CRITERION-GUIDED EVIDENCE CHAIN COMPOSITION AND SYMBOLIC REASONING FOR PARENT-CHILD INTERACTION VIDEO ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Parent-child interaction (PCI) video assessment requires determining whether caregiver and child behaviors jointly satisfy standardized criteria governing participant roles, target identity, and temporal relations. Relevant cues are distributed across time and modalities, so retrieving relevant segments alone is insufficient to establish whether these constraints are jointly satisfied. Moreover, existing video understanding models may miss relevant behavioral cues in long videos, risking the misinterpretation of missing observations as evidence that a behavior did not occur. To address these challenges, we introduce EviChain, a framework for criterion-guided evidence chain composition and symbolic reasoning over PCI videos. EviChain organizes multimodal observations into a reusable pool of interaction chains, supporting multiple assessment criteria without repeating the initial perception stage. Each criterion is compiled into a Question Logic Program (QLP) that specifies the required evidence roles and logical constraints, guiding retrieval and incremental composition of complementary chains to fill gaps in the evidence required for execution. Sufficiency-aware three-valued symbolic execution distinguishes support, refutation, and insufficient evidence, mitigating the risk of missing observations being treated as negative evidence. Only unresolved cases are passed to an evidence-conditioned video-language verifier. We also introduce PCI-Bench, a benchmark for standardized PCI assessment comprising 243 real-world videos averaging 153 seconds in duration and 17,739 criterion-level assessments. On this benchmark, EviChain achieves 69.90% accuracy and 63.75% Macro-F1, exceeding the strongest evaluated baseline for each metric by 2.40 and 4.52 percentage points, respectively.

1 INTRODUCTION

Many fine-grained video assessment tasks require determining whether observed behaviors satisfy predefined criteria. Parent–child interaction (PCI) assessment is a representative setting, where standardized behavioral criteria support developmental evaluation (Barnard & Eyres, 1979). Each criterion specifies an observable behavior and the conditions for judging its presence, which may involve participant roles, target identity, or temporal relations. For example, one criterion asks whether the caregiver smiles or touches the child within five seconds of the child's smile or vocalization; another asks whether the child looks at the caregiver's face or the task materials after a verbal or nonverbal cue (Barnard et al., 1989). These judgments require linking observations across time and modalities (Yu & Smith, 2013; Suarez-Rivera et al., 2019): the criterion determines which behaviors constitute relevant evidence and which relations must be verified. Computational advances provide useful foundations: HARMONI extracts fine-grained 3D and acoustic measurements from naturalistic caregiver–child videos (Weng et al., 2025), while evaluations of multimodal foundation models report persistent difficulties with child-initiated eye contact, joint attention, and autism-related micro-behaviors (Shi et al., 2025; Kommineni et al., 2025). In practice, one multi-minute session is assessed against dozens of criteria with distinct evidence requirements. This setting calls for reusable behavioral observations that can be assembled and verified according to each criterion.

Recent long-video systems address context length through temporal compression, retrieval, or adaptive frame selection (Shen et al., 2025; Ma et al., 2025; Luo et al., 2025; Wang et al., 2026); other

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.

approaches ground reasoning in frames or inspect clips on demand (Ghazanfari et al., 2026; Zhang et al., 2025; Yang et al., 2026). These advances improve evidence access, but reliable PCI assessment requires more than locating relevant behaviors. Evidence distributed across time and modalities must jointly satisfy participant, target, and temporal constraints. Retrieving a caregiver's prompt and a child's response, for example, does not establish that the response addresses that prompt within the required time window. Without assembling complementary observations and verifying these relations, a system risks crediting an interaction that does not satisfy the criterion (Gandhi et al., 2022). A further challenge is determining when observations justify a negative judgment. Because cues may be missed in long videos, an unobserved response can reflect either a genuine lack of response or incomplete observation, risking false negative scores without explicit evidence sufficiency checks. Reliable assessment therefore requires both joint verification of complementary evidence and an explicit distinction between insufficient observations and evidence that refutes a criterion.

To address these challenges, we introduce EviChain, a framework for criterion-guided evidence chain reasoning over parent-child interactions. The framework assembles complementary observations across time and modalities into a coherent evidence set, allowing participant roles, target identity, and temporal relations to be checked jointly against the assessment criterion. EviChain first organizes timestamped multimodal observations into a question-independent pool of candidate interaction chains, preserving event provenance and making the representation reusable across assessment criteria. Each criterion is compiled offline into a validated Question Logic Program (QLP) that specifies the required evidence roles and decision conditions. At inference time, the QLP guides chain retrieval and role binding. When a single chain is insufficient, the system searches neighboring events and incrementally combines complementary chains to fill missing roles, subject to consistency checks on participants, targets, and temporal relations. Assembly stops once the evidence is sufficient and execution is conclusive, or no admissible addition remains.

To distinguish insufficient observations from evidence that refutes a criterion, EviChain employs three-valued symbolic execution with criterion-specific evidence sufficiency checks. The execution returns support, refutation, or an unresolved state, rather than interpreting missing observations as negative evidence. Conclusive symbolic results determine the prediction directly. Otherwise, an evidence-conditioned video-language verifier receives the selected evidence, role bindings, and unresolved predicates to produce the binary decision. This design makes evidence sufficiency explicit within the symbolic stage and directs additional verification to cases that it cannot resolve. Each prediction retains a trace of the selected evidence and symbolic execution outcome for inspection.

Our contributions are twofold:

1. We introduce EviChain, a framework that integrates criterion-guided evidence composition with sufficiency-aware symbolic reasoning. It organizes multimodal observations into a reusable pool of interaction chains and incrementally composes complementary evidence under participant, target, and temporal constraints. Coupling evidence assembly with sufficiency-aware three-valued execution enables conclusive decisions when the evidence is sufficient while distinguishing incomplete observations from genuine refutation.
2. We introduce PCI-Bench, comprising 243 parent–child interaction videos and 17,739 assessments across 73 standardized behavioral criteria. EviChain achieves 69.90% accuracy and 63.75% Macro-F1 on this benchmark. Comparative experiments and ablation studies demonstrate the strong performance of EviChain and the effectiveness of its key components.

2 Related Work

Computational Parent-Child Interaction Analysis. Computational PCI research studies how observable behavior can support interaction assessment. HARMONI combines 3D reconstruction and audio processing to measure caregiver–child behavior in naturalistic videos (Weng et al., 2025). Other studies examine foundation models for autism-related behavioral observation (Kommineni et al., 2025; Feng et al., 2024). Closely related to our setting, Shi et al. (2025) evaluated multimodal models on clinician-annotated PCI clips and identified difficulties with child-initiated eye contact and joint attention. These studies motivate richer behavioral representations and careful evaluation of perceptual capabilities. Our focus is on assembling such observations into evidence for standard-

ized assessment criteria, whose decisions depend on participant roles, temporal contingencies, and complementary modalities.

Evidence-Grounded Long-Video Understanding. Video understanding methods address temporal localization and long-context processing through several complementary approaches. SeViLA, TimeChat, VTimeLLM, and LITA study relevant-frame selection or temporally grounded video-language modeling (Yu et al., 2023; Ren et al., 2024; Huang et al., 2024a;b). LongVU compresses spatiotemporal representations, while VideoITG selects frames conditioned on instructions (Shen et al., 2025; Wang et al., 2026). Retrieval-based methods such as DrVideo and Video-RAG gather relevant textual or multimodal information (Ma et al., 2025; Luo et al., 2025); TRACE and VideoTree introduce structured event outputs and hierarchical video representations, respectively (Guo et al., 2025; Wang et al., 2025). More recent approaches make evidence acquisition part of inference: Deep Video Discovery and LongVT support iterative video inspection, while Chain-of-Frames grounds reasoning in cited frames (Zhang et al., 2025; Yang et al., 2026; Ghazanfari et al., 2026). These approaches provide mechanisms for locating and organizing relevant information (Fan et al., 2024). We address the additional requirement of assembling observations that jointly satisfy an assessment criterion: a shared candidate chain pool supports criterion-guided composition, with explicit checks on role bindings, target identity, and temporal compatibility.

Executable Reasoning and Evidence Sufficiency. Program-aided and neuro-symbolic methods connect visual perception to executable reasoning (Yi et al., 2020; Gupta & Kembhavi, 2023). ViperGPT composes perception APIs through generated Python programs (Surís et al., 2023), while NeuS-QA translates questions into temporal logic and checks them against a long-video automaton (Shah et al., 2026). These methods establish executable representations as an alternative to purely textual inference (Pan et al., 2023). Our emphasis is on criterion-guided evidence assembly: cached Question Logic Programs specify the roles and constraints used to retrieve, combine, and evaluate interaction chains. Related to the reject option studied in selective prediction (Geifman & El-Yaniv, 2019), our symbolic stage retains an unknown state (Bauer et al., 2011) when observations do not justify a conclusion. Missing observations remain unresolved rather than becoming automatic refutations. Such cases are passed to an evidence-conditioned video-language verifier; the complete system produces binary decisions rather than final abstentions.

