NDD-R1: Reinforcement Learning for Multimodal LLMs to Understand Communication in Non- and Minimally-Speaking Autistic Individuals

Siddhant Bikram Shah[1] **Kristina T. Johnson**[1,2]
[1]Department of Electrical and Computer Engineering
[2]Department of Communication Sciences and Disorders
Northeastern University, Boston, USA

Abstract

Over 30% of autistic individuals are non- or minimally-speaking, relying on a complex, idiosyncratic repertoire of multimodal behaviors to communicate. Recognizing these intentional communicative acts in naturalistic environments remains a significant challenge for current multimodal large language models (MLLMs), which often struggle with the inference of subtle nonverbal cues in this domain. We introduce NDD-R1, a multi-stage reasoning framework that adapts a general-purpose MLLM for multi-view communication understanding for this population. Our approach employs a four-stage curriculum alternating between supervised finetuning and reinforcement learning ($\text{SFT}_1 \rightarrow \text{GRPO}_1 \rightarrow \text{SFT}_2 \rightarrow \text{GRPO}_2$), integrating label-guided distillation, temperature-escalated self-distillation, and difficulty-tiered reward optimization. We evaluate NDD-R1 on an initial dataset collected through the Rapid Online Sample of Communication (ROSCO) protocol, comprising 2,582 caregiver-annotated multi-view videos of 27 non- and minimally-speaking autistic individuals. NDD-R1 outperforms baselines on action and joint classification tasks, achieving competitive function classification performance relative to Gemini, exhibiting strong multi-view scaling and generalization across unseen sessions. Our code is available at https://github.com/SiddhantBikram/NDD-R1.

1 Introduction

Over 30% of individuals with Autism Spectrum Disorder (ASD) remain non- or minimally-speaking throughout their lives (Rose et al., 2016). Many of these individuals also have overlapping intellectual disabilities and rare genetic neurodevelopmental disorders (NDDs) that affect their global development. Despite limited spoken language, these individuals communicate richly and intentionally through a multimodal repertoire of gestures, vocalizations, facial expressions, gaze shifts, body movements, and Augmentative and Alternative Communication (AAC) devices (Johnson et al., 2023; Valle et al., 2021). These behaviors convey complex needs, emotional states, and social intent, yet often diverge substantially from neurotypical interaction patterns and can be highly individualized across people and contexts (Tager-Flusberg and Kasari, 2013). Consequently, meaningful communicative acts are frequently misunderstood or dismissed entirely by broader society, leaving a large and vulnerable population without essential communicative agency (Lebenhagen, 2020).

Machine learning (ML) research efforts have largely failed to address this interactional reality. Prior work in this domain has largely focused on automated diagnosis—the binary classification of individuals as autistic or neurotypical based on deficit-oriented behavior (Wei et al., 2023; Simeoli et al., 2024). While valuable for early screening, these approaches offer limited support to millions of already-diagnosed individuals whose communicative acts remain unacknowledged.

We reframe the problem around communication understanding, operationalized through two complementary multi-label classification tasks: **communicative action**—the physical modality employed (e.g., gesture, vocalization, AAC use)—and **communicative function**—the pragmatic intent conveyed (e.g., requesting, rejecting, social engagement). A single communicative act often involves simultaneous actions (e.g., vocalizing while gesturing) serving overlapping functions (e.g., requesting while expressing emotion). This annotation schema, grounded in caregiver expertise, captures this complexity with up to three concurrent action labels and two function labels per sample.

Caregivers infer communicative intent by jointly integrating subtle visual cues, auditory signals, environmental context, and interpersonal dynamics using knowledge built over hundreds of thousands of hours of lived experience. Therefore, multi-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 34681–34697
October 24-29, 2026 ©2026 Association for Computational Linguistics

modal large language models (MLLMs) (Yin et al., 2024) are a natural fit for modeling this problem since they can jointly reason over video, audio, and language, while leveraging broad world knowledge about social interaction—enabling inference beyond surface motion recognition, such as distinguishing whether a vocalization reflects self-directed behavior or social engagement when interpreted alongside gaze and body orientation.

The majority of these behaviors occur in naturalistic home environments, where occlusions, clutter, and individuals moving in and out of sight introduce substantial ambiguity. Critically, many communicative acts cannot be reliably interpreted from a single viewpoint alone: the target of a gesture may be outside the frame, gaze direction may be ambiguous, or the spatial relationship between a child and surrounding objects may be unclear. Human caregivers naturally resolve these uncertainties by integrating information across perspectives and context over time, yet most computer vision systems operate under restrictive single-view assumptions that fail to capture the complexity of real-world communication.

To address the limitations of single-view perception, we introduce a synchronized multi-view framework for communicative behavior understanding. Multi-view input provides several critical advantages in this domain: improved robustness to occlusions and partial visibility, disambiguation of gaze direction and gestural trajectories through complementary perspectives, richer environmental grounding by capturing objects and interaction targets outside a single frame, and reduced sensitivity to transient disruptions such as camera motion or individuals entering and exiting view. Rather than relying on explicit 3D reconstruction or handcrafted pose pipelines, our framework enables MLLMs to directly reason over complementary spatial observations across views. To our knowledge, this is the first work to investigate multi-view videos with MLLMs for communication understanding in non- and minimally-speaking autistic individuals.

Even with multi-view inputs, general-purpose MLLMs struggle to decode the subtle, idiosyncratic behaviors of this population due to their reliance on training data that overwhelmingly reflects neurotypical communication norms. Inspired by DeepSeek-R1 (Guo et al., 2025), we propose NDD-R1[1] (**N**euro**D**evelopmental **D**ynamics-R1), a multi-stage reasoning framework that adapts MLLMs for this domain through alternating rounds of Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) via Group Relative Policy Optimization (GRPO). First, SFT_1 grounds the model using synthetic multi-task reasoning data. Next, GRPO_1 reinforces reasoning chains that yield accurate classifications. SFT_2 then distills GRPO_1's successful rollouts acquired via a temperature-escalation cascade, utilizing SFT_1 synthetic descriptions as a fallback for unrescued clips. Finally, GRPO_2 applies difficulty-tiered rewards derived from the cascade's rescue stage, concentrating the gradient signal on the hardest samples. This curriculum effectively equips the model to jointly reason over kinematic evidence and abstract pragmatic intent, yielding competitive performance in multi-view communication understanding for non- and minimally-speaking autistic individuals.

Our contributions are summarized as follows:

- We propose NDD-R1, a four-stage framework that adapts a general-purpose MLLM to reason about and classify nonverbal communicative behaviors from non- and minimally-speaking autistic individuals in multi-view video.
- We evaluate NDD-R1 on an initial dataset collected via the Rapid Online Sample of Communication (ROSCO) protocol, comprising 2,582 caregiver-annotated multi-view video samples from 27 non- and minimally-speaking individuals with profound autism and other complex neurodevelopmental conditions.
- We provide comprehensive evaluations including multi-view ablations, zero-shot and cross-session generalization challenge sets, and per-stage contribution analysis, establishing baselines for future work in this underserved domain.

2 Related Work

RL for Multimodal Reasoning. A growing body of work has extended GRPO (Shao et al., 2024) to MLLMs through temporal contrastive rewards (Feng et al., 2026), step-wise dense rewards (Zhang et al., 2025a), and consistency-aware objectives (Chen et al., 2025), often via iterative SFT-RL pipelines with rejection sampling (Yuan et al., 2023) and difficulty-aware learning (Tong et al., 2024). NDD-R1 is methodologically aligned

[1]We refer to our framework as NDD-R1 to reflect its target population of individuals with complex neurodevelopmental conditions, while acknowledging that empirical validation is currently restricted to the syndromes in this initial ROSCO cohort.

with this family but reuses the rescue tier of a temperature-escalation rejection-sampling cascade as a per-sample difficulty multiplier for subsequent RL rewards.

