InterventionLens: A Multi-Agent Framework for Detecting ASD Intervention Strategies in Parent-Child Shared Reading

Anonymous CVPR submission

Paper ID 15

Abstract

Home-based interventions like parent-child shared reading provide a cost-effective approach for supporting children with autism spectrum disorder (ASD). However, analyzing caregiver intervention strategies in naturalistic home interactions typically relies on expert annotation, which is costly, time-intensive, and difficult to scale. To address this challenge, we propose InterventionLens, an end-to-end multi-agent system for automatically detecting and temporally segmenting caregiver intervention strategies from shared reading videos. Without task-specific model training or fine-tuning, InterventionLens uses a collaborative multi-agent architecture to integrate multimodal interaction content and perform fine-grained strategy analysis. Experiments on the ASD-HI dataset show that InterventionLens achieves an overall F1 score of 79.44%, outperforming the baseline by 19.72%. These results suggest that InterventionLens is a promising system for analyzing caregiver intervention strategies in home-based ASD shared reading settings. Additional resources will be released on the project page.

1. Introduction

Autism Spectrum Disorder (ASD) is a prevalent neurodevelopmental disorder that substantially impairs social communication and daily functioning [4, 8]. While early intervention can improve core symptoms [7], many families lack consistent access to effective services due to long-term costs and uneven resource distribution [17]. This gap has motivated scalable, caregiver-deliverable home-based interventions [18, 22, 26, 27], among which shared book reading (SBR) is a particularly effective and accessible routine [6, 25]. In practice, families often record SBR sessions at home and submit the videos to speech-language pathologists (SLPs), who manually conduct video-by-video review to identify caregiver intervention strategies and provide guidance [2]. However, such expert-driven annotation is labor-intensive, costly, and difficult to scale. Therefore, a key technical need is to automatically detect intervention strategies from home-based videos. In this work, we focus on three caregiver intervention strategies used in parent-implemented ASD interventions: *Modeling*, where the caregiver demonstrates the target word or phrase for the child; *Mand-Model*, where the caregiver prompts the child to respond and then provides a model when needed; and *Time Delay*, where the caregiver deliberately pauses to create an opportunity for the child to initiate or complete the response [19–21]. These strategies originate from the Parent-Implemented Communication Strategies (PiCS) framework and were later adapted to shared reading and telepractice-based interventions for children with ASD [1–3].

Achieving this goal is inherently difficult because shared book reading (SBR) is highly dynamic and interactive, typically involving a tightly coupled, triadic interaction among a caregiver, a child, and a book. Caregivers drive this process through continuous verbal scaffolding while simultaneously coordinating multi-modal social cues —such as gaze, facial expressions, and gestures —to regulate child engagement [16]. Fine-grained recognition and reasoning of these interaction segments become even more challenging in unstructured, in-the-wild home settings. In these environments, non-fixed camera viewpoints, ambient noise, and variable interaction rhythms are compounded by the inherent acoustic differences between adult and child voices [15]. These factors drastically degrade the performance of mainstream Automatic Speech Recognition (ASR) systems like Whisper [5], severely undermining the reliability of subsequent interaction modeling. Finally, these technical hurdles are further magnified by a severe scarcity of high-quality annotated data. Although the ASD-HI dataset [16] partially addresses the gap in home-based SBR, its limited scale cannot support the massive data demands of modern data-intensive model training. Consequently, developing a robust and generalizable automated analysis system under these compounding constraints remains a core open challenge in the field.

To address these limitations, we propose InterventionLens, a hierarchical multi-agent framework for caregiver in-

CVPR #15

tervention detection without parameter updating. Our contributions are three-fold:

- We propose a two-layer multi-agent architecture that decouples multimodal perception from intervention detection and segmentation, combining structured transcript and book-context modeling with coarse-to-fine expert-based temporal localization.
- We introduce a progressive knowledge refinement mechanism that uses a small amount of labeled data to iteratively update decision rules and refine the agent's structured guidance. By translating clinical standard operating procedures (SOPs) into explicit agent roles and decision constraints, our framework enables structured strategy analysis without data-intensive fine-tuning.
- We demonstrate strong empirical performance on the ASD-HI dataset, where InterventionLens substantially outperforms existing baselines under strict evaluation criteria, improving the overall F1 score from 59.72% to 79.44% and Precision from 50.21% to 82.36%.

2. Related Work

LLM-based Multi-Agent Systems: Large language model (LLM)-based multi-agent systems (MAS) have emerged as an effective paradigm for complex reasoning by decomposing problems into modular subtasks handled by specialized agents [11, 23, 29]. Compared with single-agent systems, which often suffer from context dilution and error accumulation over long interactions, MAS can better isolate subproblems, reduce interference across reasoning stages, and enable targeted optimization for heterogeneous tasks [9, 12, 14, 24, 30]. While prior MAS research has mainly focused on general planning, coding, and task-solving settings, the use of MAS for ASD intervention strategy detection in shared book reading interactions remains underexplored.