3 METHOD

Given a video V of duration T_V and an assessment criterion q, EviChain constructs a reusable representation of observable behavior offline and turns the criterion into an executable evidence query. We denote the validated Question Logic Program (QLP) by $P_q = (\mathcal{R}_q, \Psi_q^+, \Psi_q^-, \Omega_q)$, where $\mathcal{R}_q$ contains the semantic evidence roles, Ψ_q^+ and Ψ_q^- specify support and refutation conditions, and Ω_q contains temporal, relational, and quantitative constraints. The online computation factorizes as

$$V \xrightarrow{f_\omega} E_V \xrightarrow{g} \mathcal{C}_V \xrightarrow{a(\cdot,P_q)} B_q^* \xrightarrow{d_\theta(\cdot,P_q)} \hat{y}_q. \tag{1}$$

Here f_ω is multimodal perception, g induces a question-independent candidate pool, a retrieves and compactly composes criterion-relevant evidence, and d_θ executes the unified decision. Criterion authoring, prompting, program compilation, and implementation details are provided in Appendix B.

3.1 EVIDENCE GRAPH & CHAIN CONSTRUCTION

Question-Independent Multimodal Evidence Graphn. Five modality-specific perceptors write to a shared temporal memory. Every native output is normalized as an atomic observation $e_i = (s_i, t_i, a_i, y_i, o_i, m_i, r_i)$, where $0 \le s_i \le t_i \le T_V$ are its temporal boundaries; a_i, y_i, o_i, and m_i are the actor, event type, optional target, and source modality; and $r_i \in [0, 1]$ is calibrated reliability. Categorical fields use finite vocabularies and each event retains a source identifier. Consequently, the same event set E_V can serve every criterion without re-running perception or exposing item labels to the perceptors.

We organize these observations into a sparse directed graph $G_V = (E_V, A_V)$. An ordered pair $i < j$ is admissible only when it satisfies the temporal gate and type-level transition rules. The retained

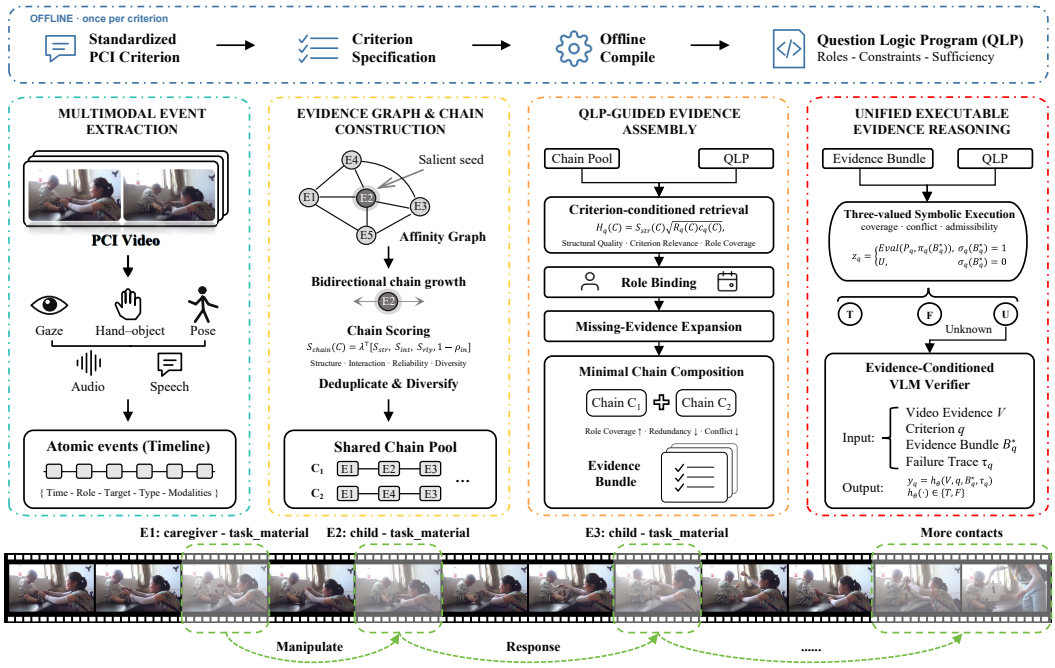


Figure 1: Overview of EviChain. Offline-compiled Question Logic Programs (QLPs) guide the retrieval and composition of reusable multimodal interaction chains. Sufficiency-aware three-valued execution resolves conclusive cases and routes unresolved cases to an evidence-conditioned video-language verifier. (See Appendix Figure A1 for an end-to-end pipeline execution demonstration.)

edge weight is

$$A_{ij} = \sqrt{r_i r_j} \sum_{k \in \{\text{tmp,act,tgt,rel,trn}\}} \alpha_k \phi_k(i,j), \qquad \alpha_k \geq 0, \quad \sum_k \alpha_k = 1. \tag{2}$$

Here $\phi_k(i,j) \in [0,1]$ measures temporal compatibility, actor continuity, target continuity, explicit relation compatibility, or behavioral transition compatibility.

Candidate discovery begins from salient observations rather than from every node. Only high-salience events are retained as seeds. Importantly, graph construction and seed selection depend on V alone: neither receives q, a question identifier, a manual rule, or a gold label. The temporal compatibility and salience definitions are provided in Appendix B.1.

Structural Candidate Chain Induction. A candidate chain $C = (e_{i_1}, \ldots, e_{i_\ell})$ is an ordered path in G_V representing one local interaction hypothesis. Starting from each salient seed, the induction procedure alternates forward and backward extensions. For an admissible event v, the marginal gain is

$$\Delta(v \mid C) = \boldsymbol{\beta}^\top [\ell(v,C), a(v,C), o(v,C), t(v,C), i(v,C)] - \beta_{\text{red}} \rho(v,C), \tag{3}$$

where the five bounded features measure the strongest retained edge, actor and target continuity, temporal compatibility, and cross-actor interaction; $\rho(v,C) \in [0,1]$ penalizes redundancy. The best positive-gain event is added until the gain, step gap, total duration, or chain length violates its bound.

Each completed chain is evaluated along four independent structural axes: temporal continuity $T(C)$, relation consistency $L(C)$, ordered dyadic transition quality $D(C)$, and generic interaction-role coverage $E(C)$. The overall chain score is

$$S_{\text{chain}}(C) = \boldsymbol{\lambda}^\top [S_{\text{str}}(C), S_{\text{int}}(C), S_{\text{rly}}(C), 1 - \rho_{\text{in}}(C)]. \tag{4}$$

Here $S_{\text{str}}(C)$ is the geometric mean of the applicable structural axes; the remaining terms measure interaction strength, event reliability, and within-chain diversity. All terms lie in $[0,1]$, and $\boldsymbol{\lambda}$ is

Algorithm 1 EviChain Evidence Assembly and Execution

Input: Event graph G_V, candidate seeds S_V, criterion program P_q, bundle cap K_B
Output: Prediction $\hat{y}_q$ and auditable trace τ_q

$\mathcal{C}_V \leftarrow \varnothing$
for all seed $s \in S_V$ **do**
    $C \leftarrow \{s\}$; grow C bidirectionally using $\Delta(v \mid C)$ until no legal extension remains
    Score C by equation 4 and add C to $\mathcal{C}_V$
Collapse near duplicates; diversify $\mathcal{C}_V$ by MMR
Rank legal chains with H_q and select the top chain C_1
$B \leftarrow \{C_1\}$; bind B to P_q and execute
if a required role is missing **then**
    Expand only compatible graph-local events; rebind and execute
while execution is insufficient or unresolved, and $|B| < K_B$ **do**
    Add chain maximizing positive marginal gain minus redundancy and conflict; rebind
$z_q, \tau_q \leftarrow \mathrm{EER}(P_q, B)$
if $z_q = \mathrm{T}$ **then**
    return Yes
else if $z_q = \mathrm{F}$ **then**
    return No
else
    return $h_\theta(V, q, B, \tau_q)$

simplex-valued. Exact definitions, aggregation weights, and implementation details are provided in Appendix B.2.

Nearby seeds often recover the same episode. We therefore collapse chains whose combined node, temporal, event-type, and actor overlap exceeds a duplicate threshold, and diversify the survivors with maximal marginal relevance (MMR) (Carbonell & Goldstein, 1998), using $S_{\mathrm{chain}}(C)$ as relevance and $\mathrm{Sim}(C, C')$ as redundancy with trade-off weight ξ. This retains strong alternatives at different times or involving different participants and targets. The resulting pool $\mathcal{C}_V = g(E_V)$ remains shared by all criteria. The pool budget is $\lceil T_V/(0.7\,\mathrm{s}) \rceil$ clipped to $[80, 250]$, where T_V is video duration. This budget caps the shared candidate pool; it is distinct from the query-specific bundle cap K_B.

3.2 QLP-Conditioned Evidence Assembly

Question dependence begins only after the candidate pool is frozen. The QLP exposes event, actor, relation, temporal, target, and semantic-role requirements. For each $C \in \mathcal{C}_V$, we form a bounded match vector $\mathbf{m}(C, q) \in [0, 1]^7$ from these typed compatibilities and the two chain quality terms. Hard violations of actor identity, direction, ownership, or time order are removed before ranking. For a legal chain, relevance, bindable role coverage, and the retrieval score are

$$\begin{aligned} R_q(C) &= \boldsymbol{\eta}^\top \mathbf{m}(C, q), \qquad c_q(C) = \frac{|\mathcal{R}_q^{\mathrm{bind}}(C)|}{|\mathcal{R}_q|}, \\ H_q(C) &= S_{\mathrm{str}}(C)\sqrt{R_q(C) c_q(C)}, \end{aligned} \tag{5}$$

where $\boldsymbol{\eta}$ is simplex-valued and $\mathcal{R}_q^{\mathrm{bind}}(C)$ contains roles with at least one legal event assignment. Let $\mathcal{C}_q$ denote the retrieved shortlist.

