# BEYOND READINESS: OPERATOR-CONDITIONED ANSWER CLOSURE IN STREAMING EVIDENCE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A supported answer can still change as a stream continues. We formalize *operator-conditioned answer closure*: a supported candidate is closed when every continuation admitted by the question operator and horizon preserves it. Four reference states distinguish absent support, provisional support, closure, and terminal unanswerability; calibrated gates derive response actions. Operator-balanced shared-prefix streams provide exact symbolic labels for comparing learned estimators across seven controllers. On 1,440 held-out symbolic trajectories from known program families, a gap-aware GRU reaches 0.9954 closure average precision and 0.9851 state macro-F1. Relative to an event-time survival baseline, it reduces premature commitment from 0.0040 to 0.0008 at identical answer coverage. Normalized delay, however, increases from 0.0612 to 0.0683, and same-target controls show no uniform ranking advantage from finer state supervision. With frozen VideoLLM features, a human-reviewed video pilot yields closure AP 0.9180 but fixed-four-class macro-F1 0.2900. Its seven held-out video records come from two source components and contain no terminal-unanswerable reference labels. These results distinguish current support from answer finality and show why closure ranking, state classification, and stopping quality require separate evaluation.

## 1 INTRODUCTION

Consider a video of a task in which blue cups appear (Figure 1). After two cups appear, *Has a blue cup appeared?* has a final positive answer. The question *How many blue cups appear during the complete task?* remains open: completing the task now gives 2; introducing another cup first gives 3. Here `DOMAIN_END` marks observed task completion, whereas `STREAM_END` ends the recording. If recording stops before task completion, the complete-task count is undetermined under this task definition. Perfect recognition of the two observed cups cannot resolve this uncertainty. We call invariance of a supported answer over task-admissible futures *answer closure*.

Streaming VideoLLMs learn response timing through adaptation to streaming inputs, respond-or-silence decisions, and evidence-window readiness (Chen et al., 2024; Wang et al., 2025b; Azad et al., 2026). Readiness labels can use later evidence on a realized trajectory without requiring invariance across admissible futures. Confidence estimates expected correctness (Guo et al., 2017; Geng et al., 2024), while selective prediction controls whether a system answers under risk (Geifman & El-Yaniv, 2019; Guo et al., 2023; Srinivasan et al., 2024). Neither quantity prevents a correct, confident candidate from remaining open to revision.

Closure depends on the question operator, horizon rule, and admissible continuation set. An answer-changing suffix disproves closure; a sound operator-specific certificate establishes it. We study this prediction target through shared-prefix groups whose futures preserve, change, or preclude an answer. Finite symbolic streams provide exact labels, while a human-reviewed video pilot supplies a preliminary test on causal VideoLLM representations.

A gap-aware GRU is one estimator of this target, alongside matched recurrent and nonrecurrent alternatives (Che et al., 2018; Hasani et al., 2022). We compare support supervision with closure supervision, then test four-state, binary-closure, and survival estimators on the same closure target.



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.

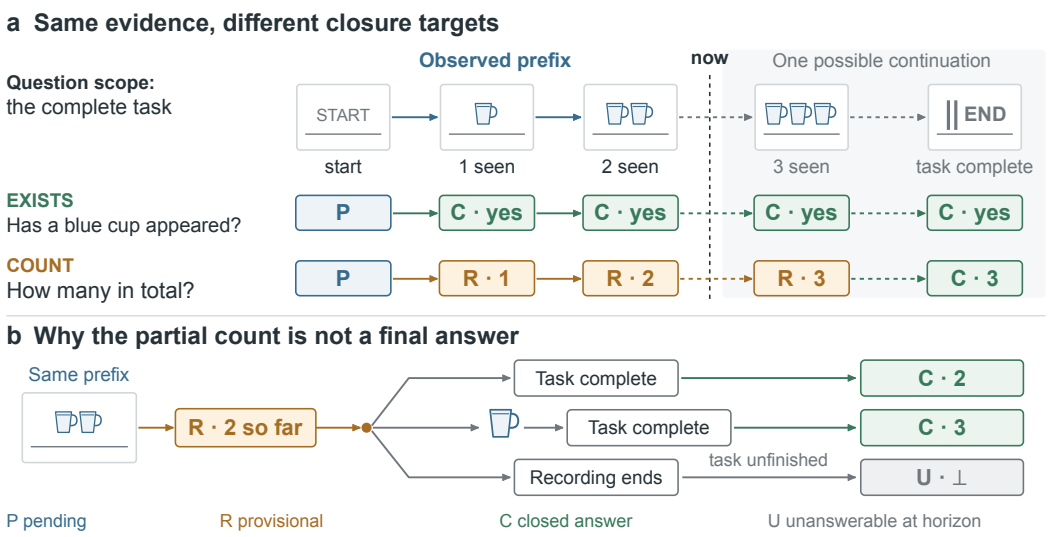


Figure 1: Operator-conditioned answer closure. **a**, The same prefix closes a positive Exists answer but leaves the complete-task Count provisional. **b**, Task completion fixes the count at 2 or 3; recording termination before completion leaves it undetermined ($\perp$). DOMAIN END closes the task domain; STREAM END alone need not do so. Boxes show reference states and candidates, not emitted actions; dashed continuations are unobserved. Drawings are illustrative, not dataset frames.

Stopping evaluation measures whether reduced premature commitment comes at the cost of answer coverage or delay.

**Contributions.**

- We formalize operator-conditioned answer closure as a finite-horizon prediction and stopping problem with four epistemic states.
- We construct operator-balanced symbolic shared-prefix groups with exact labels and an exploratory video pilot with human-reviewed, AI-assisted annotations and source-grouped splits.
- We measure target and controller effects separately, including symbolic stopping trade-offs, held-out operator expressions, and the gap between closure ranking and four-state classification on video.

## 2 Related Work

**Streaming response and readiness.** Dispider and StreamBridge separate proactive decisions from perception or memory; Proact-VL regularizes response timing, while STRIDE predicts structured activation sequences (Qian et al., 2025; Wang et al., 2025a; Yan et al., 2026; Kim et al., 2026). Evidence-window readiness, evidence localization, and temporal-logic questions address related aspects of support (Azad et al., 2026; Wang et al., 2026; Swetha et al., 2025). Closure instead requires answer invariance across admissible continuations. Appendix A.1 discusses complementary streaming and abstention benchmarks.

**Runtime monitoring and early classification.** Runtime verification assigns conclusive or inconclusive verdicts to finite prefixes and seeks the earliest sound verdict for a temporal specification (Bauer et al., 2011). TEASER gives an early-classification alternative: release a reliable current prediction or continue observing (Schäfer & Leser, 2020). Our formulation applies good-prefix monitoring to the question-dependent proposition $A(q, \cdot) = a$ for a declared continuation set $\mathcal{C}_o$. It distinguishes unsupported ($P$) from supported-but-revisable ($R$) prefixes, and answer closure ($C$) from terminal unanswerability ($U$). We study learning these verdicts from exact symbolic labels and causal video representations.

**Stopping and temporal controllers.** Closure induces a finite-horizon stopping problem related to deep stopping and FIRMBOUND's risk-earliness trade-off (Becker et al., 2019; Ebihara et al., 2025). GRU-D, LTC, and CfC model irregular observations (Che et al., 2018; Hasani et al., 2021; 2022). Our matched-controller comparisons separate the choice of temporal architecture from the closure supervision target.

# 3 Operator-Conditioned Answer Closure

## 3.1 Streaming setting

Let $X = (x_1, \ldots, x_T)$ be an ordered event stream with nondecreasing physical timestamps $s_1 \leq \cdots \leq s_T$. Throughout, $t$ indexes *consumed events*; $s_t$ is time in seconds. A frame and a later-consumed end token may share a timestamp. The horizon rule $H$ determines a terminal event index $J_H(y)$ and time $T_H(y) = s_{J_H(y)}$ on trajectory $y$. Write $J_X = J_H(X)$ and $H_X = T_H(X)$. After event $t \leq J_X$, the available information is $\mathcal{F}_t = \sigma(q, \eta_H, (x_i, s_i) : i \leq t)$, where $q$ is the question and $\eta_H$ identifies the horizon rule without revealing its future realization. $H$ may specify video end, a time limit, or an observable boundary, which affects the label only when consumed. Future events and unknown $J_X, H_X$ are excluded from $\mathcal{F}_t$; event order resolves simultaneous timestamps. Let $A(q, y) \in \mathcal{Y} \cup \{\perp\}$ be the semantic answer functional for a complete trajectory $y$. The symbol $\perp$ denotes no unique supported answer; paraphrases share one answer class.

## 3.2 Operators and admissible continuations

An operator $o \in \mathcal{O}$ maps event evidence to an answer. Table 1 defines the six base operators for real video. We reserve operator trees for symbolic composition tests. For prefix $x_{\leq t}$, the nonempty set $\mathcal{C}_o(x_{\leq t}; H)$ contains all task-admissible suffixes; at the observed horizon it is the singleton $\{\epsilon\}$, where $\epsilon$ is the empty suffix. Each suffix preserves the prefix and task constraints and terminates at its own $T_H$. The dataset fixes this set before observing the realized suffix, so shared prefixes have identical continuation sets. Symbolic suffixes are enumerable; video interventions sample suffixes and operator rules supply conditional certificates. We require continuation sets to be prefix-consistent: every suffix admissible after observing a longer prefix must extend a suffix that was already admissible before that observation. The operator, horizon, and task constraints cannot be changed after the new evidence is observed.

A certificate is an observable predicate $\mathrm{Cert}_o(x_{\leq t}, a; H)$. It is sound only if it entails $\mathsf{K}_t(a) = 1$ over the declared continuation set; Appendix A.3 states this condition explicitly. Table 1 lists the observable events from which the six base-operator certificates are constructed. Domain exhaustion is valid only when an observed task boundary excludes later in-domain evidence.

Table 1: Sufficient certificate families. A domain boundary may precede the global horizon; early closure means that no admissible in-domain evidence can change the answer.

| Operator | Local witness certificate | Domain-closing certificate |
|---|---|---|
| Exists | observed witness: positive | domain end without a witness: negative |
| For all | observed counterexample: negative | domain end without a counterexample: positive |
| Count | none in general | observed end of the counted domain |
| Last | none in general | relevant sequence termination |
| Accumulate | individual monotone claims | observed end of the answer domain |
| Complete | observed DONE: positive | domain end without DONE: negative |

## 3.3 Closure and four states

Let $a_t$ be the operator's current-evidence candidate, which may be absent. Define $\mathrm{Supp}_t(a) = 1$ when some consumed prefix supplies a registered evidence witness for candidate class $a$. This historical relation persists; it does not entail that $a$ answers the completed question or remains the current candidate. For Count, observing two objects supports the partial count 2, even if a third later

appears. Appendix A.2 makes the distinction executable for every operator. **Definition 1 (answer closure).** A candidate $a \in \mathcal{Y}$ is closed after event $t$, relative to $(q, o, H)$, when

$$\mathsf{K}_t(a) = \mathbf{1}[\mathrm{Supp}_t(a) = 1 \ \wedge \ \forall c \in \mathcal{C}_o(x_{\le t}; H),\ A(q, x_{\le t} \oplus c) = a] = 1. \tag{1}$$

The first closure event and its physical timestamp are

$$j^* = \min\{t : \exists a \in \mathcal{Y},\ \mathsf{K}_t(a) = 1\}, \qquad \tau^* = s_{j^*}. \tag{2}$$

If no such event $t \le J_X$ exists, both are undefined ($\varnothing$). Confidence such as $p(a \mid x_{\le t}, q) = 0.99$ aggregates likely outcomes. Equation 1, by contrast, requires invariance over the admissible set. A verified existential witness can therefore close a positive answer even when the generator is poorly calibrated.

Let $\mathcal{A}_t^{\mathrm{sup}} = \{a : \mathrm{Supp}_t(a) = 1\}$ and $\mathcal{A}_t^{\mathrm{cl}} = \{a : \mathsf{K}_t(a) = 1\}$. The four states form a disjoint decision rule: $S_t = C$ if $|\mathcal{A}_t^{\mathrm{cl}}| = 1$; otherwise $S_t = U$ at $t = J_X$, $S_t = P$ before $J_X$ if $\mathcal{A}_t^{\mathrm{sup}} = \varnothing$, and $S_t = R$ in the remaining pre-horizon case. Thus PENDING has no supported candidate, PROVISIONAL has support without answer closure, COMMITTED has a unique supported and closed answer, and UNANSWERABLE is the terminal residual state. Here $C$ denotes answer finality, not an action already emitted by the system. Although $P$ and $R$ both lead to continued observation, they distinguish missing support from supported evidence that remains revisable. Three-state controls test whether this finer supervision improves prediction or stopping. **Proposition 1 (absorption).** Under prefix consistency, $\mathsf{K}_t(a) = 1$ implies $\mathsf{K}_{t'}(a) = 1$ on every admissible complete extension $y$ at indices $t < t' \le J_H(y)$. Thus $C$ is absorbing. State $U$ is assigned only after the declared horizon is observed and is terminal by construction. A proof is given in Appendix A.3. For an accumulative answer $a = \{a_j\}$, each atomic claim has a closure time $\tau_j^*$. The set closes only when every admissible continuation preserves its contents and adds no required claim.

### 3.4 DECISION QUANTITIES

Let $(\hat{j}, \hat{a})$ be the first ANSWER event and candidate, with timestamp $\hat{\tau} = s_{\hat{j}}$. Define $j^\dagger = j^*$ if closure occurs and $J_X + 1$ otherwise; prematurity is $\hat{j} < j^\dagger$. Thus an answer before an end token can be premature even if both have the same physical timestamp.

The first terminal action is ANSWER, UNANSWERABLE, or abstention; absence of output is not recoded as $\perp$. A closable trajectory with no ANSWER by $H_X$ is a missed commitment. Appendix A.4 specifies the outcome accounting and optional application-weighted risk; no cost setting is treated as universal.

## 4 SHARED-PREFIX CLOSURE DATA

### 4.1 EXACT SYMBOLIC AND REAL-VIDEO SETTINGS

**Label semantics.** Symbolic labels exhaustively enumerate $\mathcal{C}_o$, making Equation 1 exact. Video interventions sample a finite subset $\widetilde{\mathcal{C}}_o \subseteq \mathcal{C}_o$. An answer-changing branch certifies non-closure, but failure to sample one does not certify closure. Positive video labels require a sufficient operator certificate and remain conditional on the continuation grammar and evidence annotation. At $H_X$, supported unique answers receive $C$; other records receive $U$. Appendix A.11 details these label categories.

