

A Deterministic Evidence Layer for Vision-Language Autism Screening from Naturalistic Home Video

Wenqi Li^{1,†} Mindi Ruan^{2,†} Chuanbo Hu¹ Shuo Wang^{3,*} Xin Li^{1,*}

¹Department of Computer Science, University at Albany, SUNY, Albany, NY, USA

²Lane Department of Computer Science and Electrical Engineering,
West Virginia University, Morgantown, WV, USA

³Department of Radiology, Washington University in St. Louis, St. Louis, MO, USA

[†]Co-first authors. *Corresponding authors: shuowang@wustl.edu, xli48@albany.edu

Abstract

Autism spectrum disorder (ASD) is diagnosed through specialist observation of a child’s social behavior, and access to that expertise is the bottleneck for early identification. Vision-language models (VLMs) describe a child’s behavior from video well; the verdict drawn from the description is unstable: at temperature 0, across eight pipeline configurations on one backbone, 16–37% of clips change their predicted label between repeated runs, and the cause lies in the serving stack. We keep the VLM frozen and move the decision out of the model. Under our grounded perception constraints the VLM writes an event table of timestamped, glossary-labeled events that names the eliciting press and logs counter-evidence; a text-only stage age-calibrates the confidence of each row; a deterministic weight-of-evidence scorer sums it into an evidence total and stratifies it into a risk category, so every decision decomposes into named per-feature contributions and can be re-scored from the saved table. On 43 caregiver-recorded, protocol-free home free-play clips of preschool children, the pipeline reaches AUC 0.851 ± 0.012 , 86.0% accuracy, and F_1 71.8 over three runs. It labels 74.4% of clips correctly in every run (60.5% for the zero-shot baseline) and flags no typically developing clip in every run (9 of 31 at zero-shot). An ablation on the same backbone attributes the gain to the grounded perception constraints read through the deterministic scorer.

1 Introduction

Autism spectrum disorder (ASD) is diagnosed by watching a child: a specialist observes social communication and repetitive behavior, usually with a standardized instrument such as the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) [38], against DSM-5 criteria [2]. That expertise is scarce, waits for evaluation are long [18], and the lifetime costs of ASD are substantial [5]. Home video is the signal families can supply without a clinician: a phone, a few minutes of play, no protocol. We ask whether a vision-language model (VLM) can turn such video into a screening decision worth acting on, positioned after a positive caregiver-report screen [47] as a triage step during the wait for evaluation.

A VLM can describe what a child does in a clip, and the description is often good. The decision drawn from it is not reliable: when we run screening pipelines built on Gemini repeatedly at temperature 0, 16–37% of clips change their label between runs across eight pipeline configurations on one backbone, five repeats of an identical request spread accuracy by 7.0 points, and the mechanism lies in the serving stack [22], out of reach of anyone building on a closed API. A screening decision that changes between identical runs is not a clinical tool, and a decision that is the same every time is useful only

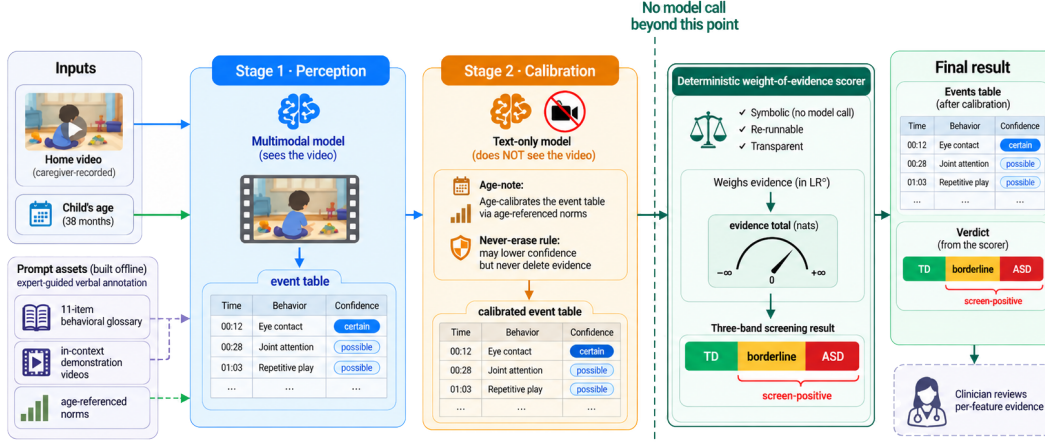


Figure 1: The screening pipeline. Stage 1, a VLM that sees the video, writes a timestamped event table under grounded perception constraints; Stage 2, a text-only model given the table and the child’s age, calibrates each row under the never-erase rule; the deterministic scorer sums the calibrated table into an evidence total and stratifies it into a low-, intermediate-, or high-risk category.

if it is also right. The quantity we target is therefore stable correctness: the share of clips a pipeline labels correctly in every run.

Two responses dominate the emerging literature. One line of work audits frozen general-purpose models on fine-grained clinical behavior tasks and finds that they handle coarse tasks such as activity recognition but fail on fine-grained ones such as activity segmentation and anxiety detection [31, 12]; another trains reliability in, fine-tuning the backbone on clinician-rated videos [23]. We take a third approach (Fig. 1): keep the model frozen and take the verdict away from it. In our pipeline the VLM is used only as a high-recall perception front-end. Prompted with a clinically aligned behavioral glossary, it must commit to timestamped, event-level observations (including explicit counter-evidence) that a human can check against the video. A text-only second stage then age-calibrates each event under a *never-erase* rule: it can lower the weight of an observation and cannot delete it. A deterministic $\ln LR^+$ weight-of-evidence scorer sums the calibrated table into an evidence total and stratifies it into a risk category. The diagnostic reasoning thus splits into a perceptual step checked against the video and a probabilistic step checked by arithmetic. Every label decomposes into named per-feature contributions, can be re-scored without a model call, and exposes a tunable recall–specificity operating point for triage. Our main contributions are:

1. **An auditable, training-free VLM screening architecture that separates grounded observation from a deterministic, evidence-based decision.** A frozen VLM writes a timestamped, glossary-labeled event table; a text-only stage age-calibrates it; a deterministic weight-of-evidence scorer turns it into an evidence total and a risk category, so every decision decomposes into named per-feature contributions and can be re-scored without a model call. On 43 caregiver-recorded, protocol-free free-play clips it reaches AUC 0.851, and its decision rule is two explicit cut-offs at interpretable evidence levels, the social-communication cap and the weight of a single certain stereotypy (Appendix B.1).
2. **Grounded perception constraints for VLM reasoning.** Stage 1 must commit to events with rationales confined to the interval; social-response deficits are coded only as failed opportunities with the eliciting press named; typical behaviors are logged as explicit counter-evidence. In the cumulative ablation the constraints lift specificity from 91.4% to 94.6% and, read through the scorer designed for these tables, raise F_1 from 50.7 to 71.8 and accuracy from 77.5% to 86.0%.
3. **A stable-correctness account of VLM-based screening.** At temperature 0, 16–37% of clips change label between repeated runs. The full pipeline labels 74.4% of clips correctly in every one of three runs, against 60.5% for the zero-shot baseline; no typically developing (TD) clip is flagged in every run (9 of 31 at zero-shot), three ASD clips are missed in every run, and the remaining variation enters through the event table.

2 Related Work

Screening ASD from video. Classical pipelines quantify engineered behavioral signals with handcrafted features and shallow statistical analyses or classifiers: gaze measured by eye tracking [30, 27] or estimated from video [63], facial dynamics [20], head and body pose [21], and speech acoustics [25]. Learned image and video models, deep [48, 49] or few-shot over handcrafted facial-dynamics descriptors [64], raise accuracy on protocol-elicited recordings, where micro-expression analysis is already limited by video quality [50]; most models were built under clinical protocols and degrade in unconstrained home scenes [9]. Caregiver-recorded home video has long been recognized as a scalable screening substrate [1, 32, 57], and the strongest current results are supervised: ensembles over pose, speech, and object features from protocolized one-minute recordings reach AUROC 0.83 (0.73 external) on 510 children [28]. Two recent neighbors bring VLMs to this substrate: zero-shot Gemini feature scoring on 50 public YouTube clips fed into pre-validated logistic classifiers [3], and LoRA fine-tuning of Gemini on 400 clinician-rated home videos [23]. We differ from the supervised line and both neighbors on three axes at once: no elicitation protocol, no training anywhere in the pipeline (including the decision layer), and a private evaluation corpus.

Decomposed and verifiable medical reasoning. Separating perception from downstream reasoning is by now a recognized design pattern in both general multimodal systems [19] and clinical applications: text-only diagnostic pipelines fuse LLM-extracted evidence with deterministic rules [39], LLMs have been applied to autism-associated language disorders [24], and a two-stage VLM analysis of parent-child interaction video finds expert alignment strong at the observation layer but weak at the judgment layer [52], supporting a separation between observation and judgment. A growing line of work likewise reserves the final say for a layer other than the LLM: Prolog-based clinical decision support [55], LLMs demoted to explainers of statistical models [60], and deterministic aggregation over expert-validated ECG observations [56]; probabilistic combination of evidence for diagnosis goes back to the earliest proposals for computer-aided diagnosis [34], and its additive form, a sum of log-likelihood-ratio weights, to early clinical decision-support systems [54]. Radiology reasoning models instead enforce evidence-diagnosis consistency through reward shaping during training [65]. A parallel line adapts or trains vision-language models for the clinical domain [43, 36]; here the backbone stays frozen. We apply this separation to raw, unprotocolized home video. Prior systems either transcribe clinician-led assessment recordings into behavioral scripts and let an LLM issue the verdict from the script [37], or train a video classifier of stereotyped behaviors with auxiliary text supervision on public clips [11]; we contribute a different interface object: a probe-gated, timestamped event table with explicit counter-evidence under a never-erase rule, consumed by a likelihood-ratio scorer, so the label is consistent with the logged evidence by construction.