The top chain is bound first by enumerating type-compatible events in temporal order and rejecting assignments that violate Ω_q. If only a brief or low-confidence role is missing, a bounded local search examines existing graph neighbors. This expansion preserves provenance: it may recover a previously unselected atomic event but cannot synthesize one.

When a single chain is insufficient or leads to an unresolved execution, the composer constructs a compact complementary bundle. Let $B = \{C_1, \ldots, C_m\} \subseteq \mathcal{C}_q$ and $E(B) = \bigcup_{i=1}^{m} E(C_i)$ denote

the union of provenance-preserving event identifiers in the selected chains. The bundle score is

$$B_q^* = \underset{B \subseteq \mathcal{C}_q,\, |B| \le K_B}{\arg\max} \left[\mathrm{cov}_q(B) - \lambda_r \,\mathrm{red}(B) - \lambda_c \,\mathrm{conf}_q(B) \right]. \tag{6}$$

Here $\lambda_r, \lambda_c \geq 0$ are fixed penalty coefficients; coverage, redundancy, and conflict are defined in Appendix B.3.

Equation 6 is implemented greedily. Starting from the highest-ranked chain, the composer adds the candidate with the largest positive marginal score only when it covers at least one previously unbound required role. Role binding, conflict detection, and executable evaluation are repeated after every addition. The formal procedure stops at sufficient and conclusive execution, the bundle cap, or the absence of an admissible positive-gain addition.

Joint execution across chains. Role-wise coverage is a retrieval surrogate: individually available role candidates need not admit a jointly consistent assignment. The event union is formed by identifier, so an observation shared by two chains is counted once. For local chain-grounded predicates, role binding and symbolic evaluation operate on this pooled evidence:

$$\begin{aligned} \pi_q(B) &= \mathrm{Bind}(P_q, E(B)), \qquad \pi_q(B)(r) \subseteq E(B), \quad r \in \mathcal{R}_q, \\ u_q^{\pm}(B) &= \mathrm{Eval}_{\mathrm{AST}}\big(\Psi_q^{\pm};\, \pi_q(B)\big) \in \{\mathrm{T}, \mathrm{F}, \mathrm{U}\}. \end{aligned} \tag{7}$$

Here $\pi_q(B)(r)$ contains the event identifiers assigned to role r and is empty when that role remains unbound; $u_q^+(B)$ and $u_q^-(B)$ are the truth values of the support and refutation conditions under the same binding. Different roles may therefore be supplied by different chains, while their relations are evaluated jointly. For example, an initiation in one chain and a response in another must be evaluated together against the required temporal relation. The item's full support or refutation condition, including any quantifier and applicability rule, must still be evaluated before a final symbolic label is accepted. Conflict and sufficiency checks control acceptance of the resulting symbolic decision (Section 3.3).

3.3 Unified Executable Evidence Reasoning

Binding produces $\pi_q(B_q^*)$, a mapping from program roles to event identifiers. EviChain evaluates the bound QLP under three-valued semantics, returning support (T), refutation (F), or an unresolved result (U).

A criterion-specific sufficiency condition $\sigma_q(B_q^*) \in \{0, 1\}$ additionally checks whether the assembled evidence permits a conclusive decision. It evaluates required-role coverage, binding admissibility, and evidence conflicts. The symbolic outcome is

$$z_q = \begin{cases} \mathrm{Eval}\big(P_q, \pi_q(B_q^*)\big), & \sigma_q(B_q^*) = 1, \\ \mathrm{U}, & \sigma_q(B_q^*) = 0. \end{cases} \tag{8}$$

Under strong Kleene semantics (Kleene, 1952), missing primitive observations remain unresolved rather than becoming refutations. Conjunction, disjunction, quantification, temporal ordering, counting, duration, and ratio predicates propagate the three truth values according to their corresponding execution rules.

The complete EviChain decision follows a single conditional rule: $\hat{y}_q = z_q$ whenever $z_q \in \{\mathrm{T}, \mathrm{F}\}$; otherwise, $\hat{y}_q = h_\theta(V, q, B_q^*, \tau_q)$, where τ_q denotes the predicate-level failure trace. Thus, h_θ functions as an internal conditional branch rather than a second answering pipeline: it remains inactive whenever symbolic execution is conclusive and otherwise receives the same criterion, selected bundle, role bindings, and unresolved predicates. Direct Video+Question inference is retained only as an ablation. For every prediction, EviChain stores the selected event identifiers, bundle scores, z_q, τ_q, and the final decision source.

4 Experiments

We evaluate overall assessment performance, evidence composition, and the contributions of perceptual inputs and conditional verification.

Table 1: Comparison with existing methods on the PCI assessment benchmark. We report accuracy (Acc.) together with macro-averaged precision (Macro-P), recall (Macro-R), and F1 (Macro-F1) to account for the imbalanced label distribution. V and A+V denote visual-only and audio-visual inputs, respectively. The best and second-best results are shown in **bold** and underline, respectively.

Method	Input	Acc. ↑	Macro-P ↑	Macro-R ↑	Macro-F1 ↑
Video understanding and reasoning methods					
SeViLA (Yu et al., 2023) [NeurIPS'23]	V	37.32	54.68	50.33	28.63
VideoITG-8B (Wang et al., 2026) [CVPR'26]	V	62.60	58.38	57.36	57.41
CoF (Ghazanfari et al., 2026) [CVPR'26]	V	44.42	57.88	54.24	41.61
LongVT-RL (Yang et al., 2026) [CVPR'26]	V	40.50	59.71	52.45	34.48
SASR (Yang et al., 2024) [EMNLP-F'24]	A+V	64.55	**77.36**	51.60	42.33
TSPM (Li et al., 2024) [ACM MM'24]	A+V	<u>67.50</u>	<u>68.96</u>	57.24	54.39
General-purpose multimodal language models					
Qwen2.5-VL-7B (Bai et al., 2025)	V	58.72	59.17	<u>59.87</u>	58.13
LLaVA-OneVision-7B (Li et al., 2025)	V	60.63	55.83	55.05	54.94
VideoLLaMA2.1-7B (Cheng et al., 2024)	A+V	65.87	62.85	56.80	55.08
Qwen2.5-Omni-7B (Xu et al., 2025)	A+V	63.15	59.62	59.08	<u>59.23</u>
EviChain (Ours)	A+V	**69.90**	67.99	**63.41**	**63.75**

4.1 Experimental Setup

Benchmark. We introduce PCI-Bench using parent–child interaction recordings from a multi-center study. The evaluation set contains 243 video sessions, each annotated for all 73 criteria of the PCI Educational Scale, yielding 17,739 binary video–criterion assessments. Labels were determined through voting by multiple domain experts and underwent an additional verification step before finalization. A separate set of 10 videos is used for method development and is excluded from evaluation. The criteria concern observable caregiver and child behaviors, including temporal responses and interactions with task materials; the labels do not represent clinical diagnoses. Dataset composition and duration statistics are provided in Appendix A.1.

Baselines and evaluation. We compare against eleven baselines spanning frame selection, temporal reasoning, audio-visual reasoning, and general video-language understanding. All baselines are evaluated using publicly released official checkpoints and their recommended preprocessing and inference configurations, without task-specific training or fine-tuning on PCI-Bench. We use the same evaluation split and binary-answer protocol across methods while preserving each model's native input interface. Table 1 reports results on the full benchmark. These are system-level comparisons rather than matched-compute experiments. For EviChain, the candidate pool is shared across criteria, and the conditional verifier is invoked only after an unknown symbolic result. The main evaluation uses the production verifier on symbolic-unknown cases; the verifier-input study evaluates protocol-trimmed views on a fixed unknown subset.

4.2 Main Results

Table 1 shows that EviChain achieves the highest accuracy and Macro-F1 on PCI-Bench, reaching 69.90% and 63.75%, respectively. It outperforms TSPM, the strongest baseline in accuracy, by 2.40 percentage points, and Qwen2.5-Omni, the strongest baseline in Macro-F1, by 4.52 points. The differing baseline results highlight the importance of class-balanced evaluation: TSPM achieves 67.50% accuracy but only 54.39% Macro-F1, whereas Qwen2.5-Omni attains 59.23% Macro-F1 despite its lower accuracy. In contrast, EviChain improves both metrics, demonstrating consistent gains in overall and class-balanced performance.

These results highlight the challenges of standardized interaction assessment, where recognizing isolated behaviors is often insufficient. As detailed in Table A2, 50 of the 73 criteria involve absence, temporal or quantitative constraints, directed responses, or composite behaviors. EviChain addresses these requirements through criterion-guided evidence composition and sufficiency-aware symbolic reasoning. The decision-source analysis in Table A6 shows that symbolic execution re-

Table 2: Perceptual evidence ablation on PCI-Bench (%). Cov.: symbolic coverage; Ans. Acc.: accuracy on determinate symbolic outputs. The remaining metrics evaluate final outputs. Audio removal excludes acoustic events and speech transcripts; transcript removal retains acoustic events.

