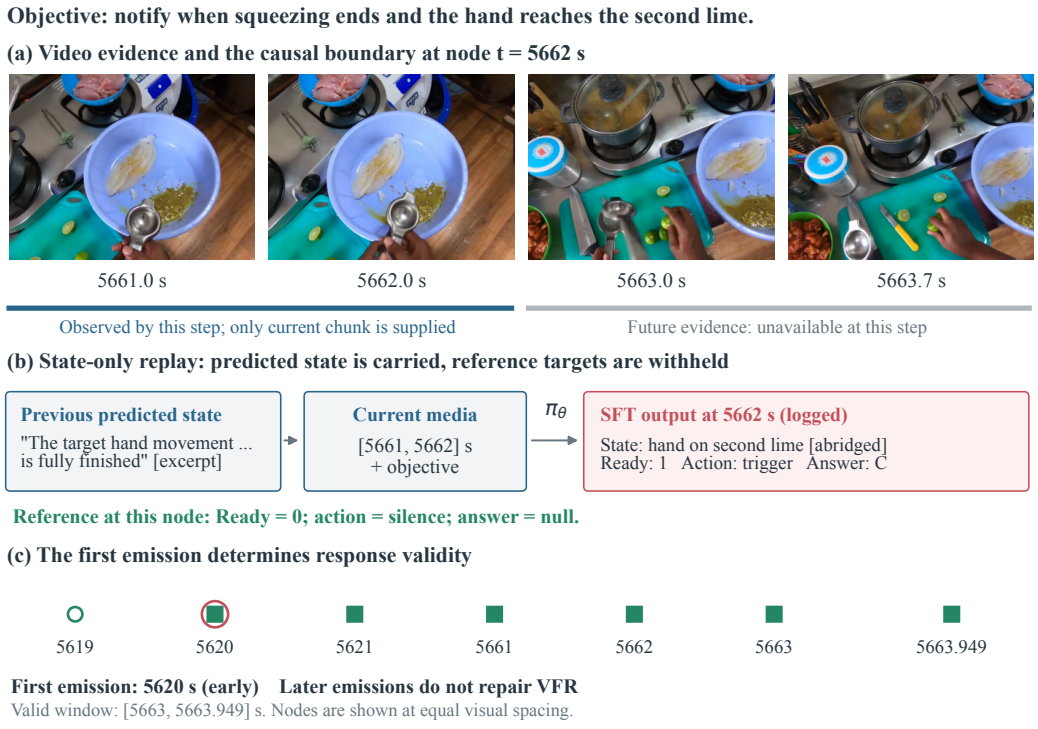


Figure 4: **From visible evidence to a scored streaming response.** A real lime-preparation clip aligns causal access, carried model state, withheld targets, and logged SFT actions. The first emission precedes the valid window and is therefore invalid, irrespective of later emissions. Display frames provide context, not a reconstruction of the model's sampled frames.

Temporal and semantic validity. Membership in $\mathcal{W}_k^+$ establishes that a response is warranted at time t, but not that its content is correct. Conversely, matching the reference answer outside that window may reflect a premature guess or a late response. Equation 13 therefore requires both conditions at the same emission: the response must occur within a supported window and be grounded in observations available by that time. Neither answer correctness nor timing alone serves as a proxy for valid streaming behavior.

Negative intervals and continued replay. Negative intervals test whether the policy preserves non-emission when evidence does not warrant a response. They cover unresolved portions of positive trajectories and trajectories in which the task condition never occurs. Continuing replay after emission keeps later state updates and actions observable, distinguishing coherent behavior from subsequent unsupported state changes or invalid emissions.

Invalid outputs and visibility. Invalid outputs receive an invalid state/action label, a null answer, and readiness on the wrong side of the gold binary label. State carryover keeps the last valid prediction. Answer text attached to WAIT/SILENCE is suppressed and recorded as inconsistent. Decoding constraints enforce schema only; they cannot repair semantic content.