**Symbolic event streams.** Dynamic programming enumerates reachable terminal answer classes under a bounded event grammar. A prefix closes exactly when its reachable set is a singleton non-$\perp$ supported answer; gold $A$, $S_t$, and $\tau^*$ require no visual perception. Table 2 reports record, source, and prefix-state counts. Each partition balances the six base operators, with four branches per intervention group. A separate 240-record test from 20 source-event components contains unseen binary AND, OR, and PAIR compositions. Base splits separate source-event components, not program families: 819 of 1,440 test records reuse a training program/token template with different timestamps. Structure holdouts and evidence-bearing prefix tests are distinguished in Appendix A.7.

Table 2: Data partitions and reference-state counts. Source components are disjoint across splits within each domain. Symbolic labels follow exhaustive finite continuations; video labels were AI-assisted and manually reviewed by an author. Prefixes within a record are correlated.

| Domain | Split | Records | Sources | $P$ | $R$ | $C$ | $U$ |
|---|---|---|---|---|---|---|---|
| Symbolic | Train | 2,880 | 120 | 4,236 | 7,052 | 7,167 | 853 |
| | Dev. | 960 | 40 | 1,428 | 2,376 | 2,442 | 288 |
| | Test | 1,440 | 60 | 2,177 | 3,892 | 3,775 | 428 |
| | Composition | 240 | 20 | 497 | 664 | 562 | 80 |
| Video pilot | Train | 12 | 6 | 49 | 267 | 43 | 1 |
| | Dev. | 3 | 2 | 3 | 69 | 3 | 0 |
| | Test | 7 | 2 | 15 | 84 | 35 | 0 |

**Real video.** The video study is an exploratory pilot with AI-assisted annotations manually reviewed and confirmed by an author. Visual and task-validity review retains 22 of 36 candidate records from 10 source components. The retained records cover all six base operators but are not operator-balanced: the counts are 4, 3, 4, 3, 6, and 2 in the order of Table 1. The source-grouped test contains seven records from two components; neither development nor test contains a $U$ target. The pilot evaluates reference-state prediction, not generated answers or end-to-end stopping.

### 4.2 Interventions and controls

Intervention groups share the question, operator, horizon declaration, timestamps, and hash-verified decoded prefix until divergence. Their suffixes preserve, change, or preclude an answer through added evidence, irrelevant content, or termination. Here *counterfactual* changes future evidence, unlike hypothetical questions about complete videos in ACQUIRED (Wu et al., 2023). Paired references agree before divergence; the video pilot contains three paired groups alongside unpaired records. Question, clock, boundary, and last-frame probes test simple cues. Appendix A.11 gives the intervention rules.

### 4.3 Annotation and splits

Each record is one branch, and all branches from a source component remain in one split. AI-assisted annotation assigns video evidence and state timelines at one frame per second; an author manually reviewed and confirmed the retained labels. The exclusions change the retained sample distribution. Prefix hashes and temporal checks verify structural consistency, not independent semantic correctness. Appendix A.11 specifies the review, exclusion, and split rules. Symbolic revision targets are uniform over all reachable terminal answer classes at each prefix, including $\perp$; they are not frequencies of the four generated branches. Video revision targets are masked.

## 5 Closure Estimation

The estimator learns reference closure states from causal prefixes (Figure 2). Its predictions approximate the formal target; a predicted $C$ is not itself a verified closure certificate.

### 5.1 Causal backbone interface

The video comparison freezes the released `wangyueqian/MMDuet2` backbone (Qwen2.5-VL-3B-Instruct), with $z_t = B_\psi(x_{\leq t}, q)$ its final-normalized last-prompt-token state. Fixed pooling uses the full causal prefix without dynamic token dropping. Reference states and closure times are training labels, never inputs. The symbolic comparison replaces $z_t$ with nine-dimensional causal event features and supplies a shared rule-based candidate interface for evaluating decisions. Thus symbolic answer accuracy measures decisions over symbolic candidates, whereas the video pilot measures reference-state prediction.

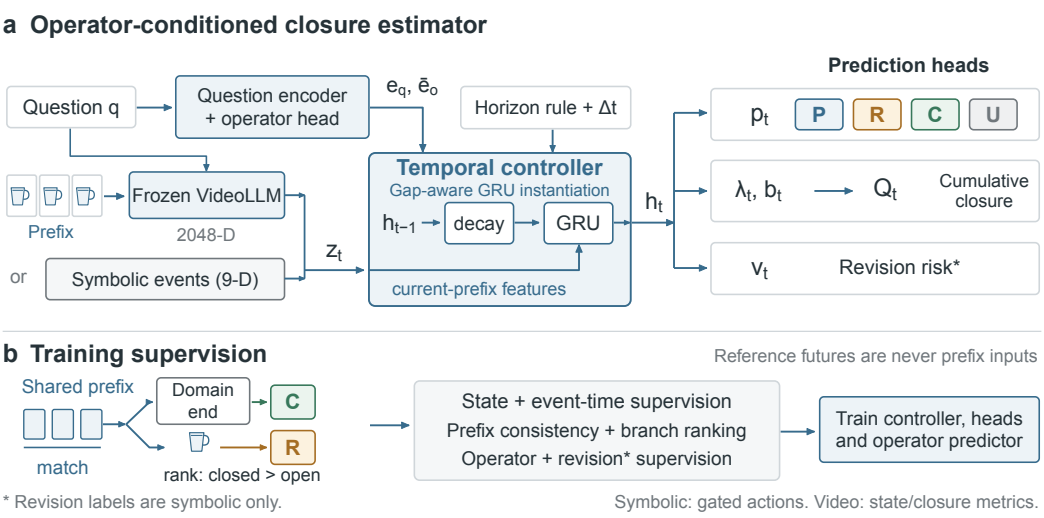


Figure 2: Operator-conditioned closure estimator. **a**, Causal symbolic or frozen-video features enter a temporal controller, illustrated with a gap-aware GRU. Question, predicted operator, horizon rule, and elapsed time condition the state, closure, and revision heads. **b**, Reference labels and paired-prefix constraints supply training supervision; future observations never enter the prefix representation. Revision labels and action evaluation are symbolic only.

### 5.2 Operator-conditioned temporal controller

The question encoder produces $e_q = g_q(q)$ and predicts $\pi = \mathrm{softmax}(W_o e_q) = p_\phi(o \mid q)$. The soft operator embedding is $\bar{e}_o = \sum_o \pi(o) e_o$. The horizon embedding is $e_H = E_H[\eta_H]$, where $\eta_H$ reveals the boundary type but no future content. For $\Delta t_t = s_t - s_{t-1}$,

$$h_t = f_\theta(h_{t-1}, [z_t, e_q, \bar{e}_o, e_H], \Delta t_t)\,. \tag{3}$$

The primary comparison uses $\bar{e}_o$ from predicted operators during both training and inference. Gold one-hot and absent operator embeddings provide two controls.

We instantiate $f_\theta$ with a gap-aware GRU inspired by GRU-D (Che et al., 2018). It embeds $\log(1 + \Delta t_t)$, decays the previous state, and then applies a GRU cell; Appendix A.4 gives the exact update. Matched alternatives include a prefix MLP, a vanilla GRU, GRU-$\Delta t$, and CfC (Hasani et al., 2022). Mamba and LIF provide state-space and spiking alternatives (Gu & Dao, 2024; Neftci et al., 2019). Appendix A.5 specifies the implementations, including the unfused Mamba-1 reference scan.

### 5.3 Outputs and loss

The controller predicts state probabilities $p_t$, closure intensity $\lambda_t$, terminal closure probability $b_t$, and revision risk $v_t$. State supervision distinguishes evidence conditions; the event-time objective learns when first closure occurs. Revision risk is an auxiliary symbolic target measuring candidate disagreement across reachable terminal answer classes. Interval closure probability is $r_t = 1 - \exp(-\lambda_t \Delta t_t)$, with $r_t = b_t$ for an observed zero-duration terminal token; $Q_t = 1 - \prod_{j \le t}(1 - r_j)$ is the cumulative probability of first closure. These predictions use the arriving observation and never permit an action inside an unobserved gap. Appendix A.4 gives the head equations, censored event likelihood, and revision targets. The total objective is

$$\mathcal{L} = \mathcal{L}_{\text{state}} + \lambda_h \mathcal{L}_{\text{haz}} + \lambda_p \mathcal{L}_{\text{pair}} + \lambda_o \mathcal{L}_{\text{op}} + \lambda_v \mathcal{L}_{\text{rev}}. \tag{4}$$

$\mathcal{L}_{\text{state}}$ is class-balanced four-state cross-entropy; $\mathcal{L}_{\text{op}}$ supervises operators and $\mathcal{L}_{\text{rev}}$ is masked binary cross-entropy. The pair loss matches pre-divergence predictions and ranks a closed branch above a time-matched open branch using $\chi_t = p_t(C) Q_t$. Development-only temperature scaling gives $\tilde{p}_t$ and $\tilde{Q}_t$ (Guo et al., 2017); their product $\tilde{\chi}_t = \tilde{p}_t(C)\tilde{Q}_t$ supplies AP, Brier and ECE, but is not a guaranteed calibrated probability. The heads have no hard consistency constraint; ranking and state classification are evaluated separately. The matched readiness baseline replaces closure-specific

losses with binary cross-entropy on current candidate support, retaining the inputs, architecture, optimizer, and training budget. Its calibrated readiness score drives the same dwell-policy grid; Appendix A.4 specifies the comparison.

### 5.4 Derived actions

The policy maps $P/R$ to CONTINUE. An available candidate is emitted only when predicted state $C$ and calibrated state/closure scores satisfy threshold and dwell gates. At the observed horizon, a failed answer gate yields UNANSWERABLE if the $U$ gate passes, or ABSTAIN otherwise. Abstention is distinct from a predicted $U$. Development data select temperatures and commitment thresholds; $\alpha_U = 0.5$ is fixed. Appendix A.4 specifies the gate, dwell-reset, and selection rules.

## 6 Experiments

### 6.1 Evaluation protocol

Comparisons share domain-specific splits and inputs, with seeds 17, 29, and 43 and 20 AdamW epochs (Appendix A.4). Checkpoint selection uses a record-weighted mean of batch-local development losses; calibration also uses development data only. The main controller matrix uses CUDA; three-state comparisons use paired CPU retrainings, and direct binary closure uses CPU. Evaluation uses frozen checkpoints; Appendices A.6 and A.7 identify training regimes.

We evaluate closure ranking with precision–recall summaries (Davis & Goadrich, 2006). Closure AUPR is non-interpolated average precision (AP), averaged equally across operators. Test positive prevalences are 0.3675 (symbolic) and 0.2418 (video). State F1 first averages the four class F1 scores, assigning zero to a class absent from both targets and predictions, and then averages over operators. Video F1 includes absent $U$; standard deviations describe seed variation. We report Brier score (Brier, 1950) and adapt ECE (Guo et al., 2017) to 15 equal-frequency bins. Symbolic action metrics balance terminal classes within intervention groups, then average groups; paired intervals use 2,000 source-component bootstrap resamples. Pilot results are descriptive, with only two test components.

**Baselines.** Table 3 pairs closure and readiness training within MLP, GRU, GRU-$\Delta t$, gap-aware GRU, CfC, Mamba, and LIF controllers. Direct binary closure uses BCE on $\mathbf{1}[S_t = C]$ and the shared operator loss; an event-time survival baseline predicts first closure without state or revision supervision. Its calibrated score is cumulative event probability, whereas the four-state estimator uses $\tilde{\chi}_t$.

### 6.2 Support supervision versus closure supervision

Closure supervision produced higher closure AP than support supervision in all seven controllers (Table 3). This diagnoses target mismatch, not superiority over a published readiness system. Gap-aware GRU reached symbolic AP 0.9954 and four-state F1 0.9851; ordinary GRU and Mamba reached similar AP (0.9955 and 0.9957). The distinction between support and closure persisted across architectures.

### 6.3 Four-state supervision on a common closure target

Direct binary closure reached symbolic AP 0.9858 and survival reached 0.9953, compared with 0.9954 for the primary four-state model (Table 4). In paired CPU comparisons with target-specific class balancing, three-state and four-state AP were 0.9955 and 0.9953 (Figure 3; Table 18). Matching state-group loss mass and the scalar policy likewise gave no uniform four-state ranking advantage (Table 24).

### 6.4 Premature commitment and the cost of waiting

At answer coverage 0.6412, prematurity was 0.0239 for binary closure, 0.0040 for survival, and 0.0008 for the primary four-state model. Their conditional delays were 0.0363, 0.0612, and 0.0683, respectively (Table 4). The four-state model's $U$ coverage of 0.3588 raised total coverage to 1 without

Table 3: Closure versus readiness across controllers. Values are three-seed means; AP is operator-macro average precision against the closure target. C and B denote closure and binary-readiness training. F1 is the four-state macro-F1 of C, averaged over operators. Symbolic test: 1,440 records, 60 components. Human-reviewed video pilot: 7 records, 2 components. Seed variation appears in Table 6. The survival baseline uses only event-time supervision and has no state head.

| | Symbolic | | | Video pilot | | |
|---|---|---|---|---|---|---|
| Controller | AP (C) | AP (B) | F1 (C) | AP (C) | AP (B) | F1 (C) |
| MLP | 0.9835 | 0.5542 | 0.9213 | 0.9232 | 0.3638 | 0.2360 |
| GRU | 0.9955 | 0.4798 | 0.9865 | 0.9212 | 0.4264 | 0.2351 |
| GRU-$\Delta t$ | 0.9953 | 0.4786 | 0.9851 | 0.9209 | 0.4236 | 0.2837 |
| Gap-aware GRU | 0.9954 | 0.4771 | 0.9851 | 0.9180 | 0.4381 | 0.2900 |
| CfC | 0.9952 | 0.4957 | 0.9850 | 0.9204 | 0.3540 | 0.3381 |
| Mamba | 0.9957 | 0.4654 | 0.9826 | 0.9375 | 0.2596 | 0.4084 |
| LIF (current) | 0.9812 | 0.4927 | 0.9205 | 0.9222 | 0.2974 | 0.1221 |
| Survival baseline | 0.9953 | n/a | n/a | 0.9258 | n/a | n/a |

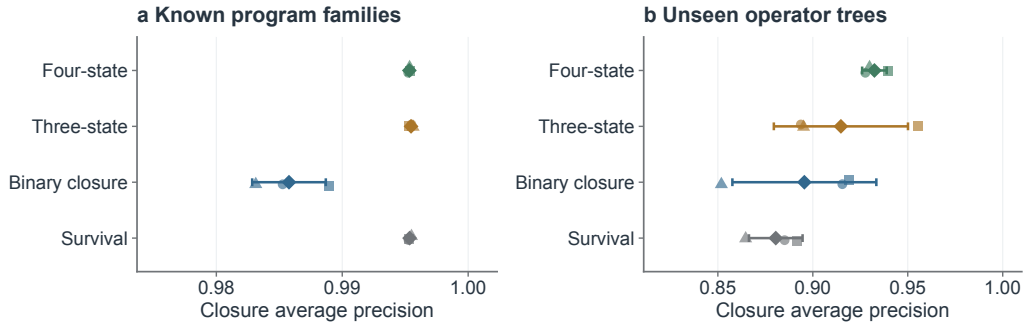


Figure 3: Same-target closure prediction. **a**, Known program families: 1,440 branches from 60 sources. **b**, Nine unseen operator trees: 1,152 branches from 32 sources. Small points are seeds, diamonds means, and bars seed SD; horizontal scales differ. Four-state and three-state results use the paired comparison described in Appendix A.7; all methods predict the same closure label.

additional answers. An observed-domain-end policy eliminated premature answers but incurred delay 0.0967. Both three-state comparisons likewise showed lower four-state prematurity with higher delay and different answer coverage.