MLLMs for Behavior Analysis. Recent work leverages MLLM-generated descriptions to supervise smaller skeleton- or vision-based action recognizers (Chen et al., 2026; Xiang et al., 2023; Qu et al., 2024), while adjacent tasks such as mimed action interpretation (Cho et al., 2025; Li et al., 2025) extend nonverbal video understanding more broadly. Within work targeting non- and minimally-speaking autistic individuals specifically, prior efforts address isolated sub-tasks: nonverbal vocalization recognition (Narain et al., 2022), skeleton-based action characterization (Shah and Johnson, 2026a), gaze detection (Deng et al., 2026), and single-view video (Shah and Johnson, 2026b). We build directly on the data collection paradigm and action-function taxonomy by the latter (Shah and Johnson, 2026b), which modeled a single view; our contribution is a synchronized multi-view task formulation together with a multi-stage SFT/RL curriculum designed for it. More broadly, skeleton- and pose-based representations discard affective signals such as facial expression, gaze dynamics, and vocalization, which are essential for holistically inferring communicative intent. As models for nonverbal communication understanding primarily focus on neurotypical behaviors, we establish general-purpose MLLMs as the most robust, appropriate, and reproducible backbones and baselines.

3 Dataset

We utilize an initial dataset collected via the Rapid Online Sample of Communication (ROSCO) protocol, a remote data collection framework designed to capture naturalistic communicative behaviors from non- and minimally speaking autistic individuals in the home environment. Unlike traditional laboratory-based assessments, which are resource-intensive and often struggle to capture the full range of a child's daily communicative repertoire, ROSCO brings observation into the home through remote recordings, facilitated via Zoom video conferencing. By recording in a familiar setting with primary caregivers, ROSCO minimizes behavioral discrepancies associated with clinical visits, allowing for the capture of authentic, spontaneous interactions.

3.1 Study Cohort

The cohort comprised 27 individuals (12 male, 15 female) aged 2-12 years old (mean = 6.04 years, SD = 2.75), diagnosed with Phelan-McDermid Syndrome (PMS), PTEN Hamartoma Tumor Syndrome (PHTS), or Tuberous Sclerosis Complex (TSC). Data collection yielded a total of 2,582 multi-view video samples (mean duration = 3.37 seconds, SD = 3.24) derived from 34 recording sessions across the 27 participants, with some participants contributing multiple sessions (e.g., TALK015 and TALK015b). Further details on data collection are provided in Section 3.3. All data were collected under the oversight of an Institutional Review Board (IRB). Informed consent was obtained from legal guardians, and assent was obtained from participants, following approved IRB guidelines.

3.2 Multi-View Acquisition

Communicative behaviors in this population are often subtle, rapid, and spatially distributed, simultaneously combining multiple communicative modalities. A single camera inevitably suffers from occlusion (e.g., the caregiver's body or objects in frame blocking the child's hand), unfavorable viewing angles (e.g., gaze direction unresolvable from a frontal view), and spatial ambiguity (e.g., whether a reach is directed at an object or a person).

To address these limitations, ROSCO's multi-view setup employs 3 synchronized cameras positioned to provide complementary spatial coverage of the interaction space. Two additional smartphones, shipped to families with researcher-guided setup instructions and mounted on tripods, provide complementary perspectives: (1) a room-wide view capturing posture, movement, and the spatial relationships between the child, caregiver, and surrounding objects, and (2) a closer over-the-shoulder view capturing fine-grained gestures, body movements, and AAC device interaction.

This multi-view setup provides three important advantages for communication understanding. First, it improves robustness to occlusion and partial visibility: when one camera is blocked by caregivers, motion, or clutter, alternative views often preserve visibility of the child's face, hands, posture, or interaction target. Second, complementary viewpoints help resolve ambiguities in gaze direction, head orientation, and gestural trajectories that are difficult to infer reliably from a sin-

gle perspective. Finally, multi-view input provides stronger spatial grounding for communicative behaviors such as reaching, pointing, hand-leading, or AAC interaction by revealing the relationship between the child, surrounding objects, and caregivers across perspectives.

Each video clip is represented as a sequence of K synchronized views $\mathbf{v}_i = \{v_i^{(1)}, \ldots, v_i^{(K)}\}$, where $K \in \{3, 4\}$ depending on the session configuration (one session has 4 views). The model receives all views as input and is prompted to integrate observations across camera angles. Clips are non-overlapping and disjoint: no frame recurs across clips, and no sliding window is used, so adjacent clips share no content.

3.3 Data Annotation

ROSCO employs a caregiver-centric annotation protocol to ensure the ground truth reflects the semantic intent of the child. Primary caregivers, who possess intimate knowledge of their child's idiosyncratic behaviors, annotated the video clips by rewatching the ROSCO sessions with a researcher in 10-second snippets, following a structured interview protocol. In each snippet, they were prompted to identify whether their child communicated, and if so, to specify the communicative behavior's action and function.

ROSCO's annotation schema is both multi-label and multi-class, designed to capture the complexity of real-world interaction. Each video sample was annotated for both Communicative Action (the physical modality used) and Communicative Function (the pragmatic intent). Recognizing that communicative acts are rarely isolated, the schema allows for up to three concurrent action labels and up to two function labels. When multiple communicative acts co-occur within a single clip, they are annotated as multiple concurrent Action or Function labels on that clip rather than split across neighboring clips, ensuring that a single behavior is never duplicated across the train/test split. Samples missing either action or function labels were filtered out, retaining only those with at least one label each across both axes. The dataset exhibits substantial class imbalance, reflecting the natural distribution of communicative behaviors in this population.

Actions. Samples are labeled with up to 3 of 6 actions, defined as follows:

1. **Augmentative and Alternative Communication (AAC)**: Use of high-tech (tablets, speech-generating devices) or low-tech (picture cards, communication boards) AAC systems.
2. **Body**: Use of primarily the body or head, including postural shifts (e.g., leaning in, turning away) or whole-body movements (e.g., rocking, walking away).
3. **Face**: Use of facial expressions to convey meaning (e.g., smiling, grimacing, frowning) that is not primarily a gaze shift.
4. **Gesture**: Use of the hands or arms to communicate, including pointing, reaching, waving, or hand leading.
5. **Looking**: Use of eye gaze or head orientation to communicate.
6. **Vocalization**: Use of any non-speech or speech-like sound made with the vocal tract to communicate (e.g., grunt, squeal, word approximation).

Functions. Samples are labeled with up to 2 of 6 functions, defined as follows:

1. **Commenting**: Sharing observations, expressing interest, or drawing attention to objects, events, or a person.
2. **Emotion**: Expressing an internal feeling state like happiness, frustration, or distress.
3. **Reject**: Refusing objects or discontinuation of a behavior or activity.
4. **Request**: Asking for objects or continuation of a behavior or activity.
5. **Self-Directed Behavior (SDB)**: Behaviors that were perceived as not being intentionally communicative or directed at another person, such as self-stimulatory behavior.
6. **Social**: Social interactions, such as cooperative play or expressing greetings.

3.4 Challenge Sets

We evaluate generalization along three axes, each testing a distinct capability:

Zero-Shot Generalization. All data from participant TALK009 is excluded from training, testing whether the model can classify communicative behaviors of a completely unseen individual. This session additionally has four camera views compared to the three views in other sessions, allowing us to test model generalization to additional views.

Cross-Session Generalization. Sessions TALK015b and TALK026b are held out entirely from training, while earlier sessions from the same participants (TALK015; TALK026 and TALK026c) remain in the training set. This

evaluation tracks an individual's communicative repertoire across different timepoints, and how performance scales with the number of prior sessions available (TALK026 contributes two training sessions, TALK015 contributes one).

