Navigating Intention Reasoning Trees to Guide Agentic Preference Optimization

Ziben Yue[1], **Hongyu Yang**[2], **Weixin Li**[1*]
[1]School of Computer Science and Engineering, Beihang University
[2]Institute of Artificial Intelligence, Beihang University
{yueziben, hongyuyang, weixinli}@buaa.edu.cn

Abstract

Understanding human intentions in videos requires deep cognitive reasoning, yet Multimodal Large Language Models (MLLMs) frequently suffer from factual hallucinations and counterfactual blindness. To address these limitations, we introduce **ASTRO**, a novel framework that navigates **Intention Reasoning Trees** to guide preference optimization. Our approach employs a collaborative multi-agent system integrated with Monte Carlo Tree Search (MCTS) to systematically explore complex reasoning spaces based on visual evidence. By navigating this tree, ASTRO dynamically extracts distinct cognitive trajectories—correct, error, and ambiguous paths—to construct high-quality preference pairs. Finally, we apply Direct Preference Optimization (DPO) using these contrastive paths, explicitly enhancing the model's factual grounding and sensitivity to causal boundaries. Extensive experiments across complex video intention benchmarks demonstrate that ASTRO achieves superior performance while exhibiting remarkable data efficiency, significantly outperforming fully supervised baselines using only a fraction of the original training data.

1 Introduction

Understanding human intentions in videos is a cornerstone of artificial social intelligence. While structured reasoning techniques—such as Chain-of-Thought (CoT) (Wei et al., 2022), Tree of Thoughts (ToT) (Yao et al., 2023a), and ReAct (Yao et al., 2023b)—have significantly enhanced the problem-solving capabilities of Large Language Models (LLMs), extending these to dynamic video contexts remains challenging. Recent efforts like Agent-of-Thoughts Distillation (Shi et al., 2025b) and R3-VQA (Niu et al., 2025) highlight the necessity of explicit cognitive pathways and social cue integration. However, current Multimodal Large Language Models (MLLMs) still heavily rely on superficial pattern matching rather than genuine causal reasoning (Li et al., 2024; Poppi et al., 2026). Specifically, they exhibit two critical vulnerabilities (Figure 1): (1) **Factual Hallucinations**, fabricating explanations disconnected from grounded spatiotemporal evidence, and (2) **Counterfactual Blindness**, failing to identify exact contextual boundaries due to statistical biases (Li et al., 2024; Poppi et al., 2026). Furthermore, addressing these flaws via traditional Supervised Fine-Tuning (SFT) is notoriously data-hungry and inefficient.

To overcome these challenges, we propose **ASTRO** (**A**gentic **S**earch **T**ree **R**easoning **O**ptimization), a novel framework that shifts MLLM alignment from token-level cloning to trajectory-level cognitive exploration (Figure 2). Inspired by counterfactual analysis (Qu et al., 2024) and leveraging a Monte Carlo Tree Search (MCTS) architecture (Xie et al., 2024a), ASTRO deploys a collaborative multi-agent system to systematically expand an Intention Reasoning Tree (IRT) driven by step-level dual-axis (factual and counterfactual) critiques. Crucially, this tree search serves as a powerful online synthetic data engine. By navigating the IRT, ASTRO dynamically extracts three contrastive trajectories: a correct path (τ_c), a factual-error path (τ_e), and a counterfactual-ambiguous path (τ_a). We then apply Direct Preference Optimization (DPO) (Rafailov et al., 2023) to explicitly penalize these on-policy hard negatives, effectively internalizing both visual grounding and causal sensitivity into the model's parameters (Chen et al., 2024a).

In summary, our main contributions are:

- We introduce ASTRO, a pioneering multi-agent MCTS framework that structures complex video intention understanding into an explorable cognitive tree.
- We propose a dual-boundary preference opti-

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 37209–37228
October 24-29, 2026 ©2026 Association for Computational Linguistics

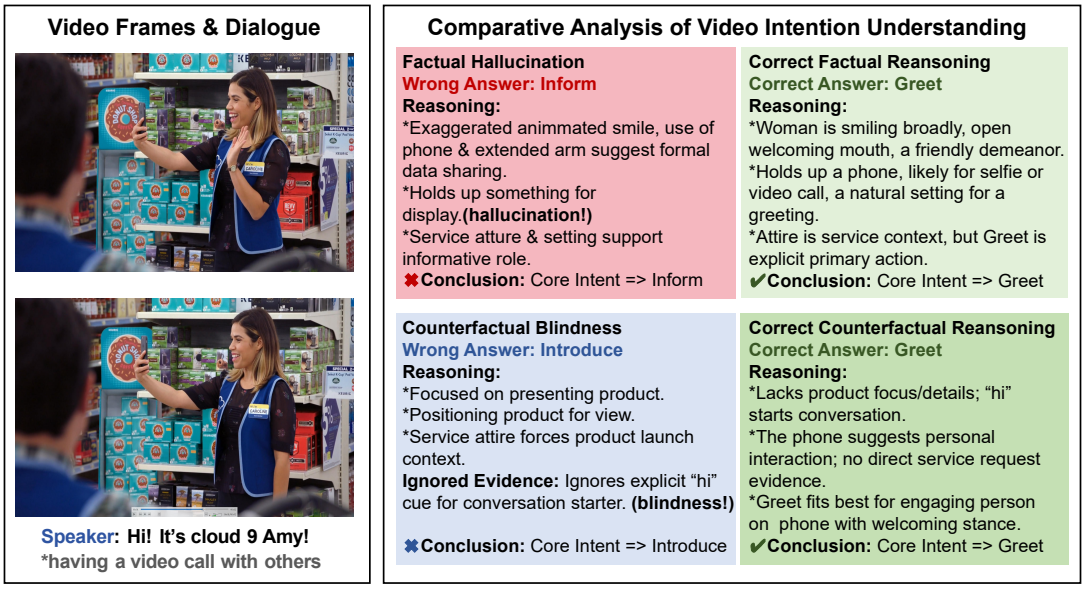


Figure 1: **An illustrative example of Factual Hallucination and Counterfactual Blindness compared with correct reasoning.** The figure visualizes how the model can provide explanations disconnected from grounded spatiotemporal visual evidence and rely on superficial statistical biases to make incorrect judgments. Factual and counterfactual reasons are dynamically presented to contrast correct and flawed reasoning processes.

mization strategy that sharply delineates factual and counterfactual reasoning boundaries.

- We achieve state-of-the-art performance on challenging benchmarks (IntentQA, Social-IQ 2.0, MIntRec2.0). Notably, ASTRO exhibits extreme data efficiency, significantly surpassing fully supervised models while utilizing a minimal fraction of the training data.

2 Related Work

Video Intention Understanding and Theory of Mind. Understanding video intentions requires deep cognitive reasoning and Theory of Mind (ToM) (Zhang et al., 2024). While early models focused on simple action recognition, recent benchmarks like EgoGazeVQA (Peng et al., 2026), IntentQA (Li et al., 2023b), Social-IQ 2.0, and MIntRec2.0 (Zhang et al., 2024) demand fine-grained multimodal alignment for complex social intelligence. Despite these advancements, Multimodal Large Language Models (MLLMs) still struggle with factual hallucinations and cognitive boundaries. Recent methods like DEL-ToM (Wu et al., 2025) and ReasVQA (Liang et al., 2025) attempt to enhance reasoning but either focus solely on text or ignore counterfactual boundaries. Our work bridges this gap by explicitly modeling visual factual and counterfactual constraints.

Multi-Agent Reasoning and Counterfactual Verification. Multi-agent debate (MAD) mitigates single-agent hallucinations through collaboration (Du et al., 2024), but can collapse into shared hallucinations (Liang et al., 2026). To address this, frameworks like MAVEN (Yao et al., 2026), CAMEL (Li et al., 2023a), and AgentVerse (Chen et al., 2024b) employ role-playing and strict role-decoupling. Furthermore, counterfactual reasoning is crucial for evaluating cognitive boundaries, as demonstrated in undercover gaming (Liang et al., 2026), clinical diagnostics (You et al., 2026), visual QA (Goyal et al., 2019), and explainability (Qu et al., 2024). Inspired by these, ASTRO leverages a decoupled multi-agent system to explicitly perform factual verification and counterfactual probing for complex video scenes.