First emission and later recovery. The first TRIGGER/ANSWER defines $\hat{\tau}$ in Equation 5. VFR requires this first response to satisfy both the temporal and semantic conditions in Equation 8. A premature or incorrect first response remains invalid even if the model later responds correctly. Such a recovery can make Miss zero while VFR remains zero; Early and Miss can therefore overlap.

Generalization beyond this release. For multiple opportunities, match predicted responses to reference windows only when both temporal support and answer correctness hold. Deterministic maximum-cardinality matching, with delay and timestamp tie-breaking, extends the definition. This opportunity-level extension is not used for the reported single-opportunity benchmark scores.

A.3 Metric Definitions and Aggregation

State is phase classification, not free-form factual-state extraction; Ready measures evidence discrimination, with Brier reported separately for calibration. Ans. evaluates the earliest target-window node (or the first evidence-sufficient node when none is scheduled in the window). It is excluded from StreamScore.

Metric	Scoring unit	Operational summary	Dir.
Answerability and the four StreamScore components			
Ans.	Target node	Answer correctness at selected target time	↑
State	State-task nodes	Macro-F1 over before / changing / after	↑
Ready	Replay nodes	AUROC of evidence-sufficiency predictions	↑
Action	Replay nodes	Macro-F1 over the four action labels	↑
VFR	Trajectories	Timely, correct first emitted response	↑
StreamScore	Components	Mean of State, Ready, Action, and VFR	↑
Calibration, timing, and full-trajectory diagnostics			
Readiness Brier	Replay nodes	MSE against binary readiness targets	↓
Silence / Trigger F1	Action labels	Per-class silence / trigger F1	↑
Early	Positive traj.	Emission before sufficient evidence	↓
Miss	Positive traj.	No correct response in the valid window	↓
Median trigger offset	First trigger	Median offset from valid-window start	$\to 0$
Strict TSR	Full traj.	All five output checks pass	↑
Schema Valid	Output nodes	Complete structured-output parsing	↑

Table 8: Metrics and scoring units. Definitions and eligible-subset aggregation are specified in the text. Ans. is excluded from StreamScore; VFR follows Equation 8. Early and Miss are non-exclusive.

Each component uses its eligible task subsets and macro-aggregation. Source-video bootstrap respects shared media. The auditable projections and their equal-weight summary are

$$\Phi = (M_{\text{state}}, M_{\text{ready}}, M_{\text{action}}, M_{\text{GR}}), \quad \text{StreamScore} = \tfrac{1}{4}(M_{\text{state}} + M_{\text{ready}} + M_{\text{action}} + M_{\text{GR}}). \tag{14}$$

Here M_{GR} is VFR under the first-response definition in Equation 8. Components describe complementary projections, not a guarantee that every node of a trajectory is valid.

Scoring units. Node-level metrics isolate recurring replay decisions, while trajectory-level metrics test whether they culminate in valid online behavior. State, Ready, and Action aggregate over eligible nodes; VFR, Early, Miss, and Strict TSR aggregate over complete trajectories. This prevents long trajectories from dominating response validity through their node count, while source-video bootstrap preserves dependence among trajectories from the same media.

Summary score and answerability. StreamScore equally weights the four behavioral projections so that no component dominates through task frequency or scale. Ans. remains separate because it measures content recovery at a selected evidence-sufficient point, whereas VFR requires the first self-timed response to be timely and correct. Reporting both preserves whether the model can answer from sufficient evidence and organize valid behavior while that evidence emerges.

Interpreting component scores

A correct state phase does not establish that every carried fact is correct. Readiness AUROC measures ranking, whereas Brier assesses probability calibration. Action macro-F1 evaluates label quality, while VFR additionally requires a valid response time and content. Their mean is a compact comparison score; it does not replace the component profile or certify every decision along a trajectory. These metrics are complementary operational projections of the theoretical trajectory variables rather than exhaustive measurements of streaming intelligence. Readiness and action are jointly generated fields: no fixed readiness threshold deterministically selects an action.