Task Ablation. To isolate the contribution of each reasoning and description type, we restrict the supervision available during training for specific participants who showed a representative variety of actions and functions:

- TALK012: Train only on Function labels
- TALK025: Train only on Action labels
- TALK028: Train only on Joint labels

For the remaining participants, we stratified an 80/20 train-test split by balancing action and function distributions across splits.

4 Methodology

We present NDD-R1, a four-stage framework for interpreting communicative behaviors of non- and minimally-speaking autistic individuals from multi-view video. Fig. 1 illustrates the overall framework.

4.1 SFT_1: Label-Guided Domain Adaptation

A cold-start SFT phase before RL seeds the model with well-formatted reasoning patterns and prevents readability degradation (Guo et al., 2025). We ground the general-purpose MLLM π_θ (Qwen3-Omni) to the communicative patterns of non- and minimally-speaking autistic individuals by using synthetic, label-guided reasoning traces. Qwen3-Omni's architecture interleaves multiple videos sequentially within the 1D text token sequence. For each training sample consisting of multi-view video v_i and ground-truth label y_i, we query a high-capability reference model $\mathcal{M}_{\text{ref}}$ (Gemini-3-Flash). Instead of asking $\mathcal{M}_{\text{ref}}$ to discover the answer blindly—risking hallucinations in this atypical domain (Feng et al., 2024)—we condition the prompt on the known label y_i (Yoo et al., 2025). This forces the reference model to articulate its thought process t_i that explicitly justifies y_i from observable evidence in $\mathbf{v}_i$:

$$\mathrm{t}_i \sim \mathcal{M}_{\text{ref}}(\cdot \mid \mathbf{v}_i, \mathrm{y}_i, \mathcal{P}_{\text{label-guided}}) \tag{1}$$

The resulting trace is formatted as r_i = `<think>` t_i `</think><answer>` y_i `</answer>` and used as the supervised target for π_θ.

To prevent the model from overfitting to imbalanced dataset distributions, we apply a square-root softened oversampling strategy. For class c with frequency n_c and majority-class frequency $n_{\max}$, the per-sample repetition factor, ρ_c, is

$$\rho_c = \min\Big(2,\ \max\Big(1,\ \text{round}\Big(\tfrac{1}{2}\sqrt{\tfrac{n_{\max}}{n_c}}\Big)\Big)\Big) \tag{2}$$

The model is then finetuned using standard next-token cross-entropy. We employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) targeting all linear modules with rank $r = 32$, scaling factor $\alpha = 64$, and dropout 0.05.

4.2 GRPO_1: Reward-Aligned Reasoning Optimization

To bridge the gap between descriptive SFT capability and classification performance, we apply reasoning-oriented RL via GRPO (Shao et al., 2024). For each video $\mathbf{v}_i$, π_θ generates G candidate responses. Each response is scored by a composite rule-based reward function combining formatting and accuracy (Wu et al., 2026).

Format Reward. A binary reward $R_{\text{format}} \in \{0, 1\}$ ensures that the response encloses reasoning strictly within `<think>` tags and concludes with the prediction in `<answer>` tags.

Class-Weighted Accuracy Reward. Standard accuracy rewards inherently favor majority classes, making it unsuitable for our naturalistic distribution. We counteract this by using a class-weighted accuracy reward, R_{acc}, where rare correct predictions yield proportionally larger rewards. For each class c, the continuous weight w_c is derived from a square-root-softened inverse-frequency criterion, normalized such that the rarest class receives a maximum weight of 1.0:

$$w_c = \frac{\sqrt{\frac{N}{C \cdot n_c}}}{\max_{c'} \sqrt{\frac{N}{C \cdot n_{c'}}}}, \tag{3}$$

where N denotes the corpus size and C the number of classes per axis. For single-task samples, if the predicted label $\hat{\mathrm{y}}$ matches any valid ground-truth annotation, $R_{\text{acc}} = w_{\hat{\mathrm{y}}}$; otherwise, 0. For joint-task samples (action and function), each axis independently contributes half of its primary-class weight if correct.

The total scaled reward for a rollout is computed as $R_{\text{total}} = R_{\text{format}} + \lambda \cdot R_{\text{acc}}$. We then maximize the standard clipped GRPO surrogate objective, utilizing these rewards to compute the group-relative advantage while penalizing Kullback-Leibler (KL) divergence against the SFT_1 checkpoint.

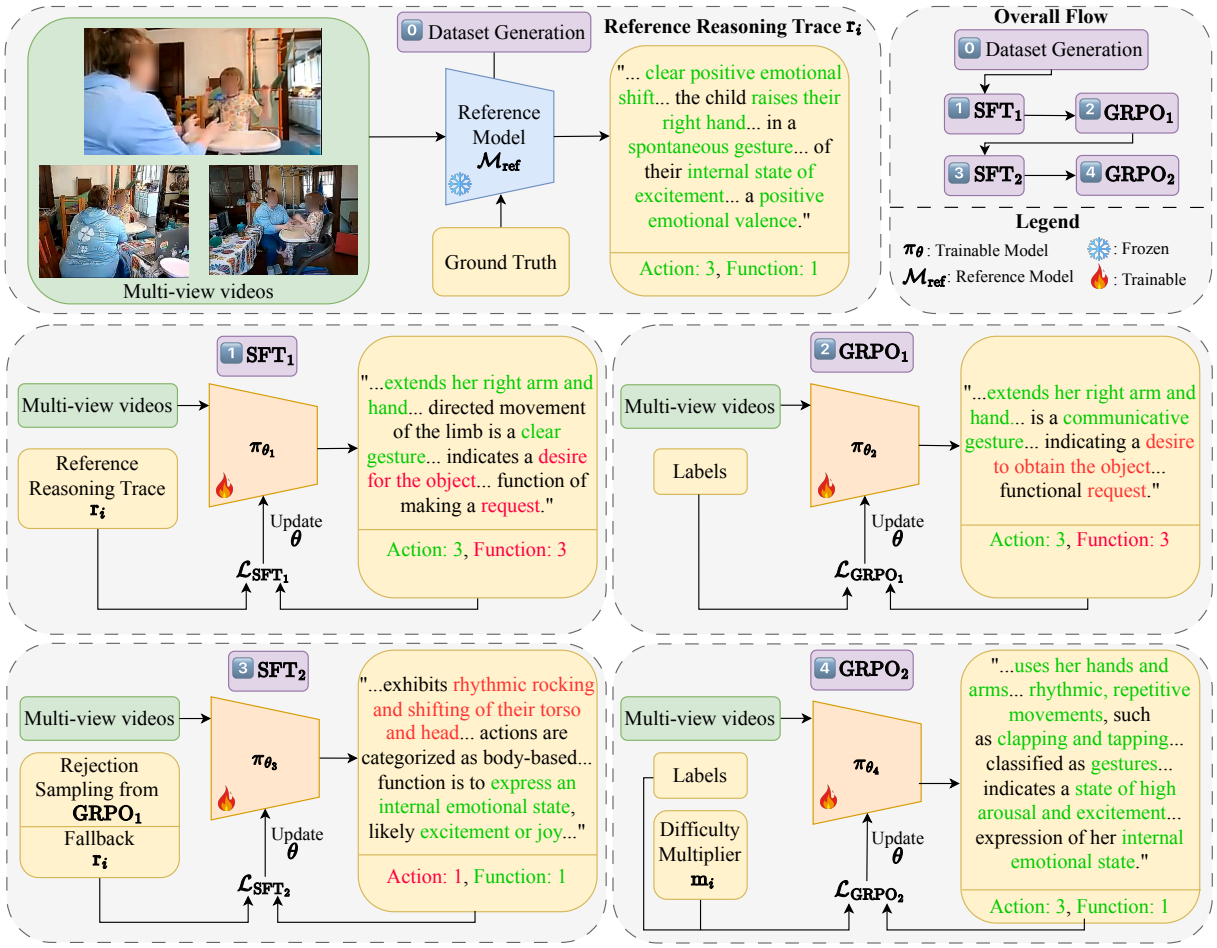


Figure 1: **Overview of the NDD-R1 framework. (0)** Data Generation: A frozen reference model ($\mathcal{M}_{\text{ref}}$) generates ground-truth reasoning traces (r_i). **(1)** SFT$_1$: Domain adaptation of the trainable model (π_θ) on r_i. **(2)** GRPO$_1$: RL via class-weighted accuracy rewards. **(3)** SFT$_2$: Self-distillation via rejection sampling, using r_i as fallback for unrescued clips. **(4)** GRPO$_2$: Difficulty-tiered policy optimization using a sample-specific multiplier (m_i). Text colors highlight the progression from reasoning mismatches (red) to alignment (green).