Against support-only readiness, the upper one-sided 95% bound for increased delay was 0.0737, exceeding the 0.05 safeguard. The joint stopping criterion therefore failed (Table 11). Appendix A.6 reports policy-grid sensitivity, paired intervals, and an equal-action-repertoire control.

Table 4: Symbolic stopping controls (three-seed means). Readiness diagnoses target mismatch; binary closure and survival predict closure. Ans. counts misses as errors but ignores timing: a correct realized answer can still be premature. A-cov. and U-cov. partition coverage; delay conditions on commitment to a closable record. No-U maps U to abstention. The exact monitor uses known semantics on the same inputs.

| Method | Ans. | A-cov. | U-cov. | Prem. | Miss | Delay |
|---|---|---|---|---|---|---|
| Binary readiness | 0.5491 | 0.9204 | 0.0000 | 0.9204 | 0.0000 | 0.0000 |
| Binary closure | 0.9859 | 0.6412 | 0.0000 | 0.0239 | 0.0000 | 0.0363 |
| Survival | 0.9995 | 0.6412 | 0.0000 | 0.0040 | 0.0000 | 0.0612 |
| Four-state closure | 1.0000 | 0.6412 | 0.3588 | 0.0008 | 0.0000 | 0.0683 |
| Four-state, no U action | 1.0000 | 0.6412 | 0.0000 | 0.0008 | 0.0000 | 0.0683 |
| Observed domain end | 1.0000 | 0.6412 | 0.0000 | 0.0000 | 0.0000 | 0.0967 |
| Exact rule monitor | 1.0000 | 0.6412 | 0.3588 | 0.0000 | 0.0000 | 0.0000 |

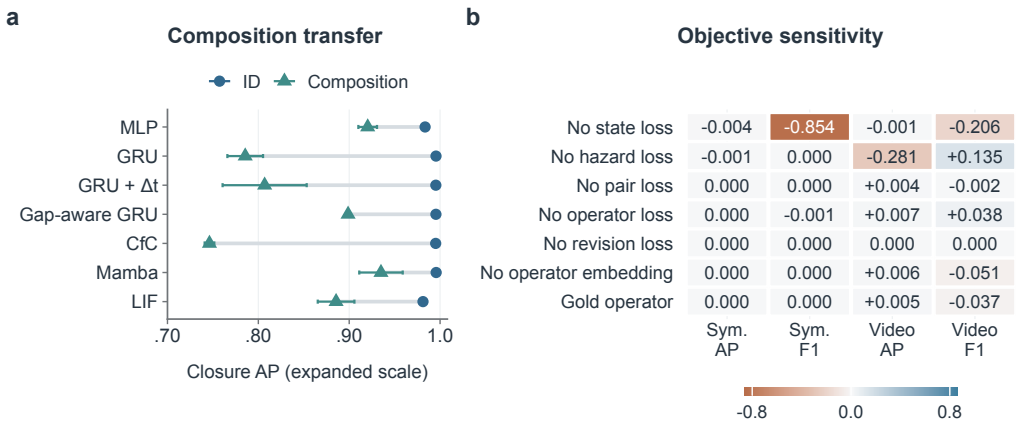


Figure 4: Composition and loss ablations. **a**, AP means and seed SD for base operators and 240 composition records from 20 sources. **b**, Changes from full gap-aware GRU in AP and state F1; colors show magnitude, not significance. Video revision targets are masked. Exact scores appear in Tables 8 and 7.

### 6.5 Generalization and Ablations

**Symbolic transfer.** On unseen compositions, gap-aware GRU obtained AP 0.8988, compared with 0.7855 for GRU, 0.7462 for CfC, and 0.9350 for Mamba. The MLP reached 0.9203 (Figure 4a; Table 8). On nine further unseen trees, four-state, three-state, binary-closure, and survival AP were 0.9325, 0.9148, 0.8956, and 0.8805. Eight trees share base-operator state semantics; the largest gains concern unfamiliar encodings of Last. All four models exceeded 0.9977 on longer bodies, although time rescaling changed their stopping policies (Appendix A.7).

On 2,304 evidence-bearing-prefix branches, the same four targets gave AP 0.9947, 0.9954, 0.9645, and 0.9957 (Table 22). For already-closed histories, four-state AP 0.9980 coexisted with delay 0.5853 and 25% missed answers (Table 23). Removing state loss left symbolic AP at 0.9915 but reduced state F1 to 0.1312. Removing pair loss or the auxiliary operator embedding barely changed symbolic AP; questions still encoded the operator tree (Table 7).

**Video representations and boundary cues.** On the human-reviewed pilot, gap-aware GRU reached AP 0.9180 versus readiness AP 0.4381, but four-state F1 was 0.2900. Mamba reached AP 0.9375 and F1 0.4084; survival reached AP 0.9258 without a state head. End-only and question/clock/end controls reached AP 0.8282 and 0.8969 (Table 9). Much of the ranking was therefore accessible from task and boundary cues without visual features. Removing hazard loss reduced AP to 0.6368 but increased F1 to 0.4245 (Figure 4b), further separating ranking from state classification. All layer probes following TRAPSBench (Pramono et al., 2026) appear in Table 10.

## 7 Discussion and Limitations

Answer closure distinguishes a supported candidate from a final answer under task-admissible futures. Shared-prefix continuations test this distinction with identical observed evidence. Four states expose different evidence conditions, yet high closure AP can coexist with incorrect states, missed answers, or long delays. Stopping gains therefore depend on the operating point and the cost of waiting.

Symbolic labels assume an exact grammar, and expression holdouts often reuse known monitoring semantics. No estimator was uniformly best; the joint delay safeguard failed. The pilot's two test components, absent $U$, and boundary shortcuts leave video generalization, unanswerability detection, and generated-answer quality unresolved.

### AI USE STATEMENT

OpenAI Codex assisted literature search, research ideation, formalization, experiment design, code drafting, manuscript drafting and revision, result interpretation, and consistency checks, including review of mathematical arguments. OpenAI image generation assisted early conceptual figure drafts; the final figures were reconstructed as editable vector graphics. AI assistance also contributed to the symbolic-data implementation and initial video annotation. An author manually reviewed and confirmed the video labels; the annotation procedure combines AI assistance with author review. All numerical results derive from recorded model executions and deterministic aggregation. The authors take responsibility for the manuscript, artifacts, and interpretation.

### ETHICS STATEMENT

Exposing a provisional state may make temporal uncertainty easier to inspect, but a committed token may also invite misplaced trust. Other risks include biased continuation design, privacy or licensing violations in source video, and annotation burden. AI-assisted annotation and author review may retain perceptual errors or reviewer bias; label consistency alone does not establish correctness. Source media remain subject to their original licenses and access conditions. High-stakes deployments should explain the task-relative meaning of commitment and retain human override. Calibration on this benchmark does not constitute a safety guarantee.

### REPRODUCIBILITY STATEMENT

The experimental record includes the symbolic generator, pilot review rationales and exclusions, source-grouped split manifests, model interfaces, seeds, configurations, and checkpoint hashes. The supplementary archive `supplementary_material.zip` provides a `README.md` with extraction and verification commands; the extracted `manifest.json` records file identities. The result tables are generated from recorded metrics with checks against per-run reports and checkpoint identities. Pilot predictions evaluate agreement with human-reviewed, AI-assisted reference labels, while the WEB baseline retains original question annotations and a frozen local judging protocol. Raw judge responses support an exact cached replay of all 500 WEB examples and their aggregate scores. Appendix A covers formal edge cases and additional implementation details.

## REFERENCES

Shehreen Azad, Vibhav Vineet, and Yogesh Singh Rawat. StreamReady: Learning what to answer and when in long streaming videos, 2026. URL https://arxiv.org/abs/2603.08620.

Andreas Bauer, Martin Leucker, and Christian Schallhart. Runtime verification for LTL and TLTL. *ACM Transactions on Software Engineering and Methodology*, 20(4):14:1–14:64, 2011. doi: 10.1145/2000799.2000800. URL https://doi.org/10.1145/2000799.2000800.

Sebastian Becker, Patrick Cheridito, and Arnulf Jentzen. Deep optimal stopping. *Journal of Machine Learning Research*, 20(74):1–25, 2019. URL https://jmlr.org/papers/v20/18-232.html.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Dibyadip Chatterjee, Edoardo Remelli, Yale Song, Bugra Tekin, Abhay Mittal, Bharat Bhatnagar, Necati Cihan Camgöz, Shreyas Hampali, Eric Sauser, Shugao Ma, Angela Yao, and Fadime Sener. Memory-efficient streaming VideoLLMs for real-time procedural video understanding, 2025. URL https://arxiv.org/abs/2504.13915.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8:6085, 2018. doi: 10.1038/s41598-018-24271-9.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. VideoLLM-online: Online video large

language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/CVPR52733.2024.01742.

Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. LiveCC: Learning video LLM with streaming speech transcription at scale, 2025. URL https://arxiv.org/abs/2504.16030.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240, 2006. doi: 10.1145/1143844.1143874.

Akinori F. Ebihara, Taiki Miyagawa, Kazuyuki Sakurai, and Hitoshi Imaoka. Learning the optimal stopping for early classification within finite horizons via sequential probability ratio test. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2501.18059.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, 2019. URL https://arxiv.org/abs/1901.09192.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 2024. doi: 10.18653/v1/2024.naacl-long.366.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2312.00752.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017. URL https://arxiv.org/abs/1706.04599.

Yangyang Guo, Fangkai Jiao, Zhiqi Shen, Liqiang Nie, and Mohan Kankanhalli. UNK-VQA: A dataset and a probe into the abstention ability of multi-modal large models, 2023. URL https://arxiv.org/abs/2310.10942.

Ramin Hasani, Mathias Lechner, Alexander Amini, Daniela Rus, and Radu Grosu. Liquid time-constant networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/16936.

Ramin Hasani, Mathias Lechner, Alexander Amini, Lucas Liebenwein, Aaron Ray, Max Tschaikowski, Gerald Teschl, and Daniela Rus. Closed-form continuous-time neural networks. *Nature Machine Intelligence*, 4:992–1003, 2022. doi: 10.1038/s42256-022-00556-7.

Junho Kim, Hosu Lee, James M. Rehg, Minsu Kim, and Yong Man Ro. STRIDE: When to speak meets sequence denoising for streaming video understanding, 2026. URL https://arxiv.org/abs/2603.27593.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs fail on unanswerable questions, 2025. URL https://arxiv.org/abs/2506.09038.

Emre O. Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019. doi: 10.1109/MSP.2019.2931595.

Fnu Pramono, John Cai, and Sourabh Kulkarni. TRAPSBench: Vision-language models encode but fail to express epistemic restraint, 2026. URL https://arxiv.org/abs/2608.13167.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video LLMs with active real-time interaction via disentangled perception, decision, and reaction, 2025. URL https://arxiv.org/abs/2501.03218.

Patrick Schäfer and Ulf Leser. TEASER: Early and accurate time series classification. *Data Mining and Knowledge Discovery*, 34:1336–1362, 2020. doi: 10.1007/s10618-020-00690-z. URL https://doi.org/10.1007/s10618-020-00690-z.

Tejas Srinivasan, Jack Hessel, Tanmay Gupta, Bill Yuchen Lin, Yejin Choi, Jesse Thomason, and Khyathi Chandu. Selective "selective prediction": Reducing unnecessary abstention in vision-language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. doi: 10.18653/v1/2024.findings-acl.767.

Sirnam Swetha, Hilde Kuehne, and Mubarak Shah. TimeLogic: A temporal logic benchmark for video QA, 2025. URL https://arxiv.org/abs/2501.07214.

Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. StreamBridge: Turning your offline video large language model into a proactive streaming assistant, 2025a. URL https://arxiv.org/abs/2505.05467.

Shijie Wang, Honglu Zhou, Ziyang Wang, Ran Xu, Caiming Xiong, Silvio Savarese, Chen Sun, and Juan Carlos Niebles. Evidence-backed video question answering, 2026. URL https://arxiv.org/abs/2607.11862.

Yueqian Wang, Songxiang Liu, Disong Wang, Nuo Xu, Guanglu Wan, Huishuai Zhang, and Dongyan Zhao. MMDuet2: Enhancing proactive interaction of video MLLMs with multi-turn reinforcement learning, 2025b. URL https://arxiv.org/abs/2512.06810.

Yueqian Wang, Xiaojun Meng, Yifan Wang, Huishuai Zhang, and Dongyan Zhao. ProactiveVideoQA: A comprehensive benchmark evaluating proactive interactions in video large language models, 2025c. URL https://arxiv.org/abs/2507.09313.

Te-Lin Wu, Zi-Yi Dou, Qingyuan Hu, Yu Hou, Nischal Reddy Chandra, Marjorie Freedman, Ralph M. Weischedel, and Nanyun Peng. ACQUIRED: A dataset for answering counterfactual questions in real-life videos. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://arxiv.org/abs/2311.01620.

Weicai Yan, Yuhong Dai, Qi Ran, Haodong Li, Wang Lin, Tao Jin, Xing Xie, Hao Liao, and Jianxun Lian. Proact-VL: A proactive VideoLLM for real-time AI companions, 2026. URL https://arxiv.org/abs/2603.03447.