Tree Search for Data Synthesis and Inference Scaling. Navigating the complex reasoning space of ToM requires deliberate search strategies. The pioneering Tree of Thoughts (ToT) framework (Yao et al., 2023a) first introduced tree-based search for LLM reasoning, enabling deliberate problem-solving through systematic exploration. Monte Carlo Tree Search (MCTS) has been extensively integrated with LLMs for complex reasoning and synthetic data generation (Xie et al., 2024a). Recent advancements focus on refining search efficiency and data quality. For instance, PRISM-MCTS

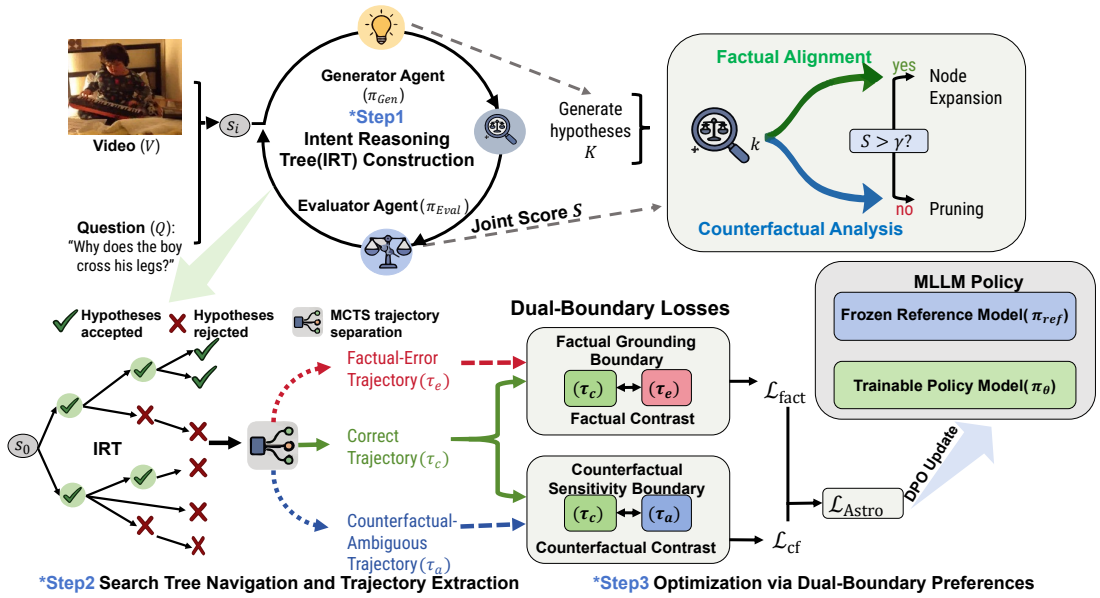


Figure 2: **The overall architectural workflow of the proposed ASTRO framework.** ASTRO conceptualizes video intention understanding as an optimization path-finding process on an Intention Reasoning Tree (IRT). **Step 1: Intent Reasoning Tree Construction.** Driven by a cooperative multi-agent system, the framework constructs an IRT utilizing step-level dual-axis evaluation. **Step 2: Search Tree Navigation and Trajectory Extraction.** The full tree search trajectories are navigated and systematically extracted into multi-level preference pairs based on their semantic validity via Monte Carlo Tree Search (MCTS). **Step 3: Optimization via Dual-Boundary Preferences.** Preference alignment is executed to permanently internalize both factual grounding and causal sensitivity into the LVLM parameters.

(Li et al., 2025) introduces reflective memory to share heuristics globally, while DITS (Shi et al., 2025a) and OPTIMA (Chen et al., 2025) emphasize data influence and token efficiency, proving that MCTS is highly effective in extracting high-quality multi-agent reasoning trajectories. Graph of Thoughts (GoT) (Besta et al., 2024) further extends this paradigm by allowing non-linear reasoning structures. In this work, we deploy MCTS as a synthetic data engine to explicitly extract contrastive reasoning paths—specifically segregating correct, factual-error, and counterfactual-ambiguous trajectories from the intention reasoning tree.

Multimodal Preference Optimization. Direct Preference Optimization (DPO) (Rafailov et al., 2023) aligns language models but faces multimodal challenges like visual condition ignorance (Wang et al., 2024; Xie et al., 2024b). Inspired by LIMA (Zhou et al., 2023), we target strong alignment with minimal data. To mitigate visual hallucinations, methods like SymMPO (Liu et al., 2026), JoMPO (Kim and Yoon, 2026), and DA-DPO (Qiu et al., 2026) introduce fine-grained constraints and difficulty-aware weighting. For video tasks, Video-DPO (Zhang et al., 2025) uses LM rewards, and CounterVid (Poppi et al., 2026) generates synthetic pairs to mitigate temporal hallucinations. Building on these, ASTRO constructs high-difficulty preference pairs directly from agentic MCTS paths, explicitly injecting dual-boundary sensitivity via DPO for data-efficient alignment.

3 ASTRO: Agentic Search Tree Reasoning Optimization

3.1 Problem Formulation and ASTRO Overview

Given a video clip $\mathcal{V}$ and a textual question $\mathcal{Q}$ regarding human intentions, the goal is to select the correct explanation A^* from an option pool $\mathcal{A} = \{a_1, \ldots, a_N\}$. Traditional paradigms directly estimate the single-pass probability $P_\theta(A \mid \mathcal{V}, \mathcal{Q})$, forcing a massive single-step cognitive leap that triggers factual hallucinations and counterfactual blindness.

To overcome this, **ASTRO** reformulates intention inference into a multi-step cognitive exploration over an **Intention Reasoning Tree (IRT)** $\mathcal{T}$. We define its key components as follows: **State** ($s_t = \{\mathcal{V}, \mathcal{Q}, \mathcal{E}_t\}$) includes the initial context and the sequence of verified historical evidence $\mathcal{E}_t$ up to step t; **Action** (a_t) represents an intermediate cognitive hypothesis step; and **Reasoning Trajectory**

($\tau = (s_0, a_0, \ldots, A)$) denotes a complete path culminating in a final prediction $A \in \mathcal{A}$.

Under this formulation, the optimization of ASTRO shifts from token-level cloning to trajectory-level preference alignment via a two-stage paradigm: **(1) Navigate and Extract:** MCTS navigates $\mathcal{T}$ using dual-axis step-level rewards to isolate three distinct trajectories: correct (τ_c), factual-error (τ_e), and counterfactual-ambiguous (τ_a). **(2) Align and Optimize:** The policy π_θ is optimized via Direct Preference Optimization (DPO) over contrastive pairs (τ_c, τ_e) and (τ_c, τ_a) to internalize factual grounding and causal sensitivity.

3.2 Agentic Reasoning Space Construction

To construct a structured reasoning space, ASTRO deploys a collaborative two-agent system that iteratively expands the IRT $\mathcal{T}$ via combined factual and counterfactual verification.

3.2.1 Hypothesis Generation via Generator Agent

At step t, given state s_t, the **Generator Agent** (π_{Gen}) performs divergent causal reasoning to propose M candidate explanatory hypotheses $\mathcal{K}_t = \{k_t^1, \ldots, k_t^M\}$:

$$k_t^m \sim \pi_{Gen}(\cdot \mid \mathcal{V}, \mathcal{Q}, \mathcal{E}_t), \quad m \in \{1, \ldots, M\}. \tag{1}$$

3.2.2 Dual-Axis Evaluation via Evaluator Agent

The **Evaluator Agent** seamlessly integrates the textual critiques and their corresponding normalized scalar rewards ($r_{\text{fact}}^m, r_{\text{cf}}^m \in [0, 1]$) across dual axes. Formally, the outputs are evaluated via contrastive cognitive dimensions as:

$$(c_{\text{fact}}^m, r_{\text{fact}}^m) = \pi_{Eval}\left(\text{fact} \mid \mathcal{V}, \mathcal{Q}, \mathcal{E}_t, k_t^m\right), \tag{2}$$

$$(c_{\text{cf}}^m, r_{\text{cf}}^m) = \pi_{Eval}\left(\text{cf} \mid \mathcal{V}, \mathcal{Q}, \mathcal{E}_t, k_t^m\right), \tag{3}$$

where π_{Eval} produces the qualitative rationales and quantitative scores for factual correctness and counterfactual robustness, respectively.

We define a joint score function $\mathcal{S}(k_t^m) = r_{\text{fact}}^m \cdot r_{\text{cf}}^m$. A candidate hypothesis is promoted to a verified evidence block e_{t+1} and spawns a new child state s_{t+1} if and only if its score surpasses a threshold γ:

$$e_{t+1} = k_t^m \iff \mathcal{S}(k_t^m) > \gamma. \tag{4}$$

The state then transitions to $s_{t+1} = \{\mathcal{V}, \mathcal{Q}, \mathcal{E}_t \cup \{e_{t+1}\}\}$. Detailed prompt templates for both agents are relegated to Appendix B.

3.3 Search Tree Navigation and Trajectory Extraction

With the state-action-reward space formalized via the collaborative two-agent system, exhaustively exploring the exponentially expanding Intention Reasoning Tree (IRT) $\mathcal{T}$ becomes computationally intractable. To achieve deliberate and scalable cognitive exploration, ASTRO deploys Monte Carlo Tree Search (MCTS) to strategically navigate $\mathcal{T}$. Instead of generating an intention explanation in a single pass, MCTS treats the inference process as an optimal path-finding problem guided by step-level rewards synthesized across the multi-agent hierarchy.

3.3.1 MCTS Iteration in Intention Space

Each complete MCTS iteration consists of four sequential phases: Selection, Expansion, Evaluation, and Backpropagation, dynamically driven by the multi-agent interaction.

Selection. Starting from the root node s_0, the search policy traverses down the existing tree by recursively choosing the most promising reasoning action a_t that maximizes the Upper Confidence Bound for Trees (UCT) score $U(s_t, a)$, defined as:

$$U(s_t, a) = Q(s_t, a) + c_{\text{puct}} \cdot \frac{P(a \mid s_t)\sqrt{N(s_t)}}{1 + N(s_t, a)}. \tag{5}$$

The optimal action a_t is then selected as follows:

$$a_t = \arg\max_{a \in \mathcal{A}(s_t)} U(s_t, a), \tag{6}$$

where $Q(s_t, a)$ represents the accumulated action-value, $N(s_t, a)$ is the visit count, and $P(a \mid s_t)$ is the prior probability, and c_{puct} is a hyperparameter balancing exploitation and exploration.