Reliability of clinical (V)LM decisions. Run-to-run instability of LLM outputs persists at temperature 0; its mechanism lies in batch-variant inference kernels and is not fixable from the user side of a closed API [22]. Clinical text measurements are accumulating: a repeatability framework over 912,000 generations [53] and management-recommendation flips across five identical queries of the same vignette [33]; the remedies these studies report are prompt design, re-prompting, and sampling or ensembling several models, with the serving-stack fix left to providers [22]. Agent benchmarks measure consistency across repeated attempts with pass^k [61]. We measure clip-level label flips across repeated runs of a clinical video pipeline and report pass³ (Section 4.4), and the remedy is architectural: the decision rule is a fixed function of the saved event table and adds no variance of its own, so the variance that remains enters through the table, where it can be measured.

3 Method

The pipeline is training-free, and its label is not produced by a model. A frozen VLM supplies perception; a deterministic evidence layer, made of a behavioral glossary, a structured event schema, and a weight-of-evidence scorer, holds the clinical knowledge and the decision (Fig. 1). A clip passes through three steps: 1) Stage 1, one call to the VLM, turns the video into a table of timestamped, glossary-labeled events under grounded perception constraints; 2) Stage 2, one call to a text-only model, re-weights each row against the child’s age without deleting anything; 3) the scorer sums the table into an evidence total and stratifies it by two fixed cut-offs into a risk category. Every clip yields three outputs: the calibrated event table, the evidence total, and the risk category. Because the label

is arithmetic on a saved table, any run can be re-scored without a model call, and a screen-positive decision traces to specific rows.

3.1 Clinical Cohort and Home-Video Corpus

The cohort consists of preschool-aged children who completed comprehensive autism evaluations at an academic medical center; participant-level ASD or typically-developing (TD) status was determined by licensed clinicians using ADOS-2 [38] and DSM-5 criteria [2] with consensus review. Evaluation labels for this study are clip-level ASD/TD ratings confirmed by independent ASD specialists. From a broader home-recording protocol we analyze only the *free-play* segment of 47 children; four clips serve as in-context demonstrations, leaving a 43-clip evaluation set (12 ASD, 31 TD; per-group ages in Table A1). Caregivers filmed one continuous 3–6 minute segment of the child at play in the home, on their own smartphones or tablets, with no standardized prompts, scripted interactions, or elicitation presses. Nothing is required of families beyond a camera, and the footage carries spontaneous social, communicative, and sensorimotor behavior.

Expert knowledge enters through *expert-guided verbal annotation*: a domain expert thinks aloud over the video, recording start/end timestamps, a behavior category, an atypical/typical judgment, and an observable-cue rationale (roughly twice real time per clip). These traces ground an 11-item glossary mapping ADOS-2 Social Affect and Restricted/Repetitive Behavior constructs and DSM-5 criteria to observable descriptors (Table A6, Appendix E.2), and supply the gold-answer exemplar tables of Section 3.3. All procedures were approved by the institutional review boards of the participating institutions; parents provided written informed consent covering recording, research use, and AI-based analysis. Direct identifiers were removed and each recording was assigned a study code; the recordings are handled as identifiable human-subjects data under the approved protocol. The recordings remain under the originating institutions’ data-use protocols and are not publicly distributed; no synthetic data are used anywhere in this work.

3.2 Problem Formulation

Let v denote a short home video clip captured during free play, a the child’s age in months, and $y \in \{\text{ASD}, \text{TD}\}$ the clip-level expert rating. We construct a screening function $f : (v, a) \mapsto (\mathcal{E}', s, \hat{y})$ that returns a calibrated event table $\mathcal{E}'$, an evidence total s , and a risk category $\hat{y}$. Instead of training a video classifier, we decompose screening into two fixed model calls, which we call stages, followed by a deterministic decision layer:

$$M_{\text{percept}}(v; \Pi_1) \rightarrow \mathcal{E} = \{e_k\}_{k=1}^K, \quad M_{\text{judge}}(\mathcal{E}, a; \Pi_2) \rightarrow \mathcal{E}',$$

$$s = \text{SCORE}(\mathcal{E}') = \sum_f w_f c_f, \quad \hat{y} = \text{stratify}(s) \in \{\text{low}, \text{intermediate}, \text{high}\}, \quad w_f = \ln \text{LR}_f^+,$$

where the perception model M_{percept} is a VLM that sees the video, the judgment model M_{judge} is text-only and never sees v , c_f is a confidence factor, the sum is subject to the domain cap of Section 3.5, and stratify assigns a risk category by two fixed cut-offs, with both *intermediate* and *high* risk screen-positive. Each event is a tuple $e_k = (t_k^{\text{start}}, t_k^{\text{end}}, g_k, r_k, q_k)$: a temporal extent, a label g_k from the behavioral glossary $\mathcal{G}$ (eleven ASD-related features plus a typical-behavior marker), a short observable-evidence rationale r_k , and a confidence $q_k \in \{\text{certain}, \text{possible}\}$.

3.3 Stage 1: Perception into a Grounded Event Table

Clinical observation separates what was seen from what it means: a rater first logs behaviors, then interprets them. Stage 1 is the logging step, a single call to the frozen VLM. Its input is the video; the child’s age is withheld, so that observation is not shaped by expectation. Its prompt has four parts, each with one job (Appendix E): (a) the behavioral glossary (Table A6, Appendix E.2), which fixes the label space and defines each feature in observable terms; (b) four demonstration videos, each paired with its expert event table (Section 3.1), which show the output format and the labeling conventions, delivered multi-turn, each video as its own user turn and its table as the assistant turn, so that every table is bound to its video; (c) a four-step scaffold that orders the work into intermediate steps, in the spirit of chain-of-thought prompting [59]: 1) propose candidate time intervals; 2) describe the observable behavior in each (mental-state attribution is discouraged); 3) assign a glossary label ($g_k \in \mathcal{G}$) or mark the behavior as typical; 4) write a rationale restricted to cues visible inside the interval; (d) the grounded perception constraints below, which state when a feature may be coded.

The output is an event table with one row per observed behavior and six columns: Start and End (timestamps), Behavior (a short observable description), Comment (the matched glossary feature, or a typical-behavior marker), Note (context such as the adult action that preceded the behavior), and Confidence (certain when clearly visible; possible when occluded, brief, or marginal). Typical behaviors (clear eye contact, responsive smiling, pointing to share, reciprocal turn-taking, pretend play) are logged as their own rows, explicit counter-evidence, so the table records what argued against a deficit as well as for it. Later steps parse the table programmatically.

Grounded perception constraints. Video VLMs hallucinate temporal content [35] and lose accuracy once an answer must be tied to a time in the clip [40], so a rationale may cite only cues inside its own interval. The six social-communication features name behaviors that are reduced or absent, and an absence can be written down without any occasion to show the behavior; the press rule and the deferral of #4 therefore count an absence only against an occasion the table records, as ADOS-2 does with its presses [38]. Feature numbers below refer to the glossary (Table A6, Appendix E.2). How a feature may be coded depends on the evidence it needs, and the eleven features fall into three classes: (a) *local features* (#1 eye contact, #2 facial expression, and #6–#11, from lack of pretend play to unusual sensory responses) are coded whenever observed in an interval; (b) *probe-gated response features* (#3 reduced gestures and joint attention, #5 impaired reciprocity and response to name) are coded only when an adult makes a social bid, a *press* such as calling the child’s name, pointing at an object, or holding out a toy, and the child fails to respond; absence without a press is not coded, the press is named in the Note column (e.g., name called $\times 2 \rightarrow$ no orient), and a #3 or #5 row without one is invalid; (c) the *pervasive feature* (#4 limited peer interest and quality of social overtures) is not coded in Stage 1, because it is a judgment about the whole clip, so Stage 1 logs the raw social material (overtures and shared bids as typical-behavior rows, sustained solitary or parallel play as its own row) and leaves the decision to Stage 2.

3.4 Stage 2: Text-Only Age Calibration Under the Never-Erase Rule

Stage 2 is the interpreting step, a second call to a text-only model. Its input has four parts, the Stage 1 table, the child’s age in months, age-referenced behavioral norms, and two hand-written demonstration exchanges of the #4 rule delivered as prior turns (Appendix E.4); it never sees the video. Most glossary features name a departure from an age expectation: pretend play is not expected before 18 months, solitary or parallel play is age-typical to about 42 months, and echoing of words is part of language learning to about 30 months (norms in Appendix E.3); ADOS-2 selects its module and algorithm by age and language level [38], and DSM-5 states that the manifestations of ASD vary with developmental level and chronological age [2]. The cohort spans 26 to 48 months, across the ADOS-2 Toddler/Module 1 boundary at 30 months (Table A1), so a table read without age could not exempt the youngest children’s age-typical behavior.