4.3 SFT$_2$: Rejection-Sampled Self-Distillation

While GRPO$_1$ improves classification accuracy, RL rollouts often contain reasoning trajectories drawn from a sampling distribution shaped by exploration under reward optimization rather than by direct supervision (Zhang et al., 2025b).

To consolidate successful reasoning trajectories into a focused reference policy, we perform a second SFT pass on data curated from the GRPO$_1$ policy via a temperature-escalation cascade.

For each clip, we sequentially sample rollouts across three generation temperatures to secure a correct prediction: standard exploration ($T = 1.0$), high-confidence exploitation ($T = 0.5$), and distribution-tail exploration ($T = 1.3$). The rescue stage s_i is recorded as the lowest cascade tier that yields a structurally valid and semantically correct trace: $s_i \in \{1, 2, 3\}$.

If a clip remains unrescued after all three cascade stages, we assign $s_i = 4$ and revert to the label-guided reference output r_i generated by $\mathcal{M}_{\text{ref}}$ in Stage 1. This fallback mechanism ensures that the SFT$_2$ corpus maintains full coverage of complex samples rather than skewing toward easy samples the model can already solve. The curated dataset is re-balanced using the oversampling factor ρ_c (Equation 2) and trained via standard SFT.

4.4 GRPO$_2$: Difficulty-Tiered Policy Optimization

The final stage applies a second GRPO pass utilizing a dynamic, difficulty-tiered reward curriculum. The underlying rationale is that video clips requiring deeper cascade stages correspond to parts of the input distribution where the policy remains weakest; therefore, concentrating reward magnitude on these regions yields larger marginal gains than uni-

form reinforcement.

We derive a deterministic difficulty multiplier m_i directly from the clip's rescue stage s_i established during the SFT_2 cascade:

$$\mathrm{m}_i = \begin{cases} 1.0, & s_i = 1 \quad \text{(easy)}, \\ 1.2, & s_i = 2 \quad \text{(medium)}, \\ 1.4, & s_i = 3 \quad \text{(hard)}, \\ 1.7, & s_i = 4 \quad \text{(label-guided fallback)}. \end{cases} \tag{4}$$

The intermediate multipliers ($1.0 \leq \mathrm{m}_i \leq 1.4$) scale linearly to provide a stable gradient for moderate errors, while the final tier introduces a nonlinear jump (1.7) to prioritize severe edge cases requiring synthetic fallback. We strictly cap this maximum scalar at 1.7 to maximize the contrastive learning signal without inducing the gradient spikes that cause catastrophic forgetting.

The class-weighted accuracy reward for GRPO_1 is multiplicatively modulated by this sample-specific difficulty scalar:

$$R_{\text{acc}}^{\text{GRPO}_2}(o_j, y_i) = \mathrm{m}_i \cdot R_{\text{acc}}^{\text{GRPO}_1}(o_j, y_i). \tag{5}$$

The format reward and GRPO advantage computations remain identical to GRPO_1. This multiplicative composition deliberately enforces that clips which are both rare (high w_c) and contextually difficult (high m_i) attract the strongest gradient reinforcement upon success, anchoring the model against regression while expanding its discriminative capabilities.

5 Experimental Results

5.1 Performance Comparison

Table 1 presents the quantitative evaluation of our proposed pipeline across single-task and joint classification settings.

Action Classification. NDD-R1 achieves the best performance in identifying communicative actions. In the single-task setting, it achieves the highest performance (63.97% Accuracy, 51.75 F1), with a substantial improvement over the zero-shot Qwen3-Omni baseline (50.70% Accuracy, 35.55 F1). This trend also holds in the joint classification setting, where NDD-R1 achieves 64.52% accuracy and 53.13 F1, indicating that the model successfully isolates kinematic features even when simultaneously burdened with pragmatic reasoning tasks.

Function Classification. Inferring pragmatic intent is inherently more abstract and subjective, reflected by lower absolute scores across all models. NDD-R1 establishes the highest accuracy in both the single- (42.47%) and joint-task (40.96%) function settings. While the massive-scale Gemini-3-Flash retains a marginal advantage in Function F1 (32.09 single-task, 32.44 joint), likely due to the frontier model's extensive pre-training on generalized abstract reasoning, NDD-R1 remains competitive (31.09 single, 31.77 joint) and strictly outperforms all its intermediate checkpoints.

Joint Classification. Performance generally increased for both action and function classification when predicting both together, indicating that MLLMs can handle complex, multi-objective prompts. This suggests that the models leverage pragmatic context to better resolve ambiguous physical actions (and vice versa), establishing a synergistic rather than competitive relationship between the two axes. However, absolute performance remains modest across all tasks, even for finetuned methods.

Table 1: **Main results.** Best results are in **bold**, and second-best results are underlined. SFT_1, GRPO_1, and SFT_2 represent different stages of NDD-R1's training.

Model	Single Task				Joint Task			
	Action		Function		Action		Function	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Gemini-3-Flash	48.84	41.14	36.61	**32.09**	51.78	46.02	37.30	**32.44**
ROSCO-Omni	53.06	36.74	38.72	27.55	53.39	37.56	40.02	29.29
Qwen3-Omni-30B	50.70	35.55	31.06	24.66	52.59	37.28	40.28	24.03
SFT_1	58.13	46.61	35.77	28.13	62.87	49.64	35.77	26.70
GRPO_1	58.90	46.81	39.45	28.45	64.38	51.86	40.27	29.28
SFT_2	62.28	48.93	38.50	27.47	61.51	50.29	38.36	30.45
NDD-R1	**63.97**	**51.75**	**42.47**	31.09	**64.52**	**53.13**	**40.96**	31.77

5.2 Multi-view Ablation

To evaluate spatial reasoning and robustness to visual information scaling, we ablate the number of synchronized camera views provided during inference in Table 2. The zero-shot baseline (Qwen3-Omni) exhibits severe modality overload across both classification axes. For action prediction, performance peaks at two views before degrading by 3.38 F1 points when all views are introduced; for function prediction, F1 degrades monotonically as more views are added (from 26.88 to 24.03 F1). In contrast, NDD-R1 shows monotonic performance scaling as the number of views in-

Table 2: **Performance (Accuracy and F1-Macro) under varying multi-view input configurations.** Best results across both labels and views are in **bold**.

Model	Action						Function					
	1 View		2 Views		All Views		1 View		2 Views		All Views	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Qwen3-Omni-30B	46.17	27.30	52.27	36.80	50.70	33.42	35.01	26.88	35.27	24.82	40.28	24.03
NDD-R1	58.69	43.32	62.60	48.73	**64.52**	**53.13**	35.02	27.64	36.86	29.70	**40.96**	**31.77**

Table 3: **Participant-wise joint classification performance (F1-Macro) on challenge sets.** Best results are in **bold**. Sample counts represent the size of the respective test sets.