Expansion. When a selected action leads to a leaf node or an unexpanded state s_t, the framework triggers the node expansion mechanism. The Generator Agent π_{Gen} is invoked to sample M distinct intermediate candidate hypotheses $\mathcal{K}_t = \{k_t^1, k_t^2, \ldots, k_t^M\}$ according to Equation 1, creating M potential child edges branching from s_t.

Evaluation. Once the candidate actions are spawned, the Evaluator Agent π_{Eval} immediately performs the dual-axis factual and counterfactual qualitative audit defined in Equation 2 and Equation 3. The instantaneous joint reward for the newly expanded node is computed as $\mathcal{S}(k_t^m) = r_{\text{fact}}^m \cdot r_{\text{cf}}^m$.

Backpropagation. The joint reward $\mathcal{S}$ obtained from the Evaluator Agent π_{Eval} is propagated backward along the path traversed in the current iteration. For each state-action pair (s, a) on the path, the visit counts and action-values are updated as follows:

$$N(s, a) \leftarrow N(s, a) + 1, \tag{7}$$

$$Q(s, a) \leftarrow \frac{(N(s, a) - 1) \cdot Q(s, a) + \mathcal{S}}{N(s, a)}. \tag{8}$$

3.3.2 Contrastive Trajectory Segregation

The core utility of navigating the IRT via MCTS extends beyond simply identifying the correct final answer at inference time; it serves as an online synthetic data engine that collects highly discriminative reasoning trajectories for subsequent preference learning. Upon completing a predefined number of search rollouts, ASTRO systematically segregates three distinct categories of paths across the traversed tree space based on the step-level feedback from the Evaluator Agent π_{Eval}:

1. **Correct Trajectory (τ_c):** This path corresponds to the sequence of reasoning states and actions that achieves the highest cumulative Q-value and successfully matches the ground-truth option A^*. It reflects a flawless cognitive chain where each step is both visually grounded and causally robust:

$$\tau_c = \arg \max_{\tau \in \mathcal{T}, A = A^*} \sum_{(s,a) \in \tau} Q(s, a). \tag{9}$$

2. **Factual-Error Trajectory (τ_e):** This trajectory is extracted from paths that terminate at incorrect answer options by minimizing the step-level factual scores r_{fact} provided by π_{Eval}. It precisely isolates the specific reasoning forks where the model suffers from severe factual hallucinations and fabricates ungrounded visual assumptions:

$$\tau_e = \arg \min_{\tau \in \mathcal{T}, A \neq A^*} \sum_{(s,a) \in \tau} r_{\text{fact}}(s, a). \tag{10}$$

3. **Counterfactual-Ambiguous Trajectory (τ_a):** This trajectory tracks reasoning paths that may appear superficially aligned with the visual scene but are inherently fragile under causal perturbations. It is identified by minimizing the cumulative counterfactual rewards r_{cf} from π_{Eval}:

$$\tau_a = \arg \min_{\tau \in \mathcal{T}, A \neq A^*} \sum_{(s,a) \in \tau} r_{\text{cf}}(s, a). \tag{11}$$

By extracting τ_e and τ_a relative to τ_c, ASTRO constructs customized, on-policy hard negatives that precisely mirror the model's inner cognitive flaws, paving the way for targeted preference alignment.

3.4 Optimization via Dual-Boundary Preferences

The ultimate goal of the ASTRO framework is to internalize the explicit, search-derived cognitive boundaries into the parametric weights of the MLLM. While traditional Supervised Fine-Tuning (SFT) solely maximizes the likelihood of the correct answer, it fails to teach the model *why* alternative paths are incorrect, leading to persistent hallucinations and cognitive blind spots. To this end, we employ Direct Preference Optimization (DPO) (Rafailov et al., 2023) on the high-quality contrastive trajectories extracted from the Intention Reasoning Tree (IRT) $\mathcal{T}$ via our multi-agent MCTS pipeline.

3.4.1 Factual and Counterfactual Alignment

Given the multi-modal context $x = \{\mathcal{V}, \mathcal{Q}\}$, standard DPO requires a preferred response y_w and a rejected response y_l. By leveraging the three distinct categories of trajectories isolated in §3.3, we construct two complementary sets of preference pairs to align the model across dual cognitive boundaries simultaneously:

Factual Grounding Alignment. We form the factual preference pair by setting $y_w = \tau_c$ and $y_l = \tau_e$. The trajectory τ_e represents a catastrophic factual failure (hallucination) generated by the model's own policy. By explicitly penalizing τ_e, the policy model is heavily discouraged from synthesizing fabricated ungrounded explanations.

Counterfactual Robustness Alignment. We establish the counterfactual preference pair by setting $y_w = \tau_c$ and $y_l = \tau_a$. The trajectory τ_a embodies a subtle logical vulnerability where the model succumbs to superficial spatial-temporal shortcuts. Aligning the policy against τ_a forces the weights to internalize precise counterfactual reasoning thresholds.

Dataset	#Train	#Test/Val	#Sampled
IntentQA	12,119	2,134	1,078
SIQ2	6,159	943	1,028
MIntRec2.0	6,165	2,033	1,101

Table 1: Statistics of the evaluated benchmarks and our sampled training sizes.

Let $\mathcal{R}_\theta(\tau \mid x) = \sum_{(s,a)\in\tau} \log \frac{\pi_\theta(a|s)}{\pi_{\text{ref}}(a|s)}$ define the implicit trajectory reward, and let $\Delta\mathcal{R}_\theta(\tau_i, \tau_j \mid x) = \mathcal{R}_\theta(\tau_i \mid x) - \mathcal{R}_\theta(\tau_j \mid x)$ denote the implicit reward margin between two competitive trajectories. We formulate our unified dual-boundary optimization loss, $\mathcal{L}_{\text{ASTRO}}$, as the joint optimization of the factual and counterfactual alignment objectives:

$$\mathcal{L}_{\text{ASTRO}}(\pi_\theta, \pi_{\text{ref}}) = \mathcal{L}_{\text{fact}}(\pi_\theta, \pi_{\text{ref}}) + \mathcal{L}_{\text{cf}}(\pi_\theta, \pi_{\text{ref}}), \tag{12}$$

where the individual boundary objectives are compactly formulated as:

$$\mathcal{L}_{\text{fact}} = -\mathbb{E}_{(x,\tau_c,\tau_e)} \left[\log \sigma \left(\beta\Delta\mathcal{R}_\theta(\tau_c, \tau_e \mid x)\right)\right], \tag{13}$$

$$\mathcal{L}_{\text{cf}} = -\mathbb{E}_{(x,\tau_c,\tau_a)} \left[\log \sigma \left(\beta\Delta\mathcal{R}_\theta(\tau_c, \tau_a \mid x)\right)\right], \tag{14}$$

where σ is the sigmoid function, and β is a hyperparameter controlling the divergence penalty from the reference policy. By jointly minimizing $\mathcal{L}_{\text{ASTRO}}$, ASTRO enforces strict factual and causal boundaries within a single data-efficient training pass.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets and Low-Resource Sampling

To comprehensively evaluate the video intention reasoning capabilities and the data efficiency of the proposed ASTRO framework, we conduct experiments on three challenging benchmarks: IntentQA (Li et al., 2023b), Social-IQ 2.0 (SIQ2), and MIntRec2.0 (Zhang et al., 2024). The detailed statistics of these datasets, including the number of video clips and the data splits, are summarized in Table 1.

Low-Resource Setting. A core objective of our framework is to break the reliance on massive annotated data. Therefore, rather than utilizing the full training sets, we strategically sample a minuscule fraction to construct our Intention Reasoning Trees (IRT). As shown in Table 1, we utilize approximately $1/11$ of the original IntentQA training set, and merely $1/6$ for both SIQ2 and MIntRec2.0. This low-resource setting rigorously validates the immense Fisher information gain provided by the trajectory-level hard negatives extracted via MCTS.

4.1.2 Baselines

We benchmark our framework against state-of-the-art methods tailored for each dataset. For IntentQA, we compare against **Magma** (Yang et al., 2025) and **CDR** (Li and Zuo, 2026), which represent strong reasoning-augmented architectures. For SIQ2, we include the powerful proprietary method **LVD** (Mori et al., 2025). For MIntRec2.0, we select **Listen Then See** (Agrawal et al., 2024) and **DialogGraph-LLM** (Liu et al., 2025) as competitive baselines. Across all benchmarks, **Accuracy (ACC %)** is uniformly adopted as the primary evaluation metric.

4.1.3 Implementation Details

Our implementation strictly separates the offline trajectory generation phase from the policy optimization phase.

MCTS Trajectory Generation. The agents orchestrating the reasoning space construction are instantiated using **InternVL2.5-8B** (Chen et al., 2024c). During expansion, the maximum candidate width M generated by the Generator Agent is dynamically aligned with the number of options in the dataset: $M = 5$ for IntentQA and $M = 4$ for SIQ2. For MIntRec2.0, which approximates a classification task, we employ a zero-shot prompt to pre-select the top-4 most probable categories, setting $M = 4$. The MCTS exploration hyperparameter is set to $c_{\text{puct}} = 1.4$. For each sample, the search conducts 20 rollouts with a maximum tree depth of 3.