Challenges in Multimodal Perception. Fine-grained analysis of parent-child shared book reading requires accurate speech recognition and reliable temporal grounding. However, child speech remains challenging for traditional ASR systems because of its substantial acoustic variability and pronunciation instability [10, 13]. In parallel, recent multimodal large language models have shown promising video understanding ability, but they still struggle to provide precise temporal localization in long-form videos [28, 31], especially when second-level timestamps are required for segment-level analysis. These limitations make robust multimodal perception a key bottleneck for automated intervention detection in naturalistic home environments.

ASD Intervention Detection: In the domain of ASD language intervention assessment, the ASD-HI dataset [16] provides the main benchmark for caregiver intervention strategy detection in home-based shared book reading videos.The strongest publicly available baseline uses Whisper for speech transcription and a single LLM agent guided by prompts designed by ASD intervention experts for intervention identification, followed by greedy sequential scanning for temporal segmentation. This benchmark establishes an important foundation for the task; however, its overall detection performance still leaves substantial room for improvement because it relies on error-prone transcripts and rigid heuristic boundary search.

3. Problem Formulation

We formulate the task as a strategy detection problem over parent–child shared book reading videos. Given an input video, the objective is to temporally localize intervention strategies and classify each detected segment into its corresponding strategy category. Let $V \in \mathcal{V}$ denote a parent-child shared book reading video, and let $\mathcal{S} = \{\textit{Modeling}, \textit{Mand-Model}, \textit{Time Delay}\}$ be the intervention strategy space. The goal is to predict a set of temporally localized intervention strategy segments:

$$\hat{\mathcal{Y}} = \{(\hat{t}^s_m, \hat{t}^e_m, \hat{s}_m)\}_{m=1}^{M}, \quad \hat{s}_m \in \mathcal{S}, \tag{1}$$

where M is the total number of predicted interventions. For the m-th predicted segment, $\hat{t}^s_m$ and $\hat{t}^e_m$ denote its predicted start and end timestamps, respectively, and $\hat{s}_m$ represents its assigned strategy label.

InterventionLens first utilizes the OPA module to process the raw video V into grounded interaction transcript features $\mathcal{X}$. Subsequently, the BMA module is applied to extract the book context representation $\mathcal{B}$ aligned with the current interaction:

$$\mathcal{X} = \textit{OPA}(V), \qquad \mathcal{B} = \textit{BMA}(\mathcal{X}). \tag{2}$$

Next, the system employs the ICSA module to generate a set of K intervention candidates along with their routing cues based on the transcript features $\mathcal{X}$:

$$\mathcal{C} = \{(c_k, q_k)\}_{k=1}^{K} = \textit{ICSA}(\mathcal{X}), \tag{3}$$

where $c_k = (t^s_k, t^e_k)$ denotes the k-th intervention candidate segment with its initial coarse start and end times, and $q_k \in \mathcal{Q} = \{\textit{MODEL-like}, \textit{MAND-like}, \textit{TD-like}\}$ is the corresponding strategy cue used for expert routing.

Each candidate is then routed to the corresponding strategy-specific expert $\textit{Expert}_{q_k}$ for verification and boundary refinement. The expert first evaluates the candidate c_k against its predefined strategy criteria in conjunction with the book context $\mathcal{B}$ to verify and assign the definitive strategy label $\hat{s}_k \in \mathcal{S}$. For accepted candidates, it subsequently

CVPR #15

adjusts the coarse boundaries (t_k^s, t_k^e) into precise boundaries $(\hat{t}_k^s, \hat{t}_k^e)$:

$$\hat{y}_k = \mathit{Expert}_{q_k}(c_k, \mathcal{B}), \tag{4}$$

where the output is explicitly defined as $\hat{y}_k \in \{(\hat{t}_k^s, \hat{t}_k^e, \hat{s}_k), \varnothing\}$. Here, $\varnothing$ denotes that the candidate is rejected (i.e., classified as background or invalid) during the strategy-specific verification.

The final prediction set is composed of all accepted and temporally refined candidates:

$$\hat{\mathcal{Y}} = \{\hat{y}_k \mid \hat{y}_k \neq \varnothing\}_{k=1}^{K}. \tag{5}$$

4. Method

As shown in Fig. 1, InterventionLens is a two-layer multi-agent framework that decouples multimodal perception and contextual grounding from downstream intervention detection and boundary refinement. Specifically, the Perception Layer converts raw shared book reading videos into time-aligned multimodal interaction transcripts and infers structured book context from the ongoing session. Conditioned on these structured representations, the Intervention Detection and Segmentation (IDS) Layer follows a coarse-to-fine pipeline that first extracts and routes candidate intervention segments, and then performs strategy-specific verification and temporal boundary refinement.