Model	TALK009 *Zero-shot* *51 samples*		TALK015b *1 train session* *122 samples*		TALK026b *2 train sessions* *93 samples*		TALK012 *Function only* *8 samples*		TALK025 *Action only* *15 samples*		TALK028 *Joint only* *7 samples*	
	Action	Function	Action	Function	Action	Function	Action	Function	Action	Function	Action	Function
Gemini-3-Flash	29.78	31.81	21.60	27.75	**48.81**	**15.28**	32.14	38.89	31.47	**36.02**	**62.96**	**46.15**
ROSCO-Omni	31.90	**37.47**	24.58	24.35	30.07	13.07	22.72	40.31	25.75	17.59	17.12	12.24
Qwen3-Omni-30B	27.90	31.45	27.48	22.75	35.80	10.44	31.43	28.00	37.41	11.43	22.22	45.45
NDD-R1	**38.72**	25.95	**33.57**	**27.94**	38.95	11.53	**53.57**	**100.00**	**68.40**	33.33	16.67	18.18

creases, achieving peak performance with all available views across both action and function axes. Furthermore, under constrained single-view conditions, NDD-R1 proves significantly more robust, retaining 81.5% of its peak Action F1 and 87.0% of its peak Function F1. These results indicate that NDD-R1 leverages better information by resolving multiple views, and extracts stable communicative signals from limited or missing views.

5.3 Challenge Set Analysis

Table 3 compares the performance of NDD-R1 against Gemini-3-Flash and the baseline Qwen3-Omni model across three distinct generalization challenges: zero-shot adaptation, cross-session longitudinal tracking, and task-specific ablation.

Zero-Shot Generalization. On the strictly held-out participant (TALK009), NDD-R1 demonstrates generalizability learned from the entire training dataset. Despite the introduction of a novel four-view camera configuration and an entirely unseen subject, NDD-R1 outperforms both the Qwen baseline and the Gemini-3-Flash reference model for action classification, but fails to infer communicative functions (purpose or pragmatic intents) without prior exposure to the individual's specific communicative baseline.

Cross-Session Generalization. NDD-R1 effectively leverages training data to understand longitudinally-differing communicative repertoires. On TALK015b (one prior training session), our model establishes the best performance across both axes. For TALK026b (two prior sessions), while the frontier Gemini model retains the highest absolute performance, NDD-R1 still improves upon the zero-shot Qwen baseline. However, this implies that more training sessions on the same person may not strictly hold transferable information, especially when the recorded behaviors exhibit high variability.

Task Ablation and Supervision. When trained exclusively on function-centric descriptions (TALK012), NDD-R1 cleanly isolates pragmatic intent without interference from action-related reasoning cues. Similarly, restricting supervision to action descriptions (TALK025) yields competitive kinematic identification. However, given the small test size, these gains may reflect memorization or overfitting rather than generalization. When supervision is restricted entirely to joint descriptions (TALK028), performance collapses catastrophically, falling significantly behind both baseline models. This indicates that forcing the model to learn entangled, multi-axis reasoning without foundational single-task grounding

Table 4: **Participant-clustered bootstrap 95% confidence intervals (10,000 resamples) for macro-F1.** Best results are in **bold**. SFT_1, $GRPO_1$, and SFT_2 represent different stages of NDD-R1's training.

Model	Single Task		Joint Task	
	Action F1	Function F1	Action F1	Function F1
Qwen3-Omni-30B	35.55 [28.88, 43.23]	24.80 [19.37, 28.83]	37.28 [31.89, 43.33]	24.03 [20.86, 27.67]
SFT_1	46.61 [40.06, 50.22]	28.13 [24.75, 31.90]	49.64 [44.26, 52.37]	26.70 [23.22, 29.56]
$GRPO_1$	46.81 [41.40, 51.50]	28.45 [24.59, 31.52]	51.86 [46.29, 57.07]	29.28 [25.17, 32.68]
SFT_2	51.48 [45.15, 57.12]	27.54 [23.06, 31.99]	50.29 [43.97, 54.25]	30.45 [25.50, 34.42]
NDD-R1	**51.75** [45.63, 57.35]	**31.09** [26.04, 35.92]	**53.13** [46.29, 57.27]	**31.77** [26.88, 36.92]

Table 5: **Aggregated participant-wise performance for joint classification.** Results report mean ± standard deviation across 27 participants. Best results are in **bold**.

Model	Action		Function	
	Acc	F1	Acc	F1
Gemini-3-Flash	55.35±16.04	38.93±15.03	40.77±13.58	22.24±8.39
ROSCO-Omni	56.91±16.05	36.94±11.41	44.07±16.88	26.83±11.35
Qwen3-Omni-30B	55.56±14.80	35.60±13.36	43.86±16.59	25.47±11.54
NDD-R1	**66.60**±13.80	**46.51**±16.91	**44.50**±17.96	**28.98**±17.84

severely destabilizes reasoning.

5.4 Statistical Significance of Curriculum Gains

In Table 4, we quantify variance via participant-clustered bootstrapping: for each model, we resample participants with replacement 10,000 times and recompute macro-F1 over the resulting draws, reporting the resulting 95% confidence intervals. We resample at the participant level rather than the clip level, since clips from the same participant are not independent.

NDD-R1's 95% confidence intervals lie strictly above the mean performance of the Qwen3-Omni baseline across all four evaluated settings, proving that the full four-stage curriculum yields significant gains over the base model. While stage-to-stage intervals overlap, the cumulative training consistently yields the strongest final configuration.

5.5 Aggregated Participant-Wise Performance

To account for the high inter-subject variability inherent in the communicative behaviors of non- and minimally-speaking autistic individuals, we compute the aggregated participant-wise performance for the joint classification task (Table 5).

NDD-R1 consistently achieves the highest mean performance across all metrics. For action classification, NDD-R1 achieves a substantial improvement over the baseline Qwen3-Omni and Gemini-3-Flash models. Furthermore, on the highly subjective Function axis, NDD-R1 yields the highest mean accuracy and F1 score. The standard deviations remain uniformly large across all evaluated models, quantitatively reflecting the intrinsically idiosyncratic and heterogeneous nature of the target population's behaviors.

6 Conclusion

In this work, we presented NDD-R1, a multi-stage reasoning framework designed to interpret the communicative actions and functions of non- and minimally-speaking autistic individuals from naturalistic, multi-view video. Through a difficulty-tiered, reasoning-aligned SFT and RL curriculum, our framework successfully adapts a general-purpose MLLM to decode subtle, idiosyncratic, and highly contextual multimodal behaviors.

NDD-R1 establishes new performance baselines, outperforming frontier models in both single-task and joint classification settings. Our qualitative and topic-modeling analyses reveal that NDD-R1 learns to effectively model both visual evidence and abstract pragmatic intent for this population.

This work demonstrates the viability of utilizing reasoning-aligned, multi-view MLLMs to recognize the communicative agency of a historically underserved population. Future work will explore the longitudinal growth of individuals in this population, enhance data collection and caregiver annotation protocols to generate more consistent and scalable labels, and leverage more detailed descriptions from videos to further disambiguate complex, overlapping communicative acts from these individuals.

Acknowledgements

We acknowledge the ongoing development and refinement of the ROSCO protocol, including substantial contributions from Dr. Lauren Thompson and input and guidance from numerous collaborators and colleagues. We are grateful to Drs. Audrey Thurm, Carol Wilkinson, and Mustafa Sahin for their guidance and support of the broader ROSCO research program. We also thank Emine Arcasoy, Bianca Booth, Isabelle Iannotti, Miranda Kannisto, Emma McGonigle, and Tsambika Rizas for their contributions to data collection and preparation, and the study participants and their families for making this research possible. This work was supported in part by the NIH NINDS TALK Supplement (3U54NS092090-10S1).

Ethical Considerations

Potential Risks. Deploying automated systems to interpret nonverbal communication carries the inherent risk of algorithmic misinterpretation. If a model incorrectly decodes a subtle communicative cue, it could lead to false assumptions regarding the individual's desires or needs, directly affecting their autonomy and the quality of care they receive. Furthermore, utilizing video-based behavioral analysis in domestic settings raises significant privacy concerns. To prevent data exploitation or non-consensual surveillance, any real-world application of this technology must enforce rigorous data security measures, ensure informed consent, and center the dignity and well-being of the users.