# A Formal and Dataset Details

## A.1 Complementary literature

ProVideLLM studies memory-efficient streaming representations; LiveCC learns from temporally aligned speech transcripts (Chatterjee et al., 2025; Chen et al., 2025). Selective prediction and unanswerability benchmarks measure restraint (Geifman & El-Yaniv, 2019; Guo et al., 2023; Srinivasan et al., 2024; Kirichenko et al., 2025). TRAPSBench identifies an encoding-expression gap on matched knowable and unknowable inputs (Pramono et al., 2026). Layer-wise decodability in the present study is a representation diagnostic rather than evidence about the generator's speaking decisions. The four states operationalize good-prefix monitoring for a question-dependent answer-equality proposition. Both $P$ and $R$ are inconclusive, but only $R$ contains an evidence-supported candidate; $C$ records answer-specific closure, whereas $U$ records terminal unanswerability. The formal contribution is the task specification and four-state refinement, not a replacement for classical monitoring semantics. Evaluation uses exact labels when possible and treats semantic matching as a separate analysis.

## A.2 Operator edge cases

Figure 5 summarizes the six certificate families. Exists-positive and For-all-negative answers close after a verified witness and counterexample, respectively; their complements require domain exhaustion. For Count and Last, an event boundary replaces video end only when later events lie outside the quantified domain. Individual accumulative claims can close before the full set. A supported claim remains provisional whenever an admissible suffix can invalidate it.

Every prefix admits its realized continuation. A branch contradicting an absorbing label requires re-examination of the record and its certificate. Composed operators are evaluated bottom-up as trees; holdout tests expression-level generalization, with monitoring-semantic equivalence assessed separately.

**Candidate, historical support, and terminal answer.** In the symbolic implementation, the current candidate $a_t$ is the operator value on the events consumed so far. It is absent at START and whenever a required value, such as the last object, is undefined. Historical support is $\mathrm{Supp}_t(a) = \mathbf{1}[\exists k \le t : a_k = a]$. This deliberately weak predicate records a partial-evidence value. It cannot be substituted for entailment of the terminal answer $A(q, X)$. For video, the analogous witness is an annotated evidence span; its validity depends on visual interpretation.

Table 5: Implemented base-operator semantics. The predicates concern in-domain events. A declared domain boundary prevents later in-domain evidence. $\perp$ means that the completed record does not support a unique terminal answer.

| Operator | Current candidate after a body event | Terminal answer |
|---|---|---|
| Exists | whether a target has appeared | true after a witness; false after domain end without a witness; otherwise $\perp$ |
| For all | whether no non-target has appeared | false after a counterexample; true after domain end without one; otherwise $\perp$ |
| Count | number of target events so far | count after domain end; otherwise $\perp$ |
| Last | most recent A/B object, if any | last object after domain end; $\perp$ if absent or domain remains open |
| Accumulate | set of A/B types seen so far | complete set after domain end; otherwise $\perp$ |
| Complete | whether DONE has appeared | true after DONE; false after domain end without DONE; otherwise $\perp$ |

For example, with target A, the Count stream START, A, A, B, DOMAIN_END, STREAM_END has candidates absent, $1, 2, 2, 2, 2$ and states $P, R, R, R, C, C$, for a budget allowing those suffixes. Historical support contains both 1 and 2; only 2 closes. On the same stream, Last has candidates absent, A, A, B, B, B and the same state sequence, but closes on B. Replacing DOMAIN_END with immediate STREAM_END yields $U$ under the explicit-completion semantics of both operators, even though the last partial candidate exists. These terminal rules are part of the task definition, not universal conventions for counting truncated videos.

**Composition with an undefined value.** All children share the target, event budget, domain, and horizon of their program. AND and OR use the following three-valued rules, with F and T denoting Boolean values rather than epistemic states:

$$\begin{array}{c|ccc} \wedge & \mathsf{F} & \mathsf{T} & \perp \\ \hline \mathsf{F} & \mathsf{F} & \mathsf{F} & \mathsf{F} \\ \mathsf{T} & \mathsf{F} & \mathsf{T} & \perp \\ \perp & \mathsf{F} & \perp & \perp \end{array} \qquad \begin{array}{c|ccc} \vee & \mathsf{F} & \mathsf{T} & \perp \\ \hline \mathsf{F} & \mathsf{F} & \mathsf{T} & \perp \\ \mathsf{T} & \mathsf{T} & \mathsf{T} & \mathsf{T} \\ \perp & \perp & \mathsf{T} & \perp \end{array} . \tag{5}$$

PAIR$(u, v)$ is the ordered pair $(u, v)$ when both children are defined and $\perp$ otherwise. These rules apply both to partial candidates and terminal values. For target A, a truncated stream A, STREAM_END gives Exists$= \mathsf{T}$ and Complete$= \perp$: their OR is T, whereas their AND and PAIR are $\perp$. Thus a Boolean short circuit can close a composition even when one child is undefined; closure is computed from the composed reachable answer set, not by requiring every child to close.

### A.3 Absorption and certificate soundness

A certificate is sound when

$$\mathrm{Cert}_o(x_{\le t}, a; H) = 1 \quad \Longrightarrow \quad \mathrm{Supp}_t(a) = 1 \ \wedge \ \forall c \in \mathcal{C}_o(x_{\le t}; H),\ A(q, x_{\le t} \oplus c) = a. \tag{6}$$

Fix an admissible complete trajectory $y$ and event indices $t < t' \le J_H(y)$; write $x_{\le t'} = x_{\le t} \oplus u$. Prefix consistency is the condition

$$v \in \mathcal{C}_o(x_{\le t'}; H) \quad \Longrightarrow \quad u \oplus v \in \mathcal{C}_o(x_{\le t}; H). \tag{7}$$

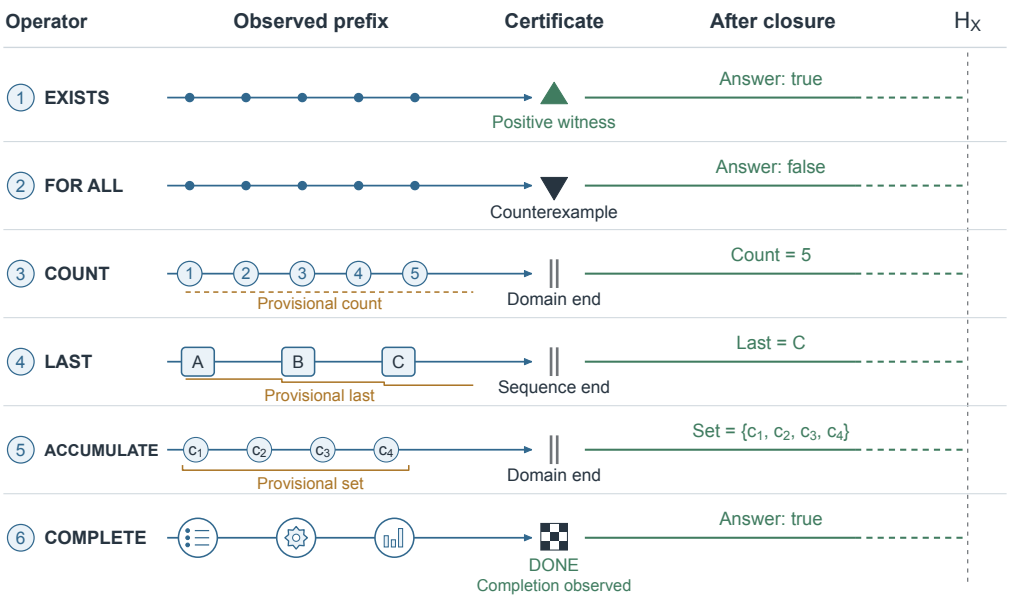


Figure 5: Operator-specific closure certificates. Blue marks observations, orange provisional candidates, and green answers fixed after a certificate. Double bars mark a domain boundary; DONE marks task completion, not the stream horizon. Each row illustrates one certificate case. Negative Exists and Complete, and positive For-all, instead require domain exhaustion without the corresponding witness, completion, or counterexample (Table 1). That boundary may precede $H_X$; these alternative answers are not later transitions on the illustrated closed trajectories.

Assume $\mathsf{K}_t(a) = 1$. The supporting evidence span remains in $x_{\le t'}$, so $\mathrm{Supp}_{t'}(a) = 1$. For any $v \in \mathcal{C}_o(x_{\le t'}; H)$, prefix consistency gives $u \oplus v \in \mathcal{C}_o(x_{\le t}; H)$. Definition 1 then gives $A(q, x_{\le t'} \oplus v) = A(q, x_{\le t} \oplus u \oplus v) = a$. Because support persists and the equality holds for every later admissible continuation, $\mathsf{K}_{t'}(a) = 1$. This proves absorption of $C$. State $U$ is evaluated only after the horizon token or boundary enters $\mathcal{F}_t$, so no later in-task transition exists.

For Exists, a valid positive witness is preserved by every admissible suffix. For For-all, a valid counterexample is preserved in the same way. For Count and Last, and for completeness of an accumulated set, the relevant value is fixed only after an observed boundary closes the quantified domain. For Complete, a positive claim closes at DONE; non-completion closes when an observed domain boundary excludes later completion, which may precede the global horizon. In each case, the record stores the rule, its evidence spans, the boundary when required, and the answer entailed by the rule. These fields make the certificate auditable; checking their presence alone does not establish Equation 6 or the correctness of a visual annotation.

### A.4 Controller equations and decision metrics

The controller predicts the state distribution $p_t = \mathrm{softmax}(W_s h_t)$. It also predicts a nonnegative endpoint-conditioned physical-time intensity $\lambda_t = \mathrm{softplus}(w_h^\top h_t) + 10^{-6}$, a terminal closure probability $b_t = \sigma(w_b^\top h_t)$, and revision risk $v_t = \sigma(w_v^\top h_t)$. Affine biases are suppressed in this notation. For ordinary intervals, $r_t = 1 - \exp(-\lambda_t \Delta t_t)$. For an observed terminal token with $\Delta t_t = 0$, $r_t = b_t$. This complementary-log-log parameterization is evaluated only after $x_t$ arrives. It models an event attached to $(s_{t-1}, s_t]$ and never permits an action inside an unobserved gap. For a current candidate $a_t$, $v_t$ estimates revision under the reachable-class measure defined below. The symbolic generator computes this target by dynamic programming over all admissible suffixes; video supervision is masked. For uncensored closure, $j^*$ is the first closed *event index* in Equation 2, not the first event with timestamp at least $\tau^*$.

$$\mathcal{L}_{\mathrm{haz}} = -\sum_{j<j^*} \log(1 - r_j) - \log r_{j^*}; \tag{8}$$

When closure does not occur by $H_X$, the objective omits the final term and enforces survival through the observed horizon. A zero-duration terminal event contributes the Bernoulli term for $b_t$ to $\mathcal{L}_{\text{haz}}$ instead of an artificial elapsed time. $\Delta t_t$ records exposure in seconds but does not assume frame-rate invariance. The state loss distinguishes UNANSWERABLE from low closure intensity. All recorded trajectories start non-closed, and any first closure is attached to a positive-duration observation or the terminal Bernoulli event. Nonterminal zero-gap observations have $r_t = 0$ and cannot initiate closure in this parameterization. Initially closed streams or simultaneous nonterminal closure events require an additional point-mass head; they are outside the evaluated protocol.

Before divergence, $\mathcal{L}_{\text{pair}}$ matches stochastic views of the same prefix in state and cumulative closure probability. Dropout supplies distinct stochastic views without changing observed evidence; branch metadata is excluded from model inputs. After divergence, the loss ranks $\chi_t$ for a certified-closed branch above an open branch matched in physical time. The experiments fix the margin at $m = 0.2$. For the matched binary-readiness baseline, $y_t^{\text{ready}} = \text{Supp}_t(a_t)$ for the current annotated candidate and zero when no candidate exists. It reuses the backbone, temporal controller, inputs, operator head, candidate interface, optimizer, training steps, seeds, and threshold/dwell grid. Binary cross-entropy replaces the four-state and closure-specific losses, and the calibrated readiness score alone drives the matched dwell policy.

For the gap-aware GRU, let $\rho_t = \text{MLP}_\Delta(\log(1 + \Delta t_t))$ and $\gamma_t = \exp[-\text{softplus}(W_\gamma \rho_t + b_\gamma)]$. The implemented update is

$$\widetilde{h}_{t-1} = \gamma_t \odot h_{t-1}, \quad u_t = W_x[z_t, e_q, \bar{e}_o, e_H, \rho_t] + b_x, \quad h_t = \text{GRUCell}(u_t, \widetilde{h}_{t-1}). \tag{9}$$

Let $\mathcal{R}_t = \{A(q, x_{\le t} \oplus c) : c \in \mathcal{C}_o(x_{\le t}; H)\}$ be the finite set of reachable terminal answer classes, including $\perp$ when reachable. For current candidate $a_t$, the implemented revision target is

$$y_t^{\text{rev}} = \frac{1}{|\mathcal{R}_t|} \sum_{y \in \mathcal{R}_t} \mathbf{1}[y \neq a_t]. \tag{10}$$

This is a uniform measure on reachable answer classes, not on suffix paths or the four sampled branches. It is a task-defined auxiliary target, not a probability of real-world futures; it is masked when $a_t$ is absent. For a pre-divergence pair $(i, j)$, the pair term is $\text{JS}(p_t^i, p_t^j) + |Q_t^i - Q_t^j|$. For a post-divergence closed/open pair it is $[m - \chi_t^i + \chi_t^j]_+$, where $\chi_t = p_t(C)Q_t$.

Losses use raw probabilities. For the policy, state logits are divided by development temperature $T_s$, and $\widetilde{Q}_t = \sigma(\text{logit}(\text{clip}(Q_t, 10^{-6}, 1 - 10^{-6}))/T_Q)$. Write the resulting state distribution as $\widetilde{p}_t$. The calibrated commitment gate and dwell update are

$$\begin{aligned} g_t &= \mathbf{1}[\arg\max_s \widetilde{p}_t(s) = C]\mathbf{1}[\widetilde{p}_t(C) \ge \alpha_C]\mathbf{1}[\widetilde{Q}_t \ge \alpha_Q]\mathbf{1}[a_t \neq \varnothing], \\ d_t &= g_t g_{t-1}(d_{t-1} + \Delta t_t), \qquad g_0 = 0. \end{aligned} \tag{11}$$

The action is ANSWER only when $g_t = 1$ *and* $d_t \ge \delta_C$, including when $\delta_C = 0$. Thus a new threshold crossing starts at zero dwell and cannot receive credit for the preceding unobserved gap; a failed gate resets the accumulator. Dwell measures persistence of the gate, not of the answer value: changing $a_t$ does not reset it. The primary symbolic operating points use zero dwell. If this condition fails, return CONTINUE for $t < J_X$; at $t = J_X$, return UNANSWERABLE only if $\arg\max_s \widetilde{p}_t(s) = U$ and $\widetilde{p}_t(U) \ge \alpha_U$, otherwise ABSTAIN. The answer gate has priority.