A.4 STRICT-CAUSAL REPLAY

The evaluator supplies the current chunk, objective, and last valid model-emitted state. Reference states, future observations, full prefixes, and native hidden caches are withheld. Thus, the causal history $\mathcal{F}_t$ in Section 3 is instantiated through the model's explicit carried state rather than by replaying $o_{1:t}$ at every node. All scheduled nodes are scored, unlike the minimal paired diagnosis, which may stop at the first response.

Interface scope. This controlled interface standardizes observable information, state carryover, and output semantics across models, making trajectory behavior directly comparable and the carried state auditable. The resulting scores measure online behavior under StreamAct's explicit-state realization. They are distinct from native-system evaluations that preserve architecture-specific KV caches, memory banks, recurrent state, or proprietary response controllers; such native mechanisms may retain additional history unavailable through the shared adapter. Accordingly, the benchmark supports controlled diagnosis under a common information contract, while native-system capability remains a separate evaluation setting.

Algorithm 1 Strict-causal trajectory replay and scoring

Require: Ordered stream O, objective q, withheld reference trajectory $\mathcal{T}^\star$
Initialize carried state $\hat{s}_0 = \emptyset$ and prediction log $\mathcal{T}_\theta = \emptyset$
for each scheduled node t **do**
  Reveal only current chunk o_t, objective q, and last valid state $\hat{s}_{t-1}$
  Parse $(\hat{s}_t, \hat{r}_t, \hat{a}_t, \hat{y}_t) \leftarrow \pi_\theta(o_t, \hat{s}_{t-1}, q)$
  if output is invalid **then**
    Log the invalid record and retain the last valid carried state
  else
    Log the decision; suppress answer text for WAIT/SILENCE
  end if
end for
Score $\mathcal{T}_\theta$ against $\mathcal{T}^\star$
return the prediction log and component scores Φ

A.5 COMPARISON SCOPE

The OVO-Bench and StreamingBench blocks in Table 2 preserve their benchmark-specific evaluation configurations and are used to compare rankings across protocols, rather than as an input-matched model comparison. Mage-VL uses its latest-four-frame setting for OVO RT/BT and StreamingBench RTVU; MOSS-VL-Realtime uses its visual-only setting. StreamingBench All applies the RTV/OS/Ctx. family aggregation shown in the main table. Doubao does not support 250 Proactive Output items through the evaluated interface, so its aggregate covers the supported set; assigning zero to all unsupported items gives a conservative full-denominator lower bound of 61.36.

By contrast, every StreamAct-Bench model is evaluated on the same ordered nodes and information boundaries defined in Section A.4. Model-specific wrappers only serialize media inputs and the common output schema. JSON constraints enforce parseability without exposing reference states, action labels, response windows, or answer semantics.

B STREAMACT-BENCH DETAILS

B.1 RELEASE SCALE AND TASK SCOPE

Release scale and counting unit. The frozen release contains 5,436 trajectories and 91,839 decision nodes. A node is a scheduled state–readiness–action–answer prediction, not an independent video or question. Models share the same ordered nodes and information boundaries. Node-level predictions support component diagnostics; answer and trajectory metrics aggregate over trajectories, with source video retained as the bootstrap unit.

Release quality control. Only media-grounded trajectories that pass causal, temporal, and executable validation enter the frozen release. A trajectory is excluded in full if any scheduled node

falls outside the decodable media boundary; invalid tails are not truncated. This keeps every model on the same intact temporal sequence and avoids changing the scoring horizon across systems.