Stage 2 performs three operations in order: 1) it synthesizes the pervasive feature #4 from the whole table, checking counter-evidence first, so that any clear overture, approach, or shared bid anywhere in the table blocks the row, and it is added only for a diffuse, consistent lack of social interest with adults available yet no overtures, with the weighed counter-evidence stated in the Note; 2) it calibrates each remaining row against the norms for the child’s exact age, lowering Confidence from certain to possible and writing an age note in the Note column when the behavior is common at that age (repeated questioning judged age-appropriate, for instance), and keeping certain when the deficit clearly exceeds age expectation in intensity, duration, or pervasiveness; 3) it re-emits the full table. Under the *never-erase rule* the only permitted changes are the lowered confidence, the age note, and the added #4 row: Stage 2 may not delete a row, raise a confidence, or relabel a coded feature as typical. The asymmetry is deliberate. Every row remains a visually grounded observation, and Stage 2’s reading, which is not grounded, can lower its weight and cannot remove it. The output is the calibrated event table; the verdict Stage 2 also writes is used only as the LLM verdict of the ablation (Section 4.3).

3.5 Deterministic Weight-of-Evidence Scorer

The scorer reads the calibrated table and produces the evidence total and the risk category in four steps. 1) Each glossary feature f carries a weight equal to its log positive-likelihood-ratio [26] (Good’s weight of evidence [16]),

$$w_f = \ln \text{LR}_f^+ = \ln \frac{\text{sens}_f}{1 - \text{spec}_f},$$

with provisional per-feature recall (sensitivity) and specificity set from published effect sizes and prevalence estimates (Appendix B); summing weights follows the Bayesian weight-of-evidence interpretation (derivation and design rationale in the same appendix). 2) Each feature counts *once*, at its highest observed confidence ($c_f = 1.0$ for certain, 0.25 for age-downgraded possible), so repeated events cannot inflate the score. 3) Because social-communication (SC) features are positively correlated, their sum saturates at the SC cap $\kappa_{SC} = 1.5$ nats, while the near-independent restricted/repetitive (RRB) features sum in full:

$$s = \min\left(\sum_{f \in SC} w_f c_f, \kappa_{SC}\right) + \sum_{f \in RRB} w_f c_f.$$

4) Two fixed cut-offs stratify s into three risk categories, low ($s < 1.5$), intermediate ($1.5 \leq s < 1.8$), and high ($s \geq 1.8$), with intermediate and high both screen-positive. The cut-offs sit at interpretable evidence levels: 1.5 nats equals the SC cap, so a saturated social-communication domain alone is screen-positive, and 1.8 nats lies just under the weight of a single certain stereotypy (feature 7, $w_7 = 1.838$), so high risk requires restricted or repetitive evidence. The decision is therefore a fixed function of the saved table [51]: any run can be re-scored, with weights or cut-offs re-tuned, without a model call, and each screen-positive total decomposes into the rows that produced it.

4 Experiments and Results

4.1 Experimental Setup

All models are used training-free (temperature 0, top_p 0.95, top_k 40, max_output_tokens 65,536). Seeds are the decoding seeds passed in the generation config; every configuration is run at three seed values and reported as mean $\pm$ s.d. The backbone is Gemini 3.1 Pro [17] (gemini-3.1-pro-preview, high thinking) unless a table names another. Evaluation uses the 43-clip test set (12 ASD, 31 TD; Section 3.1); we report recall, specificity, F_1 , and accuracy with ASD as the positive class, AUC where a configuration yields a continuous score, and count unparseable predictions as errors.

Three configurations recur through the section: (a) The *zero-shot baseline* is a single VLM call over the video and the child’s age with the glossary, the clinician persona, the four-step scaffold, and the structured output schema, without exemplars or age norms; its label is the VLM’s own verdict. (b) The *two-stage split* separates perception from judgment as in Section 3: Stage 1 writes the event table with the multi-turn exemplars, and the text-only Stage 2 applies the age calibration and returns a verdict; it has neither the grounded perception constraints nor the scorer, and its label is the LLM verdict of Stage 2. (c) The *full pipeline* adds the grounded perception constraints to Stage 1 and replaces the LLM verdict with the deterministic scorer, intermediate and high risk both counting as screen-positive.

4.2 Classification Performance

On the 43-clip cohort the full pipeline attains an AUC of 0.851 ± 0.012 over the scorer’s evidence total, with $86.0\% \pm 0.0\%$ accuracy, $94.6\% \pm 1.9\%$ specificity, $63.9\% \pm 4.8\%$ recall, and F_1 of 71.8 ± 1.6 across three runs, committing one or two false positives per run (Fig. 2A). At the two-stage split’s false-positive rate of 8.6% the curve gives 63.9% recall against its 41.7%; the zero-shot baseline’s operating point (86.1% recall at a false-positive rate of 37.6%) lies on the curve. An age-only logistic baseline fitted in-sample reaches AUC 0.706 (older children skew ASD here).

Fig. 2B shows the margin behind these numbers: each clip’s evidence total under the full pipeline, averaged over the three runs and with its run-to-run range, against the scorer’s two fixed cut-offs. Ten of the 31 TD clips score zero in every run, 26 never reach the screen-positive cut-off, and the five that cross it do so in one run each, so no TD clip is flagged in more than one run. Two of the five stop exactly at the cut-off, at the social-communication cap of 1.5 nats with no repetitive-behavior evidence; the other three exceed it once, by 0.2 to 0.9 nats. Of the 12 ASD clips, five exceed the high-risk cut-off in every run, and three stay below the screen-positive cut-off in every run with near-empty event tables, which holds recall at or below 75.0%. The run-to-run range of the total has a median of 0.38 nats for TD clips and 1.41 nats for ASD clips, and eight clips cross the screen-positive cut-off in some runs only. The cap and cut-offs were set on an earlier run over this cohort; re-selecting the cap and the cut-off by leave-one-out lowers pooled recall to 58–61% and specificity to 91–95%, and the cap, equal to the cut-off, cannot change a screen-positive decision (Appendix B.1).

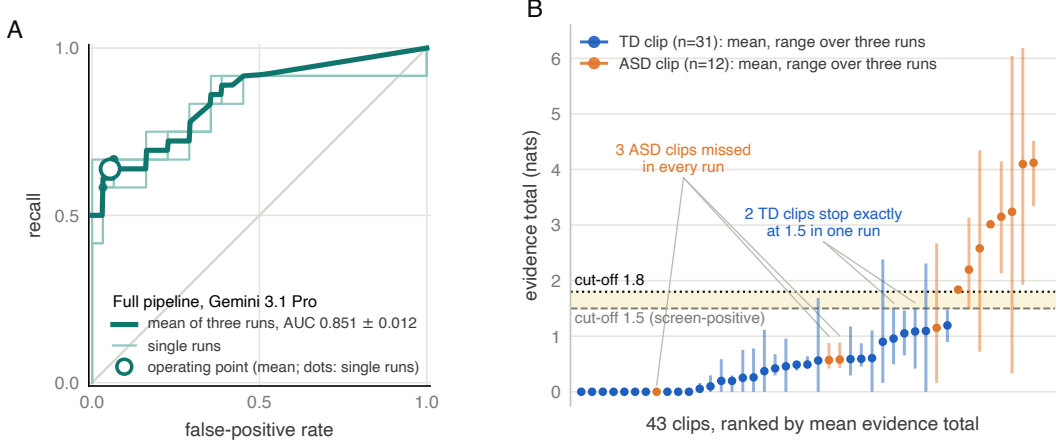


Figure 2: Classification performance of the full pipeline of our method. **A**: ROC curve of the full pipeline, mean over the three runs (bold) and single runs (thin), with its operating point (large marker: mean; small markers: single runs). **B**: evidence total of every clip under the full pipeline, mean over the three runs (dot) and range (bar). Annotations mark the three ASD clips missed in every run and the two TD clips whose total stops exactly at 1.5 in one run.

Table 1: Cumulative ablation of the full pipeline on Gemini 3.1 Pro (temperature 0, three seeds each, mean $\pm$ s.d.; positive class ASD). Each row adds one module to the row above. *Verdict* names the source of the label: VLM (single call), LLM (the text-only Stage 2), or scorer.

Configuration	Verdict	Acc (%)	Recall (%)	Spec (%)	F ₁ (%)
Zero-shot baseline	VLM	69.0 $\pm$ 5.4	86.1 $\pm$ 4.8	62.4 $\pm$ 6.7	60.9 $\pm$ 4.8
+ Few-shot exemplars	VLM	66.7 $\pm$ 1.3	55.6 $\pm$ 4.8	71.0 $\pm$ 0.0	48.1 $\pm$ 3.2
+ Age calibration	VLM	65.9 $\pm$ 2.7	91.7 $\pm$ 0.0	55.9 $\pm$ 3.7	60.0 $\pm$ 1.9
+ Two-stage split	LLM	77.5 $\pm$ 3.6	41.7 $\pm$ 8.3	91.4 $\pm$ 1.9	50.7 $\pm$ 9.0
+ Grounded perception constraints	LLM	81.4 $\pm$ 2.3	47.2 $\pm$ 4.8	94.6 $\pm$ 4.9	58.7 $\pm$ 1.4
+ Deterministic scorer (our full pipeline)	scorer	86.0 $\pm$ 0.0	63.9 $\pm$ 4.8	94.6 $\pm$ 1.9	71.8 $\pm$ 1.6

4.3 Ablation Study

The full pipeline is the endpoint of five cumulative additions to the zero-shot baseline, re-run on the same backbone (Table 1). Accuracy moves from 69.0% at zero-shot to 77.5% with the two-stage split, 81.4% with grounded perception constraints, and 86.0% once the deterministic scorer reads the same runs; over the last three rows recall rises from 41.7% to 47.2% to 63.9% and F₁ from 50.7 to 58.7 to 71.8 while specificity stays at 91.4–94.6%. The two-stage split raises accuracy by 11.6 points, mainly through specificity (55.9% $\rightarrow$ 91.4%) at a cost in recall (91.7% $\rightarrow$ 41.7%). The prompt-level additions shift the operating point without improving it: accuracy stays at 65.9–69.0% over the first three rows, and the exemplars, added to an otherwise byte-identical prompt, trade 30.6 points of recall for 8.6 of specificity. The architectural additions raise specificity from 62.4% at zero-shot to 91.4% with the two-stage split and 94.6% with the constraints, on the same backbone.