For event-order comparisons, set $\bar{j} = \hat{j}$ after a commitment and $\bar{j} = J_X + 1$ otherwise. Let $a^* = A(q, X) \in \mathcal{Y} \cup \{\perp\}$. At $H_X$, the predicted outcome is ANSWER, UNANSWERABLE, or ABSTAIN. The absence of output is not recoded as $\perp$. The first non-CONTINUE action terminates a trajectory. Its outcome is correct answer, incorrect answer, correct unanswerability, false unanswerability, or abstention; answer actions are additionally marked premature when $\hat{j} < j^\dagger$. Always-wait and always-commit policies define the two degenerate endpoints. For an optional application-weighted risk, define $D = (\hat{\tau} - \tau^*)_+$ only when both timestamps exist, and $D = 0$ otherwise. Define $Y = \ell(\hat{a}, a^*)$ on emitted answers to answerable records, and $Y = 0$ otherwise. Then, for nonnegative costs,

$$\mathcal{R} = c_e \Pr(\bar{j} < j^\dagger) + c_l \mathbb{E}[D] + c_m \Pr(\hat{j} = \varnothing,\ j^* \neq \varnothing) + c_y \mathbb{E}[Y]. \tag{12}$$

This optional risk is defined even for always-wait; the reported tables retain its components separately rather than selecting costs after evaluation.

For a committed, closable trajectory, normalized conditional delay is $(\hat{\tau} - \tau^*)_+/(H_X - t_q)$, where $t_q$ is question time and $H_X > t_q$. Per-trajectory coverage indicates an ANSWER or UNANSWERABLE action by $H_X$, and prematurity is $m_i^{\text{prem}} = \mathbf{1}[\hat{j}_i \neq \varnothing, \hat{j}_i < j_i^\dagger]$. Missed commitment indicates no ANSWER on a trajectory with $j_i^* \neq \varnothing$; false UNANSWERABLE decisions on such trajectories also count as missed. Reported action metrics use a nested macro average, not an unweighted trajectory mean. For a metric $m$, let $I_{g,a}^m$ contain eligible branches in group $g$ with exact terminal answer class $a = A(q, X)$, including $\perp$. Let $\mathcal{A}_g^m$ and $\mathcal{G}^m$ retain only nonempty classes and groups. Then

$$\overline{m} = \frac{1}{|\mathcal{G}^m|} \sum_{g \in \mathcal{G}^m} \frac{1}{|\mathcal{A}_g^m|} \sum_{a \in \mathcal{A}_g^m} \frac{1}{|I_{g,a}^m|} \sum_{i \in I_{g,a}^m} m_i. \tag{13}$$

Coverage and prematurity include all branches; answer accuracy includes answerable branches, missed commitment includes closable branches, and delay includes committed, closable branches of positive duration. Missing conditional values are omitted at each level rather than assigned zero. Answer coverage and $U$ coverage apply the same unconditional weights to the respective action indicators and sum to total coverage. The development policy uses this total coverage. For example, the seed-17 four-state run has unweighted prematurity 0.0006944 but reported macro prematurity 0.0004630; its answer coverage is 0.7028 unweighted and 0.6412 under Equation 13. For pre-divergence pairs $\mathcal{P}$, distributional prefix consistency is

$$1 - \frac{1}{|\mathcal{P}| \log 2} \sum_{(i,j,t) \in \mathcal{P}} \text{JS}(p_t^i, p_t^j). \tag{14}$$

**Ranking and calibration estimators.** For each operator, non-interpolated AP is $\sum_k (r_k - r_{k-1}) p_k$, where $r_k$ and $p_k$ are recall and precision after including every prefix at the $k$th distinct score in descending order, with $r_0 = 0$. All tied scores enter together; operators with no positive reference prefix have undefined AP and are omitted from its macro average. Binary Brier score is the mean of $(s_i - y_i)^2$ for closure score $s_i$ and target $y_i = \mathbf{1}[z_i^* = C]$. ECE uses up to 15 approximately equal-frequency bins, extending each boundary to keep score ties together.

**Connections to other decision metrics.** Closure-specific scores distinguish temporal finality from the accuracy and earliness summarized in neighboring evaluation protocols. For trajectory $i$, let $e_i = (\widetilde{\tau}_i - t_{q,i})/(H_{X,i} - t_{q,i})$, where $\widetilde{\tau}_i$ is the physical action time, or $H_{X,i}$ when no action is emitted. The early-classification summary is $E = N^{-1} \sum_i e_i$ and $\text{HM} = 2(1 - E)\text{Acc}/[(1 - E) + \text{Acc}]$ (Schäfer & Leser, 2020). At empirical coverage $\kappa$, selective risk is the mean answer loss among the top-$\kappa$ trajectories ranked by the selection score (Geifman & El-Yaniv, 2019).

**Joint stopping diagnostic.** For the gap-aware closure head minus same-head binary readiness, success requires the upper one-sided 95% component-bootstrap bound for premature commitment to be below zero. It also requires, as an intersection–union test, lower bounds above $-0.01$ for answer accuracy and coverage and upper bounds below $0.01$ and $0.05$ for missed commitment and normalized delay. A fixed threshold grid minimizes absolute development-coverage error from 90%; ties choose the higher commitment threshold and then longer dwell. These settings are frozen, and achieved test coverage is reported without retuning. The comparison is exploratory because it follows development of the task and estimator. All conditions must hold jointly; a reduction in premature commitment cannot compensate for a failed delay safeguard.

The threshold grid is $\alpha_C = \alpha_Q \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ with $\delta_C \in \{0, 0.5, 1\}$ seconds and fixed $\alpha_U = 0.5$. The readiness and direct binary-closure policies use the same grid on their scalar scores and abstain at the horizon if they have not answered; neither has a $U$ head. Temperatures are chosen from $\{0.5, 0.75, 1, 1.5, 2, 3\}$, minimizing development state cross-entropy or closure-target binary cross-entropy for $Q_t$ separately; readiness uses its own support labels. Equal losses favor temperature closest to one. For all three primary symbolic seeds, the four-state head selects state and cumulative temperatures $0.75$, threshold $0.9$, and zero dwell. Readiness selects temperature $0.5$, threshold $0.9$, and zero dwell. Coverage selection does not optimize premature error or guarantee the nominal 90% on test.

Training uses AdamW for 20 epochs, learning rate 0.003, weight decay 0.01, and gradient-norm clipping at 1. Hidden widths are 64 for symbolic inputs and 32 for video features; batches contain

32 and 8 intervention groups, respectively. The implementation resets $h_0$ at each trajectory and applies masks before all sequence reductions. It treats $\Delta t_t \geq 0$ as physical seconds, not frame indices. The state, hazard, terminal-event, and revision heads have separate projections of $h_t$; the operator head projects the encoded question, not $h_t$. Sequence losses ignore padded times; the revision loss also ignores absent candidates and video records without reachable-class targets. The terminal-event head is evaluated only on an observed zero-duration boundary. All input dimensions, hidden width, dropout, loss weights, pair margin, and inference thresholds are serialized with the checkpoint. The full objective uses $(\lambda_h, \lambda_p, \lambda_o, \lambda_v) = (1, 1, 0.25, 0.5)$, margin $0.2$, and dropout $0.1$. The video representation and question embedding each have 2,048 dimensions; the symbolic question representation has 66 dimensions. Both domains use context width 16, operator width 8, horizon width 4, and gap width 8. With $n_s$ training prefixes in state $s$ and $N_{\mathrm{tr}} = \sum_s n_s$, state weights are $w_s = N_{\mathrm{tr}}/(Kn_s)$ for nonempty classes and zero otherwise, where $K = 4$ (or 3 after merging $P/R$). Weighted state CE is summed over valid prefixes and divided by their count, not by the sum of their weights. Revision BCE uses the same reduction over its valid targets; operator CE averages records. Hazard NLL sums the at-risk events through and including first closure, or through the horizon for censored records, then averages trajectories. Probabilities in BCE are clipped to $[10^{-6}, 1 - 10^{-6}]$. Each pre-divergence pair contributes JS divergence plus $|Q_t^i - Q_t^j|$; each post-divergence pair contributes the margin hinge. Each set is averaged separately, then the available set means are averaged. An empty pair loss is zero. These reductions are batch-local. Development selection averages the combined batch losses with weights equal to their record counts. Groups are sorted by identifier and packed without shuffling; symbolic development batches contain 32 groups, except the final 16-group batch. This selection criterion is not a globally prefix-normalized objective and depends on the stated batch partition. Bootstrap comparisons first average jointly eligible group-level values within each source component and seed, then resample source components with all paired seeds retained. Eligibility is matched at the group level; delay intervals do not condition on an identical set of individual committed branches.

### A.5 Additional measured comparisons

Tables 6–13 report seed variation, objective ablations, composition transfer, nuisance cues, representation probes, calibration, and model size. All controller and ablation rows use the same split within each domain. The survival baseline uses 20 epochs and the same optimizer settings, with 32 symbolic groups or 4 video groups per batch; its development-calibrated cumulative event probability does not predict $P$, $R$, or $U$ separately. Its trajectory-mean event NLL is 0.1598 on symbolic data and 1.5181 on the video pilot. The four-state objective is therefore assessed for its richer decision target, not presumed to dominate event-time estimation.

**Controller specifications.** For width $d$, controller cost is $O(d^2 + d \dim z)$ per step and constant in history length; this excludes full-prefix backbone encoding. The video head has 106,651 trainable parameters (Table 13). The input width before temporal processing is $\dim z + 16 + 8 + 4$. The MLP uses Linear–SiLU–Linear–Tanh with the domain's hidden width and no recurrent state. GRU projects this input to hidden width before one GRUCell; GRU-$\Delta t$ additionally concatenates $\log(1 + \Delta t_t)$. Gap-aware GRU uses the two-layer width-8 SiLU gap encoder and decay in Equation 9. CfC uses one hidden-width Linear–SiLU layer on concatenated input and prior state, two tanh branches, and interpolation $\sigma(a_t \Delta t_t + b_t)$ between them; these $a_t, b_t$ are internal gates, not the answer candidate or terminal head. LIF has learned per-unit time constants $\mathrm{softplus}(\theta) + 10^{-4}$ initialized at $\theta = 0$, threshold 1, subtractive detached spike reset, and a triangular surrogate derivative of width 1. The membrane is recurrent; the output heads read spikes. The primary LIF is a continuous-current cell: $u_t^- = \exp(-\Delta t_t/\tau) u_{t-1} + [1 - \exp(-\Delta t_t/\tau)] I_t$, with affine input $I_t$. It cannot incorporate a new input at $\Delta t_t = 0$, including every STREAM_END in these datasets. An event-input comparison instead uses $u_t^- = \exp(-\Delta t_t/\tau) u_{t-1} + I_t$, retaining the same spike and reset rules (Appendix A.8). Mamba uses one Mamba-1 block with state dimension 16, causal convolution width 4, expansion 2, and automatic time-step rank; its internal selective step is not physical $\Delta t_t$. A differentiable reference scan replaces fused kernels. The pinned implementation is v2.2.6.post3, revision `10b5d6358f27`. All variants apply dropout 0.1 to the temporal representations passed to the heads, with zero initial recurrent state. The controller heads are trained in float32.

Table 6: Closure AP, mean ± sample standard deviation over seeds 17, 29, and 43. Seed standard deviations describe optimization variation on a fixed split, not uncertainty over new source videos.

| | Symbolic | | Video pilot | |
|---|---|---|---|---|
| Controller | Closure | Readiness | Closure | Readiness |
| MLP | $0.9835 \pm 0.00050$ | $0.5542 \pm 0.00118$ | $0.9232 \pm 0.00084$ | $0.3638 \pm 0.02743$ |
| GRU | $0.9955 \pm 0.00012$ | $0.4798 \pm 0.00341$ | $0.9212 \pm 0.00960$ | $0.4264 \pm 0.05448$ |
| GRU-$\Delta t$ | $0.9953 \pm 0.00008$ | $0.4786 \pm 0.00359$ | $0.9209 \pm 0.00312$ | $0.4236 \pm 0.05802$ |
| Gap-aware GRU | $0.9954 \pm 0.00001$ | $0.4771 \pm 0.00478$ | $0.9180 \pm 0.00865$ | $0.4381 \pm 0.12044$ |
| CfC | $0.9952 \pm 0.00006$ | $0.4957 \pm 0.03753$ | $0.9204 \pm 0.00365$ | $0.3540 \pm 0.03372$ |
| Mamba | $0.9957 \pm 0.00039$ | $0.4654 \pm 0.00387$ | $0.9375 \pm 0.03328$ | $0.2596 \pm 0.00926$ |
| LIF | $0.9812 \pm 0.00179$ | $0.4927 \pm 0.02831$ | $0.9222 \pm 0.00087$ | $0.2974 \pm 0.07411$ |
| Survival | $0.9953 \pm 0.00013$ | n/a | $0.9258 \pm 0.00524$ | n/a |

**Video prompt and representation.** The frozen backbone runs in bfloat16 under PyTorch 2.5.1+cu124 and Transformers 4.49.0, using SDPA. The processor permits $4 \times 28^2$ to $256 \times 28^2$ pixels per image. A prefix is rejected if it exceeds 128 frames or 16,384 tokens; it is not silently truncated. Vision encoding batches two independent still-image grids at a time, then retains all image tokens in the causal language prefix. There is no KV reuse, teacher-forced answer, generation, or dynamic token dropping. The user-message content follows this template; braces denote record fields and timestamps have six decimal places:

```
Question at {question_time}s:  {question}
The declared horizon is the observed end of the stream.
Observed frame at {source_time}s; available at
{availability}s.
```
*[image; repeat the frame line and image for each visible frame]*
```
Current time:  {time}s.
```
<STREAM_END> *[only after the terminal event]*

The processor's chat template appends the generation prompt. We pool the final-normalized state at the last valid prompt token before the language-model head. The separate question vector uses the same pooling on a user message containing only the question, without frames, timestamps, or gold operator IDs. The first cache event is a question-only $P$ prefix. Source-time annotations become eligible only when their frames are available; an explicit end token is consumed after the last frame. This prevents a frame timestamp from revealing evidence before its arrival. If the last frame arrives before the recorded end, an additional clock-only prefix advances time to that endpoint before the zero-gap end token. It retains the same frames and is an offline integration convention, not an independently scheduled pre-end action opportunity. The valid-prefix mask includes it in training, calibration, and the primary prefix metrics. Training, development, and test contain 10, 1, and 6 such nodes. The test has 134 prefixes: 114 frame, 7 question-only, 6 clock-only, and 7 end-token prefixes. Appendix A.8 reports scoring-only exclusion of the six clock nodes without changing recurrent predictions or calibration.