Biases. This training data inherently mirrors the natural distribution of daily communicative acts, leading to significant class imbalances. Because of the profound heterogeneity in how these individuals communicate, a model that performs well on average might still systematically fail for someone with rare or unique expression styles. Relying on such a system could produce inequitable outcomes, disproportionately disadvantaging individuals whose specific interaction styles are underrepresented in the training corpus.

Reproducibility Statement. We include implementation details and hyperparameter settings for all models in Appendix A. Our code is available at https://github.com/SiddhantBikram/NDD-R1.

Limitations

Despite the advancements presented in this work, we recognize several key limitations. First, the absolute predictive performance of our model remains modest. The system requires substantial refinement and reliability testing before it can be considered for actual clinical or domestic deployment.

Second, while this initial ROSCO dataset is pioneering in its naturalistic approach, it is fundamentally constrained in size, comprising 2,582 samples from 27 individuals. Furthermore, the cohort is limited to participants diagnosed with Phelan-McDermid Syndrome (PMS), PTEN Hamartoma Tumor Syndrome (PHTS), and Tuberous Sclerosis Complex (TSC). Gathering data from this vulnerable demographic is extraordinarily complex due to ethical constraints, recruitment hurdles, and specialized study design requirements. As a result, our findings may not generalize to the wider autism spectrum or to other complex neurodevelopmental disorders. Nevertheless, our work lays the groundwork for more inclusive, equitable assistive technologies that center the communicative realities of non- and minimally-speaking individuals.

Third, nonverbal communication and its subsequent interpretation are highly dependent on cultural norms. This initial ROSCO data stems from a US-based demographic, meaning the behaviors observed might not translate globally. Additionally, while the ROSCO study protocol ensures consistency in timing and naturalistic activities, the uncontrolled nature of home recordings means that background noise, available objects, room layout, and individual caregiver dynamics heavily influence the captured behaviors. These contextual variables contribute to the high inter-participant variance seen in our results, and may complicate generalization to new environments.

Fourth, ROSCO's reliance on primary caregivers for ground-truth annotations ensures high ecological validity, but it intrinsically limits the data to a single interpretive lens. Clinicians, educators, and the wider population may interpret the same interaction differently, and human interpreters are susceptible to misinterpretation just like machines.

Finally, while ROSCO's action-function annotation schema provides a strong foundation for testing algorithmic frameworks in this underserved domain, it is still preliminary and is being continuously refined to better match developmental constructs alongside even more consistent and rigorous labeling procedures. Future research must tackle the broader interaction pipeline, alongside expanded data collection.

References

Yang Chen, Jingcai Guo, Miaoge Li, Zhijie Rao, and Song Guo. 2026. Star++: Region-aware conditional semantics via interpretable side information for zero-shot skeleton action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*.

Yi Chen, Yuying Ge, Rui Wang, Yixiao Ge, Junhao Cheng, Ying Shan, and Xihui Liu. 2025. Grpo-care: Consistency-aware reinforcement learning for multimodal reasoning. *arXiv preprint arXiv:2506.16141*.

Hyundong Justin Cho, Spencer Lin, Tejas Srinivasan, Michael Saxon, Deuksin Kwon, Natali T Chavez, and Jonathan May. 2025. Can vision language models understand mimed actions? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26744–26759.

Shijian Deng, Erin E Kosloski, Siva Sai Nagender Vasireddy, Jia Li, Randi Sierra Sherwood, Feroz Mohamed Hatha, Siddhi Patel, Pamela R Rollins, and Yapeng Tian. 2026. Toward gaze target detection of young autistic children. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38367–38375.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2026. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Kristina T Johnson, Jaya Narain, Thomas Quatieri, Pattie Maes, and Rosalind W Picard. 2023. Recanvo: A database of real-world communicative and affective nonverbal vocalizations. *Scientific Data*, 10(1):523.

Chandra Lebenhagen. 2020. Including speaking and nonspeaking autistic voice in research. *Autism in Adulthood*, 2(2):128–131.

Daniel Lee and H Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13.

Hengzhi Li, Megan Tjandrasuwita, Yi R Fung, Armando Solar-Lezama, and Paul Pu Liang. 2025. Mimeqa: Towards socially-intelligent nonverbal foundation models. *arXiv preprint arXiv:2502.16671*.

Jaya Narain, Kristina T Johnson, Thomas F Quatieri, Rosalind W Picard, and Pattie Maes. 2022. Modeling real-world affective and communicative nonverbal vocalizations from minimally speaking individuals. *IEEE Transactions on Affective Computing*, 13(4):2238–2253.

Haoxuan Qu, Yujun Cai, and Jun Liu. 2024. Llms are good action recognizers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18395–18406.

V Rose, David Trembath, Deb Keen, and Jessica Paynter. 2016. The proportion of minimally verbal children with autism spectrum disorder in a community-based early intervention programme. *Journal of Intellectual Disability Research*, 60(5):464–477.

Siddhant Bikram Shah and Kristina T Johnson. 2026a. Axon: Action characterization through cross-modal knowledge distillation for neurodiverse individuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 39228–39238.

Siddhant Bikram Shah and Kristina T Johnson. 2026b. Rosco-omni: Multimodal llm-based communication understanding for non-and minimally-speaking autistic individuals. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 40453–40469.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Roberta Simeoli, Angelo Rega, Mariangela Cerasuolo, Raffaele Nappo, and Davide Marocco. 2024. Using machine learning for motion analysis to early detect autism spectrum disorder: A systematic review. *Review Journal of Autism and Developmental Disorders*, pages 1–20.

Helen Tager-Flusberg and Connie Kasari. 2013. Minimally verbal school-aged children with autism spectrum disorder: The neglected end of the spectrum. *Autism research*, 6(6):468–478.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. 2024. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. *Advances in Neural Information Processing Systems*, 37:7821–7846.

Chelsea La Valle, Karen Chenausky, and Helen Tager-Flusberg. 2021. How do minimally verbal children and adolescents with autism spectrum disorder use communicative gestures to complement their spoken language abilities? *Autism & Developmental Language Impairments*, 6:23969415211035065.

Qiuhong Wei, Huiling Cao, Yuan Shi, Ximing Xu, and Tingyu Li. 2023. Machine learning based on eye-tracking data to identify autism spectrum disorder: A systematic review and meta-analysis. *Journal of biomedical informatics*, 137:104254.

Jiulong Wu, Yucheng Shen, Lingyong Yan, Haixin Sun, Deguo Xia, Jizhou Huang, and Min Cao. 2026. Facial-r1: Aligning reasoning and recognition for facial emotion analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 26939–26947.

Wangmeng Xiang, Chao Li, Yuxuan Zhou, Biao Wang, and Lei Zhang. 2023. Generative action description prompts for skeleton-based action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10276–10285.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Cheol-Hwan Yoo, Jang-Hee Yoo, and Jaeyoon Jang. 2025. Care-vl: A domain-specialized vision-language model for early asd screening. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 57–66. Springer.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025a. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1859–1869.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, and 1 others. 2025b. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, and 1 others. 2025. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29733–29735.

A Experimental Settings

Implementation Details. Model finetuning and inferencing were conducted on multiple NVIDIA H200 GPUs depending on shared resource availability, using the ModelScope Swift (Zhao et al., 2025) framework. We used a batch size of 8 and accumulated gradients over 4 steps. Gemini API was used to generate descriptions for the instruction dataset and for Gemini-based model inference. We processed video inputs at a resolution of $384 \times 384 = 147,456$ pixels with a sampling rate of 2 frames per second (FPS). For NDD-R1, SFT_1 and SFT_2 used a learning rate (LR) of $1e-4$ for 2 epochs, and GRPO_1 and GRPO_2 used an LR of $1e-6$ for 1 epoch with a λ of 5 and $G = 8$ generations.