4.4 Stable Correctness Across Repeated Runs

Reasoning models sample an internal trace before the final answer, so a VLM verdict is stochastic even at temperature 0: the single-call baseline (Gemini 2.5 Pro, five repeated runs of the identical request) averages 73.0% $\pm$ 3.5% accuracy with a best-to-worst spread of 7.0 points (per-run detail in Table A5). The instability is not confined to one configuration. With the cross-seed flip rate defined as the fraction of the 43 clips whose label differs across a configuration’s repeated runs, the eight configuration–decision-rule combinations measured on Gemini 3.1 Pro span 16.3% to 37.2% (median 19.8%; per-configuration rates in Table A4). The metric counts binary verdicts near a decision boundary, where nondeterminism concentrates; the high run-to-run consistency reported for ordinal feature ratings on curated public clips [3] measures a different quantity. Fig. 3 reports two quantities

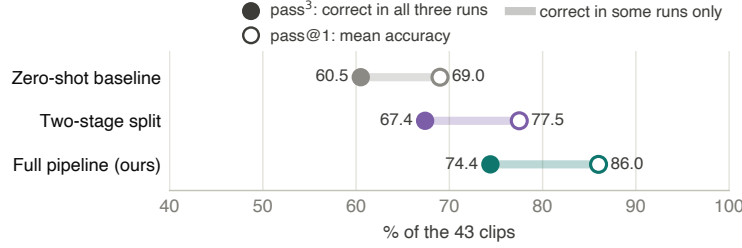


Figure 3: Stable correctness of the three configurations of Section 4.1. For each configuration, $\text{pass}@1$ and pass^3 as defined in Section 4.4; the shaded segment is the share of clips correct in some runs only.

per configuration: $\text{pass}@1$, the accuracy of a single run averaged over the three runs, and pass^3 , the share of clips labeled correctly in all three runs [61]. pass^3 rises from 60.5% at zero-shot to 67.4% with the two-stage split and 74.4% for the full pipeline, 26, 29, and 32 of the 43 clips; the gap to $\text{pass}@1$ stays between 8.5 and 11.6 points in all three, so a single run overstates what each configuration delivers in every run. Nine TD clips are flagged in every zero-shot run and none from the two-stage split onward, while five ASD clips are missed in every run of the two-stage split, with or without the constraints, and three by the full pipeline. The scorer recovers two of the five: their event tables carry at least 1.5 nats of evidence in each run while the LLM verdict reads TD in each run. The three clips missed in every run have near-empty event tables, so the remaining variation and the remaining misses both enter through the event tables.

5 Discussion and Conclusion

The operating point is consistent with confirmatory triage in an already-referred population: at this cohort’s 27.9% ASD prevalence, 94.6% specificity gives precision 0.80 (8 of 10 flagged clips in one run), which would fall sharply at general-population prevalence, and 63.9% recall means a negative result cannot exclude ASD. In the ablation, grounded perception constraints read through the deterministic scorer carry the gain; the three ASD clips missed in every run set the current recall ceiling at 75.0%.

The deterministic decision layer changes what a reviewer receives. Each flagged row carries timestamps, a glossary label, and a rationale, and the verdict follows from those rows by arithmetic, so a disagreement can be traced to a specific event. Labels flipped across repeated runs in every configuration tested at temperature 0 (Section 4.4); the scorer removes the decision rule from that variance, and the clips labeled correctly in every run rise from 26 to 32 of 43 along the ablation. The scorer’s gain over the LLM verdict holds when its cap and cut-off are re-selected by leave-one-out: pooled recall is 58–61% at 91–95% specificity (Appendix B.1), against 47.2% at 94.6% on the same tables (Table 1), and the ranking beats age alone (Section 4.2). The cut-off is also a parameter a site can set; raising it to the high-risk cut-off is the rule-in end of the dial (Appendix B.1). The zero-shot operating point, 86.1% recall at a false-positive rate of 37.6%, lies on the pipeline’s curve (Fig. 2A), so a site that needs a rule-out screen can lower the cut-off on the same saved tables and read off the specificity it gives up. A model verdict offers no such dial, and in the ablation the prompt-level additions move recall by 30 to 36 points in either direction (Table 1).

The study includes 43 clips from one site, with only 12 ASD cases, and the DeLong interval $[0.72, 1.00]$ is wide enough to admit a mediocre screen. Labels are clip-level expert ratings; one clip per child under-samples behavior; the scorer’s cap and cut-offs were set on this cohort and cannot be validated on it; Stage 2’s age norms are coarse; and nothing here speaks to ages, languages, or cultures outside the preschool, single-country cohort, so fairness across sex, language, and culture is untested. A multi-seed decomposition of the stages would sharpen the reliability account (Section 4.4). The main remaining bottleneck is perception: evidence missed in Stage 1 cannot be recovered downstream, by the scorer or by a lower cut-off. The three clips missed in every run are the concrete target. The next steps are better grounded behavioral observation, including prosody and caregiver speech for the under-detected reciprocity features, and external validation of the evidence layer at a second site and in prospective studies in which clinicians act on the tables.

Acknowledgments and Disclosure of Funding

This material is based upon work supported by the UAlbany Start-up funds for Prof. Xin Li. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author contributions. W.L. and M.R. are co-first authors and contributed equally to this work. C.H. contributes to the data collection and analysis of this project. S.W. and X.L. are responsible for securing the funding and supervising the project.

References

- [1] Halim Abbas, Ford Garberson, Eric Glover, and Dennis P. Wall. Machine learning approach for early detection of autism by combining questionnaire and home video screening. *Journal of the American Medical Informatics Association*, 25(8):1000–1007, 2018. doi: 10.1093/jamia/ocy039.
- [2] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*. American Psychiatric Publishing, Washington, DC, 2013.
- [3] Parnian Azizian, Mohammadmahdi Honarmand, Aditi Jaiswal, Aaron Kline, Kaitlyn Dunlap, Peter Washington, and Dennis P. Wall. Multimodal LLM vs. human-measured features for AI predictions of autism in home videos. *Algorithms*, 18(11):687, 2025. doi: 10.3390/a18110687.
- [4] Marianne L. Barton, Diana L. Robins, Dasal Jashar, Laura Brennan, and Deborah Fein. Sensitivity and specificity of proposed DSM-5 criteria for autism spectrum disorder in toddlers. *Journal of Autism and Developmental Disorders*, 43(5):1184–1195, 2013. doi: 10.1007/s10803-013-1817-8.
- [5] Ariane V. S. Buescher, Zuleyha Cidav, Martin Knapp, and David S. Mandell. Costs of autism spectrum disorders in the united kingdom and the united states. *JAMA Pediatrics*, 168(8):721–728, 2014. doi: 10.1001/jamapediatrics.2014.210.
- [6] Catherine A. Burrows, James W. Bodfish, Jason J. Wolff, Elayne P. Vollman, Melody R. Altschuler, Kelly N. Botteron, Stephen R. Dager, Annette M. Estes, Heather C. Hazlett, John R. Pruett, Robert T. Schultz, Lonnie Zwaigenbaum, Joseph Piven, Jed T. Elison, and IBIS Network. Cataloguing and characterizing interests in typically developing toddlers and toddlers who develop ASD. *Autism Research*, 14(8):1710–1723, 2021. doi: 10.1002/aur.2543.
- [7] Xuening Chang, Lingli Sun, and Ruizhen Li. Application of symbolic play test in identification of autism spectrum disorder without global developmental delay and developmental language disorder. *BMC Psychiatry*, 23(1):138, 2023. doi: 10.1186/s12888-023-04647-6.
- [8] Geraldine Dawson, Karen Toth, Robert Abbott, Julie Osterling, Jeff Munson, Annette Estes, and Jane Liaw. Early social attention impairments in autism: Social orienting, joint attention, and attention to distress. *Developmental Psychology*, 40(2):271–283, 2004. doi: 10.1037/0012-1649.40.2.271.
- [9] Ryan Anthony J. de Belen, Tomasz Bednarz, Arcot Sowmya, and Dennis Del Favero. Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019. *Translational Psychiatry*, 10(1):333, 2020. doi: 10.1038/s41398-020-01015-w.
- [10] Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, 1988. doi: 10.2307/2531595.
- [11] Andong Deng, Taojiannan Yang, Chen Chen, Qian Chen, Leslie Neely, and Sakiko Oyama. Language-assisted deep learning for autistic behaviors recognition. *Smart Health*, 32:100444, 2024. doi: 10.1016/j.smhl.2023.100444.