Preference Optimization. The target policy model to be optimized is **InternVL3-8B** (Zhu et al., 2025). We utilize the `modelscope-swift` framework for all fine-tuning stages. To prevent formatting collapse during DPO, we first conduct a format-alignment Supervised Fine-Tuning (SFT) for 2 epochs using solely the extracted Correct Trajectories (τ_c). Subsequently, we apply the dual-boundary DPO in two successive stages: a factual alignment DPO utilizing the (τ_c, τ_e) pairs, followed by a counterfactual alignment DPO utilizing the (τ_c, τ_a) pairs.

4.2 Main Results

We evaluate the comprehensive video intention reasoning capabilities of our proposed ASTRO framework against existing state-of-the-art methods. The overall experimental results, measured in terms of Accuracy (ACC %), are summarized in Table 2.

Method	IntentQA	SIQ2	MIntRec2.0
CDR	72.10	–	–
Magma	88.60	62.99†	65.02†
LVD	–	75.20	–
Listen Then See	76.99†	82.10	60.10†
IntPro	–	–	67.37
DialogGraph-LLM	–	–	70.90
ASTRO (Ours)	**90.01**	**83.56**	**71.47**

Table 2: Main results on IntentQA, Social-IQ 2.0 (SIQ2), and MIntRec2.0 benchmarks. Performance is reported in Accuracy (ACC %). Results marked with † indicate our own reproduced evaluation on the respective benchmark, as the original papers did not report results there. The best results are highlighted in bold. All reproduced models were trained using the full dataset.

As demonstrated in Table 2, ASTRO consistently outperforms all baseline methods across all three evaluated benchmarks, establishing new state-of-the-art performances and validating the efficacy of our trajectory-level preference learning paradigm.

Analysis on IntentQA. On the IntentQA benchmark, which heavily tests context-aware causal and temporal intention reasoning, ASTRO achieves a remarkable accuracy of 90.01%. This performance significantly surpasses the strong baseline Magma by 1.41% and drastically outperforms CDR by 17.91%. This substantial leap indicates that performing optimization over the tree-search-extracted factual-error trajectories (τ_e) effectively penalizes ungrounded visual hallucinations. By learning from these on-policy hard negatives, the MLLM successfully anchors its explanations to verifiable spatiotemporal visual evidence.

Analysis on Social-IQ 2.0. For the Social-IQ 2.0 (SIQ2) dataset, which demands sophisticated social intelligence and Theory of Mind (ToM) modeling, ASTRO establishes a top-tier accuracy of 83.56%. It yields a 1.46% improvement over Listen Then See and remarkably beats the LLM-driven baseline LVD by 8.36%. This prominent advancement directly highlights the impact of our counterfactual sensitivity alignment. By explicitly penalizing counterfactual-ambiguous trajectories (τ_c), ASTRO empowers the model to delineate fine-grained cognitive boundaries, enabling it to decode subtle facial expressions, micro-actions, and contextual shifts within intricate human interactions.

Analysis on MIntRec2.0. On the fine-grained multimodal intent recognition task of MIntRec2.0, ASTRO achieves an accuracy of 71.47%, outperforming the highly competitive IntPro (70.09%) and beating DialogGraph-LLM by a wide margin of 4.10%. The steady gains in multi-party conversational contexts demonstrate that our trajectory-level preference optimization generalizes robustly across diverse intention understanding scenarios, transitioning seamlessly from interactive VQA to structural intent classification tasks.

4.3 Ablation Studies and Analysis

We conduct ablation studies to analyze the contribution of each component in ASTRO, including the optimization modules, data efficiency, and MCTS iteration dynamics.

4.3.1 Effect of Core Optimization Modules

We evaluate different variants of our trajectory-level preference optimization pipeline on IntentQA using InternVL3-8B as the backbone. Results are shown in Table 3.

Model Configuration	IntentQA
InternVL3-8B (Zero-shot Baseline)	76.99
+ SFT (Full 12,119 data)	85.05
+ SFT (Sampled 1,078 data, format only)	79.00
+ DPO w/o SFT (No format alignment)	79.89
+ DPO w/o Counterfactual Pair (τ_a)	85.09
+ DPO w/o Factual Pair (τ_e)	83.08
ASTRO (Full 1,078 data)	**90.01**

Table 3: Ablation of optimization modules on IntentQA.

The results lead to three observations: **(1) Necessity of SFT Alignment:** DPO without format-alignment SFT achieves only 79.89%, while SFT alone on the sampled 1K data reaches 79.00%, showing that DPO is the major source of performance gains. **(2) Importance of Dual Boundaries:** Removing factual pairs or counterfactual pairs reduces accuracy to 83.08% and 85.09%, respectively, demonstrating that both grounded reasoning and logical robustness are essential. **(3) ASTRO vs. Full SFT:** ASTRO trained with only 1,078 samples achieves 90.01%, outperforming full-data SFT (85.05%) by nearly 5 points, highlighting the effectiveness of trajectory-level hard negatives.

4.3.2 Data Scaling and Efficiency

We further study the impact of training data scale on IntentQA. Results are summarized in Figure 3.

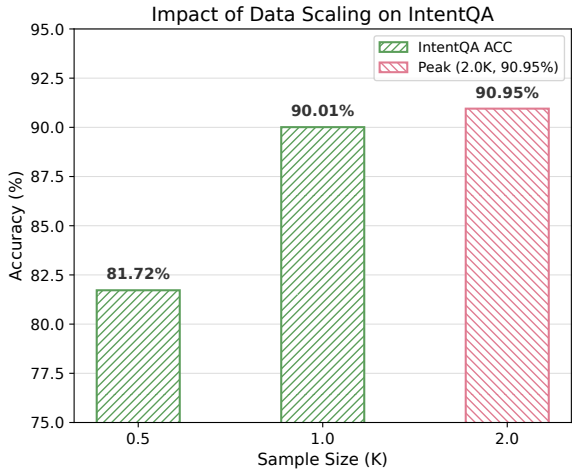


Figure 3: Impact of training data scaling on IntentQA.

Increasing the training size from 1.0K to 2.0K yields only a marginal improvement (+0.94%) while doubling the MCTS trajectory construction cost. Therefore, we identify ∼1.0K samples as the best trade-off between performance and efficiency, demonstrating the strong data efficiency of ASTRO.

4.3.3 Dynamics of MCTS Iterations

We analyze the effect of MCTS rollout iterations on trajectory construction efficiency using SIQ2. The average time cost per path is shown in Figure 4.

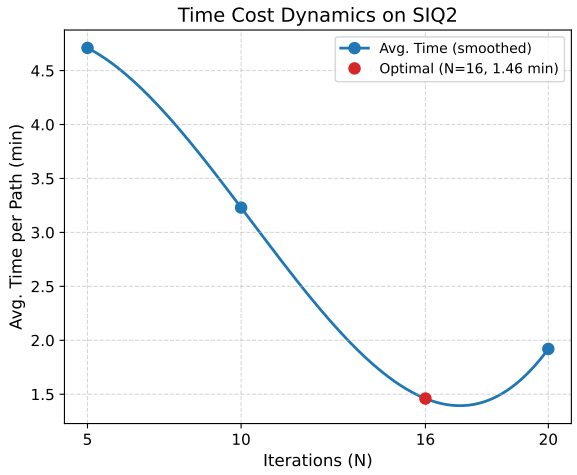


Figure 4: Time cost dynamics of MCTS iterations on SIQ2.

The results exhibit a non-linear U-shaped trend. With very few iterations ($N = 5$), the search frequently fails to find valid reasoning paths, causing repeated re-sampling overhead. In contrast, larger iteration counts ($N \geq 20$) increase tree complexity and node evaluation cost. We finally adopt $N = 20$, which balances computational efficiency (1.92 min/path) and trajectory quality for downstream DPO training.

5 Conclusion

In this paper, we introduced **ASTRO** (**A**gentic **S**earch **T**ree **R**easoning **O**ptimization), a data-efficient framework designed to elevate deep video intention understanding in Multimodal Large Language Models (MLLMs). ASTRO tackles the persistent challenges of factual hallucinations and counterfactual blindness by reformulating black-box video prediction into an explorable, search-guided cognitive path-finding process.

Central to our approach, a collaborative multi-agent system driven by Monte Carlo Tree Search (MCTS) constructs Intention Reasoning Trees through rigorous dual-axis evaluations. ASTRO dynamically extracts contrastive reasoning trajectories—specifically, correct, factual-error, and counterfactual-ambiguous paths—which are subsequently optimized via a Direct Preference Optimization (DPO) pipeline to internalize precise visual grounding and causal sensitivity.

Extensive evaluations on IntentQA, Social-IQ 2.0, and MIntRec2.0 demonstrate that ASTRO consistently establishes new state-of-the-art performance using merely a minuscule fraction of the training data. Ultimately, this work catalyzes a paradigm shift from token-level behavioral cloning toward trajectory-level cognitive preference optimization under data-constrained conditions. Future research will explore accelerating MCTS rollouts, extending to long-form narratives, and deploying in real-time embodied AI applications.

Limitations

While data-efficient, ASTRO presents three limitations. First, offline MCTS tree construction is computationally heavy (∼1.9 mins/trajectory), hindering scalability. Second, synthesized trajectory quality relies strictly on the base MLLM's visual and instruction-following capabilities. Finally, its complex multi-agent architecture is computationally overkill for standard, low-level action recognition tasks.