	Symbolic stage		Complete pipeline	
Configuration	Cov.	Ans. Acc.	Acc.	Macro-F1
All modalities	86.21	69.72	69.90	63.75
w/o audio evidence	70.37	70.07	66.22	55.42
w/o speech transcripts	77.25	71.15	67.52	57.48
w/o hand–object interactions	85.02	68.29	66.78	57.34
w/o gaze	85.43	68.69	67.52	59.79
w/o body movements	86.24	69.36	67.93	60.01

Table 3: Evidence assembly ablation (%). $\overline{n}_{\text{chain}}$ is the mean chain count; Cov. and Ans. Acc. are symbolic coverage and accuracy on determinate outputs. All variants use the same perception.

			Symbolic stage		Final prediction	
ID	Assembly strategy	$\overline{n}_{\text{chain}}$	Cov.	Ans. Acc.	Acc.	Macro-F1
A	Top-1 retrieval	0.94	28.11	74.17	60.19	59.29
C	Top-3 union	2.80	30.63	**74.78**	61.32	60.25
B	Local role expansion	3.64	**86.77**	66.59	66.05	54.01
D	QLP composition w/o conflict control	1.76	86.21	69.72	68.67	60.89
E	Full QLP composition	1.75	86.21	69.72	**69.90**	**63.75**

solves 86.21% of video–criterion pairs, with conditional verification handling the remaining 13.79%. Further analysis by evidence requirement (Appendix C.3, Table A9) shows that EviChain consistently outperforms TSPM in both accuracy and macro-F1 across all three criterion groups. The largest gains occur on criteria requiring global coverage and exclusion, with improvements of 6.61 percentage points in accuracy and 16.51 points in macro-F1.

4.3 Perceptual and Evidence Assembly Analysis

Perceptual inputs. Table 2 separates the effects of perceptual evidence on symbolic coverage and final prediction quality. Removing acoustic and speech evidence causes the largest drops in coverage and Macro-F1 (15.84 and 8.33 percentage points), while removing speech transcripts alone reduces them by 8.96 and 6.27 points, respectively, highlighting the contribution of linguistic content. Among visual sources, hand–object interactions have the largest Macro-F1 impact (6.41 points), despite reducing coverage by only 1.19 points. Removing body movements even slightly increases coverage while lowering Macro-F1, showing that more determinate decisions do not necessarily imply better predictions. These results indicate that perceptual evidence supports both symbolic decision availability and recognition of unsatisfied criteria.

Evidence assembly. Table 3 compares assembly strategies with perception, QLPs, and verification held fixed. Increasing retrieval from one chain to three raises symbolic coverage by only 2.52 points (28.11% to 30.63%), whereas full composition reaches 86.21% coverage with fewer chains on average than top-3 union (1.75 versus 2.80). Local role expansion attains similarly high coverage (86.77%), but full composition improves answered accuracy from 66.59% to 69.72% while using roughly half as many chains. The difference persists in final predictions: Macro-F1 increases from 54.01% to 63.75%. Together, these comparisons suggest that completing evidence roles is sufficient to make many cases decidable, whereas selecting compatible, complementary chains improves the reliability of those decisions. Conflict control adds a further 2.86 points of final Macro-F1 over the conflict-free composer, despite unchanged aggregate symbolic coverage and answered accuracy.

Table 4: Verifier inputs on 2,446 fixed unknown cases (%). M3/M4 share events. The production reference uses its original full input package and is outside the controlled sequence.

Verifier input	Acc.	Macro-F1
M1: Video + question	43.21	31.12
M2: M1 + QLP	45.95	45.50
M3: M2 + flat events	<u>55.93</u>	**55.33**
M4: M2 + chains	50.29	48.54
M5: M4 + AST trace	**57.03**	<u>53.40</u>
M6: M5 w/o video	52.66	51.25
Full verifier (prod.)	71.01	69.92

Table 5: System configurations (%). Upper rows use a different verification protocol; lower rows retain the full symbolic stage. Random closure assigns labels to unknowns.

Configuration	Acc.	Macro-F1
Alternative evidence		
Flat events	55.09	55.08
Retrieved events	55.08	55.07
Full symbolic stage		
Binary output	66.97	60.87
Omni verifier	67.87	59.13
Full EviChain	**69.90**	**63.75**

4.4 Verifier and System-Level Analysis

Verifier inputs. Table 4 evaluates protocol-trimmed input views on the same 2,446 symbolic-unknown cases, with a shared system prompt and reasoning rules. M1–M5 vary access to the QLP, atomic events, event organization, and AST traces; M3/M4 use the same events, and M6 removes video from M5. Adding flat events improves accuracy by 9.98 percentage points, whereas reorganizing them into chains reduces it by 5.64 points. Adding AST traces yields the highest controlled accuracy (57.03%), while flat events achieve the highest controlled Macro-F1 (55.33%). Removing video lowers accuracy by 4.37 points. The full verifier is reported separately as a production reference (71.01% accuracy). It uses the original pipeline-selected evidence bundle and production input package, including support/refute organization, conflict cues, and criterion-dependent textual audio observations. Main end-to-end results combine conclusive symbolic decisions with the production verifier; Appendix C.4 details the protocols and class-wise results.

System configurations. Table 5 shows nearly identical performance for flat and retrieved atomic evidence, suggesting that retrieval alone provides little benefit without richer evidence assembly in these configurations. Full EviChain achieves substantially higher accuracy and Macro-F1, although its distinct verification protocol precludes attributing the entire gain to assembly. Preserving U for conditional verification improves accuracy and Macro-F1 by 2.93 and 2.88 percentage points over forced binary output without a verifier (Appendix C.2). Holding the symbolic stage fixed, replacing the verifier with Omni reduces accuracy by 2.03 percentage points and Macro-F1 by 4.62 points, showing that verification remains consequential for unresolved cases.

5 Conclusion

We introduced EviChain, a framework for criterion-guided evidence chain composition and symbolic reasoning in parent–child interaction video assessment. Rather than treating retrieved observations as independent evidence, EviChain organizes multimodal events into a reusable pool of interaction chains and uses compiled Question Logic Programs to assemble complementary observations under participant, target, and temporal constraints. Sufficiency-aware three-valued execution distinguishes supported or refuted criteria from unresolved cases, which are routed to an evidence-conditioned video-language verifier. We also introduced PCI-Bench, comprising 243 videos and 17,739 assessments across 73 standardized criteria. On this benchmark, EviChain achieves 69.90% accuracy and 63.75% Macro-F1, with 86.21% of assessments receiving a determinate symbolic decision. Ablation studies further demonstrate the importance of compatible evidence composition beyond single-chain retrieval. More broadly, this framework may extend to other clinical and behavioral video assessment tasks. Once compiled into QLPs, standardized assessment criteria could guide evidence composition and reasoning across multiple videos.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, drafting and restructuring portions of the manuscript, literature discovery, and feedback on the presentation of the methodology and experimental analysis. AI-assisted literature suggestions were manually verified against the original sources, and all AI-assisted text, technical descriptions, experimental interpretations, and citations were reviewed and revised by the authors. Generative AI tools were not used to generate synthetic datasets or to prove mathematical claims. The authors take full responsibility for the final content of the paper, including all methods, experiments, results, claims, and citations.

ETHICS STATEMENT

The parent–child interaction recordings used in this study were collected under approval from an institutional ethics committee for a multicenter study. The committee name and approval identifier are withheld to preserve anonymity during review and will be provided in the final version.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide the code and implementation details for the complete experimental pipeline in an anonymous repository: `https://anonymous.4open.science/r/anoymous666123/`. Appendix A describes the dataset composition, development split, and evaluation protocol, while Appendix B details evidence construction, composition, and executable reasoning. Appendix C provides additional experimental results and analyses of the evaluation protocols. Due to privacy restrictions on the parent-child interaction recordings, PCI-Bench is not publicly downloadable. Researchers interested in reproducing the reported results using the dataset should contact the authors to discuss access arrangements.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025.

Kathryn E. Barnard and Sandra J. Eyres (eds.). *Child Health Assessment. Part 2: The First Year of Life*. DHEW Publication No. (HRA) 79-25. U.S. Government Printing Office, Washington, D.C., 1979. ERIC Document Reproduction Service No. ED176864.

Kathryn E. Barnard, Mary A. Hammond, Cathryn L. Booth, Helen L. Bee, Sandra K. Mitchell, and Susan J. Spieker. Measurement and meaning of parent-child interaction. In Frederick J. Morrison, Catherine E. Lord, and Daniel P. Keating (eds.), *Applied Developmental Psychology*, volume 3, pp. 39–80. Academic Press, San Diego, CA, 1989.

Andreas Bauer, Martin Leucker, and Christian Schallhart. Runtime verification for LTL and TLTL. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 20(4), 2011.

Jaime G Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pp. 335–336, 1998.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. VideoAgent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pp. 75–92. Springer, 2024.

Tiantian Feng, Anfeng Xu, Rimita Lahiri, Helen Tager-Flusberg, So Hyun Kim, Somer Bishop, Catherine Lord, and Shrikanth Narayanan. Can generic llms help analyze child-adult interactions involving children with autism in clinical observation? *arXiv preprint arXiv:2411.10761*, 2024.