- [12] Shijian Deng, Erin E. Kosloski, Siddhi Patel, Zeke A. Barnett, Yiyang Nan, Alexander Kaplan, Sisira Aarukapalli, William T. Doan, Matthew Wang, Harsh Singh, Pamela R. Rollins, and Yapeng Tian. Hear me, see me, understand me: Audio-visual autism behavior recognition. *IEEE Transactions on Multimedia*, 27:2335–2346, 2025. doi: 10.1109/TMM.2024.3521838.
- [13] Thomas W. Frazier, Eric A. Youngstrom, Leslie Speer, Rebecca Embacher, Paul Law, John Constantino, Robert L. Findling, Antonio Y. Hardan, and Charis Eng. Validation of proposed DSM-5 criteria for autism spectrum disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, 51(1):28–40, 2012. doi: 10.1016/j.jaac.2011.09.021.
- [14] Riccardo Fusaroli, Anna Lambrechts, Dan Bang, Dermot M. Bowler, and Sebastian B. Gaigg. Is voice a marker for autism spectrum disorder? a systematic review and meta-analysis. *Autism Research*, 10(3):384–407, 2017. doi: 10.1002/aur.1678.
- [15] Michel Godel, François Robain, Fiona Journal, Nada Kojovic, Kenza Latrèche, Ghislaine Dehaene-Lambertz, and Marie Schaer. Prosodic signatures of ASD severity and developmental delay in preschoolers. *npj Digital Medicine*, 6(1):99, 2023. doi: 10.1038/s41746-023-00845-4.
- [16] I. J. Good. *Probability and the Weighing of Evidence*. Charles Griffin, London, 1950.
- [17] Google DeepMind. Gemini 3.1 pro model card. Google DeepMind model card, <https://deepmind.google/models/model-cards/gemini-3-1-pro/>, 2026. Published February 19, 2026.
- [18] Eliza Gordon-Lipkin, Jessica Foster, and Georgina Peacock. Whittling down the wait time: Exploring models to minimize the delay from initial concern to diagnosis and treatment of autism spectrum disorder. *Pediatric Clinics of North America*, 63(5):851–859, 2016. doi: 10.1016/j.pcl.2016.06.007.
- [19] Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Xin Jin, Zhenguo Li, James T. Kwok, and Yu Zhang. Reasoning-aligned perception decoupling for scalable multi-modal reasoning. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2506.04559.
- [20] Tanaya Guha, Zhaojun Yang, Ruth B. Grossman, and Shrikanth S. Narayanan. A computational study of expressive facial dynamics in children with autism. *IEEE Transactions on Affective Computing*, 9(1):14–20, 2018. doi: 10.1109/TAFFC.2016.2578316.
- [21] Jordan Hashemi, Thiago Vallin Spina, Mariano Tepper, Amy Esler, Vassilios Morellas, Nikolaos Papanikolopoulos, and Guillermo Sapiro. A computer vision approach for the assessment of autism-related behavioral markers. In *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 1–7, 2012. doi: 10.1109/DevLrn.2012.6400865.
- [22] Horace He and Thinking Machines Lab. Defeating nondeterminism in LLM inference. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20250910.
- [23] Mohammadmahdi Honarmand, Parnian Azizian, Aaron Kline, Kae Nurge, Zerin Nasrin Tumpa, Saimourya Surabhi, et al. Fine-tuning a multimodal large language model for clinician-grade autism behavioral scoring from short home videos. arXiv preprint arXiv:2606.27484, 2026.
- [24] Chuanbo Hu, Wenqi Li, Mindi Ruan, Xiangxu Yu, Shalaka Deshpande, Lynn K. Paul, Shuo Wang, and Xin Li. Exploiting large language models for diagnosing autism associated language disorders and identifying distinct features. *npj Digital Medicine*, 8:763, 2025. doi: 10.1038/s41746-025-02133-9.
- [25] Chuanbo Hu, Jacob Thrasher, Wenqi Li, Mindi Ruan, Xiangxu Yu, Lynn K. Paul, Shuo Wang, and Xin Li. Speech pattern disorders in verbally fluent individuals with autism spectrum disorder: a machine learning analysis. *Frontiers in Neuroinformatics*, 19:1647194, 2025. doi: 10.3389/fninf.2025.1647194.
- [26] Roman Jaeschke, Gordon H. Guyatt, and David L. Sackett. Users’ guides to the medical literature. III. how to use an article about a diagnostic test. B. what are the results and will they help me in caring for my patients? *JAMA*, 271(9):703–707, 1994. doi: 10.1001/jama.1994.03510330081039.

- [27] Warren Jones and Ami Klin. Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature*, 504(7480):427–431, 2013. doi: 10.1038/nature12715.
- [28] Dong Yeong Kim, Ryemi Do, Youmin Shin, Hewoen Sim, Hanna Kim, Sungchul Cho, et al. Automated AI based identification of autism spectrum disorder from home videos. *npj Digital Medicine*, 8:607, 2025. doi: 10.1038/s41746-025-01993-5.
- [29] So Hyun Kim and Catherine Lord. Restricted and repetitive behaviors in toddlers and preschoolers with autism spectrum disorders based on the autism diagnostic observation schedule (ADOS). *Autism Research*, 3(4):162–173, 2010. doi: 10.1002/aur.142.
- [30] Ami Klin, Warren Jones, Robert Schultz, Fred Volkmar, and Donald Cohen. Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry*, 59(9):809–816, 2002. doi: 10.1001/archpsyc.59.9.809.
- [31] Aditya Kommineni, Digbalay Bose, Tiantian Feng, So Hyun Kim, Helen Tager-Flusberg, Somer Bishop, Catherine Lord, Sudarsana Kadiri, and Shrikanth Narayanan. Towards child-inclusive clinical video understanding for autism spectrum disorder. arXiv preprint arXiv:2409.13606, 2024.
- [32] Pradeep Raj Krishnappa Babu, J. Matias Di Martino, Rachel Aiello, Brian Eichner, Steven Espinosa, Jennifer Green, Jill Howard, Sam Perochon, Marina Spanos, Saritha Vermeer, Geraldine Dawson, and Guillermo Sapiro. Validation of a mobile app for remote autism screening in toddlers. *NEJM AI*, 1(10), 2024. doi: 10.1056/AIcs2400510.
- [33] Susan Landon, Thomas Savage, S. Ryan Greysen, and Eric Bressman. Variation in large language model recommendations in challenging inpatient management scenarios. *Journal of General Internal Medicine*, 41(5):1286–1291, 2026. doi: 10.1007/s11606-025-09888-7.
- [34] Robert S. Ledley and Lee B. Lusted. Reasoning foundations of medical diagnosis. *Science*, 130(3366):9–21, 1959. doi: 10.1126/science.130.3366.9.
- [35] Chaoyu Li, Eun Woo Im, and Pooyan Fazli. VidHalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13723–13733, 2025. doi: 10.1109/CVPR52734.2025.01281.
- [36] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, pages 28541–28564, 2023.
- [37] Wenxing Liu, Yueran Pan, Dong Zhang, Hongzhu Deng, Xiaobing Zou, and Ming Li. Detecting children with autism spectrum disorder based on script-centric behavior understanding with emotional enhancement. *Neurocomputing*, 670:132434, 2026. doi: 10.1016/j.neucom.2025.132434.
- [38] Catherine Lord, Michael Rutter, Pamela C. DiLavore, Susan Risi, Katherine Gotham, and Somer Bishop. *Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) Manual (Part I): Modules 1–4*. Western Psychological Services, Torrance, CA, 2012.
- [39] Yiqing Lyu, Xianbing Zhao, Buzhou Tang, and Ronghuan Jiang. Dep-LLM: Training-free depression diagnosis via evidence-guided structured multi-factor with reliable LLM reasoning. arXiv preprint arXiv:2606.10796, 2026.
- [40] Jiahao Meng, Tan Yue, Qi Xu, Haochen Wang, Zhongwei Ren, Weisong Liu, et al. VideoZeroBench: Probing the limits of video MLLMs with spatio-temporal evidence verification. arXiv preprint arXiv:2604.01569, 2026.
- [41] Meghan Miller, Ana-Maria Iosif, Monique Hill, Gregory S. Young, A. J. Schwichtenberg, and Sally Ozonoff. Response to name in infants developing autism spectrum disorder: A prospective study. *The Journal of Pediatrics*, 183:141–146.e1, 2017. doi: 10.1016/j.jpeds.2016.12.071.