Task family	Evidence-to-action requirement	Primary diagnostics
OnlineQA	Answer from sufficient causal evidence.	Ans., VFR
EvidenceQA	Wait until later evidence resolves uncertainty.	Ready, Early
Trigger	Act when the instructed condition occurs.	Trigger F1, offset, Miss
State	Maintain and revise facts across chunks.	State, Strict TSR
AV-Sync	Use temporally aligned audio and video.	Ans., VFR
Negative	Suppress responses to insufficient evidence.	Silence F1, false triggers

Table 9: StreamAct-Bench task scope. Conceptual task families emphasize distinct requirements under the shared trajectory schema; exported task-label counts are shown separately.

Task coverage and evaluation balance. Figure 5 summarizes task coverage and support along the principal evaluation axes. Proactive triggering, online QA, and state tracking form the main task mixture. Audio requirements and answer positions are broadly balanced, while every action class and both readiness labels receive substantial node-level support. The benchmark therefore tests more than a majority non-emission policy without requiring every task family to occur equally often.

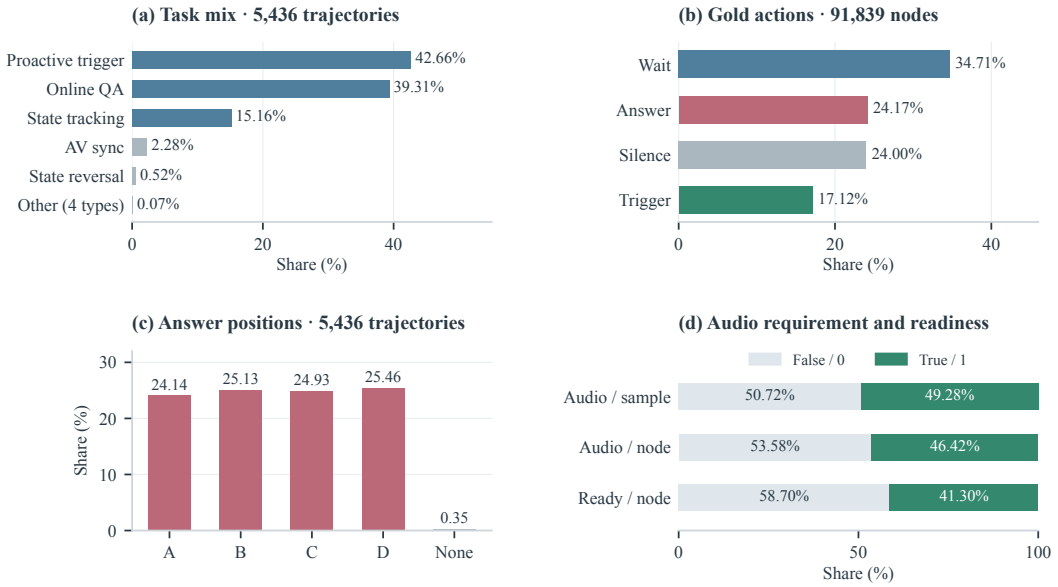


Figure 5: **Task coverage across evaluation axes.** (a) Task mixture, with rare auxiliary labels grouped as Other. (b) Node-level action coverage. (c) Answer positions, with non-emission trajectories shown separately. (d) Audio dependence and readiness. Percentages use the stated trajectory or node denominator.

Source diversity and temporal coverage. The release spans 1,705 source videos, with no source contributing more than 12 trajectories. It includes answer, proactive-trigger, and non-emission trajectories, together with before, changing, and after state phases. A trajectory contains a median of 15 decision nodes, and its annotated response window has a median duration of 2.0 s. Source caps, shared endpoint conventions, and identical replay timing prevent individual videos or model-specific schedules from dominating the comparison.

B.2 Media-Grounded Construction Pipeline

StreamAct-Bench uses staged annotation rather than direct question generation from source metadata. Every retained field must be traceable to causally available visual or native-audio evidence;

source questions, subtitles, filenames, and metadata may locate candidates but never serve as factual evidence.