Mona Gandhi, Mustafa Omer Gul, Eva Prakash, Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Measuring compositional consistency for video question answering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5036–5045. IEEE, 2022.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pp. 2151–2159. PMLR, 2019.

Sara Ghazanfari, Francesco Croce, Nicolas Flammarion, Prashanth Krishnamurthy, Farshad Khorrami, and Siddharth Garg. Chain-of-frames: Advancing video understanding in multimodal llms via frame-aware reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2746–2755, 2026.

Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. TRACE: Temporal grounding video llm via causal event modeling. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 89594–89618, 2025.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14953–14962. IEEE, 2023.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14271–14280. IEEE, 2024a.

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *European Conference on Computer Vision*, pp. 202–218. Springer, 2024b.

Stephen Cole Kleene. *Introduction to metamathematics*. Bibliotheca Mathematica. North-Holland, 1952.

Aditya Kommineni, Digbalay Bose, Tiantian Feng, So Hyun Kim, Helen Tager-Flusberg, Somer Bishop, Catherine Lord, Sudarsana Kadiri, and Shrikanth Narayanan. Can multimodal foundation models help analyze child-inclusive autism diagnostic videos? In *Proc. Interspeech 2025*, pp. 3050–3054, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Guangyao Li, Henghui Du, and Di Hu. Boosting audio visual question answering via key semantic-aware cues. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5997–6005, 2024.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, et al. Video-RAG: Visually-aligned retrieval-augmented long video comprehension. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. Drvideo: Document retrieval based long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18936–18946. IEEE, 2025.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3806–3824, 2023.

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

Sahil Shah, SP Sharan, Harsh Goel, Minkyu Choi, Mustafa Munir, Manvik Pasula, Radu Marculescu, and Sandeep Chinchali. Neus-qa: Grounding long-form video understanding in temporal logic and neuro-symbolic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8805–8813, 2026.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54582–54599, 2025.

Weiyan Shi, Hai Viet Le, and Kenny Tsu Wei Choo. Towards multimodal large-language models for parent-child interaction: A focus on joint attention. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–6, 2025.

Catalina Suarez-Rivera, Linda B Smith, and Chen Yu. Multimodal parent behaviors within joint attention support sustained attention in infants. *Developmental psychology*, 55(1):96, 2019.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11854–11864. IEEE, 2023.

Shihao Wang, Guo Chen, De-an Huang, Zhiqi Li, Minghan Li, Guilin Liu, Jan Kautz, Jose M Alvarez, Lei Zhang, and Zhiding Yu. VideoITG: Multimodal video understanding with instructed temporal grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24640–24650, 2026.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3272–3282. IEEE, 2025.

Zhenzhen Weng, Laura Bravo-Sánchez, Zeyu Wang, Christopher Howard, Maria Xenochristou, Nicole Meister, Angjoo Kanazawa, Arnold Milstein, Elika Bergelson, Kathryn L Humphreys, et al. Artificial intelligence–powered 3D analysis of video-based caregiver-child interactions. *Science advances*, 11(8):eadp4422, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report, 2025.

Tianyu Yang, Yiyang Nan, Lisen Dai, Zhenwen Liang, Yapeng Tian, and Xiangliang Zhang. SaSR-Net: Source-aware semantic representation network for enhancing audio-visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15894–15904, 2024.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, et al. Longvt: Incentivizing "thinking with long videos" via native tool calling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33816–33826, 2026.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: CoLlision events for video REpresentation and reasoning. In *International Conference on Learning Representations*, 2020.

Chen Yu and Linda B Smith. Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PloS one*, 8(11): e79659, 2013.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *Advances in Neural Information Processing Systems*, volume 36, pp. 76749–76771, 2023.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

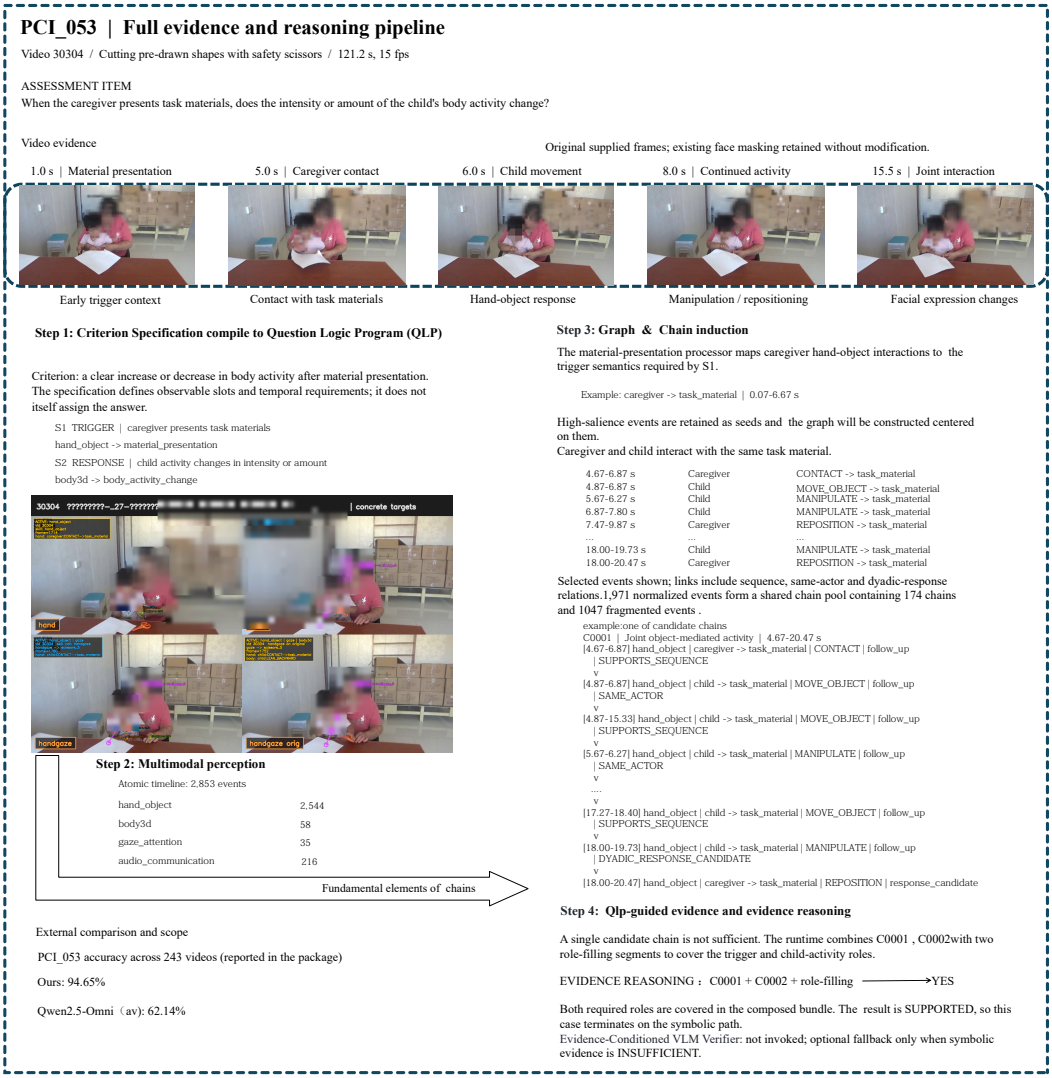


Figure A1: **Full-pipeline execution workflow of EviChain.** The figure traces atomic multimodal event extraction, QLP-guided evidence composition, and conclusive symbolic termination on a benchmark instance.

A PCI-Bench and Evaluation Protocol

A.1 Dataset Composition and Development Split

The source annotation table contains 255 sessions. We exclude two sessions with incomplete criterion labels and reserve 10 sessions for development, leaving 243 evaluation videos. Each evaluation video has binary reference labels for all 73 criteria, giving 17,739 video–criterion pairs: 11,253 positive and 6,486 negative labels. The development videos and their families are disjoint from the evaluation set.

Table A1 summarizes the evaluation recordings. The 15 instructional tasks describe the activity performed in a session; they are distinct from the 73 assessment criteria applied to every session. Bead threading accounts for 44.4% of the videos, and most recordings last between one and three minutes. These distributions characterize the benchmark's scope and should not be interpreted as a balanced sample of instructional activities or video durations.

Table A1: PCI-Bench evaluation set. All statistics exclude the 10 development videos. Every evaluation video is assessed against all 73 criteria.

Evaluation size and labels	
Video sessions	243
Criteria per video	73
Video–criterion pairs	17,739
Positive / negative labels	11,253 / 6,486
Video duration	
Total duration	10.34 h
Mean $\pm$ standard deviation	153.2 ± 103.8 s
Median [Q1, Q3]	126.9 $[84.5, 196.8]$ s
Minimum / maximum	$19.7/722.8$ s
≤ 1 min	31 (12.8%)
$>$ 1–3 min	143 (58.8%)
$>$ 3–5 min	45 (18.5%)
$>$ 5 min	24 (9.9%)
Recorded instructional task	
Bead threading	108 (44.4%)
Copying a straight line with a crayon	34 (14.0%)
Completing a three-piece puzzle	26 (10.7%)
Stacking blocks	20 (8.2%)
Other 11 tasks	55 (22.6%)

A.2 Criterion Structure and Representative Examples

The criteria are organized into three mutually exclusive and exhaustive groups describing their evidence requirements (Table A2): atomic evidence (C1), compositional and temporal reasoning (C2), and global coverage and exclusion (C3). This grouping reflects the primary behavioral and structural constraints required to verify each criterion. Fifty of the 73 criteria belong to Groups C2 and C3, extending beyond atomic event or state recognition to require joint verification of temporal contingencies, multi-actor coordination, or session-wide behavioral patterns.