4.1. Perception Layer

The Perception Layer consists of two components: the Omni-Perception Agent (OPA), which constructs time-aligned multimodal interaction transcripts, and the Book Modeling Agent (BMA), which reconstructs a structured book-context representation.

Omni-Perception Agent (OPA): The Omni-Perception Agent (OPA) constructs time-aligned multimodal interaction transcripts from caregiver-child Shared Book Reading (SBR) videos for downstream intervention analysis. It captures verbal exchanges between caregivers and children together with observable actions and facial expressions from both participants. We use *gemini-3-flash-preview* as the backbone multimodal audio-video understanding model to parse dialog and visible interaction cues. For temporal grounding, we follow an OCR-assisted timestamping strategy in which a red timestamp is overlaid in the bottom-right corner of each frame and used as an explicit temporal reference to predict start and end times [28]. For long recordings, each reading session is processed as a sequence of 15-second clips. In parallel, the audio stream is transcribed using *Whisper*. We then align Gemini-generated caregiver utterances with Whisper transcriptions under semantic and temporal consistency constraints. When a Gemini-parsed caregiver utterance has a semantically matched Whisper counterpart with closely aligned timestamps, we replace the Gemini-predicted temporal boundaries with the corresponding Whisper timestamps. The resulting transcript serves as perceptual input to the downstream BMA and IDS Layer.

Book Modeling Agent (BMA): Shared book reading forms a triadic interaction among the caregiver, the child, and the book. As a result, caregiver interventions are often grounded in the specific context of the book being read rather than in dialogue alone. To incorporate this contextual information, we introduce the Book Modeling Agent (BMA), which infers a Structured Book-Context Representation from the parent–child interaction transcript. Instead of reconstructing the full book content, BMA extracts contextual cues such as recurring sentence patterns and potential target words that are relevant to the current reading interaction. This inferred book context is then used as semantic grounding for downstream expert modules, enabling more accurate intervention detection by aligning caregiver utterances with the underlying reading context.

4.2. Intervention Detection & Segmentation Layer

The Intervention Detection and Segmentation Layer detects and temporally segments caregiver intervention strategies by leveraging the enriched interaction transcripts and reconstructed book context produced by the Perception Layer. It is implemented through task-specific agents whose decision logic is derived from the ASD-HI coding manual and refined on the training split. Rather than training model parameters, we translate the operational definitions of Modeling, Mand-Model, and Time Delay into reusable decision rules.

Intervention Candidate Scanner Agent (ICSA) ICSA is a module for candidate intervention extraction. Based on the caregiver-child interaction transcripts produced by the Perception Layer, it identifies interaction segments that may constitute intervention events. Specifically, ICSA extracts two clinically meaningful candidate patterns: Complete Intervention Loops, which follow a caregiver stimulus, child response, and caregiver reinforcement sequence, and Intervention Attempts, in which a caregiver provides a stimulus but the child does not respond. Based on these candidates, ICSA further assigns a coarse-grained strategy cue, such as MODEL-like, MAND-like, or TD-like, and uses it as a routing signal to dispatch each candidate to the corresponding strategy-specific expert for subsequent fine-grained verification and boundary refinement.

Strategy-Specific Experts (SSEs) The Strategy-Specific Experts (SSEs) perform strategy-specific verification and