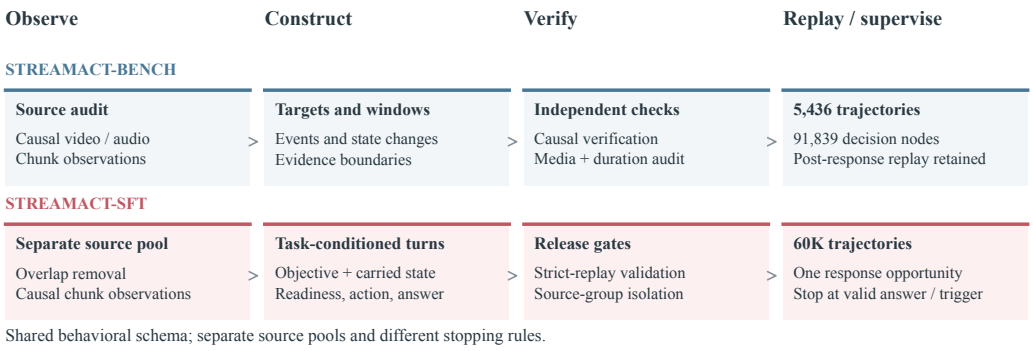


Figure 6: **Aligned construction paths with separate source pools.** Both paths apply causal observation, temporal grounding, and independent validation. Bench retains post-response replay; SFT terminates after a valid response and uses source-group splits.

Causal observation. Media checks first establish a valid chunk timeline. Annotation then uses only the visual content and native audio available within each causal chunk to record entities, actions, text, speech or sounds, state changes, and uncertainty. Source questions, subtitles, filenames, and metadata may help locate candidate events but cannot supply target facts or enter the final evidence record.

Temporal grounding. Chunk observations are consolidated into event and state transitions of *before* $\rightarrow$ *delta* $\rightarrow$ *after*. Readiness identifies the first point at which the available evidence supports a reliable decision, separately from event onset and completion. Boundary-sensitive cases are reviewed at finer temporal resolution with native audio; retained windows must be temporally valid, causally supported, and non-overlapping with negative intervals.

Independent acceptance. Candidate trajectories are checked against the underlying media for factual grounding, causal sufficiency, response timing, answer uniqueness, state continuity, action–answer consistency, and future leakage. They must also pass executable checks for fields, media references, timestamps, and response windows. Material revisions trigger renewed media review, and trajectories that do not satisfy both semantic and executable requirements are excluded from the frozen benchmark. Validation and revision records remain linked to every accepted trajectory for audit.

B.3 Provenance, Review, and Splits

Every released trajectory retains a link to its source video, media interval, causal chunk observations, and temporal annotations, allowing each target and response window to be traced back to the supporting audio-visual evidence. Boundary-sensitive triggers, hard negatives, audio-dependent cases, and state reversals receive additional media review, while the remaining records are covered by stratified audit. Source-video identity is used for deduplication, per-video caps, data partitioning, and bootstrap sampling; trajectories from the same source are never separated across partitions. Only records that pass the grounding, timing, and executable checks in Section B.2 enter the frozen release.

C StreamAct-SFT Details

StreamAct-SFT shares the benchmark's behavioral schema but uses a separately audited source pool. Its 60K replayable trajectories are serialized into structured causal turns only after chunk-level observation, timing calibration, objective writing, trajectory construction, and causal verification. Source records serve only as localization seeds, and videos overlapping StreamAct-Bench or other evaluation blacklists are excluded before splitting and expansion.

C.1 Source Construction and Quality Gates

The source pool is frozen by parent identity and content hash before trajectory generation. Retained sources must have valid video, native audio, duration, and timestamps; source questions, subtitles, ASR, filenames, and metadata are used only to locate candidate evidence. Two-second causal chunks provide the default observation grid, with finer review for boundary-sensitive events. Source-level deduplication, duration-aware caps, and benchmark blacklists prevent overlap with StreamAct-Bench and limit concentration from individual videos.