Only representative paraphrases are shown because the full scale wording is subject to copyright restrictions. The examples identify the relevant actors and conditions without reproducing the complete 73-item instrument. These descriptive groups characterize the assessment evidence requirements and are distinct from the executable logic families used by the sufficiency checker in Appendix B.4.

Table A2: Criterion structure groups and representative paraphrases. Counts sum to 73; the item identifier in parentheses refers to the original scale.

Group	Evidence requirement	Items	Representative condition
C1	Atomic evidence	23	(25) The caregiver smiles or laughs at the child during teaching.
C2	Compositional and temporal reasoning	34	(64) The child vocalizes or babbles within 5 seconds after the caregiver speaks.
C3	Global coverage and exclusion	16	(11) The caregiver avoids physically forcing the child to complete the task.

A.3 Evaluation Units and Metric Interpretation

Complete-pipeline metrics pool predictions over the 17,739 video–criterion pairs. Accuracy is the fraction of correct binary predictions. Macro-precision, macro-recall, and Macro-F1 are the unweighted averages of the corresponding positive- and negative-class metrics. Thus, Macro-F1 averages the two class-specific F1 scores; it is neither the harmonic mean of macro-precision and macro-recall nor an average of 73 separately computed criterion scores.

For symbolic execution, let N be the total number of pairs, N_{det} the number with determinate outputs, and N_{correct} the number of correct determinate outputs. Coverage is N_{det}/N, answered accuracy is $N_{\mathrm{correct}}/N_{\mathrm{det}}$, and symbolic end-to-end accuracy is N_{correct}/N, counting unresolved cases as incorrect. Symbolic end-to-end accuracy therefore equals coverage times answered accuracy when both are expressed as fractions. The complete pipeline instead evaluates the final binary predictions after conditional verification. Conditional accuracies from different decision subsets are not directly comparable without their associated coverage.

All methods answer the same binary assessment task. The main comparison uses each model's supported input interface and is a system-level comparison rather than a matched-compute evaluation.

B Implementation and Executable Reasoning

B.1 Development Procedure and Candidate Construction

We use a separate set of 10 videos for method development and parameter selection, excluding all evaluation videos. Method-specific weights, thresholds, search budgets, and stopping criteria are selected on this development set and frozen before evaluation. Most parameters are manually selected, while family-conditioned sufficiency thresholds are determined by grid search on the same set. No task-specific gradient-based training, fine-tuning, or reinforcement learning is performed. Behavioral constraints prescribed by individual criteria, such as response time windows, are encoded in their QLPs rather than tuned as model hyperparameters.

Temporal compatibility and seed salience. We define the temporal gap between observations i and j as $\delta_{ij} = \max(0, s_j - t_i)$ and compute their temporal compatibility as $\phi_{\mathrm{tmp}}(i, j) = \exp(-\delta_{ij}/\tau)$, where $\tau > 0$ controls the rate of decay with increasing temporal separation. Actor continuity, target continuity, explicit relation compatibility, and behavioral transition compatibility provide four additional terms $\phi_k(i, j) \in [0, 1]$. The resulting edge weight satisfies $0 \leq A_{ij} \leq \sqrt{r_i r_j} \leq 1$. This reliability-dependent upper bound limits the weight assigned to connections involving low-confidence observations, even when their structural compatibility is high.

We rank e_i by the weighted combination $\mathrm{sal}(e_i) = \boldsymbol{\gamma}^\top [r_i, \widetilde{d}_i, x_i, \nu_i]$, where $\widetilde{d}_i, x_i$, and ν_i are normalized graph degree, cross-modal support, and transition novelty, and $\boldsymbol{\gamma}$ lies on the probability simplex.

Table A3 collects the numerical settings for candidate construction, retrieval, and local expansion. The event graph and candidate pool are constructed once per video without access to the assessment query or reference labels. Seeds must satisfy both the salience rank and absolute threshold. Bidirectional growth selects the legal extension with the largest gain and stops at the stated gain, gap, duration, or event-count limit. These limits govern candidate construction; they do not replace criterion-specific temporal conditions during execution.

B.2 Chain Scores and Pool Diversification

Growth and structual aggregation. In chain growth, $\rho(v, C) \in [0, 1]$ penalizes redundancy which stops the chain from growing with near-duplicate events so informative extensions are kept. Bidirectional growth is important because a reliable response may be detected before the weaker initiating event is selected as a seed.

Each completed chain is evaluated along four independent structural axes: temporal continuity $T(C)$, relation consistency $L(C)$, ordered dyadic transition quality $D(C)$, and generic interaction-role coverage $E(C)$. Each quantity is a normalized mean or ratio in $[0, 1]$; the role template used by $E(C)$ describes generic transitions such as initiation–response–follow-up and is not derived from

Table A3: Candidate construction and retrieval settings. Entries in each weight vector follow the feature order shown. Durations are in seconds.

Component	Parameter or feature order	Value
Event graph	Temporal decay τ; minimum affinity	3.0; 0.36
	Time, actor, target, relation, transition weights	$(.30, .15, .15, .20, .20)$
Seed selection	Reliability, degree, cross-modal support, novelty weights	$(.30, .30, .20, .20)$
	Rank fraction; minimum salience; seed cap	28%; 0.28; 200
Chain growth	Link, actor, target, time, interaction weights	$(.30, .18, .18, .18, .16)$
	Redundancy penalty; minimum gain	0.20; 0.30
	Maximum step gap; duration; event count	6; 20; 16
Chain scoring	Minimum structural score	0.25
	Structure, interaction, reliability, diversity weights	$(.40, .20, .30, .10)$
Pool selection	Duplicate similarity threshold	> 0.96
	MMR quality / similarity weights	$0.85/0.15$
Retrieval	Event, actor, relation, time, target, structure, chain weights	$(.20, .15, .15, .15, .10, .15, .10)$
	Minimum relevance $R_q(C)$	0.15
Local expansion	Beam width; iterations; retained completions	8; at most 3; at most 5

q. For the applicable axes $\mathcal{D}(C) \subseteq \{T, L, D, E\}$, we summarize structural consistency by the geometric mean

$$S_{\text{str}}(C) = \left(\prod_{d \in \mathcal{D}(C)} d(C) \right)^{1/|\mathcal{D}(C)|}, \tag{A1}$$

so that one strong dimension cannot mask a structural failure.

The structural dimensions used in the main text are defined below. For a chain with n ordered events, let gap_k denote the gap between consecutive events, b_k their explicit relation compatibility, and A_k their retained affinity. Let $\mathcal{X}(C)$ contain the cross-actor transitions, with transition scores d_x. Then

$$T(C) = \frac{1}{n-1} \sum_{k=1}^{n-1} \exp\left(-\frac{\text{gap}_k}{\tau_T} \right), \qquad L(C) = \frac{1}{n-1} \sum_{k=1}^{n-1} (0.55 b_k + 0.45 A_k), \tag{A2}$$

$$D(C) = \frac{1}{|\mathcal{X}(C)|} \sum_{x \in \mathcal{X}(C)} d_x, \qquad E(C) = \frac{|\mathcal{R}_{\text{obs}}(C) \cap \mathcal{R}_{\text{exp}}(C)|}{|\mathcal{R}_{\text{exp}}(C)|}.$$

Here $\mathcal{R}_{\text{exp}}(C)$ contains generic structural roles, such as initiation, response, and follow-up, rather than criterion-specific roles. Only applicable dimensions enter the geometric structural score defined in the main text. A putative response already present before its trigger receives transition score 0.2. This is a candidate-ranking penalty.

Interaction strength $S_{\text{int}}(C)$ measures the average affinity of retained cross-actor transitions, while $S_{\text{rly}}(C)$ robustly aggregates event reliability using the mean, median, and minimum reliability among core events. The final term $1 - \rho_{\text{in}}(C)$ rewards within-chain diversity by penalizing redundant observations.

$$S_{\text{int}}(C) = \frac{1}{|\mathcal{A}_{\text{cross}}(C)|} \sum_{(i,j) \in \mathcal{A}_{\text{cross}}(C)} A_{ij}, \tag{A3}$$

$$S_{\text{rly}}(C) = 0.50 \, \text{mean}_{e \in C} \, r_e + 0.30 \, \text{median}_{e \in C} \, r_e + 0.20 \min_{e \in \mathcal{E}_{\text{core}}(C)} r_e. \tag{A4}$$

We set $S_{\text{int}}(C) = 0$ if no valid cross-actor edge exists. The set $\mathcal{E}_{\text{core}}(C)$ contains events occupying activated core structural roles. The remaining diversity term is $1 - \rho_{\text{in}}(C)$, where $\rho_{\text{in}}(C)$ is the normalized within-chain redundancy ratio. The four aggregation weights are listed in Table A3.