CVPR #15

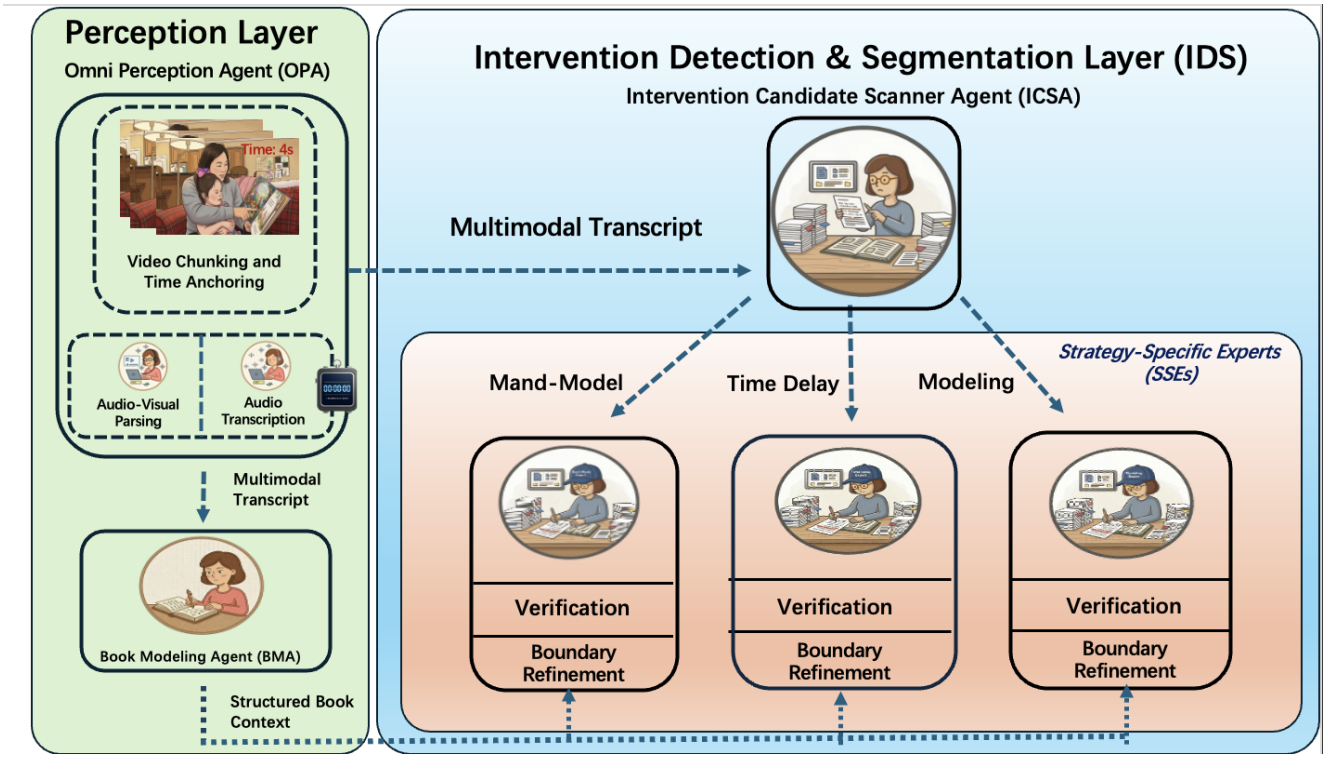

Figure 1. Overview of InterventionLens. The framework consists of a Perception Layer and an Intervention Detection & Segmentation Layer (IDS). The Perception Layer first converts the raw shared book reading video into a time-aligned multimodal transcript and structured book context. The IDS Layer then proposes candidate intervention segments and routes them to strategy-specific experts, each of which performs strategy-specific verification and boundary refinement to produce the final temporally localized intervention segments.

temporal boundary refinement for the candidate intervention segments routed from the Intervention Candidate Scanner Agent (ICSA). To account for the substantial differences in linguistic patterns and interaction structures among Modeling, Mand-Model, and Time Delay, we design SSE as a parallel strategy-specific expert architecture. This architecture consists of three independent expert systems, each dedicated to one intervention strategy. All three expert systems follow the same two-stage cascade paradigm, where Stage 1 performs strategy-specific verification and Stage 2 refines temporal boundaries.

Strategy-Specific Verification: Strategy-Specific Verification is the core logical filtering stage for candidate intervention screening. Because the candidate segments proposed by ICSA still contain noisy or weak matches, this stage verifies each candidate using strategy-specific constraints induced from the ASD-HI dataset together with book-context cues from the Book Modeling Agent (BMA). The verification considers utterance form, interaction structure, local discourse context, and book-related patterns. Only candidates consistent with the target strategy are retained for subsequent boundary refinement.

Temporal Boundary Refinement: After strategy verification, the retained candidates proceed to the temporal boundary refinement stage, where their temporal extent is localized more precisely. Guided by local dialogue flow and expert annotation conventions, this stage refines the start and end boundaries of each intervention segment. Its goal is to narrow the relatively broad candidate region from the previous stage into a precise intervention interval by removing adjacent turns before and after the core intervention that do not belong to the intervention itself. As a result, the refined segments are more closely aligned with expert annotation standards and more accurately capture the actual temporal span of the intervention.

4.3. Progressive Knowledge Refinement for SSEs

We initialize the agent's structured guidance using the ASD-HI annotation protocol, which specifies strategy definitions, temporal boundary judgment, and key interaction cues. We then iteratively refine the guidance on the training split with F1 score as the optimization objective. In each iteration, we run the full system, analyze recurrent failure cases, and update the guidance rules based on the most frequent error patterns. We stop the refinement process when the F1 score does not improve for three consecutive iterations.