Removing the pair term changes symbolic AP from 0.9954 to 0.9954 at four-decimal precision. Replacing predicted operator embeddings with gold or absent embeddings likewise yields similar symbolic AP. The question representation still exposes the exact symbolic operator tree in all three cases. On the pilot, removal of the revision loss is exactly unchanged because all revision targets are masked. These ablations constrain component-level claims rather than supporting an additive improvement from every loss.

The observed-end control's pilot AP of 0.8282 is far above the 0.2418 positive prevalence. Question-only AP is 0.2373, whereas question plus clock and observed end reaches 0.8969. The combination shows that task identity and boundary timing can explain much of the pilot's closure ranking without visual features. Layer probes are also variable: block 18 reaches AP 0.9708, while the final-normalized representation gives 0.8169. All layers are reported as exploratory probes; the best test layer is not substituted into the primary comparison.

Figure 6 decomposes four-state prediction errors for the primary controller. Figure 7 displays all measured layer and nuisance probes on a shared AP scale.

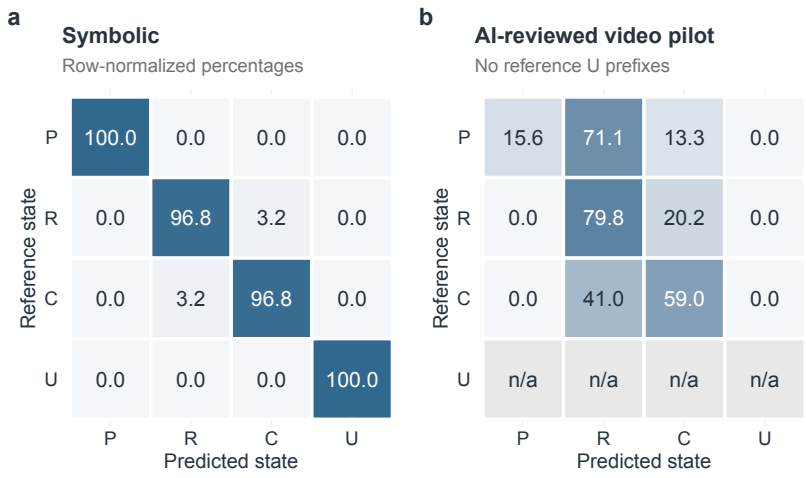


Figure 6: Four-state errors of the gap-aware GRU. Entries are row-normalized percentages, computed within seed and averaged over three seeds. **a**, Symbolic reference counts: 2,177 $P$, 3,892 $R$, 3,775 $C$, and 428 $U$ prefixes. **b**, Video reference counts: 15 $P$, 84 $R$, 35 $C$, and no $U$ prefixes. Gray cells denote an absent reference class, not zero error. These matrices pool prefixes within each class; Table 3's operator-macro F1 uses different weighting.

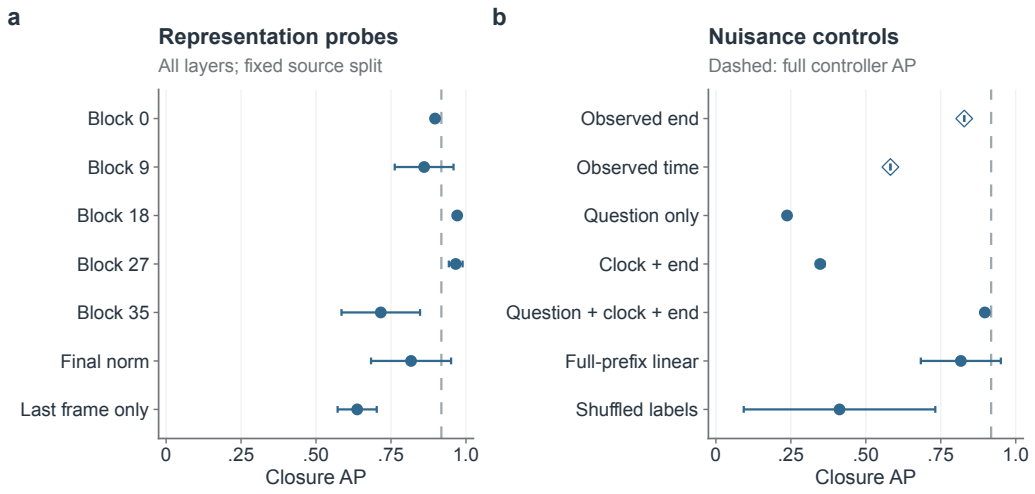


Figure 7: Representation depth and nuisance cues on the video pilot. **a**, All tested layer probes, including separately encoded last-frame-only input. **b**, Boundary, question, and shuffled-label controls. Circles and bars show three-seed means and standard deviations; hollow diamonds denote deterministic controls. The dashed line marks the full gap-aware controller's mean AP. Each probe uses the same seven test records from two source components. The best test layer is not substituted into the primary comparison.

Table 7: Gap-aware GRU ablations, three-seed means on the same splits. AP and F1 follow Table 3. The video pilot has no eligible revision targets, so removing the revision term leaves its objective unchanged. With no state loss, F1 evaluates a state projection without direct four-class cross-entropy supervision.

| | Symbolic | | Video pilot | |
|---|---|---|---|---|
| Variant | AP | F1 | AP | F1 |
| Full objective | 0.9954 | 0.9851 | 0.9180 | 0.2900 |
| No state loss | 0.9915 | 0.1312 | 0.9173 | 0.0839 |
| No hazard loss | 0.9948 | 0.9854 | 0.6368 | 0.4245 |
| No pair loss | 0.9954 | 0.9849 | 0.9216 | 0.2883 |
| No operator loss | 0.9952 | 0.9843 | 0.9251 | 0.3277 |
| No revision loss | 0.9955 | 0.9850 | 0.9180 | 0.2900 |
| No operator embedding | 0.9954 | 0.9854 | 0.9238 | 0.2394 |
| Gold operator embedding | 0.9955 | 0.9854 | 0.9226 | 0.2526 |

Table 8: Unseen symbolic compositions: 240 records from 20 held-out source-event components, using binary AND, OR, and PAIR trees absent from training. Composition AP is mean ± seed SD; F1 is a seed mean. ID and composition use their own operator sets and prevalences, so their difference is descriptive rather than a matched-distribution effect.

| Controller | ID AP | Composition AP | Composition F1 |
|---|---|---|---|
| MLP | 0.9835 | 0.9203 ± 0.01025 | 0.8022 |
| GRU | 0.9955 | 0.7855 ± 0.01962 | 0.8286 |
| GRU-$\Delta t$ | 0.9953 | 0.8069 ± 0.04627 | 0.8286 |
| Gap-aware GRU | 0.9954 | 0.8988 ± 0.00138 | 0.8315 |
| CfC | 0.9952 | 0.7462 ± 0.00530 | 0.8308 |
| Mamba | 0.9957 | 0.9350 ± 0.02390 | 0.8286 |
| LIF | 0.9812 | 0.8856 ± 0.02017 | 0.6869 |

Table 9: Nuisance and shuffled-label controls: operator-macro closure AP. Learned probes use three seeds (mean ± SD); end and observed-time scores are deterministic. These controls retain the grouped split and do not use unobserved video duration.

| Input / control | Symbolic | Video pilot |
|---|---|---|
| Observed-end indicator | 0.4704 | 0.8282 |
| Observed time | 0.8167 | 0.5820 |
| Question only | 0.3622 ± 0.00043 | 0.2373 ± 0.00000 |
| Clock + observed end | 0.7411 ± 0.02064 | 0.3479 ± 0.01406 |
| Question + clock + end | 0.8172 ± 0.00159 | 0.8969 ± 0.00537 |
| Final-prefix linear | 0.8849 ± 0.00252 | 0.8169 ± 0.13347 |
| Shuffled training labels | 0.5108 ± 0.19990 | 0.4123 ± 0.31940 |

Table 10: Exploratory linear probes on the human-reviewed video pilot. Decoder block indices are zero-based. Every probe uses the same source split; last-frame-only inputs are separately encoded. Test-layer differences are descriptive and do not select the representation used in Table 3.

| Representation | Closure AP | State F1 |
|---|---|---|
| decoder block 0 | 0.8969 ± 0.00504 | 0.4876 ± 0.02651 |
| decoder block 9 | 0.8605 ± 0.09800 | 0.5051 ± 0.02501 |
| decoder block 18 | 0.9708 ± 0.00104 | 0.4977 ± 0.06005 |
| decoder block 27 | 0.9658 ± 0.02296 | 0.4358 ± 0.05322 |
| decoder block 35 | 0.7159 ± 0.13090 | 0.3236 ± 0.06599 |
| final norm | 0.8169 ± 0.13347 | 0.3458 ± 0.06204 |
| last frame final norm | 0.6372 ± 0.06525 | 0.3525 ± 0.01396 |

Table 11: Exploratory symbolic stopping safeguards for closure minus readiness. One-sided 95% intervals use 2,000 bootstrap resamples of 60 source components, retaining all three paired seeds. The joint criterion is not met because the delay upper bound exceeds 0.05. This post-development diagnostic is not a preregistered confirmatory test.

| Metric | $\Delta$ | One-sided interval | Safeguard |
|---|---|---|---|
| Premature | -0.9196 | $(-\infty, -0.9174]$ | Met |
| Answer accuracy | +0.4509 | $[0.4319, +\infty)$ | Met |
| Coverage | +0.0796 | $[0.0778, +\infty)$ | Met |
| Missed | +0.0000 | $(-\infty, 0.0000]$ | Met |
| Delay | +0.0683 | $(-\infty, 0.0737]$ | Not met |

Table 12: Operator-macro discrimination and calibration, three-seed means. All metrics use closure labels. Readiness is calibrated on support labels, so its Brier/ECE also reflect target mismatch. Closure scores multiply separately temperature-scaled committed-state and cumulative-event probabilities; the product need not be calibrated. Survival uses cumulative first-closure probability. ECE uses 15 equal-frequency bins with tied scores kept together.

| Domain | Objective | AUROC | Brier | ECE |
|---|---|---|---|---|
| Symbolic | Closure | 0.9957 | 0.0197 | 0.0184 |
| Symbolic | Readiness | 0.6694 | 0.4081 | 0.4081 |
| Symbolic | Survival | 0.9956 | 0.0152 | 0.0056 |
| Video pilot | Closure | 0.9093 | 0.2138 | 0.2261 |
| Video pilot | Readiness | 0.6237 | 0.7143 | 0.7293 |
| Video pilot | Survival | 0.9175 | 0.1848 | 0.1902 |

Table 13: Trainable head parameters in the measured closure configurations. The frozen 3B video backbone is excluded. Width is 64 for symbolic inputs and 32 for video representations. Counts include the question projection and operator head, so these comparisons hold width rather than parameter count fixed.

| Controller | Symbolic | Video pilot |
|---|---|---|
| MLP | 8,348 | 100,739 |
| GRU | 29,148 | 106,019 |
| GRU-$\Delta t$ | 29,212 | 106,051 |
| Gap-aware GRU | 30,324 | 106,651 |
| CfC | 24,924 | 104,931 |
| Mamba | 36,828 | 109,603 |
| LIF | 4,252 | 99,715 |

### A.6 SAME-TARGET STOPPING AND JOINT-HEAD DIAGNOSTICS

The direct binary-closure control replaces support labels with $\mathbf{1}[S_t = C]$ in the scalar BCE head, retaining the gap-aware GRU, question-only operator head, inputs, seeds, widths, training batches, optimizer, and 20-epoch budget. Its scalar head is independently temperature-scaled using development closure labels. It has no four-state or unanswerability head. The direct binary controls use server CPU, while the primary controller runs use CUDA; Table 15 gives both domains. The pilot's mean AP is 0.3947 for this control, but its seven-record test cannot establish a stable architecture or loss ranking.

For survival, saved cumulative probabilities are temperature-scaled with the frozen development temperature and passed to the same scalar threshold/dwell policy as binary closure. All policies select thresholds on development coverage before test aggregation. Mapping four-state $U$ outputs to abstention and reselecting on development leaves its answer coverage, prematurity, and delay unchanged here. The observed-domain-end policy answers the shared candidate only after DOMAIN_END; it otherwise abstains. The exact rule monitor applies hand-coded certificates to the same observed symbols and declared program available to the symbolic learners, returning $U$ at the horizon if no terminal answer is supported. Its advantage is exact semantic knowledge, not additional raw information or future events. Such error-free symbols and certificates are unavailable to a general video system.

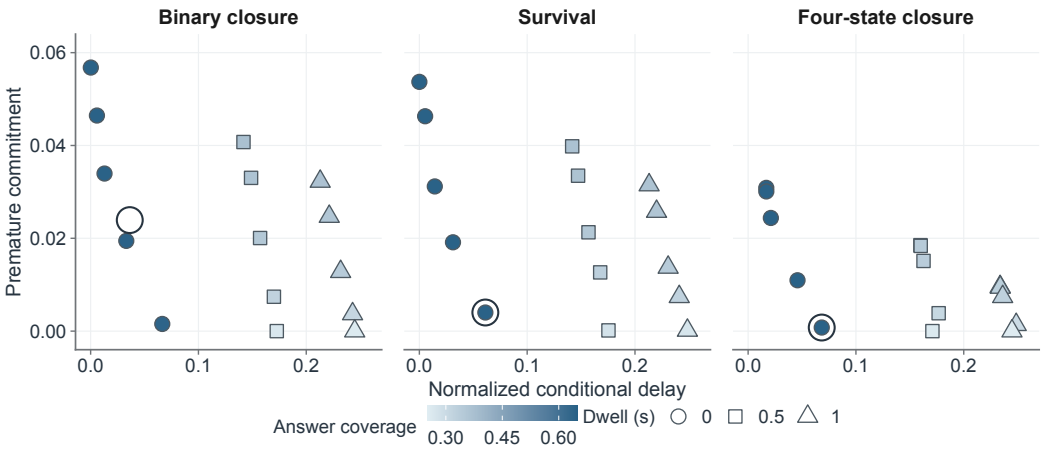


Figure 8: Symbolic stopping sensitivity on the fixed threshold/dwell grid. Each filled point averages three seeds at one of five thresholds and three dwell settings; darker fill denotes greater answer coverage. Open rings average each seed's development-selected policy, whose threshold may differ across seeds. All test settings are shown, without selecting a test-optimal frontier. Delay conditions on commitment, so lower delay at reduced coverage is not an unconditional improvement.

Table 17 separates local state decisions from the cumulative gate. Their disagreement is measured on the same symbolic test prefixes; the product score is not constrained to equal either posterior.