For single-task evaluation, we score a prediction as accurate if it matches any ground-truth label assigned to the sample. For joint-task evaluation, we score the action and function axes independently as individual single tasks.

Baselines. We used Gemini-3-Flash (Team et al., 2023) through Gemini API, Qwen3-Omni-30B-A3B-Thinking (Xu et al., 2025) through HuggingFace, and ROSCO-Omni (Shah and Johnson, 2026b) through its official Github repository[2]. We use Qwen3-Omni as the foundational architecture and benchmark for NDD-R1 because it is one of the few open-source MLLMs capable of native, synchronized audio-visual processing, which is essential given the central role of vocalizations in our task.

B Per-class performance

Table 6 presents the per-class metrics for the NDD-R1. For action classification, the model excels at identifying explicit physical modalities, achieving strong performance for AAC, Vocalization, and Gesture. Conversely, subtle behaviors such as Looking and Face remain challenging, likely due to the difficulty of extracting fine-grained facial landmarks and gaze vectors in naturalistic environments (Deng et al., 2026).

[2]https://github.com/SiddhantBikram/ROSCO-Omni

A similar pattern can be observed for function classification–direct intents, such as Request and Emotion, are classified with high reliability. In contrast, ambiguous or rare internal states, such as Social and SDB, exhibit significantly lower performance.

While NDD-R1 successfully interprets prominent physical actions and pragmatic intents, disambiguating low-arousal communicative signals remains a frontier challenge for multimodal alignment.

Table 6: **Per-class joint classification performance by NDD-R1.**

Class	Precision	Recall	F1
Action			
AAC	60.67	85.71	71.05
Body	72.15	36.77	48.72
Face	38.46	38.46	38.46
Gesture	61.67	78.97	69.26
Looking	30.43	15.56	20.59
Vocalization	77.71	64.82	70.68
Macro avg	56.85	53.38	53.13
Micro avg	64.52	62.06	63.26
Function			
Commenting	27.78	37.74	32.00
Emotion	47.40	46.91	47.15
Reject	45.00	23.28	30.68
Request	46.47	60.67	52.63
SDB	18.33	12.22	14.67
Social	14.71	12.50	13.51
Macro avg	33.28	32.22	31.77
Micro avg	40.96	40.85	40.90

C Analysis of Reasoning Traces

C.1 Quantitative Trace Evaluation

To assess the internal logic and physical grounding of the generated reasoning traces, we evaluated the textual contents of the `<think>` blocks generated by the models across the test set, using deterministic lexical and structural heuristics to measure reasoning quality.

Physical Density. Quantifies the descriptive visual grounding of the reasoning trace. It is calculated as the ratio of specific anatomical and kinematic terms (e.g., "hand," "face," "gaze," "torso") per 10 words. Higher density indicates that the model grounds its reasoning in observable physical reality rather than generalized, abstract assertions.

Target Overlap. Measures the lexical alignment between the reasoning trace and the ground-truth labels. It calculates the percentage of traces that successfully utilize the predefined vocabulary associated with the clip's true action and function classes, indicating whether the model logically deduced the correct class versus guessing.

Our results are shown in Table 7. Our evaluation demonstrates that NDD-R1 consistently produces physically grounded and structurally sound reasoning traces compared to both the zero-shot baseline and the Gemini reference model.

Table 7: **Evaluation of joint inference reasoning trace quality across models.** NDD-R1 achieves the highest physical density and lexical alignment with ground-truth targets.

Model	Avg. Physical Density	Target Overlap (%)	
		Action	Function
Gemini-3-Flash	0.156	63.66	57.24
Qwen3-Omni-30B	0.111	76.23	70.49
ROSCO-Omni	0.070	71.17	65.98
NDD-R1	**0.160**	**77.19**	**71.45**

C.2 Qualitative Progression Analysis

To illustrate the evolution of the model's reasoning capabilities across the NDD-R1 curriculum, we analyze the outputs of a joint classification sample from the baseline through all training stages.

Table 8 demonstrates how Gemini and early stages of the trainable model suffer from feature blindness and label-reasoning misalignment. In the final stage, NDD-R1 reliably produces an accurate and coherent reasoning trace.

D Topic Modelling

To analyze the linguistic representations learned by NDD-R1, we applied Non-Negative Matrix Factorization (NMF) (Lee and Seung, 2000) to the reasoning traces generated by the model for single-task action and function classification. Traces were preprocessed to remove generic domain stopwords. Further, we applied orthogonal stopword masking by explicitly removing function-related terminology from the Action corpus and action-related terminology from the Function corpus, preventing topic representations from being biased by imbalanced action–function co-occurrence patterns. For each predicted action and function class, we extracted the top two latent subtopics (ranging from

Stage	Action	Function	Reasoning Trace	Analysis
Ground Truth	**5 (Vocalization)**	**1 (Emotion)**	–	*Reference labels derived from primary caregiver annotation.*
Gemini-3-Flash	4 (Look)	5 (Social)	"...child's gaze is directed towards the laptop screen... use of eye gaze to attend to a shared object of interest... constitutes joint attention... The primary modality observed is looking."	Fails to detect the auditory vocalization, incorrectly attributing the primary communicative modality to visual gaze.
ROSCO-Omni	5 (Vocalization)	3 (Request)	"...child is making vocalizations—sounds like 'Eh' or similar non-speech sounds... child's vocalizations might be a request for attention or action."	Predicts the correct vocal modality but reads an affective vocalization as a bid for action.
Qwen3-Omni	5 (Vocalization)	3 (Request)	"...child is making vocalizations—sounds like 'Eh'... maybe trying to get attention or request something."	Isolates the correct vocal modality but relies on untrained priors, inaccurately inferring a pragmatic request without contextual substantiation.
SFT_1	5 (Vocalization)	0 (Comment)	"...produces a clear, single-syllable vocalization... sharing her attention and interest... characterizes this as a vocalization used for commenting."	Adopts appropriate clinical terminology and accurately identifies the action, but misclassifies the pragmatic intent as a comment.
$GRPO_1$	5 (Vocalization)	0 (Comment)	"...produces a clear, high-pitched vocalization that sounds like a squeal of excitement... sharing her interest... serves as a comment."	Exhibits explicit label-reasoning misalignment–correctly observes a "squeal of excitement" but assigns a conflicting classification.
SFT_2	3 (Gesture)	3 (Request)	"...child extends her right arm and hand toward the laptop's keyboard... directed movement... clear example of a gesture... indicates a desire to access... communicative request..."	Misses the vocalization and fabricates a physical reaching gesture to justify an incorrect request classification.
NDD-R1	5 (Vocalization)	1 (Emotion)	"...brief, high-pitched vocalization... squeal. ...not a request... but rather an expression of engagement and excitement... positive emotional valence..."	Accurately grounds the acoustic evidence with the correct internal emotional valence.

Table 8: **Qualitative reasoning analysis for one sample.** NDD-R1 successfully resolves prior modality hallucinations and label-reasoning misalignments to correctly predict the sample's action and function as Vocalization and Emotion, respectively. Within the reasoning traces, green highlights factually accurate observational grounding and correct predictions, whereas red denotes structural hallucinations, feature blindness, and classification errors.

unigrams to trigrams) to identify the primary lexical clusters driving the model's classifications. As detailed in Table 9, the model successfully learns specialized, decoupled vocabularies for distinct behaviors.

E Prompt templates

Our framework employs distinct prompt configurations for data distillation and model inference, both utilizing `<think>` and `<answer>` formatting protocols. During dataset creation, the reference model ($\mathcal{M}_{\text{ref}}$) receives label-guided prompts containing ground-truth annotations and cognitive hints to enforce accurate multimodal grounding and prevent visual hallucinations (Figure 2). Throughout the SFT and RL phases, the model (π_θ) utilizes unhinted inference templates. These configurations provide an expert preamble, complete class definitions, and explicit formatting instructions without exposing the target labels (Figures 3, 4, and 5).