CVPR #15

5. Experiments

This section presents the experimental evaluation of InterventionLens on the ASD-HI benchmark. To ensure direct comparability, we follow the same evaluation protocol as prior work. We first introduce the dataset, evaluation metrics, and implementation details. We then compare InterventionLens with the strongest reported baseline and conduct an ablation on the Book Modeling Agent (BMA). Finally, we provide additional analysis on the unguided pre-feedback subset, where caregivers had not yet received expert feedback, and discuss the main error patterns observed in the benchmark.

5.1. Experimental Setup

Dataset We evaluate InterventionLens on the ASD-HI [16] benchmark, a multimodal parent-child shared book reading dataset collected in home-based intervention settings for families of children with ASD. The dataset records natural caregiver child interactions in real home environments. It contains 48 complete shared-reading sessions from 3 families, together with expert-provided fine grained annotations of caregiver intervention segments. For each intervention segment, the dataset provides not only the corresponding intervention strategy label and its temporal boundaries, but also fine-grained annotations of the caregiver-child interaction process within the segment, including dialogue content, behavioral actions, and other interaction cues. In addition, the samples in the dataset can be further divided into two groups: those collected before caregivers received expert feedback, and those collected after caregivers received expert feedback.

Due to variations in home environments and recording conditions, camera placement is not consistent across families. Moreover, caregivers differ in prompting style, and children vary in engagement and responsiveness, resulting in substantial cross-family variability and distribution shift.The dataset provides 239 annotated segments for training and 120 annotated segments for evaluation. We note that 3 segments in the evaluation split contain incomplete annotations, specifically missing strategy labels or missing temporal boundaries. We therefore conduct our final evaluation on 117 valid test segments. All reported results follow the same evaluation split used by the original dataset authors.

Evaluation Protocol We employ Precision, Recall, and F1-score as the primary evaluation metrics. A predicted intervention is counted as a true positive only if it simultaneously satisfies three core criteria: (1) Strategy Correctness, where the predicted strategy category must exactly match the expert-annotated ground truth; (2) Strategy Completeness, where the prediction must capture the complete caregiver child interaction process underlying the intervention, including all necessary interactional details, rather than only a fragmented or isolated portion of it; and (3) Temporal Accuracy, where the predicted time interval must align with the annotated reference boundaries within a reasonable tolerance window of 1.0 second. To assess the stability of model generation, we conduct three independent runs on the same evaluation split. For each run, we compute the results for the three strategy categories Mand-Model, Modeling, and Time Delay and report the macro-averaged performance. The final results are presented as the mean $\pm$ standard deviation across the three runs.

Table 1. Overall results on ASD-HI

Method	Prec.	Rec.	F1
Baseline	50.21	73.68	59.72
Ours	82.36$\pm$3.44	76.75$\pm$0.49	79.44$\pm$1.84
Gain	+32.15	+3.07	+19.72

Implementation Details In the perception layer, we use *Whisper-1* for audio transcription and *text-embedding-3-small* to support semantic alignment between Gemini parsed caregiver utterances and Whisper transcription segments. Specifically, alignment is performed by computing the cosine distance in the embedding space. We replace the Gemini-predicted temporal boundaries with the corresponding Whisper boundaries only when the cosine distance is below 0.1 and the boundary discrepancy is within 1.0 second; otherwise, the original Gemini timestamps are retained. For multimodal understanding and downstream reasoning, all agent components use *gemini-3-flash-preview* as the backbone model, with the temperature set to 1.0.

During the Progressive Knowledge Refinement for SSEs stage, all rule induction and prompt refinement are performed on the subset collected after caregivers received expert feedback, consisting of 197 annotated segments. We use this subset because intervention behaviors at this stage are more standardized, with more stable strategy boundaries and interaction patterns, which makes it better suited for extracting reusable decision rules. In contrast, the interactions collected before expert feedback are more natural and variable, and are used in our subsequent analysis. We conduct nine rounds of refinement and stop the process when the F1 score fails to improve for three consecutive iterations, yielding a final training-set F1 score of 0.774.

5.2. Main Results

We first compare InterventionLens with the baseline reported in the original ASD-HI paper under the same evaluation protocol. Since the original ASD-HI paper reports only overall performance without strategy-wise Precision,

Table 2. Strategy wise results on ASD-HI.

Strategy	Prec.	Rec.	F1
Mand-Model	91.87±1.19	87.88±0.75	89.82±0.31
Modeling	57.34±12.72	26.67±6.67	36.28±8.25
Time Delay	65.03±6.44	72.84±5.66	68.65±5.63

Note: The evaluation contains 117 annotated intervention strategies, including 74 *Mand-Model*, 27 *Time Delay*, and 16 *Modeling* instances.

Table 3. Overall results under BMA ablation.