Ethical Considerations

This paper presents work intended to advance the field of Artificial Intelligence. While we acknowledge the potential societal implications inherent to

such research, we believe no specific consequences arising directly from our contributions require explicit highlighting here.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62276018, the Beijing Nova Program under Grant 20230484297, and the Fundamental Research Funds for the Central Universities.

References

Aviral Agrawal, Carlos Mateo Samudio Lezcano, Iqui Balam Heredia-Marin, and Prabhdeep Singh Sethi. 2024. Listen then see: Video alignment with speaker attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2018–2027.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. 2024a. Step-level value preference optimization for mathematical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7889–7903.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, and 1 others. 2024b. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *International Conference on Learning Representations*, volume 2024, pages 20094–20136.

Weize Chen, Jiarui Yuan, Chen Qian, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2025. Optima: Optimizing effectiveness and efficiency for llm-based multi-agent system. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11534–11557.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024c. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR.

Jiwon Kim and Hyunsoo Yoon. 2026. Joint multimodal preference optimization for fine-grained visual-textual alignment. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 79–94.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems*, 36:51991–52008.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. 2023b. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Nanxi Li, Zhengyue Zhao, G Edward Suh, Marco Pavone, and Chaowei Xiao. 2025. Prism: Robust vlm alignment with principled reasoning for integrated safety in multimodality. *arXiv preprint arXiv:2508.18649*.

Xuan Li and Haoran Zuo. 2026. Contextualized diverse reasoning: Enhancing video question answering with multi-perspective mllm pathways. *Preprints.org*.

Yian Li, Wentao Tian, Yang Jiao, Jingjing Chen, and Yu-Gang Jiang. 2024. Eyes can deceive: Benchmarking counterfactual reasoning abilities of multi-modal large language models. *ArXiv*, abs/2404.12966.

Dayong Liang, Xiao-Yong Wei, and Changmeng Zheng. 2026. Multi-agent undercover gaming: Hallucination removal through counterfactual test for multimodal reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 6807–6815.

Jianxin Liang, Xiaojun Meng, Huishuai Zhang, Yueqian Wang, Jiansheng Wei, and Dongyan Zhao. 2025. Reasvqa: Advancing videoqa with imperfect reasoning process. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1696–1709.

HongYu Liu, Junxin Li, Changxi Guo, Hao Chen, Yaqian Huang, Yifu Guo, Huan Yang, and Lihua Cai. 2025. Dialoggraph-llm: Graph-informed llms for end-to-end audio dialogue intent recognition. *arXiv preprint arXiv:2511.11000*.

Wenqi Liu, Xuemeng Song, Jiaxi Li, Yinwei Wei, Na Zheng, Jianhua Yin, and Liqiang Nie. 2026. Mitigating hallucination through theory-consistent symmetric multimodal preference optimization. *Advances in Neural Information Processing Systems*, 38:111259–111284.

Erika Mori, Yue Qiu, Hirokatsu Kataoka, and Yoshimitsu Aoki. 2025. Leveraging llms with iterative loop structure for enhanced social intelligence in video question answering. *arXiv preprint arXiv:2503.21190*.

Lixing Niu, Jiapeng Li, Xingping Yu, Shu Wang, Ruining Feng, Bo Wu, Ping Wei, Yisen Wang, and Lifeng Fan. 2025. R3-vqa: "read the room" by video social reasoning. *arXiv preprint arXiv:2505.04147*.

Taiying Peng, Jiacheng Hua, Miao Liu, and Feng Lu. 2026. In the eye of mllm: Benchmarking egocentric video intent understanding with gaze-guided prompting. *Advances in Neural Information Processing Systems*, 38.

Tobia Poppi, Burak Uzkent, Amanmeet Garg, Lucas Porto, Garin Kessler, Yezhou Yang, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara, and Florian Schiffers. 2026. Countervid: Counterfactual video generation for mitigating action and temporal hallucinations in video-language models. *arXiv preprint arXiv:2601.04778*.

Longtian Qiu, Shan Ning, Chuyu Zhang, Jiaxuan Sun, and Xuming He. 2026. Da-dpo: Cost-efficient difficulty-aware preference optimization for reducing mllm hallucinations. *arXiv preprint arXiv:2601.00623*.

Zhan Qu, Daniel Gomm, and Michael Färber. 2024. Cody: Counterfactual explainers for dynamic graphs. *arXiv preprint arXiv:2403.16846*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Wentao Shi, Zichun Yu, Fuli Feng, Xiangnan He, and Chenyan Xiong. 2025a. Efficient multi-agent system training with data influence-oriented tree search. *arXiv preprint arXiv:2502.00955*.

Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. 2025b. Enhancing video-llm reasoning via agent-of-thoughts distillation. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, pages 8523–8533.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024. mdpo: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8078–8088.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yuheng Wu, Jianwen Xie, Denghui Zhang, and Zhaozhuo Xu. 2025. Del-tom: Inference-time scaling for theory-of-mind reasoning via dynamic epistemic logic. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11394–11408.

Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P Lillicrap, Kenji Kawaguchi, and Michael Shieh. 2024a. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024b. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13258–13273.

Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, and 1 others. 2025. Magma: A foundation model for multimodal ai agents. In *Proceedings of the computer vision and pattern recognition conference*, pages 14203–14214.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Yinsheng Yao, Jiehao Tang, Zhaozhen Yang, and Dawei Cheng. 2026. Maven: Multi-agent verification-elaboration network with in-step epistemic auditing. *arXiv preprint arXiv:2605.07646*.

Zhiwen You, Xi Chen, Aniket Vashishtha, Simo Du, Gabriel Erion-Barner, Hongyuan Mei, Hao Peng, and Yue Guo. 2026. Improving clinical diagnosis with counterfactual multi-agent reasoning. *arXiv preprint arXiv:2603.27820*.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Wenrui Li, Yanting Chen, and 1 others. 2024. Mintrec2. 0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. *arXiv preprint arXiv:2403.10943*.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander G Hauptmann, Yonatan Bisk, and 1 others. 2025. Direct preference optimization of video large multimodal models from language model reward. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 694–717.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Evaluation Benchmarks and Dataset Details

In this section, we provide comprehensive descriptions of the three video intention understanding benchmarks evaluated in our experiments.

IntentQA : IntentQA is a large-scale video question-answering benchmark explicitly tailored for human intention understanding. It requires models to perform non-trivial temporal and causal reasoning over diverse human actions. The dataset comprises 4,426 open-source web videos with 16,297 multiple-choice question-answer pairs. Each question is accompanied by 5 candidate options, where only one is factually correct. The evaluation primarily relies on top-1 Accuracy (ACC %).

Social-IQ 2.0 (SIQ2): Social-IQ 2.0 is a benchmark dedicated to assessing social intelligence and Theory of Mind (ToM) in multimodal contexts. It contains complex real-world social scenarios (*e.g.*, debates, group negotiations, vlogs) where the characters' intentions are deeply implicit. Models must decode subtle social cues, facial expressions, micro-actions, and multi-party acoustic-visual interactions. The dataset contains over 1,000 videos from youtube with over 6,000 multiple-choice questions. Each question is paired with 4 multiple-choice options, and the performance is evaluated using standard Accuracy (ACC %).

MIntRec2.0 : Unlike the interactive VQA format of IntentQA and SIQ2, MIntRec2.0 is a fine-grained multimodal intent recognition benchmark formulated as a structural classification task. It contains 1,245 high-quality video clips extracted from multi-party conversational dramas, spanning 30 distinct fine-grained intent categories (*e.g.*, *complaining, comforting, threatening*). The performance is measured via Accuracy (ACC %) and macro F1-score across all 30 classes.

B Agentic Search Tree Construction Details

In this section, we provide the comprehensive implementation details, exact system prompts, and hyperparameter configurations of the collaborative two-agent system (§3.2) used to construct the Intention Reasoning Tree (IRT) via MCTS.

B.1 Hyperparameter Configurations

To ensure complete reproducibility, the exact hyperparameters governing the online search and node evaluation phases across IntentQA, Social-IQ 2.0, and MIntRec2.0 are summarized in Table 4. Both the Generator Agent (π_{Gen}) and the Evaluator Agent (π_{Eval}) are instantiated using the pre-trained **InternVL2.5-8B** backbone.

Hyperparameter	**Value**
Base MLLM Backbone	InternVL2.5-8B
Generation Temperature	0.2
Generation Top-p	0.9
MCTS Rollout Iterations (N)	20
Maximum Tree Depth (T)	3
Exploration Constant (c_{puct})	1.4
Confidence Threshold (γ)	0.6
Hypotheses per Expansion	3
Factual Aggregation Weight (W_{fact})	0.7
Counterfactual Aggregation Weight (W_{cf})	0.3

Table 4: Detailed hyperparameters for MCTS-driven Intention Reasoning Tree construction based on our implementation.

B.2 Verbatim System Prompts

The precise interactions within the dual-agent framework are guided by highly structured system prompts. Based on our implementation pipeline, the qualitative audit and quantitative scoring are decoupled into specific role-playing contexts to prevent formatting collapse.

Table 5 presents the verbatim prompt for the **Generator Agent**, which acts as a *Causal Hypothesis Planner* to propose divergent hypotheses. Table 6 presents the prompts for the **Evaluator Agent**. To ensure rigorous dual-axis critique, the Evaluator executes two distinct evaluation passes: a *Factual Verification* pass assessing appearance, action, and state, followed by a *Counterfactual Refutation* (Devil's Advocate) pass assessing visual contradictions and competitive superiority.