- [42] Maite Montagut-Asunción, Sarah Crespo-Martín, Gemma Pastor-Cerezuela, and Ana D’Ocon-Giménez. Joint attention and its relationship with autism risk markers at 18 months of age. *Children*, 9(4):556, 2022. doi: 10.3390/children9040556.
- [43] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023. doi: 10.1038/s41586-023-05881-4.
- [44] Caitlin McMahon Nichols, Lisa V. Ibañez, Jennifer H. Foss-Feig, and Wendy L. Stone. Social smiling and its components in high-risk infant siblings without later ASD symptomatology. *Journal of Autism and Developmental Disorders*, 44(4):894–902, 2014. doi: 10.1007/s10803-013-1944-2.
- [45] Elena Maria Riboldi, Elena Capelli, Chiara Cantiani, Carolina Beretta, Massimo Molteni, and Valentina Riva. Differentiating early sensory profiles in toddlers at elevated likelihood of autism and association with later clinical outcome and diagnosis. *Autism*, 28(7):1654–1666, 2024. doi: 10.1177/13623613231200081.
- [46] Jennifer Richler, Somer L. Bishop, Jennifer R. Kleinke, and Catherine Lord. Restricted and repetitive behaviors in young children with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 37(1):73–85, 2007. doi: 10.1007/s10803-006-0332-6.
- [47] Diana L. Robins, Karís Casagrande, Marianne Barton, Chi-Ming A. Chen, Thyde Dumont-Mathieu, and Deborah Fein. Validation of the modified checklist for autism in toddlers, revised with follow-up (M-CHAT-R/F). *Pediatrics*, 133(1):37–45, 2014. doi: 10.1542/peds.2013-1813.
- [48] Mindi Ruan, Paula J. Webster, Xin Li, and Shuo Wang. Deep neural network reveals the world of autism from a first-person perspective. *Autism Research*, 14(2):333–342, 2021. doi: 10.1002/aur.2376.
- [49] Mindi Ruan, Xiangxu Yu, Na Zhang, Chuanbo Hu, Shuo Wang, and Xin Li. Video-based contrastive learning on decision trees: from action recognition to autism diagnosis. In *Proceedings of the 14th ACM Multimedia Systems Conference (MMSys ’23)*, pages 289–300, 2023. doi: 10.1145/3587819.3590988.
- [50] Mindi Ruan, Na Zhang, Xiangxu Yu, Wenqi Li, Chuanbo Hu, Paula J. Webster, Lynn K. Paul, Shuo Wang, and Xin Li. Can micro-expressions be used as a biomarker for autism spectrum disorder? *Frontiers in Neuroinformatics*, 18:1435091, 2024. doi: 10.3389/fninf.2024.1435091.
- [51] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x.
- [52] Weiyan Shi and Kenny Tsu Wei Choo. Towards aligning multimodal LLMs with human experts: A focus on parent–child interaction. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2026. doi: 10.1145/3772318.3791267.
- [53] Cathy Shyr, Boyu Ren, Chih-Yuan Hsu, Chao Yan, Rory J. Tinker, Thomas A. Cassini, et al. A statistical framework for evaluating the repeatability and reproducibility of large language models. medRxiv preprint, 2025.
- [54] David J. Spiegelhalter and Robin P. Knill-Jones. Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *Journal of the Royal Statistical Society. Series A (General)*, 147(1):35–77, 1984. doi: 10.2307/2981737.
- [55] Xiaoyu Tan, Bin Li, Weidi Xu, Chao Qu, Wei Chu, Yinghui Xu, Yuan Qi, and Xihe Qiu. Prolog-driven rule-based diagnostics with large language models for precise clinical decision support. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15969 of *LNCS*, pages 413–423, 2025. doi: 10.1007/978-3-032-05127-1_40.
- [56] Jialu Tang, Hung Manh Pham, Ignace De Lathauwer, Henk S. Schipper, Yuan Lu, Dong Ma, and Aaqib Saeed. Interpretable multimodal zero-shot ECG diagnosis via structured clinical knowledge alignment. *npj Cardiovascular Health*, 3:1, 2026. doi: 10.1038/s44325-025-00099-x.

- [57] Qandeel Tariq, Jena Daniels, Jessey Nicole Schwartz, Peter Washington, Haik Kalantarian, and Dennis P. Wall. Mobile detection of autism through machine learning on home video: A development and prospective validation study. *PLOS Medicine*, 15(11):e1002705, 2018. doi: 10.1371/journal.pmed.1002705.
- [58] Cristina Tecar, Lacramioara Eliza Chiperi, Bianca-Elena Iftimie, Livia Livint-Popa, Emanuel Stefanescu, Maria Lucia Sur, Nicu Catalin Draghici, and Dafin Fior Muresanu. Eye-tracking as a screening tool in the early diagnosis of autism spectrum disorder: A systematic review and meta-analysis. *Journal of Clinical Medicine*, 14(24):8801, 2025. doi: 10.3390/jcm14248801.
- [59] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.
- [60] Bowen Xu, Xinyue Zeng, Jiazhen Hu, Tuo Wang, and Adithya Kulkarni. DiagnoLLM: A hybrid Bayesian neural language framework for interpretable disease diagnosis. arXiv preprint arXiv:2511.05810, 2025.
- [61] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ -bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.12045.
- [62] W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.
- [63] Xiangxu Yu, Mindi Ruan, Chuanbo Hu, Wenqi Li, Lynn K. Paul, Xin Li, and Shuo Wang. Video-based analysis reveals atypical social gaze in people with autism spectrum disorder. arXiv preprint arXiv:2409.00664, 2024.
- [64] Na Zhang, Mindi Ruan, Shuo Wang, Lynn K. Paul, and Xin Li. Discriminative few shot learning of facial dynamics in interview videos for autism trait classification. *IEEE Transactions on Affective Computing*, 14(2):1110–1124, 2023. doi: 10.1109/TAFFC.2022.3178946.
- [65] Kun Zhao, Guodong Liu, Hui Ji, Siyuan Dai, Pan Wang, Jifeng Song, et al. HERO: Hierarchical evidential reasoning optimization for radiology report generation via reason-then-summarize. arXiv preprint arXiv:2601.03321, 2026.
- [66] Feng-Lei Zhu, Shi-Huan Wang, Wen-Bo Liu, Hui-Lin Zhu, Ming Li, and Xiao-Bing Zou. A multimodal machine learning system in early screening for toddlers with autism spectrum disorders based on the response to name. *Frontiers in Psychiatry*, 14:1039293, 2023. doi: 10.3389/fpsy.2023.1039293.

Appendix

A Cohort Composition

Table A1: Composition of the home-video corpus. Ages are the children’s ages at recording. The four demonstration clips supply the in-context exemplars (Section 3.3) and are excluded from every reported metric.

Group	N	Age, mean $\pm$ s.d. (months)	Age range (months)
ASD (evaluation)	12	40.9 $\pm$ 6.6	30–48
TD (evaluation)	31	36.6 $\pm$ 6.1	26–47
Demonstration clips (2 ASD, 2 TD)	4	35.8 $\pm$ 5.2	32–43

B Scorer: Derivation and Design Rationale

Derivation. Let $y \in \{\text{ASD}, \text{TD}\}$ and let E be the set of glossary features observed in a clip. Under conditional independence of the features given y , Bayes’ rule gives

$$\ln \frac{P(\text{ASD} | E)}{P(\text{TD} | E)} = \ln \frac{P(\text{ASD})}{P(\text{TD})} + \sum_{f \in E} \ln \frac{\text{sens}_f}{1 - \text{spec}_f} + \sum_{f \notin E} \ln \frac{1 - \text{sens}_f}{\text{spec}_f}.$$

The scorer keeps only the second term, the sum of positive log-likelihood ratios $w_f = \ln \text{LR}_f^+$ over observed features [16, 26]. The third term is dropped: in a 3–6 minute unprompted segment a behavior may never have the opportunity to occur, so absence is treated as uninformative and the total is a lower bound on the evidence for ASD. Bayesian combination of evidence is the oldest formalism of computer-aided diagnosis [34]; the additive weights-of-evidence form used here follows [54].

Confidence. Events coded possible (occluded, brief, or downgraded by age calibration) contribute a quarter of their weight, a fixed discount for uncertain observations; each feature counts once, at its highest observed confidence, so repeated events cannot accumulate.

Domain saturation. Social-communication features co-occur because they express one underlying deficit, so their conditional independence fails and summing their weights double-counts; the social-communication total is therefore capped at $\kappa_{\text{SC}} = 1.5$ nats, about one strong deficit ($\text{LR}^+ \approx 4.5$). Restricted and repetitive behaviors are near-independent of each other and are summed in full.

Provenance. The eleven sensitivity/specificity pairs (Table A2) are author estimates, fixed before any reported run and not fitted to this cohort. One pair is taken directly from a published diagnostic study; feature 6 uses the symbolic play test’s reported 0.72/0.62 at its optimal cut-off, estimated against a developmental-language-disorder comparison group rather than typically developing children [7]. The remaining pairs were set from published effect sizes, prevalence contrasts, and classifier accuracies for the corresponding behavior; feature 7, for instance, from a prevalence contrast of 96–100% (autism) against 25–63% (typically developing, assessed up to 30 months) for at least one restricted or repetitive behavior in brief semi-structured observation [29]. Feature 4 has no single anchoring study. The Basis column names the evidence consulted for each pair. Insistence on sameness (feature 8) is the least secure pair: it is uncommon before age three and did not separate groups in a toddler cohort [46]. The cap and the two cut-offs (low risk < 1.5 , intermediate 1.5–1.8, high ≥ 1.8 nats) are cohort-derived constants: they were set on an earlier run over this cohort, and Appendix B.1 estimates the optimism of the cap and the screen-positive cut-off by leave-one-out re-selection.

B.1 Leave-one-out re-selection of the cap and the cut-offs

The cap and the cut-offs were fixed on an earlier run over this cohort, before the three reported runs: the screen-positive cut-off at the social-communication cap (1.5 nats) and the high-risk cut-off at 1.8 nats, just under the weight of a single certain stereotypy ($w_7 = 1.838$). We quantify the optimism of this choice on the saved tables, without a model call, in three steps (Table A3). 1) Re-selecting a single cut-off by the Youden index [62] at the fixed cap returns 1.838 in 42 of 43 folds on the first run, 1.5 in 42 of 43 folds on the third, and alternates between 1.838 and 0.33 on the second: with 12

Table A2: Per-feature weights of the deterministic scorer. Sensitivity and specificity are the assumed discriminative strength of the feature when present; $w_f = \ln(\text{sens}_f / (1 - \text{spec}_f))$ in nats; the basis column names the type of evidence consulted for each pair.