Method	Prec.	Rec.	F1
InterventionLens	82.36±3.44	76.75±0.49	79.44±1.84
w/o BMA	70.22±3.49	67.79±1.28	68.97±2.35
Drop	-12.14	-8.96	-10.47

Table 4. Strategy-wise F1 under BMA ablation.

Strategy	InterventionLens	w/o BMA	ΔF1
Mand-Model	89.82±0.31	80.27±2.90	-9.55
Modeling	36.28±8.25	16.96±9.06	-19.32
Time Delay	68.65±5.63	64.93±2.17	-3.72

Recall, or F1 scores, we restrict our comparison to the overall metrics. As shown in Table 1, InterventionLens substantially improves overall performance over the reported baseline [16]. Specifically, our method increases Precision from 50.21 to 82.36, Recall from 73.68 to 76.75, and F1 from 59.72 to 79.44. The most notable gain comes from the large improvement in **Precision** (**+32.15**), while **Recall** is also slightly improved (**+3.07**). Consequently, the overall **F1** score increases by (**+19.72**), indicating a substantially better balance between prediction accuracy and coverage. This suggests that InterventionLens does not achieve better performance simply by producing more aggressive predictions; instead, it substantially reduces false positives while maintaining strong coverage of true intervention segments.

Table 2 further reports the per-strategy performance of InterventionLens. Our method achieves the strongest results on *Mand-Model*, with an F1 of 89.82, suggesting that this strategy can be detected reliably under our framework. Performance on *Time Delay* remains competitive, reaching an F1 of 68.65. In contrast, *Modeling* is the most challenging category, with an F1 of 36.28. This gap suggests that different intervention strategies pose substantially different levels of difficulty, and that *Modeling* remains the major bottleneck in the current benchmark.

In addition, as a supplementary analysis, we further examine the performance of InterventionLens on the unguided interaction subset collected before caregivers received expert feedback. On this subset, the system achieves an overall F1 of 85.09±2.24, suggesting that the rules induced from post-feedback data remain applicable to more natural family shared-reading interactions. However, this result should not be over interpreted, since the evaluation subset contains only 24 annotated segments and is highly imbalanced across strategy categories, with only 15, 6, 3 instances for Mand-Model, Time Delay, and Modeling, respectively. We therefore treat this result as a supplementary observation on system applicability, rather than as a definitive validation of cross-distribution generalization.

5.3. Ablation Study

Ablation on the Book Modeling Agent (BMA). To evaluate the contribution of book-conditioned semantic grounding, we remove the Book Modeling Agent (BMA) while keeping all other components and experimental settings unchanged. As shown in Table 3, removing BMA leads to a clear overall performance drop: Precision decreases from 82.36 to 70.22, Recall decreases from 76.75 to 67.79, and F1 decreases from 79.44 to 68.97. This result indicates that explicit book modeling is a key component of the full system rather than an optional auxiliary module.

Table 4 further shows that the impact of removing BMA is not uniform across strategy categories. In particular, the F1 of Modeling drops from 36.28 to 16.96, while Mand-Model also decreases from 89.82 to 80.27. In contrast, Time Delay is relatively less affected, with F1 decreasing from 68.65 to 64.93. These results suggest that BMA contributes broadly to intervention strategy detection, and may be particularly important for strategies that require aligning caregiver utterances with book content and target-word semantics. We note that the Modeling category has relatively few instances in the evaluation split, and its category-level results should therefore be interpreted with caution; however, the clear degradation in both the overall results and Mand-Model still consistently indicates that explicit book modeling provides critical contextual grounding for the system. Overall, this ablation supports our central assumption that shared book reading should be modeled as a triadic interaction among the caregiver, the child, and the book.

5.4. Error Analysis

A notable observation is that the detection performance of *Modeling* is substantially lower than that of the other two strategy categories. We believe that this result is influenced by at least two factors. First, the number of *Modeling* instances in the evaluation split is very small, with only 16 samples, making the corresponding metrics more sensitive to fluctuations caused by a few individual cases. Second, we observe that there may be some inconsistency in labeling criteria between the training and evaluation splits. Specifically, when the same caregiver reads the same story

book, *Brown Bear, Brown Bear, What Do You See?*, utterances that are highly similar in both form and semantics are not always annotated consistently across splits. For example, in the training split, repeated target-word expressions such as "black sheep . . . black sheep" are labeled as *Modeling*, whereas in the evaluation split, similarly structured expressions from the same caregiver, such as "red bird . . . red bird", are not labeled as *Modeling*. This suggests that the *Modeling* category is challenged not only by limited sample size, but also potentially by ambiguous labeling boundaries or inconsistent annotation criteria.