System Prompt: Generator Agent (π_{Gen})

Role: [You are a Causal Hypothesis Planner, acting as a top-tier video reasoning analyst.]
Task: Produce {n_hypotheses} short, concrete hypotheses that would SUPPORT the candidate answer: "<OPTION>".
Video context: <VIDEO_CONTEXT>
Question: <QUESTION>
Instructions: Keep each hypothesis concise (1-2 sentences), observable or inferable from the video frames.

[Output Format]:
You must return only a JSON object, formatted as follows:

```
{
 "hypotheses": [
 "Hypothesis 1...",
 "Hypothesis 2...",
 "Hypothesis 3..."
 ]
}
```

Table 5: Verbatim system prompt for the Hypothesis Generator Agent (π_{Gen}).

C MCTS Construction of Intention Reasoning Tree Inference Paths

In this section, we detail the operational pipeline of the Monte Carlo Tree Search (MCTS) utilized to construct the inference paths within the Intention Reasoning Tree (IRT). The overall algorithmic workflow is illustrated in Algorithm 1. The IRT is designed to systematically explore the cognitive space of candidate intents and break down high-level hypotheses into verifiable sub-steps.

C.1 Taxonomy of Inference Paths

During the stochastic exploration of the IRT, the collaborative dual-agent framework evaluates and classifies the traversed trajectories into three distinct categories based on their semantic validity and dual-axis scores. This classification serves as the foundational data filter for the downstream preference alignment phase.

- **Correct Path ($\mathcal{P}_{corr}$):** The correct path represents the reasoning trajectory that successfully converges on the ground-truth intention option. It is characterized by high factual grounding ($S_{pos} \geq 0.5$) and robust resistance against counterfactual refutation ($S_{neg} \geq 0.60$). In our data pipeline, $\mathcal{P}_{corr}$ is extracted to formulate the *Chosen Response* (y_{chosen}) for DPO, encapsulating pristine causal logic and flawless multi-frame visual verification.

- **Error Path ($\mathcal{P}_{err}$):** An error path occurs when the search branches into an incorrect candidate option that is easily debunked by either a low factual verification score or high vulnerability to counterfactual falsification. Specifically, we extract specialized error trajectories where the agent uncovers contradicting evidence but still stumbles into a wrong decision. These paths are utilized to construct the *Rejected Response* ($y_{rejected}$) in DPO, explicitly training the policy model to penalize superficial visual heuristics.

- **Ambiguous Path ($\mathcal{P}_{amb}$):** Ambiguous paths typically arise from the inherent semantic overlap among complex human intentions (*e.g.*, `Inform` vs. `Introduce`). These trajectories often score a high macro-level factual plausibility (S_{pos}) based on macro-visual scenes, but fail to maintain logical consistency during the rigorous counterfactual refutation pass. Consequently, their final composite reward R falls below the quality threshold ($\gamma_{min} = 0.60$). These paths are systematically pruned and discarded by our quality failsafe to eliminate gradient confusion and guarantee data purity during alignment.

C.2 State Representation and Initialization

Each node v in the IRT represents either a macro-hypothesis (a candidate option provided in the dataset) or a micro-hypothesis (a specific visual verification step). A node stores its visit count $N(v)$, cumulative reward $Q(v)$, factual score S_{pos}, counterfactual score S_{neg}, and the extracted textual

evidence. During initialization, the root node v_{root} is expanded to create the first layer of the tree, consisting of the predefined multiple-choice candidate options (*e.g.*, A, B, C, D).

C.3 The Four Phases of MCTS

The tree is dynamically constructed through a defined number of simulations ($N = 20$). Each simulation iteration consists of four canonical phases:

1. Selection (UCT Strategy): Starting from the root, the algorithm traverses down the tree. At each fully expanded node, the selection policy is governed by the Upper Confidence Bound applied to Trees (UCT) formula:

$$UCT(v) = \frac{Q(v)}{N(v)} + c_{puct} \cdot \sqrt{\frac{\ln N(v_{parent})}{N(v)}}, \quad (15)$$

where c_{puct} is the exploration constant set to 1.414. The algorithm selects the child node that maximizes the UCT value until it reaches a leaf node that is not fully expanded or has reached the maximum depth ($T_{max} = 3$).

2. Expansion (Causal Planning): If the selected node v_{leaf} has been visited before and is not a terminal node, it is expanded. To generate children, the *Generator Agent* (π_{Gen}) receives the hierarchical context chain (from root to v_{leaf}) and proposes distinct, concrete "Verification Steps". These steps act as child nodes, probing specific spatial-temporal details necessary to confirm or deny the parent hypothesis.

3. Simulation (Dual-Axis Evaluation): Once a node is selected or newly expanded, the *Evaluator Agent* (π_{Eval}) conducts a rigorous dual-axis evaluation to calculate the reward R. The agent first scans the video to extract visual evidence specifically related to the node's hypothesis. Then, the agent assesses the appearance, action, and state to yield a factual plausibility score $S_{pos} \in [0, 1]$. If S_{pos} is extremely low (*i.e.*, $S_{pos} < 0.5$), the hypothesis is deemed factually implausible, and S_{neg} is aggressively pruned to 0. Otherwise, the Generator proposes hard-negative counterfactuals. The Evaluator attempts to falsify them, yielding a score S_{neg}. The final composite reward for the simulation is computed as the product of the dual boundaries: $R = S_{pos} \times S_{neg}$.

4. Backpropagation: The simulated reward R is backpropagated up the trajectory to the root. For each node v along the path, its visit count and cumulative value are updated as:

$$N(v) \leftarrow N(v) + 1, \quad Q(v) \leftarrow Q(v) + R. \quad (16)$$

C.4 Final Decision

After the designated number of simulations is exhausted, the optimal intent is determined by selecting the child node in the first layer with the highest visit count $N(v)$, which indicates the most heavily explored and validated hypothesis space.

Algorithm 1 Intention Reasoning Tree Path Construction via MCTS

Input: Video context $\mathcal{V}$, Question $\mathcal{Q}$, Candidate Options $\mathcal{O}$, Max Simulations K, Max Depth T_{max}
Output: Selected Intent O_{best}, Extracted Reasoning Trajectory

Initialize root node v_{root} with children corresponding to $O_i \in \mathcal{O}$
for $k = 1$ **to** K **do**
  $v \leftarrow v_{root}$
  /* 1. Selection Phase */
  while v is fully expanded **and** $v.depth < T_{max}$ **do**
    $v \leftarrow \arg\max_{v'} UCT(v')$
  end while
  /* 2. Expansion Phase */
  if $N(v) > 0$ **and** $v.depth < T_{max}$ **then**
    $ctx \leftarrow$ Trajectory path from v_{root} to v
    $\mathcal{S}_{steps} \leftarrow \pi_{Gen}(\mathcal{V}, \mathcal{Q}, ctx)$
    Append new child nodes from $\mathcal{S}_{steps}$ to v
    $v \leftarrow$ RandomChild(v)
  end if
  /* 3. Simulation Phase (Dual-Axis) */
  $\mathcal{E} \leftarrow$ ExtractEvidence$(\mathcal{V}, v.name)$
  $S_{pos} \leftarrow \pi_{Eval}$(Fact Check, $\mathcal{E}, v.name$)
  if $S_{pos} \geq 0.5$ **then**
    $\mathcal{C} \leftarrow \pi_{Gen}$(Counterfactuals, $v.name$)
    $S_{neg} \leftarrow$ Mean(π_{Eval}(Refute, $\mathcal{E}, c \in \mathcal{C}$))
  else
    $S_{neg} \leftarrow 0.0$
  end if
  $R \leftarrow S_{pos} \times S_{neg}$
  /* 4. Backpropagation Phase */
  while $v \neq$ null **do**
    $N(v) \leftarrow N(v) + 1$
    $Q(v) \leftarrow Q(v) + R$
    $v \leftarrow v.parent$
  end while
end for
return $\arg\max_{v \in v_{root}.children} N(v)$

D Inference and Evaluation Details

In this section, we outline the exact inference configurations, the structured Chain-of-Thought (CoT) prompting strategy, and the metric parsing mechanism used to evaluate the DPO-aligned policy model (**InternVL3-8B**) on the benchmark datasets.

D.1 Inference Hyperparameters and Decoding Strategy

During the evaluation phase, we aim to measure the model's deterministic reasoning capabilities without the variance introduced by probabilistic sampling. Therefore, we employ a **Greedy Decoding** strategy for all text generation tasks. The specific inference hyperparameters are set as follows:

- **Precision & Attention:** The model is loaded in `bfloat16` precision utilizing FlashAttention-2 for accelerated sequence processing.
- **Decoding Parameters:** Temperature is strictly set to $\tau = 0.0$ (`do_sample=False`), and the maximum number of newly generated tokens is capped at 1024.
- **Dense Temporal Sampling:** To capture fine-grained visual evidence during inference, we increase the uniform temporal sampling density to $T = 32$ frames per video clip, dynamically resized and processed via the identical spatial pipeline described in Appendix A.

D.2 Structured CoT Evaluation Prompt

To fully activate the dual-boundary cognitive capabilities instilled during the ASTRO-DPO training phase, we design a rigorous system prompt for inference. As shown in Table 7, this prompt enforces a 4-step Chain-of-Thought format: (1) Observation, (2) Verification Steps, (3) Reasoning (with mandatory counterfactual analysis), and (4) Final Answer.