Grounded event and state tracks cover persistent facts, long dependencies, reversals, hard negatives, distractors, and non-events. Black frames, credits, occlusion, or feed loss are not treated as semantic negatives. Audio-dependent candidates are checked against visual-only, audio-only, and joint audio-visual evidence. Timing review distinguishes event onset, completion, and the first evidence-sufficient decision point; only causally supported objectives and valid response windows proceed to trajectory assembly.

C.2 Structured Supervision and Causal Visibility

A validated trajectory is expanded chronologically into causal turns that preserve the objective and carried-state chain while exposing only the new audio-visual chunk. State targets are conditioned on the active objective rather than copied from a generic scene summary; readiness, action, and answer placement follow the verified evidence boundary. Each StreamAct-SFT trajectory contains one response opportunity and terminates after its valid answer or trigger, preventing post-response observations from entering supervision. StreamAct-Bench instead continues replay after emission to evaluate the complete trajectory.

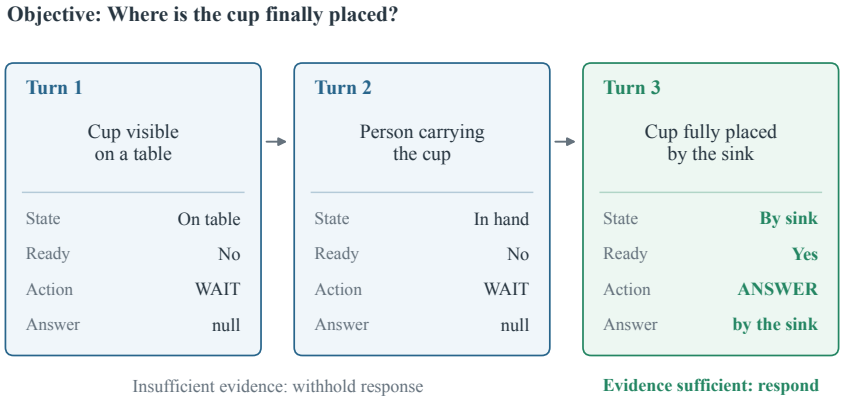


Figure 7: **Serialization of a causal training trajectory.** This schematic example aligns observations with state, readiness, action, and answer targets. Green marks the first evidence-sufficient response; supervision ends after the valid answer.

Each causal training turn contains the objective, the model's previous carried state, and the current audio-visual chunk. The assistant target serializes

$$\big(\texttt{state_delta}, \texttt{state_after}, \texttt{evidence}, \texttt{action}, \texttt{answer}\big)_t. \tag{15}$$

The answer field is null for `wait` and `silence`, and populated only for a valid `trigger` or `answer`. Inputs exclude future trajectories, final answers, and post-t summaries. Training uses the standard autoregressive objective; field-level masks or weights, when used, belong in the training configuration.

Why supervise non-emission turns? Non-emission turns directly supervise evidence insufficiency and warranted silence. Task-conditioned states specify which facts must persist across observations. This design does not guarantee correct timing or state maintenance; both require separate evaluation.

C.3 Validation and Isolation

Turn-level validation covers media integrity, timestamps, causal visibility, state continuity, evidence–action consistency, and action–answer consistency. Replay hides future turns and reference answers, removes source metadata, tests distractor-induced triggers and stale states after reversals, and requires verified native audio for audio-dependent turns. Records that fail semantic grounding or causal consistency are excluded rather than repaired without renewed review.

Source isolation precedes trajectory expansion. Parent identities and content hashes exclude overlap with StreamAct-Bench and keep all trajectories from the same source in one split. Final audits check duplicates, packed-example leakage, repeated objectives or answers, source concentration, readiness timing, and capability, action, and modality coverage. The 60K release is frozen only after source-group splits, media review, and manifest validation are complete.

D Additional Experimental Results

This appendix details the matched comparison used to isolate which supervision targets improve trajectory behavior.