6. Conclusion and Limitations

In this paper, we presented InterventionLens, a hierarchical multi-agent framework for detecting and temporally segmenting caregiver intervention strategies in home-based shared reading videos. Unlike approaches that rely on task-specific parameter updating, we decompose the task into a sequence of interconnected subprocesses, including multimodal perception, book-context modeling, intervention candidate discovery, strategy-specific verification, and temporal boundary refinement. This task decomposition allows the system to organize the overall reasoning pipeline from perception to decision-making in a modular manner, while enabling fine-grained temporal localization of intervention segments on the current benchmark.

Experiments on the ASD-HI benchmark show that InterventionLens substantially outperforms the strongest reported baseline, achieving clear gains in overall Precision and F1. Further ablation results demonstrate that explicit book-conditioned semantic grounding is a key component of the full system rather than an optional auxiliary design. Modeling shared reading as a triadic interaction among the caregiver, the child, and the book provides important contextual support for intervention detection, and is particularly helpful for strategies that depend on aligning caregiver utterances with recurring story patterns and target-word semantics.

It should also be noted that InterventionLens was developed based on the ASD-HI training set and its corresponding task formulation. As a result, the method may still exhibit a certain degree of dependence on the benchmark's annotation conventions, strategy boundaries, and data distribution. Although the experimental results show that InterventionLens achieves substantial improvements under the current evaluation protocol, its transferability and robustness across datasets, family settings, and annotation schemes still require more systematic validation.

7. Ethical Considerations

This study utilizes the ASD-HI benchmark dataset, which was collected through a prior Institutional Review Board (IRB)-approved study. The current research does not involve any new human-subject data collection; the released dataset is used strictly for academic research purposes in line with the approved data use guidelines.

For multimodal analysis, we accessed the Gemini model through Google Cloud Vertex AI. Customer data is not used to train Google's AI models as stated in their terms of use, and all API communications are protected via encryption in transit using TLS. Furthermore, we have strictly limited the use of data to research analysis, with no attempts to identify or re-identify participants, and no redistribution of raw videos or identifiable materials.

References

[1] Yusuf Akemoglu and Kimberly R Tomeny. A parent-implemented shared-reading intervention to promote communication skills of preschoolers with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 51 (8):2974–2987, 2021. 1

[2] Yusuf Akemoglu, Vanessa Hinton, Dayna Laroue, and Vanessa Jefferson. A parent-implemented shared reading intervention via telepractice. *Journal of Early Intervention*, 44 (2):190–210, 2022. 1

[3] Yusuf Akemoğlu, Dayna Laroue, Carolina Kudesey, and Mary Stahlman. A module-based telepractice intervention for parents of children with developmental disabilities. *Journal of Autism and Developmental Disorders*, 52(12):5177–5190, 2022. 1

[4] American Psychiatric Association et al. *Diagnostic and statistical manual of mental disorders*. American psychiatric association, 2013. 1

[5] Ahmed Adel Attia, Jing Liu, Wei Ai, Dorottya Demszky, and Carol Espy-Wilson. Kid-whisper: Towards bridging the performance gap in automatic speech recognition for children vs. adults. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 74–80, 2024. 1

[6] Susannah A Boyle, David McNaughton, and Shelley E Chapin. Effects of shared reading on the early language and literacy skills of children with autism spectrum disorders: A systematic review. *Focus on Autism and Other Developmental Disabilities*, 34(4):205–214, 2019. 1

[7] Annette Estes, Jeffrey Munson, Sally J Rogers, Jessica Greenson, Jamie Winter, and Geraldine Dawson. Long-term outcomes of early intervention in 6-year-old children with autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, 54(7):580–587, 2015. 1

[8] Elizabeth A Fuller and Ann P Kaiser. The effects of early intervention on social communication outcomes for children with autism spectrum disorder: A meta-analysis. *Journal of autism and developmental disorders*, 50(5):1683–1700, 2020. 1

[9] Mingyan Gao, Yanzi Li, Banruo Liu, Yifan Yu, Phillip Wang, Ching-Yu Lin, and Fan Lai. Single-agent or multi-agent systems? why not both? *arXiv preprint arXiv:2505.18286*, 2025. 2

[10] Matteo Gerosa, Diego Giuliani, and Fabio Brugnara. Acoustic variability and automatic recognition of children's speech. *Speech communication*, 49(10-11):847–860, 2007. 2

[11] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024. 2

[12] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *The twelfth international conference on learning representations*, 2023. 2

[13] Koharu Horii, Naohiro Tawara, Atsunori Ogawa, and Shoko Araki. Why is children's asr so difficult? analyzing children's phonological error patterns using ssl-based phoneme recognizers. In *Proc. Interspeech*, 2025. 2

[14] Jiangping Huang, Wenguang Ye, Weisong Sun, Jian Zhang, Mingyue Zhang, and Yang Liu. Tracecoder: A trace-driven multi-agent framework for automated debugging of llm-generated code. *arXiv preprint arXiv:2602.06875*, 2026. 2