#	Feature	Domain	sens / spec	w_f	Basis
1	Unusual / avoidant eye contact	SC	0.85 / 0.80	1.447	eye-tracking meta-analysis [58]
2	Limited facial expressions & social smiling	SC	0.78 / 0.70	0.956	high-risk sibling study; specificity limited [44]
3	Reduced gestures & joint attention	SC	0.87 / 0.82	1.576	joint-attention discriminability [8]; early joint-attention deficits predict 18-month screen risk [42]
4	Preference for solitude / limited peer interest	SC	0.80 / 0.79	1.338	author estimate; multifactorial
5	Impaired social reciprocity	SC	0.88 / 0.84	1.705	response-to-name marker studies [66, 41]
6	Lack of imaginative / pretend play	SC	0.72 / 0.62	0.639	published pair, symbolic play test, ASD vs. language disorder [7]
7	Stereotyped motor movements	RRB	0.88 / 0.86	1.838	RRB prevalence contrast, ASD vs. TD [29]
8	Inflexible adherence to routines	RRB	0.87 / 0.85	1.758	DSM-5 criterion-set specificity [13, 4]
9	Highly restricted / fixated interests	RRB	0.83 / 0.82	1.528	unusual-content specificity [6]; repetitive-interest prevalence contrast [29]
10	Echolalia & atypical speech prosody	RRB	0.81 / 0.75	1.176	acoustic meta-analysis, ASD vs. TD [14]; prosody tracks severity within ASD [15]
11	Sensory hyper-/hypo-reactivity & unusual sensory interests	RRB	0.85 / 0.84	1.670	sensory-symptom prevalence, ~90% ASD vs. 5–15% non-ASD, as reviewed in [45]

ASD clips the Youden point is not stable across runs, and the choice between the two levels turns on four clip-runs, two TD clips that stop exactly at the cap in one run each, one TD clip at 1.69 nats, and one ASD clip that reaches the cap in one run. Counting only high risk as screen-positive gives pooled recall 61.1% and specificity 97.8% over the three runs, against 63.9% and 94.6% at the deployed cut-off. 2) Re-selecting the cap (0.5 nats to no cap) and the cut-off jointly in every leave-one-clip-out fold, pooled over the three runs, gives recall 58.3–61.1% and specificity 91.4–94.6% under three selection criteria; the selected cap ranges from 0.5 to 1.75 nats across folds and criteria. 3) Because the screen-positive cut-off equals the cap, the cap cannot change a screen-positive decision; removing it moves two TD clip-runs from intermediate to high risk and the AUC by at most 0.008 (0.863 to 0.855, 0.839 to 0.839, 0.851 to 0.853); the AUC of the first run is 0.863 with DeLong 95% CI [10] [0.72, 1.00].

Table A3: Optimism of the fixed decision rule. Confusion counts, recall, specificity, and accuracy (%) are pooled over the 43 clips of the three runs (129 clip-runs). The leave-one-out rows re-select the cap and the cut-off on 42 clips in every fold, under the named criterion, and apply them to the held-out clip.

Rule	Selection	TP/FN/TN/FP	Recall	Spec.	Acc.
Deployed rule (1.5)	set on an earlier run	23/13/88/5	63.9	94.6	86.0
High risk only (1.8)	set on an earlier run	22/14/91/2	61.1	97.8	87.6
Re-selected by leave-one-out	Youden index	21/15/86/7	58.3	92.5	82.9
	maximum F_1	21/15/88/5	58.3	94.6	84.5
	recall first, spec. $\geq 90\%$	22/14/85/8	61.1	91.4	82.9

C Cross-Seed Flip Rates by Configuration

Table A4 lists the configurations behind the flip range quoted in Section 4.4.

Table A4: Cross-seed flip rates on Gemini 3.1 Pro (three repeated runs at temperature 0). *Flip* is the fraction of the 43 clips whose label differs across the runs; *Verdict* as in Table 1.

Configuration	Verdict	Flip (%)
Zero-shot baseline	VLM	16.3
+ Few-shot exemplars	VLM	18.6
Single call, multi-turn exemplars, no age calibration	VLM	37.2
+ Age calibration	VLM	23.3
+ Two-stage split	LLM	20.9
Two-stage split, read by the scorer	scorer	27.9
+ Grounded perception constraints	LLM	16.3
Full pipeline	scorer	18.6

D Run-to-run Stability of the Single-Call Baseline

Per-run detail of the five-run experiment summarized in Section 4.4: the single-call baseline (Gemini 2.5 Pro, full prompt and four in-context videos, VLM verdict) repeated five times as the identical request at temperature 0 with a shared context cache. In three runs the API returned an empty response for the same clip; it counts as an error, as in Section 4.1.

Table A5: Run-to-run stability of the single-call baseline (Gemini 2.5 Pro, full prompt, four in-context videos, VLM verdict) over five repeated runs of the identical request at temperature 0 (43 clips; an empty API response counts as an error).

Run	Accuracy (%)	Recall (%)	Specificity (%)	F ₁ (%)
run 1	72.1	58.3	77.4	53.8
run 2	69.8	58.3	74.2	51.9
run 3	76.7	75.0	77.4	64.3
run 4	69.8	66.7	71.0	55.2
run 5	76.7	75.0	77.4	64.3
mean $\pm$ s.d.	73.0 $\pm$ 3.5	66.7 $\pm$ 8.3	75.5 $\pm$ 2.9	57.9 $\pm$ 6.0

E Assets: Glossary, Prompt Structure, Norms, and Demonstrations

E.1 Prompt structure

Fig. A1 summarizes the two prompts block by block; the rules are the prompts’ own, shortened.

E.2 Behavioral glossary

The eleven features of Table A6 condense a 22-behavior, nine-task home-video protocol that the clinical team designed in 2020 and mapped item by item to ADOS-2 algorithm codes and M-CHAT-R questions; we kept the behaviors codeable from unprompted free play.

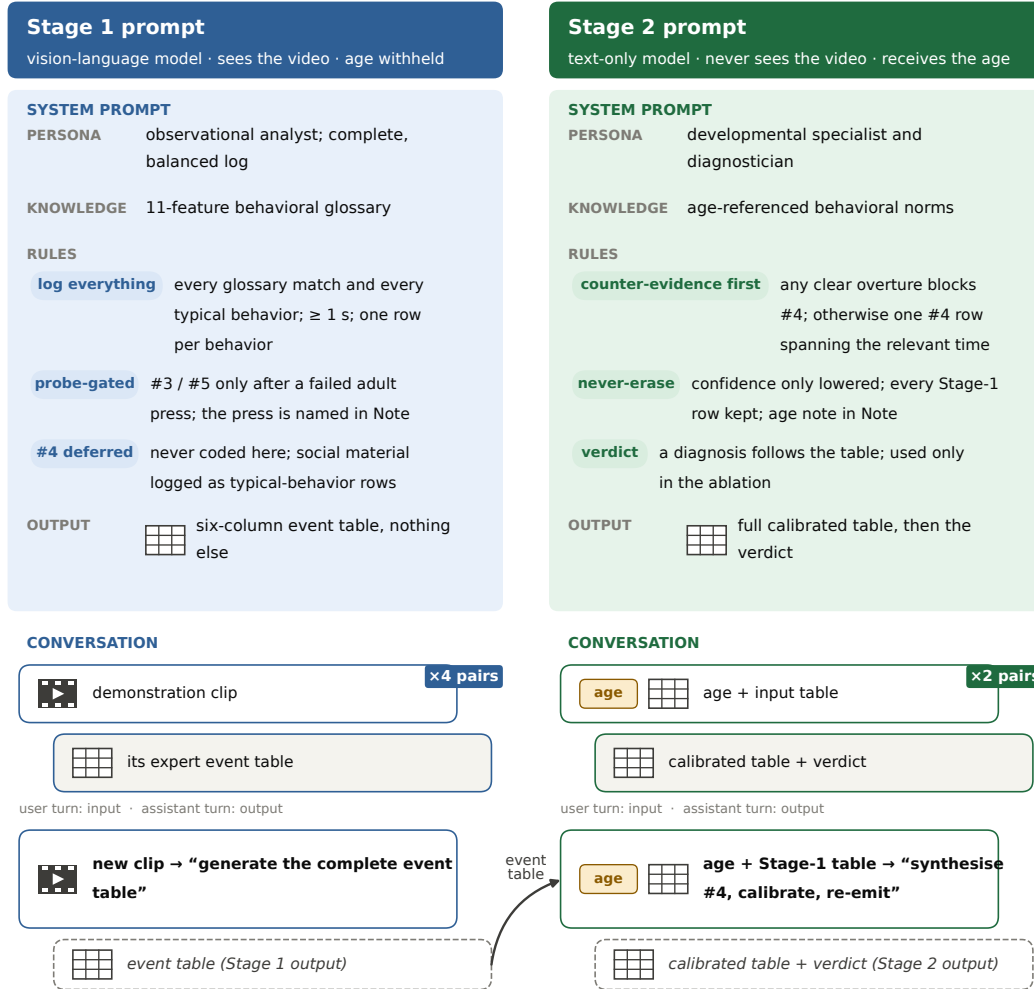


Figure A1: Anatomy of the Stage 1 and Stage 2 prompts. Top: the system prompt of each stage (persona, knowledge, rules, output mandate); the knowledge blocks are the glossary of Table A6 and the norms of Fig. A2. Bottom: the conversation the model sees, with demonstrations as prior user/assistant turn pairs (four corpus clips with their expert tables for Stage 1; the two hand-written exchanges of Fig. A3 for Stage 2) and the final user turn; the dashed bubble is the stage’s output, and Stage 1’s event table becomes the content of Stage 2’s final turn.

Table A6: The 11-item behavioral glossary: feature (with ADOS-2 / DSM-5 mapping where applicable), behavioral description, typical contexts, and an example. It constrains both expert annotation and model prompting.