D.3 Answer Parsing and Metric Calculation

To ensure rigorous and reproducible accuracy (ACC %) calculation, the final choice is extracted from the model's generated CoT text using strict Regular Expression (Regex) matching. We primarily search for the exact prefix pattern `r'Answer:\s*(a[0-4])'`. If the model occasionally omits the "Answer:" prefix due to format degeneration, a fallback boundary regex `r'\b(a[0-4])\b'` is applied to capture the terminal decision. Predictions that fail both parsing stages are strictly marked as incorrect (`error`).

E Qualitative Analysis of Failure Cases and Boundary Ambiguity

In this section, we provide a detailed qualitative analysis of a failure case from the MIntRec2.0 dataset to explicitly discuss the cognitive boundaries of Multimodal Large Language Models (MLLMs) and demonstrate the robustness of the ASTRO framework's quality failsafe mechanism.

E.1 Case Context and Search Disorientation

We analyze a specific video clip (`dia0_utt0.mp4`) where the ground-truth intent label is **Introduce**. As illustrated in Figure 5a, in this scene, a speaker in a boardroom announces: *"So without further ado, to welcome our newest part-time employee, Emma Dubanowski."*

During the Intention Reasoning Tree (IRT) construction, the MCTS explored four candidate intents: `Inform`, `Prevent`, `Greet`, and `Introduce`. According to the search logs, the model became disoriented and incorrectly selected `Inform` as the highest-value intention.

E.2 Root Causes of the Failure

Tracing the qualitative critiques of the Evaluator Agent, we identify two primary causes for this reasoning failure:

- **Inherent Semantic Overlap of Intents:** Human intentions are rarely mutually exclusive. The act of "introducing a new employee" inherently involves "informing" the team of new information. As shown in the upper portion of Figure 5b, the Evaluator Agent yielded a high factual verification score (0.70) for `Inform`, reasoning that *"The speaker's friendly facial expression, holding a clipboard, and the attentive audience support the hypothesis..."* This macro-semantic validity caused the MCTS to assign excessive confidence to `Inform` early in the search.

- **Visual Domination over Fine-grained Text:** Despite explicit textual cues in the dialogue (highlighted in green in Figure 5a, e.g., "welcome", "Emma Dubanowski") that strongly point to `Introduce`, the MLLM's joint reasoning was heavily biased by macro visual priors (*e.g.*,. the boardroom setting, the presentation posture). During the counterfactual scoring phase, the model failed to utilize the fine-grained text to draw a sharp cognitive boundary between `Inform` and `Introduce`.



Video Transcript (MIntRec2.0 - dia0_utt0)
Speaker: "So without further ado, to **welcome** our newest part-time employee, Emma Dubanowski."

(a) Spatiotemporal Context and Key Dialogue.

Agent B (Evaluator) - mode='fact'

```
{
  "winner_intent": "Inform",
  "factual_reasoning": "The speaker's friendly
    facial expression, holding a clipboard,
    and the attentive audience support the
    hypothesis...",
  "verification_score": 0.70
}
```

DPO Trajectory Extractor - Final Audit

```
WARNING - [Samples: 200/200]
 Winner logic is too weak (Score 0.53 < 0.60).
 Extracted 0 DPO pairs for this utterance.
 [Status]: Discarded.
```

(b) Curated Audit Logs. Although the factual score (0.70) exceeds the threshold, the weak counterfactual score (0.53 < 0.60) triggers the failsafe, discarding the sample.

Figure 5: Visualization of ASTRO's cognitive process on a failure case from MIntRec2.0 (`dia0_utt0.mp4`). (a) Shows the visual context and the explicit textual cues. (b) Presents the curated logs where the model incorrectly prefers `Inform` based on visual priors, but is ultimately stopped by the quality failsafe due to weak counterfactual refutation.

E.3 Robustness and Quality Failsafe of ASTRO

Although the MCTS failed to predict the correct intent, this sample perfectly validates the robustness of ASTRO as an offline synthetic data engine. At the trajectory extraction stage, the log explicitly records the data failsafe trigger (see the red bounding box in Figure 5b):

```
WARNING - [...] Winner logic is too weak (Score 0.53 < 0.60). Extracted 0 DPO pairs... [Status]: Discarded.
```

Because the `Inform` node could not withstand rigorous counterfactual refutation, its final aggregated logic score (0.53) fell below the stringent high-quality winner threshold ($\gamma_{min} = 0.60$). Consequently, the DPO Extractor triggered the **quality failsafe mechanism**, decisively discarding the sample. This ensures that only trajectories with sharp, reliable logical boundaries are forwarded to the downstream preference alignment phase, fundamentally safeguarding the purity of the DPO training.

F Showcase of Constructed DPO Preference Pairs

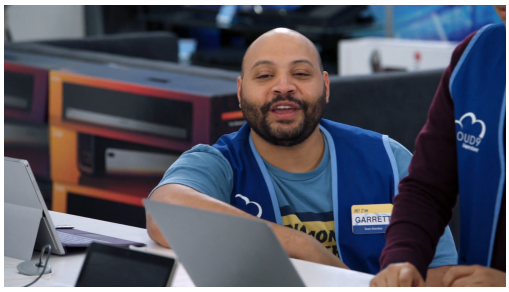
To intuitively demonstrate how the Intention Reasoning Tree (IRT) translates its stochastic search trajectories into highly effective preference data for Direct Preference Optimization (DPO), we present a comprehensive walkthrough of a verified success case from the MIntRec2.0 dataset (`dia89_utt8.mp4`). The holistic cognitive process and the resulting trajectory status are visualized in Figure 6.

As illustrated in Figure 6a, in this scene, a speaker sits in front of a laptop and casually suggests a joint action to a colleague: *"Hey, let's see if they have a page for buying a boogie board on there, huh?"* The ground-truth communicative intent is **Advise**.

During the trajectory extraction phase, the DPO Extractor leverages the taxonomy of inference paths defined in Appendix C to construct pairs with varying difficulty levels. Crucially, as audited in Figure 6b, the optimal node `Advise` achieves high scores across both dimensions ($S_{pos} = 0.70$, $S_{neg} = 0.92$), yielding a robust aggregated score of 0.64, which successfully bypasses the quality threshold ($\gamma_{min} = 0.60$). The rejected responses ($y_{rejected}$) are intentionally formulated to mimic a "lazy" reasoning model—one that makes superficial visual observations but ignores contradicting evidence, thereby jumping to a flawed conclusion.

F.1 Level 1: Factuality Penalty (Correct vs. Error Path)

The Level 1 DPO pair targets basic factuality and visual hallucinations, pairing the *Correct Path* ($\mathcal{P}_{corr}$) with an *Error Path* ($\mathcal{P}_{err}$). In this example, the MCTS explores the `Greet` intent. Although the speaker is actively talking, there are no physical greeting gestures (*e.g.*, waving, shaking hands), resulting in a low factual verification score ($S_{pos} = 0.30$).



Video Transcript (MIntRec2.0 - dia89_utt8)

Speaker: "hey, let's see if they have a page for buying a boogie board on there, huh?"

(a) Spatiotemporal Context and Key Dialogue.

Agent B (Evaluator) - mode='fact'

```
{
  "winner_intent": "Advise",
  "factual_reasoning": "The speaker's engaged
    expression, leaning forward, and hand gestures
    towards the screen support the hypothesis...",
  "verification_score": 0.70,
  "counterfactual_score": 0.92
}
```

DPO Trajectory Extractor - Final Audit

```
SUCCESS - [Sample Quality Verified]
 Winner logic is robust (Score 0.64 > 0.60).
 Extracted Level 1 & Level 2 DPO pairs.
 [Status]: Added to Training Pool.
```

(b) Curated Audit Logs. Both the factual score (0.70) and counterfactual robustness (0.92) are high, successfully passing the quality failsafe.

Figure 6: Visualization of ASTRO's cognitive process on a success case (`dia89_utt8.mp4`). (a) Shows the visual context and dialogue where the speaker suggests an action. (b) Presents the curated logs where the `Advise` intent is confidently validated across both factual and counterfactual dimensions, allowing the extraction of high-quality preference pairs.

As shown in Table 8, the rejected response correctly observes the lack of greeting gestures during the verification steps, but illogically concludes that the intent is `Greet`. Penalizing this trajectory via DPO teaches the model to align its final decision strictly with its own verification findings instead of forcing a prediction based on superficial visual presence.

F.2 Level 2: Robustness Penalty (Correct vs. Ambiguous Path)

The Level 2 DPO pair is designed to address subtle semantic overlap and cognitive boundary ambiguity, pairing the *Correct Path* ($\mathcal{P}_{corr}$) with an *Ambiguous Path* ($\mathcal{P}_{amb}$). In this example, the MCTS explores the `Inform` intent. Because `Inform` and `Advise` share similar macro-visual priors (a person talking calmly while gesturing), `Inform` receives a deceptions-high factual score ($S_{pos} = 0.70$).

However, as shown in Table 9, `Inform` is a weak logical fit for the specific dialogue (*"let's see if..."*), which represents a prospective suggestion rather than a retrospective disclosure. The blue success status box in Figure 6b signifies that the DPO Extractor successfully mined this hard negative pair. The DPO loss objective leverages this subtle contrast to force the model to look beyond macro-visual similarities and utilize fine-grained text to draw sharp cognitive boundaries between semantically adjacent intents.