For duplicate suppression, similarity combines Jaccard overlap of event identifiers, event types, and actors with temporal interval overlap:

$$\text{Sim}(C_1, C_2) = 0.55 J_{\text{node}} + 0.20\,\text{IoU}_t + 0.10 J_{\text{type}} + 0.15 J_{\text{actor}}. \tag{A5}$$

After near-duplicate collapse, let S denote the selected set. MMR (Carbonell & Goldstein, 1998) ranks each remaining chain by $0.85 S_{\text{chain}}(C) - 0.15 \max_{C' \in S} \text{Sim}(C, C')$.

B.3 Local Expansion and Consistent Composition

Criterion conditioning begins after the shared pool has been constructed. Retrieval compares event, actor, relation, temporal, and target requirements, together with structural and chain quality. The feature weights and relevance threshold are given in Table A3; the coverage-adjusted ranking function is defined in the main text.

Coverage-adjusted retrieval. Multiplying by coverage prevents a semantically similar chain that omits the trigger or response from outranking a complete interaction. Role- and time-diverse retrieval then keeps complementary hypotheses in the shortlist.

Local expansion features. If only a brief or low-confidence role is missing, a bounded local search examines existing graph neighbors using temporal compatibility, relation compatibility, chain relevance, and missing-role match.

For a currently unbound role, local expansion scores graph-neighbor events by

$$\Phi(v) = \left[\phi_{\text{tmp}}(v)\phi_{\text{rel}}(v) R_q(C) m_{\text{miss}}(v)\right]^{1/4}, \qquad S_{\text{exp}}(C) = \left[S_{\text{str}}(C) R_q(C) c_q(C)\right]^{1/3}, \tag{A6}$$

where $m_{\text{miss}}(v)$ measures compatibility with the missing role, and S_{exp} reranks completed candidates. Expansion only adds observations already present in the event graph. It can recover an event omitted by retrieval, but cannot recover a behavior that perception never represented.

Search limitations. The greedy search has no global-optimality guarantee. Exhausting its budget without a witness is a search outcome, not evidence that the behavior did not occur.

Bundle score components. Let $B = \{C_1, \ldots, C_m\}$ and $E(B) = \bigcup_{i=1}^{m} E(C_i)$. The three terms in Eq. equation 6 are defined as

$$\begin{aligned}
\text{cov}_q(B) &= \frac{1}{|\mathcal{R}_q|} \sum_{r \in \mathcal{R}_q} \mathbb{I}\left[\max_{e \in E(B)} s(e, r) \geq \theta_{\text{bind}}\right], \\
\text{red}(B) &= \frac{2}{m(m-1)} \sum_{1 \leq i < j \leq m} \frac{|E(C_i) \cap E(C_j)|}{|E(C_i) \cup E(C_j)|}, \qquad m > 1, \\
\text{conf}_q(B) &= \text{clip}_{[0,1]}\left(\sum_{\phi \in \Phi_q} w_\phi\, \mathbb{I}[\phi(B) = 1]\right).
\end{aligned} \tag{A7}$$

For a single-chain bundle, we set $\text{red}(B) = 0$. Here $s(e, r) \in [0, 1]$ denotes event–role compatibility and θ_{bind} is the binding threshold. Thus, $\text{cov}_q(B)$ measures the fraction of required roles supported by accepted event bindings, while $\text{red}(B)$ penalizes repeated evidence across chains. Finally, Φ_q contains the applicable role, participant, and temporal conflict predicates, with $w_\phi \geq 0$ denoting their penalties. These predicates include incompatible reuse of an event across roles, inconsistent actor or target bindings, shared responses across distinct triggers, and violations of the required temporal order. Table A4 specifies the conflict families used to test a binding; only those applicable to the criterion contribute to its conflict score.

Incremental composition. Starting from the highest-ranked chain, the composer admits a chain only if it gives positive marginal bundle gain and covers a previously unbound required role. It updates bindings, conflicts, and sufficiency after accepted additions. The formal procedure stops at sufficient

Table A4: Conflict families in evidence composition. A failed candidate binding does not itself establish that the video-level criterion is false.

Conflict	Binding condition
Incompatible role reuse	One event occupies mutually incompatible semantic roles.
Actor or target inconsistency	Assigned events violate participant identity, ownership, or target constraints.
Shared-response conflict	One response is reused across distinct triggers when separate instances are required.
Temporal violation	Assigned events violate the required order or latency condition.

and conclusive execution, the bundle cap, or the absence of an admissible positive-gain addition. The greedy search has no global-optimality guarantee. Exhausting its budget without a witness is a search outcome, not evidence that the behavior did not occur.

B.4 QLP Schema and Structural Sufficiency

A QLP specifies required and optional roles, admissible event types, participant and target constraints, temporal or quantitative predicates, support and refutation conditions, and a logic family. The compiler checks event vocabularies, role references, value ranges, and predicate argument compatibility. These checks establish program validity; fidelity to the assessment rubric is a separate semantic requirement.

B.5 QLP Generation Prompt

To compile natural language criteria and standardized assessment rubrics into executable Question Logic Programs (QLPs), we prompt foundation models with explicit grounding and structural constraints. The partial prompt template is provided below.

Prompt Template for Question Logic Program (QLP) Synthesis

[SYSTEM PROMPT]

You are synthesizing an executable Question Logic Program (QLP) for a structured video evidence system.
Convert the question and any supplied assessment rubric into the minimal executable program that preserves their decision criteria. Do not answer the question or output a YES/NO prediction.
Follow these rules:

1. **Preserve the criterion.** Include all roles, relations, and conditions required by the question or rubric. Do not add participant states, response stages, target constraints, durations, or aggregation conditions that they do not require.
2. **Ground each role.** Specify the participant and observable predicate candidates separately. Use only the supplied event and resolver registries. Do not invent predicates or map an unavailable behavior to a different observable action.
3. **Encode required relations explicitly.** Represent temporal order using the supplied operators. For a response required to occur after an initiation, include both roles and `AFTER(response, initiation)`. Event occurrence alone does not establish their relation.
4. **Preserve temporal constraints.** Retain explicitly stated timing requirements. Otherwise, use supplied symbolic temporal profiles where applicable. Do not invent numerical thresholds or extend response windows.
5. **Use aggregation only when required.** Include `COUNT`, `RATIO`, or `DURATION` only when the question requires it and the relevant quantities and reference interval can be grounded.
6. **Separate support from refutation.** The support program specifies evidence establishing the criterion. A refute program is optional and must describe evidence establishing its negation. Missing support evidence alone is not refutation.

7. **Validate the program.** Every required role must have a registered grounding mechanism, and every operator must have valid arguments. Do not silently remove a required condition to make the program executable.

Return one JSON object conforming to the supplied output schema. Do not include explanations outside the JSON or Markdown fences.

Evidence sufficiency checks. The family-conditioned checker receives the QLP, role-to-event bindings, the conflict score, and chain scores. It computes the features in Table A5. In particular, evidence quality is the mean score of the participating chains. Required-role coverage, conflict, and evidence quality are compared with the active family thresholds; additional trigger, response, stage-count, or critical-role requirements apply when enabled in that family's configuration.

Table A5: Features available to the sufficiency checker. Not every feature is an active constraint for every logic family.

Feature	Definition
Role coverage	Fraction of QLP-required roles with event bindings.
Trigger / response coverage	Coverage within trigger-role and response-role subsets.
Critical-role coverage	$0.5\times$ trigger coverage $+0.5\times$ response coverage.
Conflict score	Criterion-conditioned conflict in $[0, 1]$.
Evidence quality	Mean score of chains in the evidence bundle.
Stage count	Number of covered role categories.
Missing-role count	Number of required roles without bindings.

Let $f(q)$ be the QLP logic family, $Q(B)$ the bundle's mean chain score, and $G_{f(q)}(B)$ the conjunction of any additional enabled requirements. The structural check can be written as

$$\sigma_q^{\mathrm{str}}(B) = \mathbf{1}\begin{bmatrix} c_q(B) \geq \theta_{f(q)}^{\mathrm{cov}}, & \mathrm{conf}_q(B) \leq \theta_{f(q)}^{\mathrm{conf}}, \\ Q(B) \geq \theta_{f(q)}^{\mathrm{qual}}, & G_{f(q)}(B) = 1 \end{bmatrix}. \tag{A8}$$

Passing the structural check does not supply a missing predicate argument or override an illegal binding. The AST still evaluates the bound program under three-valued semantics. The checker returns the sufficiency result, family identifier, feature values, missing roles, and failed constraints, making the reason for withholding a symbolic decision explicit.

Three-valued execution and decision acceptance. A provisional AST evaluation may be computed to obtain the current binding and unresolved state before the checker is called. The gate controls whether a symbolic result may be accepted; it need not precede every internal AST computation. After additional evidence is assembled, binding and checking are repeated. In the production route described here, a failed final gate yields an unresolved result and routes the case to conditional verification. Passing the gate permits a determinate AST result to be used but does not force one if required predicates remain unresolved.The evaluated configuration includes a restricted temporal-relaxation heuristic for some questions, allowing some otherwise unresolved cases to be accepted as support subject to structural sufficiency checks.

Scope of evidence sufficiency. The implemented gate checks extracted-evidence structure and quality. It does not directly measure camera or microphone coverage, or whether perception detected every relevant behavior. Three-valued execution preserves unavailable predicate inputs as unknown; structural sufficiency does not certify perceptual completeness.