#	Feature (merged diagnostic items)	Expanded behavioral description	Typical situations / contexts	Representative examples
1	Unusual / Avoidant Eye Contact (ADOS-2 SA “Unusual Eye Contact”, DSM-5 A2 component)	Limited, fleeting or sidelong gaze; may avoid looking at faces even while interacting.	Feeding, play, greetings, clinical exams.	<i>Infant does not look eyes during feeding; toddler continues playing when a visitor enters and does not look up despite name being called.</i>
2	Limited Facial Expressions & Social Smiling	Flat or restricted affect; smiles and emotive expressions are infrequent, delayed, or context-inappropriate.	Peek-a-boo, praise moments, gift giving, celebrations.	<i>Child watches peek-a-boo with neutral face; preschooler receives a gift with minimal facial change.</i>
3	Reduced Gestures & Joint Attention	Pointing, showing, waving, and nodding are rare; may not coordinate gaze with gesture or follow others’ pointing.	Requesting help, picture-book reading, playground pointing.	<i>Child reaches toward a toy but does not point or look to a caregiver to share the request.</i>
4	Preference for Solitude / Limited Peer Interest	Appears content playing alone; seldom approaches peers or adults to share enjoyment or comfort.	Day-care free-play, playdates, family gatherings.	<i>Child ignores a peer offering a toy and continues solitary play.</i>
5	Impaired Social Reciprocity	Reduced back-and-forth interaction; may not respond to name or may show limited verbal turn-taking.	Name-call tests, greetings, Q&A exchanges.	<i>Child does not orient when name is called after multiple attempts.</i>
6	Lack of Imaginative / Pretend Play	Symbolic play is absent or limited; play remains concrete, repetitive, or focused on object properties.	Doll/kitchen sets, toy-vehicle play, cooperative make-believe.	<i>Child spins toy-car wheels and does not use the car in pretend scenarios.</i>
7	Stereotyped Motor Movements (DSM-5 B1; colloquially “stimming”)	Stereotyped motor acts such as hand-flapping, rocking, or spinning; toe walking is noted separately as an associated motor sign.	Idle moments, excitement peaks, transitions.	<i>Child flaps hands repeatedly while watching a preferred stimulus.</i>
8	Inflexible Adherence to Routines	Distress with small changes to routines, objects, routes, or expected sequences.	Morning routines, classroom agendas, travel routes.	<i>Child refuses a meal when a preferred item or arrangement is changed.</i>
9	Highly Restricted / Fixated Interests	Narrow interests of unusually high intensity that dominate play or attention.	Free-choice play, conversations, preferred media.	<i>Child focuses on one object category and rejects alternative toys.</i>
10	Echolalia & Atypical Speech Prosody	Repetition of words or phrases; scripted speech; flat, sing-song, or atypical intonation.	Adult questioning, solitary play scripting.	<i>Child repeats a question verbatim rather than answering it.</i>
11	Sensory Hyper-/ Hyporeactivity & Unusual Sensory Interests	Over- or under-response to sensory input; unusual fascination with sensory aspects of objects.	Noise/lights, swings, mealtimes, medical visits.	<i>Child covers ears and cries in response to common sounds; or repeatedly seeks spinning visual stimuli.</i>

E.3 Age-referenced norms

For every feature row, Stage 2 checks the behavior against the norm for the child’s age. A behavior that is age-typical keeps its feature number, drops to **possible**, and receives an age note; **certain** is reserved for behaviors that clearly exceed the age expectation in intensity, duration, or pervasiveness. Pervasiveness overrides the exemptions: the age-typical conditions apply only to occasional, isolated instances, and a behavior that recurs or dominates the clip stays **certain** even when each instance would be exemptible. Fig. A2 places the eleven features on the age axis.

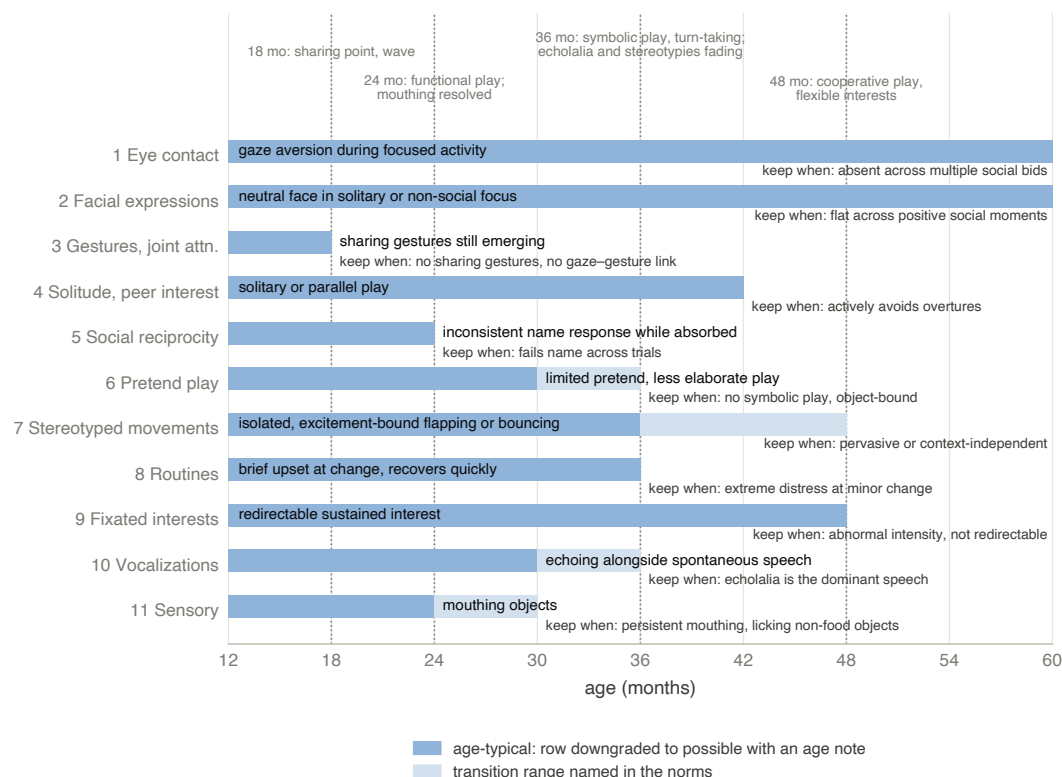


Figure A2: The age-referenced norms of Stage 2, read from the norms text. For each glossary feature the bar marks the ages at which the named behavior is age-typical, so the row is downgraded to **possible** with an age note; the light segment is a transition range the norms name; the line under the bar is the condition under which the feature is kept. Dotted lines are the developmental anchors the norms list. Pervasiveness overrides the window at any age.

E.4 Stage 2 demonstrations

Two hand-written exchanges precede the clip’s turn, each a user turn (the user prompt with the age and the input table) followed by an assistant turn (the output). Fig. A3 shows the two outputs with Stage 2’s changes marked; the input tables are identical except for those changes.

A Demonstration 1 · 30 months				
a clear social overture is present, so no #4 row is added				
Start-End Behavior	Comment	Note	Conf.	
0:10-0:18	Plays alone lining up blocks while mother sits nearby.	Typical behavior: solitary play.	certain	
1:05-1:08	Mother calls his name; he glances up briefly then returns to the blocks.	5 Response to Name: partial / delayed response.	possible	age-typical: kept, downgraded to possible, age note added
		name called → brief glance; <i>Age-note: a brief delayed glance during absorbed play is plausibly typical at 30 months.</i>		
2:30-2:40	Brings a block to his mother, looks at her, and smiles to show it.	Typical behavior: shares an item with eye contact.	certain	clear overtture: #4 is not synthesised
3:00-3:05	Flaps hands while excited by a spinning toy.	7 Stimming: hand-flapping.	possible	age-typical: kept, downgraded to possible, age note added
		<i>Age-note: brief excitement-bound flapping is common at this age.</i>		
Diagnosis: TD (confidence possible)				

B Demonstration 2 · 42 months				
no overtture anywhere, so one clip-spanning #4 row is added				
Start-End Behavior	Comment	Note	Conf.	
0:05-0:40	Plays alone with a toy car, repeatedly spinning its wheels, while both parents try to engage him.	Typical behavior: solitary play.	certain	
1:10-1:14	Father holds up a bear and says "look!"; the child does not look.	3 Reduced Joint Attention: fails to follow the show bid.	certain	press note kept; age note appended; confidence stays certain (exceeds age)
		adult shows bear → no look; <i>Age-note: response to joint attention should be well established by 42 months.</i>		
2:00-2:30	Continues spinning the wheels alone; ignores his mother offering a different toy.	Typical behavior: solitary play; ignores an offered toy.	certain	
3:00-3:20	Spins the wheels and flaps his hands repeatedly.	7 Stimming: repetitive hand-flapping.	certain	
0:05-3:20	Across the whole clip the child stays in solitary play and never responds to or initiates social contact, despite both parents' attempts.	4 Limited Peer Interest: pervasive, whole-clip limited social interest.	certain	row added by Stage 2 after the counter-evidence check found no overtture
		overtures across clip: none (ignored a show bid and an offered toy; no approach or share anywhere).		
Diagnosis: ASD (confidence certain)				

Figure A3: The two hand-written Stage 2 demonstrations, drawn as the output tables (A: 30 months; B: 42 months). Yellow cells are the age notes Stage 2 wrote into the Note column; the green outline in A marks the social overtture that blocks the pervasive feature #4; the green row in B is the clip-spanning #4 row Stage 2 added after the counter-evidence check found no overtture. Feature numbers refer to Table A6.