G Training Hyperparameters and Computational Cost

In recent years, disclosing training details and computational costs has become a critical standard for evaluating the reproducibility of Large Vision-Language Model (LVLM) research. In this section, we provide a transparent and comprehensive breakdown of the hyperparameters and hardware configurations used during the Direct Preference Optimization (DPO) alignment phase of our ASTRO framework.

G.1 Hyperparameter Configurations

The alignment training was conducted on the InternVL3-8B architecture utilizing the Swift training framework. To efficiently optimize the model while preserving its pre-trained visual capabilities, we employed Low-Rank Adaptation (LoRA) specifically targeting the language decoder's linear layers, while explicitly freezing the vision encoder (`vision_model`) and the cross-modal MLP projector (`mlp1`).

For the DPO objective, we set the KL divergence penalty coefficient ($beta$) to 0.1. The model was trained with a peak learning rate of 5×10^{-6} using a Cosine learning rate scheduler and a warm-up ratio of 0.03. To maintain stable optimization and prevent memory overflow, we applied DeepSpeed ZeRO-2 optimization in `bfloat16` precision. The complete list of hyperparameter configurations and computational expenditures is compactly summarized in Table 10.

G.2 Computational Resources

The DPO training pipeline was deployed on a SLURM-managed cluster utilizing 2 NVIDIA A800 80GB GPUs in parallel (nproc_per_node $=$ 2). With a per-device batch size of 1 and 8 gradient accumulation steps, the model was trained for 1 epoch (127 global optimization steps) using a maximum sequence length of 2048 tokens. The entire efficient alignment process was completed in approximately 1.5 hours.

System Prompts: Evaluator Agent (π_{Eval})

Part 1: Factual Verification Prompt (`mode='fact'`)

 # Role

You are an Intelligent Visual Verification Expert with access to a commonsense knowledge base. Your task is to evaluate the likelihood of a "Hypothesis" given the "Video Context" and "Question".

 # Task: Multi-Dimensional Verification

Analyze the hypothesis based on visual information and commonsense knowledge. You must evaluate three specific dimensions:

1. **Appearance**: Do the target objects match the description? (e.g., color, shape, texture).
2. **Action**: Is the specific movement or interaction clearly visible and physically logical?
3. **State**: Is the resulting state consistent with the action and context?

 # Output Format

Output a single JSON object:

```
{
 "factual_reasoning": "Cite specific visual evidence regarding Appearance, Action, and State...",
 "verification_score": <float between 0.00 and 1.00>
}
```

Part 2: Counterfactual Refutation Prompt (`mode='counterfactual'`)

 # Role

You are an Intelligent Logic Critic and "Devil's Advocate" with access to a commonsense knowledge base. Your task is to evaluate the "Falsification Likelihood" of a "Candidate Answer".

 # Task: Multi-Dimensional Refutation

Systematically attempt to debunk the Candidate Answer by evaluating three specific dimensions of failure:

1. **Visual Contradiction**: Does specific visual evidence directly contradict the candidate?
2. **Competitive Superiority**: Is there an alternative hypothesis in the context that fits the scene *better* or *more specifically* than the candidate?
3. **Commonsense Violation**: Does the candidate answer violate logical cause-and-effect or functional intent?

 # Output Format

Output a single JSON object:

```
{
 "counterfactual_reasoning": "Cite specific contradictions or superior alternatives...",
 "falsification_score": <float between 0.00 and 1.00>
}
```

Table 6: Verbatim system prompts for the Evaluator Agent (π_{Eval}). The evaluation is decoupled into two distinct passes: Factual Verification and Counterfactual Refutation (Devil's Advocate), enabling fine-grained, dimension-specific scoring.

Inference Prompt Template (Zero-Shot CoT Evaluation)

[Input Structure]

<VIDEO_TOKENS>

System:

You are a question-answering assistant that combines video understanding with commonsense reasoning. Given a question and several candidate answers, you must follow these steps to provide your final answer:

1. **Observation**: Extract the observable facts from the video.
2. **Verification Steps**: Carefully check the visual evidence to verify or debunk candidate options.
3. **Reasoning**: Provide causal logic and score the best option. You MUST include a counterfactual analysis by explicitly stating an alternative incorrect option you considered and explaining why detailed inspection debunked it.
4. **Answer**: Finally, output the chosen option exactly in the format 'Answer: aX: [option text]'.

———

User:

Question: <QUESTION_TEXT>

Options:

a0. <OPTION_0>

a1. <OPTION_1>

a2. <OPTION_2>

a3. <OPTION_3>

a4. <OPTION_4>

Table 7: The exact system prompt utilized during the inference and evaluation phase. It explicitly demands a structured reasoning process incorporating visual evidence extraction and counterfactual debunking before yielding the final prediction.

Level 1 DPO Pair (Factuality / Easy Negative)	
Chosen Response (y_{chosen}): `Advise` ($\mathcal{P}_{corr}$)	**Rejected Response ($y_{rejected}$): `Greet` ($\mathcal{P}_{err}$)**
Observation: The speaker's facial expression is one of curiosity and engagement, suggesting they are actively seeking information or advice. The speaker is leaning slightly forward, indicating interest in the conversation. **Verification Steps:** - Checked: Check the facial expression of the speaker to see if it matches the tone of advising... *Result: The speaker's tone is casual and conversational, which is consistent with suggesting an action.* - Checked: Observe the body language... *Result: The speaker appears to be gesturing with their hands, looking at a laptop screen, implying they are searching for something online.* **Reasoning:** Visually, this intent aligns perfectly (Score: 0.70). I considered alternatives like checking for a skateboard, but detailed inspection debunked them. The logic is solid. **Answer: Advise**	**Observation:** The speaker's facial expression is friendly and engaging, suggesting a greeting. *Contradicting:* The context of the conversation involves checking for a specific item, which does not align with a greeting. **Verification Steps:** - Checked: Observe the facial expression to see if it matches a greeting. Look for a smile, raised eyebrows... *Result: There is no visible smile, raised eyebrows, or an open mouth that would clearly indicate a greeting.* - Checked: Check if they are making hand gestures typical of greeting someone... *Result: No greeting gestures observed.* **Reasoning:** Based on the visual observation, this seems plausible. **Answer: Greet**

Table 8: An example of a Level 1 DPO preference pair extracted from `dia89_utt8.mp4`. The rejected trajectory demonstrates a "logical disconnect" where the model observes contradicting evidence (no greeting gestures) but still incorrectly forces the `Greet` prediction.

Level 2 DPO Pair (Robustness / Hard Negative)	
Chosen Response (y_{chosen}): `Advise` ($\mathcal{P}_{corr}$)	**Rejected Response ($y_{rejected}$): `Inform` ($\mathcal{P}_{amb}$)**
Observation: The speaker's facial expression is one of curiosity and engagement, suggesting they are actively seeking information or advice. The speaker is leaning slightly forward, indicating interest in the conversation. **Verification Steps:** - Checked: Check the facial expression of the speaker to see if it matches the tone of advising... *Result: The speaker's tone is casual and conversational, which is consistent with suggesting an action.* - Checked: Observe the body language... *Result: The speaker appears to be gesturing with their hands, looking at a laptop screen, implying they are searching for something online.* **Reasoning:** Visually, this intent aligns perfectly (Score: 0.70). I considered alternatives like checking for a skateboard, but detailed inspection debunked them. The logic is solid. **Answer: Advise**	**Observation:** The speaker's facial expression is neutral to slightly inquisitive, indicating they are seeking information. The speaker is leaning slightly forward, which often indicates engagement. **Verification Steps:** - Checked: Check the facial expression of the speaker to see if it matches the tone of informing... *Result: The tone of voice is casual and conversational, matching the intent to inform or check something. Contradicting: There is no direct action taken by the speaker to check a page, and the context appears to be more about planning rather than directly informing.* - Checked: Observe the body language to see if it indicates an informative communication style... *Result: The speaker's body language is open and relaxed.* **Reasoning:** Based on the visual observation, this seems plausible. **Answer: Inform**

Table 9: An example of a Level 2 DPO preference pair. The rejected `Inform` trajectory represents a hard negative: it possesses high visual plausibility but lacks the nuanced logical and contextual support required to outcompete the ground-truth `Advise` intent.

Configuration / Hyperparameter	Value	Configuration / Hyperparameter	Value
Model & Infrastructure		***Optimization Parameters (Cont.)***	
Base Model	InternVL3-8B	Learning Rate	5×10^{-6}
Training Framework	Swift / DeepSpeed ZeRO-2	LR Scheduler	Cosine
Precision	`bfloat16`	Warm-up Ratio	0.03
Frozen Modules	`vision_model`, `mlp1`	Per-device Batch Size	1
Training Objective	DPO (`rlhf_type=dpo`)	Gradient Accumulation Steps	8
Optimization Parameters		***LoRA Settings & Hardware Cost***	
DPO Penalty Coefficient (β)	0.1	LoRA Rank (r)	16
Max Sequence Length	2048	LoRA Alpha (α)	32
Epochs	1	LoRA Dropout	0.05
Total Global Steps	127	Number of GPUs	2
Total Elapsed Time	$\sim$ 1 hour 31 mins	GPU Model	NVIDIA A800

Table 10: Detailed training hyperparameters, LoRA configurations, and computational resource expenditures for the DPO alignment phase.