Under strong Kleene semantics Kleene (1952), false dominates conjunction and true dominates disjunction; otherwise an unresolved operand propagates unless the other operands determine the result. Negation preserves unknown. For existential criteria, a valid witness can establish support, whereas rejecting one binding or finding no witness in a bounded search cannot by itself establish refutation. Universal support likewise requires the relevant domain to be covered, while an admissible counterexample can establish refutation. Empty trigger sets must follow the criterion's applicability rule. These are requirements on predicate semantics, not additional raw-video observability tests performed by the structural gate.

B.6 Illustrative Binding Across Two Chains

Consider the trigger–response relation in item 64 of Table A2. The following schematic fragment isolates its timing relation; it is not a complete item-level QLP or a benchmark example. Suppose chain C_1 contains caregiver speech ending at 10.0 s and chain C_2 contains child vocalization beginning at 12.0 s. The fragment requires a caregiver speech role r_t and a child vocalization role r_r, with the local constraint $0 \leq s(e_r) - t(e_t) \leq 5$ s.

Each chain alone supplies only one of these two roles. Their union permits the binding $\pi(r_t) = e_t$, $\pi(r_r) = e_r$ and a two-second response latency. If the identities and other active conditions are admissible, this pair supplies a local witness. A child vocalization at 9.0 s does not satisfy this relation for the same trigger. That rejected pair does not establish a negative video-level label: a different response or trigger may still satisfy the criterion. If the response event is unavailable, the local predicate remains unresolved. The item's full support or refutation condition, including any quantifier and applicability rule, must still be evaluated before a final symbolic label is accepted.

C Additional Results and Protocol Analysis

Complete-system metrics are computed over all 17,739 video–criterion pairs, with macro metrics averaging the positive and negative classes rather than the 73 criteria. Symbolic coverage measures the fraction of determinate outputs, and answered accuracy measures accuracy on that subset. Counting unknowns as incorrect, symbolic accuracy equals coverage times answered accuracy (in fractional form).

C.1 Decision Sources and the Coverage Gain from Composition

The pipeline relies primarily on symbolic decisions, with conditional verification resolving the remaining unknown cases (Table A6). Since the verifier handles only symbolically unresolved cases, the two branch accuracies describe different subsets and are not directly comparable.

Table A6: Decision-source decomposition on 17,739 pairs. Share and subset accuracy are percentages; correct predictions are counts.

Decision source	Pairs	Share	Correct	Subset acc.
Symbolic execution	15,293	86.21	10,663	69.72
Conditional verifier	2,446	13.79	1,737	71.01
Full pipeline	17,739	100.00	12,400	69.90

Shared and newly resolved cases. Composition expands symbolic coverage while retaining all top-1-resolved cases and preserving aggregate accuracy on this shared subset (Table A7). With perceptual observations and criterion programs fixed, its lower overall answered accuracy reflects the inclusion of newly resolved cases with lower accuracy, rather than a decline on the shared subset. Equal aggregate accuracy does not imply identical pairwise predictions.

Table A7: Symbolic decision-set breakdown. Ans. Acc. is full composition's accuracy within each resolved subset; top-1 retrieval also achieves 74.17% on the shared subset. Shares use all 17,739 pairs. Rates are percentages.

Decision subset	Pairs	Ans. Acc.
Resolved by both	4,986	74.17
Resolved only by composition	10,307	67.58
Unresolved by both	2,446	–

Table A8: Binary versus three-valued decision policies on PCI-Bench (%). Both policies retain conclusive symbolic decisions and differ in how they resolve U cases. Bold indicates the better result for each metric.

Decision policy	**Acc.**	**Macro-P**	**Macro-R**	**Macro-F1**
Binary (random U)	66.97	63.69	60.75	60.87
Three-valued + verifier	**69.90**	**67.99**	**63.41**	**63.75**

Table A9: Performance by evidence requirement on PCI-Bench (%). Groups C1–C3 follow Appendix A.2. Δ denotes the absolute improvement of EviChain over TSPM in percentage points. Bold indicates the better result for each metric within each group.

	Acc.			m-F1		
Group	TSPM	EviChain	Δ	TSPM	EviChain	Δ
C1	69.14	**69.53**	+0.39	55.34	**63.01**	+7.67
C2	60.32	**62.10**	+1.78	50.66	**58.62**	+7.96
C3	80.40	**87.01**	+6.61	60.61	**77.12**	+16.51

C.2 Binary and Three-valued Decision Policies

We compare two policies on the same 17,739 video–criterion pairs, holding conclusive symbolic T/F decisions fixed. The binary ablation disables verification and randomly labels unresolved (U) cases, whereas the three-valued system preserves U and routes these cases to the conditional verifier. This comparison evaluates the combined strategy of preserving and verifying unresolved cases, rather than three-valued semantics alone. Table A8 reports accuracy and precision, recall, and F1 macro-averaged over the YES and NO classes.

Preserving unresolved cases for conditional verification outperforms random binary assignment across all reported metrics. This suggests that further verification is more effective than forced guessing when symbolic evidence is inconclusive, supporting the combined preserve-and-verify strategy.

C.3 Performance by Evidence Requirement

Table A9 compares EviChain with TSPM across the criterion groups defined in Appendix A.2, reporting group-wise accuracy and macro-F1 averaged over the YES and NO classes.

EviChain outperforms TSPM across all three groups, with the largest gains in C3. This pattern is consistent with the value of assembling distributed evidence for criteria requiring global coverage and exclusion. In C1 and C2, larger gains in macro-F1 than in accuracy suggest benefits in class-balanced performance that overall accuracy alone does not capture.

C.4 Verifier Inputs and Their Full-System Effects

Richer verifier inputs do not uniformly improve performance: M5 achieves the highest accuracy, whereas M3 yields the highest Macro-F1, suggesting that evidence format and execution traces affect class-wise decisions differently. All variants share the same 2,446 symbolic-unknown cases, system prompt, and reasoning rules. M1–M3 progressively add the QLP and flat events to video and question inputs; M4 reorganizes the same events into chains, M5 adds only the AST trace, and M6 removes video from M5.

production verifier provides a configuration-level reference, not a single-factor ablation: unlike M5, it retains the original evidence bundle and production package, including support/refute organization, conflict cues, and textual audio observations. All full-system variants retain the same 15,293 conclusive symbolic predictions and differ only in verifier outputs on the remaining 2,446 cases. Thus, $\Delta\mathrm{Acc}_{\mathrm{full}} = (2,446/17,739)\Delta\mathrm{Acc}_{\mathrm{unknown}}$, so verifier accuracy gains are diluted at the system level. Macro-F1 is recomputed from pooled predictions and does not obey this linear relation.

Table A10: Verifier-input ablation (%). Results cover 2,446 fixed unknown cases and the full system with 15,293 fixed symbolic decisions. M1–M6 share prompts and reasoning rules; M3/M4 share events, and M5 adds only AST traces. $R^{\pm}$ and $\mathrm{F1}^{\pm}$ denote YES/NO recall and F1; M-F1 is their class-averaged F1. Bold indicates the best Acc. and M-F1 among M1–M6. The production verifier uses its full input package and is excluded from highlighting.

	Fixed unknown subset						Full system	
Verifier input	Acc.	R_+	R_-	$\mathrm{F1}_+$	$\mathrm{F1}_-$	M-F1	Acc.	M-F1
M1: Video + question	43.21	1.15	98.67	2.25	59.98	31.12	66.07	61.98
M2: M1 + QLP	45.95	32.42	63.79	40.56	50.45	45.50	66.45	61.02
M3: M2 + flat events	55.93	59.31	51.47	60.48	50.18	**55.33**	67.82	61.63
M4: M2 + chains	50.29	28.04	79.62	39.08	58.01	48.54	67.04	**62.13**
M5: M4 + AST trace	**57.03**	74.69	33.74	66.41	40.39	53.40	**67.97**	60.94
M6: M5 w/o video	52.66	61.25	41.33	59.54	42.96	51.25	67.37	60.81
Full verifier	71.01	79.22	60.19	75.66	64.17	69.92	69.90	63.75

C.5 Computational Cost and Cross-Criterion Reuse

Cross-criterion reuse amortizes perception and candidate-pool construction across all 73 criteria, leaving evidence assembly, symbolic execution, and conditional verification as criterion-specific work. Experiments use eight NVIDIA RTX 3090 GPUs with 24 GB each. Under full module concurrency, estimated perception time averages 131.3 s per video; assembly and symbolic execution add 6.62 s, or 90.7 ms per criterion.

The combined estimate is approximately 138 s per video, excluding candidate-pool construction and conditional verification. The evaluated Qwen2.5-VL-7B-Instruct configuration takes approximately 398 s for 73 independent requests. These figures illustrate the potential benefit of shared processing, but do not establish an end-to-end speedup because the accounted stages and concurrency assumptions differ.

D Limitations

Our approach has several limitations. Its effectiveness depends on the quality and coverage of upstream perceptual observations, and missed behavioral cues may leave the assembled evidence incomplete. Evaluation is currently limited to PCI-Bench, while system-level comparisons are not compute-matched and statistical uncertainty remains unquantified. Cross-dataset evaluation is constrained by the limited availability of comparable datasets, as recordings involving children and families are subject to privacy and data-sharing restrictions. Future work will investigate generalization across assessment settings, more reliable multimodal perception, and calibrated selective prediction for unresolved cases.