D.1 Matched Supervision Comparisons

All variants use the same Qwen3-Omni backbone, 60K training pool, optimization budget, and evaluation protocol. Answer-only retains response targets; format-only teaches the output schema without reference behavior; $s{+}r{+}a$ supervises the three process fields while masking answers; Full SFT optimizes all targets in Equation 15. This controls model capacity and training exposure so that differences can be attributed to the supervision target.

Answer-only and format-only each improve a narrow aspect of behavior but do not recover the complete trajectory profile. Process targets produce the larger Action and StreamScore gains, while Full SFT achieves the strongest Action, VFR, and aggregate results. The comparison therefore shows that answer and process targets provide complementary supervision for emitted content and the decisions governing its use.

Supervision	**State**	**Ready**	**Action**	**VFR**	**Stream.**
Base	39.59	59.28	36.58	6.68	35.53
Answer-only	30.20	58.00	37.40	8.20	33.45
Format-only	30.90	58.80	42.20	7.50	34.85
$s{+}r{+}a$	37.55	59.95	50.45	8.95	39.23
Full SFT	39.49	60.55	53.93	11.11	41.27

Table 10: **Matched supervision targets.** All variants use the same backbone, data scale, optimization budget, and evaluation. State is phase macro-F1; Ready is AUROC; Action is four-class macro-F1; VFR evaluates the first response.

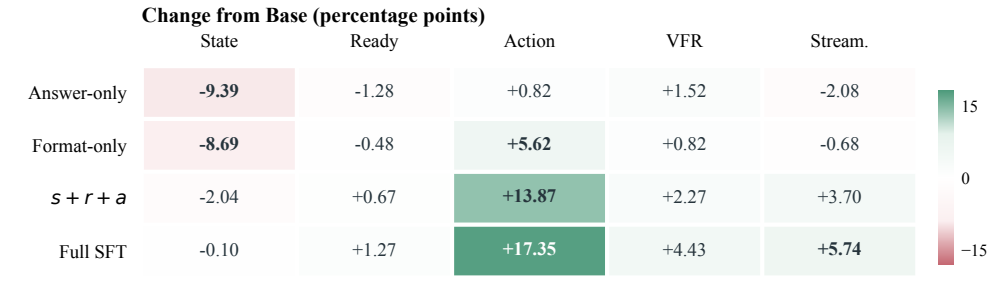


Figure 8: **Which behaviors change under each target?** SFT-minus-Base changes in Table 10; green/red denotes increase/decrease. Bold marks magnitudes of at least five points; all cells share one scale.

E RUNTIME INTERFACE AND REPRODUCIBILITY

At each replay node, the model receives the objective, question, options, current media interval, and its previously emitted state. Reference annotations, future observations, full media prefixes, and native hidden caches remain outside the input. The shared explicit-state channel makes causal history auditable and keeps the information boundary consistent across models.

Model-visible output contract. The node instruction follows the sample-specific objective, question, and option mapping; the previous predicted state is supplied separately.

Evaluation output instruction

```
Observe only the current media chunk.  Return one compact JSON object
with state_phase (before|changing|after|na), state (at most 20
words), readiness (number from 0 to 1), action
(wait|silence|trigger|answer), and answer (A-D or null).
```

Exported field	Interpretation and audit requirement
`state_phase`	Discrete phase prediction; `na` is distinct from an incorrect phase.
`state`	Model-written state text. A fluent assertion is not proof of media grounding.
`readiness`	Numeric model output in $[0, 1]$; this format alone does not establish calibration.
`action`	Predicted behavior; the first trigger or answer defines the first emission.
`answer`	Option label or null. Decode using the options supplied at that node.

Table 11: Evaluation fields and their interpretation under the shared runtime contract.

Option identities and time bases. Option labels are decoded through each node's stored mapping, which retains both original and remapped options. Response and reference timestamps use evaluation-media time; cropped presentation clips preserve their offsets to that timeline. Display frames provide visualization and need not equal the model's sampled frames.