Axis	Class	Sub-Topic 1	Sub-Topic 2
Action	**AAC**	speech, icons, navigates, icons screen	icon, communication, makes, clear example
	Body	movement, leaning, shift, forward	movements, body movements, shifts, postural shifts
	Face	bubbles, face, resting face, neutral resting face	positive state, state, clear, positive
	Gesture	arm, arm hand, right arm hand, right arm	tablet, point, uses, board
	Looking	book, orientation, head orientation, orientation eye gaze	face, direct, attention, focus
	Vocalization	vocalization, sound, vocal, vocal tract	vocalizations, sounds, series, produces series
Function	**Commenting**	screen, book, demonstrates, activity	toy, sharing, holding, small
	Emotion	positive, bubbles, smile, excitement	distress, expression internal, spontaneous expression internal, feeling
	Reject	hand, rejection, table, current	feeding, food, mouth, mouth immediately
	Request	toy, hand, physical, reaches	communication, screen, use, tool
	SDB	sensory, mouth, focused, self regulatory	movements, table, engaged, focus remains internal
	Social	attention, reciprocity, social reciprocity, contact	expression, state, internal, positive

Table 9: **Latent subtopics extracted via NMF from NDD-R1 reasoning traces for single-task action and function classification.**

Label-Guided Reference Model Prompt

You are an expert behavioral analyst observing a short video clip of a non- or minimally-speaking child with profound autism and complex neurodevelopmental disorders interacting with their caregiver. You are shown {num_views} camera angle(s) of the same moment. Focus ONLY on the child – ignore adults except as context for the child's actions. Integrate observations from all camera views. Write in a clinical, objective tone. Be concise but specific.

CONTEXT: The child's communicative behavior in this clip is known to be:
ACTION: class {action_id} – {action_name}: {action_def}
FUNCTION: class {function_id} – {function_name}: {function_def}

ACTION GUIDANCE HINT: {action_cognition_hint}
FUNCTION GUIDANCE HINT: {function_cognition_hint}

YOUR TASK: Produce an integrated reasoning trace that explains why these labels fit the observable physical behavior and communicative intent in the video. Reference what the child's hands, gaze, body, and posture are doing, and how they relate to the underlying communicative intent. Make the reasoning specific to this clip – do not write generic descriptions.

OUTPUT FORMAT – output this structure, with NO text before or after:

```
<think>
[5-8 sentences of integrated clinical reasoning based on the provided hints]
</think>
<answer>Action: {action_id}, Function: {function_id}</answer>
```

The `<answer>` tag must contain only "Action: {action_id}, Function: {function_id}" with no other text. Do not write anything outside the `<think>` and `<answer>` tags.

Figure 2: **Contextual prompt template utilized during Step 4 to generate label-guided reasoning traces from the reference model ($\mathcal{M}_{\text{ref}}$).**

Action Inference Prompt

You are observing a short video clip with multiple camera angles of a non- or minimally-speaking child with profound autism and complex neurodevelopmental disorders interacting with their caregiver.

Classify the child's communicative ACTION. Focus only on the child – ignore any adults present.

ACTION Class Definitions:

0: AAC: Augmentative and Alternative Communication: Use of high-tech (tablets, speech-generating devices) or low-tech (picture cards, communication boards) AAC systems.

1: Body: Use of primarily the body or head. This includes more holistic movements like postural shifts (e.g., leaning in, turning away), or whole-body movements (e.g., rocking, walking away).

2: Face: Use of facial expressions to convey meaning (e.g., smiling, grimacing, frowning) that is not primarily a gaze shift.

3: Gesture: Use of the hands, arms, or limbs to communicate. This includes specific, directed movements like pointing, reaching, waving, or hand leading.

4: Looking: Use of eye gaze or head orientation to direct another person's attention to a specific subject, person, or location.

5: Vocalization: Use of any non-speech or speech-like sound made with the vocal tract to communicate (e.g., grunt, squeal, laugh, word approximation).

First, reason about what you observe inside `<think>` tags. Then provide your answer inside `<answer>` tags as a single digit 0-5.

Example format:

```
<think>
[Your reasoning about what the child is physically doing]
</think>
<answer>X</answer>
```

Figure 3: **Action classification inference prompt template.**

Function Inference Prompt

You are observing a short video clip with multiple camera angles of a non- or minimally-speaking child with profound autism and complex neurodevelopmental disorders interacting with their caregiver.

Classify the child's communicative FUNCTION. Focus only on the child – ignore any adults present.

FUNCTION Class Definitions:

0: Commenting: Sharing observations, expressing interest, or drawing attention to objects or events.

1: Emotion: Expressing an internal feeling state like happiness, frustration, excitement, or distress.

2: Reject: Refusing items, activities, pushing away, head shaking, or vocal protests indicating a desire for an activity to stop.

3: Request: Asking for objects, actions, continuation, or assistance through any modality.

4: SDB: Self-Directed Behavior: Behaviors that were not perceived as being intentionally communicative or directed at another person, like self-stimulatory behavior.

5: Social: Initiating social interaction, maintaining social reciprocity, greeting, responding to a name, or joint attention.

First, reason about what you observe inside `<think>` tags. Then provide your answer inside `<answer>` tags as a single digit 0-5.

Example format:

```
<think>
[Your reasoning about why the child is communicating]
</think>
<answer>X</answer>
```

Figure 4: **Function classification inference prompt template.**

Joint Inference Prompt

You are observing a short video clip with multiple camera angles of a non- or minimally-speaking child with profound autism and complex neurodevelopmental disorders interacting with their caregiver.

Classify the child's communicative ACTION and FUNCTION. Focus only on the child – ignore any adults present.

ACTION Class Definitions:

- **0: AAC:** Augmentative and Alternative Communication: Use of high-tech (tablets, speech-generating devices) or low-tech (picture cards, communication boards) AAC systems.
- **1: Body:** Use of primarily the body or head. This includes more holistic movements like postural shifts (e.g., leaning in, turning away), or whole-body movements (e.g., rocking, walking away).
- **2: Face:** Use of facial expressions to convey meaning (e.g., smiling, grimacing, frowning) that is not primarily a gaze shift.
- **3: Gesture:** Use of the hands, arms, or limbs to communicate. This includes specific, directed movements like pointing, reaching, waving, or hand leading.
- **4: Looking:** Use of eye gaze or head orientation to direct another person's attention to a specific subject, person, or location.
- **5: Vocalization:** Use of any non-speech or speech-like sound made with the vocal tract to communicate (e.g., grunt, squeal, laugh, word approximation).

FUNCTION Class Definitions:

- **0: Commenting:** Sharing observations, expressing interest, or drawing attention to objects or events.
- **1: Emotion:** Expressing an internal feeling state like happiness, frustration, excitement, or distress.
- **2: Reject:** Refusing items, activities, pushing away, head shaking, or vocal protests indicating a desire for an activity to stop.
- **3: Request:** Asking for objects, actions, continuation, or assistance through any modality.
- **4: SDB:** Self-Directed Behavior: Behaviors that were not perceived as being intentionally communicative or directed at another person, like self-stimulatory behavior.
- **5: Social:** Initiating social interaction, maintaining social reciprocity, greeting, responding to a name, or joint attention.

First, reason about what you observe inside `<think>` tags. Then provide your answer inside `<answer>` tags in the format "Action: X, Function: Y" where X and Y are single digits 0-5.

Example format:

```
<think>
[Your reasoning about what the child is doing and why]
</think>
<answer>Action: X, Function: Y</answer>
```

Figure 5: **Joint classification inference prompt template.**