Table 14: Symbolic decisions with a shared causal rule-based candidate interface, not generated video answers. Values balance terminal classes within each intervention group, then average groups and seeds. Accuracy is correctness over answerable records; missed answers count as errors. Delay is horizon-normalized and conditional on commitment to a closable record; early answers contribute zero delay. Coverage counts answer or unanswerable decisions.

| Objective | Accuracy | Coverage | Premature | Missed | Delay |
|---|---|---|---|---|---|
| Binary readiness | 0.5491 | 0.9204 | 0.9204 | 0.0000 | 0.0000 |
| Closure | 1.0000 | 1.0000 | 0.0008 | 0.0000 | 0.0683 |

Table 15: Direct binary-closure control: BCE on the same C target, plus operator CE, without state, hazard, revision, or pair losses. Gap-aware GRU widths, inputs, splits, 20 epochs and optimizer match the primary runs; these six additional runs use server CPU while the primary models used CUDA. AP SD is seed variation, not source uncertainty. No four-state F1 or U prediction is defined.

| Domain | AP | AP SD | AUROC | Brier | ECE |
|---|---|---|---|---|---|
| Symbolic | 0.9858 | 0.0029 | 0.9949 | 0.0172 | 0.0087 |
| Video pilot | 0.3947 | 0.0566 | 0.6726 | 0.2126 | 0.2442 |

Table 16: Exploratory paired stopping differences, four-state closure minus each same-target reference. Two-sided 95% percentile intervals resample the 60 source-event components 2,000 times while retaining paired seeds. These comparisons are not preregistered or multiplicity-adjusted. Conditional metrics first retain jointly eligible groups within each component and seed; individual committed branches need not coincide.

| Reference | Metric | Difference | Interval |
|---|---|---|---|
| Binary closure | Premature | -0.0231 | $[-0.0256, -0.0208]$ |
| Binary closure | Delay | +0.0320 | $[+0.0287, +0.0352]$ |
| Survival | Premature | -0.0032 | $[-0.0060, -0.0011]$ |
| Survival | Delay | +0.0071 | $[+0.0054, +0.0087]$ |

Table 17: Calibrated symbolic head conjunctions for the full gap-aware GRU. Counts pool 10,272 test prefixes over three seeds (30,816 seed-prefix observations, not independent samples). The cumulative gate is $\tilde{Q}_t \geq 0.9$, and $\tilde{\chi}_t = \tilde{p}_t(C)\tilde{Q}_t$. Means pool observations, unlike operator-macro calibration. An empty conjunction has no defined mean.

| State argmax $C$ | Cumulative gate | Count | Mean $\tilde{\chi}$ | Closure frequency |
|---|---|---|---|---|
| no | no | 19476 | 0.0038 | 0.0184 |
| no | yes | 0 | n/a | n/a |
| yes | no | 1313 | 0.5205 | 0.7197 |
| yes | yes | 10027 | 0.9975 | 0.9995 |

### A.7 Trained three-state control and structural stress tests

The three-state control merges $P$ and $R$ into $O$, while retaining the same recurrent backbone, hazard, terminal-event, revision, operator, and pair objectives. Its state head has three trainable logits; a constant $-10^4$ fills the unused $R$ slot solely for compatibility with the shared $C/U$ policy interface. The initial $O$ row is the mean of the four-state head's $P/R$ rows; $C/U$ and all shared weights are copied from the same seeded initialization without advancing the random-number generator. The merged state weights use $K = 3$, increasing the $C/U$ state-loss weights by $4/3$ relative to the four-class objective. This comparison changes both target granularity and its induced weighting; it does not isolate the $P/R$ distinction independently of weighting. Temperature selection uses merged development labels, while closure labels, candidate rules, calibration grids and the stopping policy are unchanged. Both targets use paired CPU training with one thread per run and deterministic algorithms, separately from the primary CUDA experiment. Table 18 reports a common three-class F1: four-state probabilities are summed for $O$ before the argmax. The absent pilot $U$ class contributes

zero in both models. On the video pilot, three-state AP is 0.9269 versus 0.9220 for its four-state control, while both have merged F1 0.2804. This does not show that finer state supervision improves pilot classification.

Table 18: Trained three-state control merging $P/R$ into open state $O$. Four-state controls use the same CPU, shared initialization, seeds and training settings; state-class weights are recomputed per target. F1 uses the same three classes $O, C, U$ for both methods, not four-state F1. Values average three seeds. Video evaluates states only.

| Domain | Target | AP | $F1_3$ | A-cov. | U-cov. | Prem. | Delay |
|---|---|---|---|---|---|---|---|
| Symbolic | Four-state | 0.9953 | 0.9834 | 0.6412 | 0.3588 | 0.0009 | 0.0668 |
| Symbolic | Three-state | 0.9955 | 0.9826 | 0.6394 | 0.3580 | 0.0017 | 0.0651 |
| Video pilot | Four-state | 0.9220 | 0.2804 | – | – | – | – |
| Video pilot | Three-state | 0.9269 | 0.2804 | – | – | – | – |

**Stress-test generation.** The stress-test design fixes seed 9202402, 32 source bodies per condition, four branches per program, and all evaluation choices before model evaluation. The structure test uses six depth-one and three depth-two ordered trees listed in Table 20, with budget six. None of these trees appears in training, development, or the original three-tree composition test. Each body is N, a shuffled middle containing one A, one B, and two independent uniform draws from A/B/N/DONE, followed by DOMAIN_END. The length test instead uses six base operators and budget eight, with four random middle draws. Sources are unique within a condition. Targets are sampled uniformly from A/B per program, gaps uniformly from $[0.5, 1.5]$ seconds, and the four interventions follow the base generator. The time-shift condition doubles all timestamps of the length-test records without changing events or reference labels. An input-only check rejects an entire source if any derived program/token template occurs in training or development; no source was rejected. No model predictions or outcome labels are used for this overlap filter. The key includes expression, target, declared budget, and token sequence. A new expression or budget therefore suffices for key disjointness; token-only novelty is not claimed. The fixed inclusion of A and B and the small grammar remain design restrictions; these stress tests do not represent unrestricted program synthesis or natural-video transfer.

Table 19: Frozen-checkpoint stress tests with 32 sources per condition. New structure uses nine unseen trees (1,152 branches); longer bodies use budget eight (768 branches), and the final column doubles their timestamps. Full program/token keys are absent from training and development; this does not imply token-only novelty. AP is operator-macro; $\pm$ denotes seed SD, not sampling uncertainty.

| Method | New structure | Longer bodies | $2\times$ time |
|---|---|---|---|
| Four-state | $0.9325 \pm 0.0065$ | $0.9979 \pm 0.0001$ | $0.9979 \pm 0.0001$ |
| Three-state | $0.9148 \pm 0.0353$ | $0.9980 \pm 0.0000$ | $0.9978 \pm 0.0002$ |
| Binary closure | $0.8956 \pm 0.0379$ | $0.9980 \pm 0.0000$ | $0.9979 \pm 0.0000$ |
| Survival | $0.8805 \pm 0.0141$ | $0.9978 \pm 0.0001$ | $0.9978 \pm 0.0001$ |

All four methods reuse their base-development temperatures and thresholds, with no stress-test fitting. AP pools prefixes within root operator and averages operators; Table 20 also reports each ordered tree separately. For unseen trees, four-state and three-state answer coverage are both 0.6027 and delay is 0.1884. Their prematurity rates are 0 and 0.0003; binary closure and survival give 0.0443 and 0.0155 at the same answer coverage, with lower delays 0.1705 and 0.1790. The higher mean AP for four-state closure is concentrated in PAIR(Last, Accumulate) and PAIR(PAIR(Count, Last), Accumulate). Both have the same four-state target as Last under this grammar; unfamiliar syntax does not imply unfamiliar monitoring semantics. For length-eight bodies, four-state prematurity is 0.0014; doubling physical time raises it to 0.0049 even though AP remains about 0.9979. Answer coverage changes from 0.6445 to 0.6377. Stable ranking therefore does not imply an invariant calibrated stopping policy. These fixed-protocol comparisons are exploratory, without confirmatory significance tests.

Table 20: AP for each unseen operator tree, 128 branches from 32 sources per row. Values average three seeds. The table exposes composition-specific failures hidden by the aggregate.

| Tree | Four-state | Three-state | Binary | Survival |
|---|---|---|---|---|
| Or(Complete,Exists) | 0.9932 | 0.9935 | 0.9859 | 0.9934 |
| And(Complete,Forall) | 0.9772 | 0.9771 | 0.9713 | 0.9762 |
| Pair(Count,Exists) | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| Or(Exists,Forall) | 0.9785 | 0.9781 | 0.9691 | 0.9771 |
| And(Forall,Exists) | 0.9800 | 0.9798 | 0.9736 | 0.9798 |
| Pair(Last,Accumulate) | 0.8635 | 0.7381 | 0.6130 | 0.5386 |
| Pair(Pair(Count,Last),Accumulate) | 0.8433 | 0.7265 | 0.5726 | 0.5433 |
| And(Or(Exists,Complete),Forall) | 0.9776 | 0.9783 | 0.9731 | 0.9775 |
| Or(And(Forall,Complete),Exists) | 0.9779 | 0.9779 | 0.9656 | 0.9774 |

**Expression versus monitoring semantics.** We enumerate all 7,279 complete event histories allowed at budget six, evaluating every prefix for targets A and B. Eight of the nine held-out expressions produce the same four-state trace function as a base operator (Table 21). In particular, Count and Accumulate share a state function even though their answer values differ. All PAIR children use the same target, domain, and horizon; combining their values therefore need not introduce a new closure rule. The only tree distinct from every base state function is OR(Complete, Exists). Its four-state and three-state AP are 0.9932 and 0.9935, respectively. Thus the largest four-state gains concern robustness to unfamiliar encodings of known monitoring functions, not evidence of superior semantic extrapolation.

Table 21: Monitoring-target equivalence for the nine expression holdouts. Exhaustive enumeration covers all 7,279 legal complete event histories at budget six for each target A/B, including every prefix. Eight expressions share their entire four-state trace function with a base operator; terminal answer values need not agree. AP values are the original three-seed means on the 128 sampled branches per tree.

| Expression | Equivalent base state target | Four-state AP | Three-state AP |
|---|---|---|---|
| And(Forall,Exists) | forall | 0.9800 | 0.9798 |
| And(Complete,Forall) | forall | 0.9772 | 0.9771 |
| Or(Exists,Forall) | exists | 0.9785 | 0.9781 |
| Or(Complete,Exists) | Distinct from all six | 0.9932 | 0.9935 |
| Pair(Last,Accumulate) | last | 0.8635 | 0.7381 |
| Pair(Count,Exists) | count/accumulate | 1.0000 | 1.0000 |
| And(Or(Exists,Complete),Forall) | forall | 0.9776 | 0.9783 |
| Or(And(Forall,Complete),Exists) | exists | 0.9779 | 0.9779 |
| Pair(Pair(Count,Last),Accumulate) | last | 0.8433 | 0.7265 |

### A.8 Evidence-bearing prefixes and matched controls

**Evidence-bearing shared histories.** The original groups share only initial N events. We separately test observed-object histories with no retraining or recalibration of the original controls. Generation seed 9202403 specifies 32 unique six-event bodies for each shared-prefix length of one, two, and three objects. A uniformly sampled target A/B is the first shared object; remaining shared objects are independent uniform A/B draws. Independent uniform A/B/N/DONE draws fill the body to five events, followed by DOMAIN_END. Gaps are uniform on $[0.5, 1.5]$ seconds. Source bodies are unique across the three length strata. Each source supplies all six base operators and four branches using the original suffix interventions. No derived program/token key occurs in training or development; all generated groups are retained irrespective of states or predictions. There are 96 sources, 576 groups, and 2,304 branches. Before divergence, 433 groups are $R$ and admit different terminal answers; 143 are already $C$ and provide positive consistency controls. The latter are not counterfactual negative pairs.

Table 22 reports unchanged checkpoints and development-selected policies from the original same-target comparisons. Four-/three-state checkpoints are the paired CPU models of Table 18; binary

and survival checkpoints are those of Table 4. The group-mass-matched models below are separate retrainings. Four-state AP is 0.9947, compared with 0.9954 for three-state, 0.9645 for binary closure, and 0.9957 for survival. The four-state model has lower prematurity than survival (0.0005 versus 0.0027), but higher delay (0.1615 versus 0.1445) and different answer coverage. Table 23 separates provisional and already-closed shared histories. In the already-closed stratum, AP 0.9980 coexists with missed-answer rate 0.2500, False-U rate 0.1556, and conditional delay 0.5853: ranking does not ensure valid or timely terminal decisions. These tests cover nontrivial observed candidates; they do not establish a training benefit from the pair loss.

Table 22: Evidence-bearing shared prefixes: 2,304 branches from 96 sources, with 32 sources per shared length. Prefixes start with a target object; they contain one, two, or three object events before divergence. All six operators are balanced. Models and calibration are frozen from the original same-target comparisons. AP is operator-macro; the overall column reports mean $\pm$ seed SD, other columns seed means. Answer coverage is unconditional; delay conditions on commitment to a closable record.

| Method | Overall AP | $AP_1$ | $AP_2$ | $AP_3$ | A-cov. | Prem. | Delay |
|---|---|---|---|---|---|---|---|
| Four-state | $0.9947 \pm 0.0011$ | 0.9947 | 0.9946 | 0.9957 | 0.6433 | 0.0005 | 0.1615 |
| Three-state | $0.9954 \pm 0.0001$ | 0.9967 | 0.9946 | 0.9957 | 0.6368 | 0.0014 | 0.1564 |
| Binary closure | $0.9645 \pm 0.0043$ | 0.9562 | 0.9639 | 0.9787 | 0.6707 | 0.0247 | 0.0696 |
| Survival | $0.9957 \pm 0.0001$ | 0.9965 | 0.9946 | 0.9965 | 0.6394 | 0.0027 | 0.1445 |

Table 23: Stratification by the exact state immediately before divergence in the evidence-bearing test. There are 433 provisional groups and 143 already-closed groups; all groups are retained in the overall results. Each stratum averages its present operators. Pre-divergence closed groups are positive consistency controls, not answer-changing negative pairs. Values are three-seed means.

| Method | Shared state | AP | A-cov. | Prem. | Miss | False-U | Delay |
|---|---|---|---|---|---|---|---|
| Four-state | R | 0.9937 | 0.6080 | 0.0006 | 0.0230 | 0.0139 | 0.0215 |
| Three-state | R | 0.9947 | 0.5994 | 0.0019 | 0.0371 | 0.0213 | 0.0198 |
| Binary closure | R | 0.9570 | 0.5825 | 0.0267 | 0.0780 | 0.0000 | 0.0093 |
| Survival | R | 0.9946 | 0.6029 | 0.0036 | 0.0314 | 0.0000 | 0.0183 |
| Four-state | C | 0.9980 | 0.7500 | 0.0000 | 0.2500 | 0.1556 | 0.5853 |
| Three-state | C | 0.9979 | 0.7500 | 0.0000 | 0.2500 | 0.1737 | 0.5689 |
| Binary closure | C | 0.9981 | 0.9376 | 0.0186 | 0.0624 | 0.0000 | 0.2520 |
| Survival | C | 0.9984 | 0.7500 | 0.0000 | 0.2500 | 0.0000 | 0.5265 |

**Matched state-group mass and decision interface.** We train three paired four-/three-state symbolic models under PyTorch 2.9.0 on CPU, using seeds 17, 29, and 43 and one thread per run. All shared initialization, data, batch order, optimizer, and 20-epoch settings follow the original paired comparison. For the merged-open target, $w_O = N_{\mathrm{tr}}/[2(n_P + n_R)]$, while $w_C = N_{\mathrm{tr}}/(4n_C)$ and $w_U = N_{\mathrm{tr}}/(4n_U)$. This preserves total training-loss mass for open, closed, and unanswerable groups. It does not preserve each P/R prefix weight or the internal state-distribution representation used by the pair term. Both models use the calibrated scalar score $\tilde{p}_t(C)\tilde{Q}_t$ with the same threshold/dwell grid and development-only selection. Both abstain when that gate fails at the horizon; neither emits a U action in this comparison. The scoring target, action repertoire, and decision interface are therefore common.