[15] Rishabh Jain, Andrei Barcovschi, Mariam Yiwere, Peter Corcoran, and Horia Cucu. Adaptation of whisper models to child speech recognition. *arXiv preprint arXiv:2307.13008*, 2023. 1

[16] Zhaohui Li, Yusuf Akemoglu, Jincheng Lyu, Qingxiao Zheng, and Jinjun Xiong. Asd-hi: A parent-child interaction dataset for automated assessment of home intervention. In *International Conference on Artificial Intelligence in Education*, pages 48–62. Springer, 2025. 1, 2, 5, 6

[17] Yingying Lin, Guozhi Chen, Huaxiang Lu, Rongfei Qin, Jinsheng Jiang, Weiwei Tan, Caibin Luo, Ming Chen, Qin Huang, Liangliang Huang, et al. Inequality and heterogeneity in medical resources for children with autism spectrum disorders: a study in the ethnic minority region of southern china. *BMC Public Health*, 25(1):1677, 2025. 1

[18] Catherine Lord, Tony Charman, Alexandra Havdahl, Paul Carbone, Evdokia Anagnostou, Brian Boyd, Themba Carr, Petrus J De Vries, Cheryl Dissanayake, Gauri Divan, et al. The lancet commission on the future of care and clinical research in autism. *The Lancet*, 399(10321):271–334, 2022. 1

[19] Hedda Meadan, Lori E Meyer, Melinda R Snodgrass, and James W Halle. Coaching parents of young children with autism in rural areas using internet-based technologies: A pilot program. *Rural Special Education Quarterly*, 32(3): 3–10, 2013. 1

[20] Hedda Meadan, Maureen E Angell, Julia B Stoner, and Marcus E Daczewitz. Parent-implemented social-pragmatic communication intervention: A pilot study. *Focus on Autism and Other Developmental Disabilities*, 29(2):95–110, 2014.

[21] Hedda Meadan, Melinda R Snodgrass, Lori E Meyer, Kim W Fisher, Moon Y Chung, and James W Halle. Internet-based parent-implemented intervention for young children with autism: A pilot study. *Journal of Early Intervention*, 38(1):3–23, 2016. 1

[22] World Health Organization et al. Training for caregivers of children with developmental disabilities, including autism. *Retrieved January*, 17:2023, 2022. 1

[23] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. Multi-agent collaboration mechanisms: A survey of llms, 2025. *URL https://arxiv. org/abs/2501.06322*, 2025. 2

[24] Xiao Wang, Lu Dong, Sahana Rangasrinivasan, Ifeoma Nwogu, Srirangaraj Setlur, and Venugopal Govindaraju. Automisty: a multi-agent llm framework for automated code generation in the misty social robot. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9194–9201. IEEE, 2025. 2

[25] Marleen F Westerveld, Rachelle Wicks, and Jessica Paynter. Investigating the effectiveness of parent-implemented shared book reading intervention for preschoolers with asd. *Child Language Teaching and Therapy*, 37(2):149–162, 2021. 1

[26] Amy M Wetherby, Juliann Woods, Whitney Guthrie, Abigail Delehanty, Jennifer A Brown, Lindee Morgan, Renee D Holland, Christopher Schatschneider, and Catherine Lord. Changing developmental trajectories of toddlers with autism spectrum disorder: Strategies for bridging research to community practice. *Journal of Speech, Language, and Hearing Research*, 61(11):2615–2628, 2018. 1

[27] Paul Wai-Ching Wong, Yan-Yin Lam, Janet Siu-Ping Lau, Hung-Kit Fok, and WHO CST Team Servili Chiara 4 Salomone Erica 5 6 Pacione Laura 4 5 Shire Stephanie 6 Brown Felicity 7 8. Adapting and pretesting the world health organization's caregiver skills training program for children with autism and developmental disorders or delays in hong kong. *Scientific Reports*, 12(1):16932, 2022. 1

[28] Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13754–13765, 2025. 2, 3

[29] Yingxuan Yang, Qiuying Peng, Jun Wang, Ying Wen, and Weinan Zhang. Llm-based multi-agent systems: Techniques and business perspectives. *arXiv preprint arXiv:2411.14033*, 2024. 2

[30] Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Arik. Chain of agents: Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37:132208–132237, 2024. 2

[31] Zijia Zhao, Haoyu Lu, Yuqi Huo, Yifan Du, Tongtian Yue, Longteng Guo, Bingning Wang, Weipeng Chen, and Jing Liu. Needle in a video haystack: A scalable synthetic evaluator for video mllms. *arXiv preprint arXiv:2406.09367*, 2024. 2