State and serialization. Training serializes state updates and evidence fields, while evaluation exports `state_phase` and compact `state` text. Reproduction therefore binds the evaluation contract to the training template, field conversion, loss configuration, optimizer, and checkpoint-selection settings.

Scope of the shared adapter. The shared adapter standardizes observable history and response semantics for controlled comparison. StreamAct scores therefore characterize online behavior through this explicit-state interface. Native-system evaluations may additionally retain architecture-specific KV caches, memory banks, recurrent state, or proprietary controllers and constitute a complementary protocol with a different information interface.

Readiness and action outputs. Readiness and action are generated together in the same autoregressive record; the evaluator does not derive action from a fixed readiness threshold. Readiness is scored for evidence discrimination and calibration, action against its semantic label, and their trajectory-level consequence through VFR, Early, and Miss.

Artifacts needed for exact reproduction. Exact reproduction requires versioned manifests, checkpoints, predictions, scorers, adapters, prompts, decoding settings, and media preprocessing. The released configuration also specifies invalid-output handling, first-emission extraction, option mapping, and the mapping from node timestamps to response windows.

F Extended Related Work

Video-language modeling. Visual-language pretraining, instruction tuning, and temporal adaptation provide the foundation for video-language models (Alayrac et al., 2022; Li et al., 2023a; Liu et al., 2023; Maaz et al., 2024; Li et al., 2023b; Zhang et al., 2023; Lin et al., 2024; Wang et al., 2024b; Ren et al., 2024; Wang et al., 2024a; Bai et al., 2025; Xu et al., 2025). Long contexts, compression, memory, and modality fusion improve access to evidence. StreamAct evaluates the online behavior built on those representations: preserving relevant state, deferring unsupported decisions, and choosing appropriate responses.

Video understanding benchmarks. Video captioning (Xu et al., 2016), dense event captioning (Krishna et al., 2017), localized QA (Lei et al., 2018), long-form and embodied understanding (Grauman et al., 2022; Fu et al., 2025; Wu et al., 2024; Li et al., 2024), and online benchmarks such as OVO-Bench (Niu et al., 2025) and StreamingBench (Lin et al., 2026) broaden task coverage and temporal access. StreamAct changes the scored object to a sequence of state updates, readiness estimates, actions, and conditional emissions, rather than inferring the whole process from an answer.

Streaming architectures and readiness. Continuous perception, bounded memory, and low-latency interaction are complementary design goals (Wang et al., 2025; Chen et al., 2025; Zhang et al., 2026; Shen et al., 2026; Yang et al., 2026e; Lu et al., 2026; Yao et al., 2026; Wang et al., 2026; Liu et al., 2026). StreamReady (Azad et al., 2026) studies answer readiness through evidence windows. StreamAct includes that axis within a broader auditable trajectory, distinguishing unresolved waiting, warranted silence, proactive triggering, and query-conditioned answering. Efficient ingestion and timely answering do not by themselves specify all of these behaviors.

Online decisions and process supervision. Partially observable decision-making motivates carrying information forward and acting as observations arrive (Kaelbling et al., 1998). StreamAct makes intermediate decisions observable instead of relying solely on task completion. Its SFT intervention connects instruction tuning (Ouyang et al., 2022; Liu et al., 2023), explicit intermediate reasoning (Wei et al., 2022), and process supervision (Lightman et al., 2024) to media-grounded state, readiness, action, and answer targets. These records support behavioral diagnosis; they do not identify the model's latent causal mechanism.

Unified representations and embodied tasks. MUSE studies visual tokenization for reconstruction and semantic perception (Yang et al., 2026a); UniHOI addresses interaction understanding in a unified token space (Yang et al., 2026d). UniBVR examines unified 3D scene understanding (Yang et al., 2026c), and InstrucRobo studies object-centric instruction decoupling for manipulation (Yang et al., 2026b). These works provide representational and embodied context, not evaluated baselines: StreamAct asks whether available evidence supports the model's current online behavior.