Four-state versus three-state AP is 0.9953 versus 0.9924 in-domain, but 0.9951 versus 0.9953 on evidence-bearing prefixes. On the latter, four-state prematurity is 0 versus 0.0077, with delays 0.1635 versus 0.1207 and answer coverage 0.6440 versus 0.6478. Unseen-tree AP is 0.9341 versus 0.8878, subject to the semantic equivalences above. The split can affect learned rankings and stopping, but these comparisons do not show uniformly better prediction or an unconditional stopping advantage.

Table 24: Group-mass-matched state supervision with a common scalar stopping interface. Paired local-CPU models use three seeds; values are seed means. Three-state weights preserve the total training-loss mass of open versus C/U groups, not individual P/R prefix weights. Both policies use the same calibrated product score, development-only threshold/dwell selection, and abstention instead of U actions. F1 uses merged O/C/U classes. These are separate retrainings from Table 18.

| Test | Target | AP | $F1_3$ | A-cov. | Prem. | Delay |
|---|---|---|---|---|---|---|
| ID | Four-state | 0.9953 | 0.9833 | 0.6412 | 0.0005 | 0.0715 |
| ID | Three-state | 0.9924 | 0.9781 | 0.6406 | 0.0116 | 0.0473 |
| Object prefix | Four-state | 0.9951 | 0.8559 | 0.6440 | 0.0000 | 0.1635 |
| Object prefix | Three-state | 0.9953 | 0.8406 | 0.6478 | 0.0077 | 0.1207 |
| Unseen tree | Four-state | 0.9341 | 0.8402 | 0.6027 | 0.0000 | 0.1884 |
| Unseen tree | Three-state | 0.8878 | 0.8401 | 0.6027 | 0.0197 | 0.1808 |

**Event-aware LIF input.** The continuous-current LIF cannot use an arriving representation at zero elapsed time. We compare it with per-event input jumps in 12 paired local-CPU runs across both domains and three seeds. Initialization shapes, parameter counts, four-state objective, and all other settings are unchanged. Unlike the primary CUDA matrix, these runs isolate input injection within the same CPU environment. Symbolic state F1 increases from 0.9139 to 0.9487, while pilot state F1 decreases from 0.1302 to 0.0699. The change removes the zero-gap input restriction but does not ensure accurate pilot classification (Table 25).

Table 25: Paired LIF input-injection control. Twelve local-CPU runs use the same initialization shapes, seeds, data and 20-epoch budget. Both read spikes and use identical decay, threshold, surrogate and reset; only input injection differs. Event jumps remain nonzero at zero elapsed time. Values are mean $\pm$ seed SD; F1 uses four classes, including absent U in the pilot. These paired runs are separate from the primary CUDA matrix.

| Domain | Input injection | Closure AP | State F1 |
|---|---|---|---|
| Symbolic | Continuous current | $0.9824 \pm 0.0019$ | $0.9139 \pm 0.0064$ |
| Symbolic | Per-event jump | $0.9844 \pm 0.0011$ | $0.9487 \pm 0.0019$ |
| Video pilot | Continuous current | $0.9227 \pm 0.0000$ | $0.1302 \pm 0.1564$ |
| Video pilot | Per-event jump | $0.9227 \pm 0.0000$ | $0.0699 \pm 0.0243$ |

**Clock-only scoring sensitivity.** We exclude the six test clock-only nodes from metric aggregation while retaining their original recurrent updates and all frozen predictions. Closure AP changes from 0.9180 to 0.9190 and state F1 from 0.2900 to 0.2891. Readiness AP changes from 0.4381 to 0.4439. This checks metric eligibility, not model behaviour without clock updates (Table 26).

Table 26: Video prefix-scoring sensitivity, three-seed means. Removing six clock-only nodes changes only metric eligibility, not their recurrent updates, frozen predictions, checkpoint or temperatures. All seven question-only and seven terminal nodes remain. Readiness has no four-state classifier.

| Target | Scored prefixes | Closure AP | State F1 |
|---|---|---|---|
| Four-state | All 134 | 0.9180 | 0.2900 |
| Four-state | No clock (128) | 0.9190 | 0.2891 |
| Readiness | All 134 | 0.4381 | – |
| Readiness | No clock (128) | 0.4439 | – |

### A.9 RELEASED MMDUET2 WEB EVALUATION

We evaluate the released `wangyueqian/MMDuet2` checkpoint of Wang et al. (2025b), using its streaming client, greedy generation, and no dynamic token dropping on the 500-question WEB split of ProactiveVideoQA (Wang et al., 2025c). Frames are reconstructed from the released media at 2 FPS and consumed at one-second model steps using the released frame-index sequence. This fixes

Table 27: Published ProactiveVideoQA reference points from Table 2 of Wang et al. (2025b). Each cell is PAUC at $\omega = 0.5$ in percentage points / reply-duplicate percentage; a dash in the second position means that the source did not report the latter. These are literature values, not results of this paper, and are directly comparable only under the unchanged official data and evaluator (Wang et al., 2025c).

| Method | WEB | EGO | TV | VAD |
|---|---|---|---|---|
| VideoLLM-Online | 25.9 / – | 25.0 / – | 18.3 / 53.9 | 25.0 / – |
| MMDuet | 38.9 / 81.3 | 46.0 / 99.4 | 21.1 / 92.8 | 27.4 / 99.2 |
| MMDuet2 SFT | 37.6 / 1.7 | 26.4 / 4.4 | 27.6 / 2.2 | 26.3 / 0.0 |
| MMDuet2 RL | 53.3 / 4.2 | 33.6 / 8.1 | 43.4 / 1.0 | 28.9 / 15.2 |

Table 28: Released MMDuet2 on the 500-question ProactiveVideoQA WEB split, with PAUC at $\omega = 0.5$ and a local Qwen3-32B Q4_K_M content judge. These are baseline measurements, not closure-controller results. Quantized judging and reconstructed frames distinguish this setting from the published reference values in Table 27.

| Model | Questions | PAUC (%) | Duplicate replies (%) |
|---|---|---|---|
| Released MMDuet2 | 500 | 54.15 | 3.84 |

temporal sampling but does not assert byte-identical JPEGs to the original authors' preprocessing. Content scoring retains the released prompts and aggregation with $\omega = 0.5$, using a local Qwen3-32B Q4_K_M judge, temperature 0, seed 17, a 32-token output limit, and `/no_think`. Invalid or truncated judge outputs are rejected rather than converted to scores. The aggregation contains 1,328 eligible answer turns and uses 2,248 valid judge responses. There are no eligible immediate-answer turns in this aggregation, so that conditional mean is undefined. The duplicate rate averages per-answer-turn duplicate fractions over non-immediate turns with replies, excluding each turn's first reply from its fraction; single-reply turns contribute zero.

The dataset revision is `7f59bce2` and the Qwen3 judge-weight revision is `938a7432`; full revisions, runtime hashes, and raw judge responses are retained with the evaluation artifact. Twenty WEB source videos were used in earlier closure-pilot development. The released baseline checkpoint is not updated by that development, but the full WEB split cannot be described as untouched development holdout for a pilot-trained closure controller. The measurements in Table 28 characterize the released baseline under this scoring setup and do not establish a gain for the proposed estimator.

### A.10 SYMBOLIC GENERATOR

Each latent program defines an event alphabet, transition automaton, operator tree, horizon, and answer functional. Memoized depth-first search returns the reachable terminal answer classes for every prefix, while event predicates determine support. A prefix is closed exactly when this set has cardinality one, excludes $\perp$, and its sole answer is supported. The generator returns $\tau^*$, prefix states, minimal decisive events, and answer-divergent futures when they exist. The generation seed is 20270924; source-event component identities are disjoint across partitions. This does not make the base operator programs disjoint. All 35 program specifications in the base test appear in training; ignoring timestamps, 819 of 1,440 test records share a program and token sequence with training. The timestamps and source identities differ. The analogous development count is 580 of 960 records (46 program specifications). All 12 specifications in the composition test are absent from training. The base test evaluates new timestamped trajectories within known program families; composition tests assess expression holdout, with semantic equivalence analysed separately. The training set contains 4,236 $P$, 7,052 $R$, 7,167 $C$, and 853 $U$ targets. The corresponding test counts are 2,177, 3,892, 3,775, and 428. The exact event representation and bounded grammar make this a controlled semantic test, not a proxy for visual-perception accuracy.

**Sampling distribution and features.** Each unique source body starts with one or two N events with equal probability, followed by two, three, or four independent draws uniformly from A, B, N, and DONE, and then DOMAIN_END. The original shared prefix contains only those initial N events, so it excludes histories with observed objects. The evidence-bearing-prefix test in Appendix A.8 addresses

this coverage restriction without changing the original training data. Each body gap is sampled uniformly from $[0.5, 1.5]$ seconds, rounded to six decimals. STREAM_END has the timestamp of the preceding body event but is consumed afterward. The target is sampled uniformly from A and B for each operator; the declared event budget is the full source-body length, shared across its branches. There are 120 training, 40 development, 60 base-test, and 20 composition source bodies. Rejection sampling prevents identical source bodies across splits, but derived branches can share token templates as quantified above. For each base operator, the four branches retain the factual body, flip the first post-prefix event to the target (or the other object if already equal), close the domain immediately and pad with N, or truncate at the shared prefix. All then consume STREAM_END. Compositions are AND(Exists, Complete), OR(For-all, Complete), and PAIR(Count, Last).

Table 29: Symbolic input schema. All dynamic features use consumed events only. Declared program information is available at question time; realized terminal answers, branch IDs, and future times are excluded.

| Field | Width | Encoding |
|---|---|---|
| Observed token | 7 | START, N, A, B, DONE, DOMAIN_END, STREAM_END |
| Observed time | 1 | $s_t/10$ |
| Consumed body-event count | 1 | body events consumed / 8; end token adds zero |
| Question operator tree | 63 | 7 ordered binary-tree positions $\times$ 9 one-hot operators |
| Question target | 2 | one-hot A/B |
| Declared event budget | 1 | maximum body events / 8 |
| Horizon type | 1 ID | STREAM_END, embedded separately |

Before domain closure, admissible next events are N, A, B, DONE, DOMAIN_END, and STREAM_END. After DOMAIN_END, only N or STREAM_END is legal; at the event budget only STREAM_END is legal. No event follows STREAM_END. The reachable-answer recursion operates on this grammar, which is broader than the four sampled branches in a group.

### A.11 Record schema and inclusion criteria

Within an intervention group, the question, operator, horizon, timestamps, and decoded observations remain identical until divergence; per-frame hashes verify the shared prefix. Suffixes preserve an answer, supply answer-changing evidence, or terminate an unfinished domain. An answer-changing suffix witnesses non-closure; an answer-preserving suffix alone is not a closure certificate.

Video references are *certificate-closed*, *sampled-non-closed*, or *horizon-unanswerable*. At $H_X$, they are $C$ only for a unique supported answer, and $U$ otherwise. These judgments remain relative to the declared grammar and the correctness of visual evidence annotations.

A source trajectory is one original video; each intervention group couples a question, operator, horizon, and shared prefix. A record is one branch, and a time target pairs it with a decision time.

AI-assisted visual annotation produced the pilot's questions, operators, support onsets, evidence intervals, closure rules, and terminal labels from sampled frames. An author subsequently reviewed and confirmed the retained annotations without label changes. The reference timeline distinguishes support from closure: an observed candidate remains $R$ while a permitted continuation can change it. Fourteen records are excluded for visual-evidence or task-validity reasons before model evaluation; this selection changes the retained sample distribution. Review rationales and exclusions are retained with the annotations. These human-reviewed, AI-assisted reference labels use subjective evidence thresholds at one-frame-per-second annotation resolution.

**Worked video certificate.** For source `Jfx2FCAYGno.6.mp4`, the existential question asks whether the woman ever holds a sign reading "PRISON", restricted to sampled frames through the observed stream end. The human-reviewed annotation identifies this sign in source frame 0. Its positive-witness certificate assumes that the sign is correctly read and that later frames cannot erase the occurrence of an earlier in-scope event. The question-only cache prefix at time 0 has no visible frame and is $P$; frame 0 arrives at time 1 and changes the reference state to $C$. Under this declared existential task, suffixes may add objects, replace the held object, or end the stream, but must preserve the observed prefix, so all retain the positive answer. The certificate is conditional on the annotated

visual evidence. Input hashes establish observation identity; the author review addresses the visual interpretation, not independent inter-annotator agreement.

Validation checks timestamps, legal transitions, horizon alignment, prefix identity, and source-component splits. The reviewed pilot has no conflicting labels on verified shared prefixes; this structural consistency does not establish visual semantic correctness. All branches from a component share one split, fixed by seed 2027. Scores concern operators present in the test split, not zero-shot operator transfer.

Pilot records store question/source identifiers, question time, operator, horizon, claims, state timeline, terminal answer, sampling, provenance, and review metadata. The original annotator field identifies OpenAI Codex as the AI visual annotator and is retained alongside the author review described here. Caches add input identities and source-component split membership. Each committed claim includes evidence and a closure rule. The supplementary archive's `annotation/README.md` locates the retained review criteria, per-record rationales, exclusion decisions, and available date metadata. An exact labelling-model version and complete interactive annotation history were not retained, so the AI labelling process is not exactly replayable.