# *GRASP*: Learning to Ground Social Reasoning in Multi-Person Non-Verbal Interactions

**Junho Kim**[1] **Xu Cao**[1] **Houze Yang**[1] **Bikram Boote**[1] **Ana Jojic**[1]

**Fiona Ryan**[2] **Bolin Lai**[3] **Sangmin Lee**[4†] **James Matthew Rehg**[1†]

[1]University of Illinois Urbana-Champaign [2]Georgia Institute of Technology
[3]Amazon AGI [4]Korea University

## Abstract

Understanding social interactions requires reasoning over subtle non-verbal cues, yet current multimodal large language models (MLLMs) often fail to identify *who interacts with whom* in multi-person videos. We introduce *GRASP*, a large-scale social reasoning dataset that connects high-level social QA with fine-grained gaze and deictic gesture events. *GRASP* contains 290K question–answer pairs over 46K videos totaling 749 hours, organized by a 16-category taxonomy spanning gaze, gesture, and joint gaze–gesture reasoning, together with *GRASP-Bench* for evaluation. Unlike prior resources that focus on either isolated cues or high-level social QA, *GRASP* builds questions from identity-consistent gaze trajectories, deictic gestures, and their joint compositions into social events. Moreover, we propose *Social Grounding Reward* (*SGR*), a learning signal that uses these social events to encourage models to reason about the participants involved in each interaction. Experiments show that *SGR* improves performance on *GRASP-Bench* while maintaining zero-shot performance on related social video QA benchmarks.

## 1 Introduction

In everyday social interactions, the meaning of a conversation is often shaped and influenced in subtle but important ways by non-verbal cues such as gaze and gesture [28, 23, 40]. A brief glance toward a social partner, pointing at an object of interest, or a moment of mutual eye contact can signal who is being addressed, what is being referenced, and how participants relate to one another. Interpreting such signals [39, 7, 81] is inherently challenging, as it requires reasoning about complex interactions between individuals from ambiguous visual cues. Humans naturally resolve this complexity by integrating multimodal information [20, 35] to infer social meaning and dynamics [17, 44]. This capacity for social understanding underpins collaborative human behavior [76] and it is key capability that can enable AI systems are to interact effectively with people, from embodied agents [83] to conversational systems [70].

Prior research has made significant progress toward perceiving individual social cues as low-level signals, with a focus on gaze and gesture modalities. Gaze target estimation methods [13, 67, 68] can predict where a person is looking and gesture recognition models [7, 34, 49] can classify hand movements into predefined categories. These works operate at the level of perceptual primitives, answering *what* cue is present. In contrast, this work is the first systematic attempt to address the question of *how* gaze and gesture should be interpreted in a social context, where meaning arises from the context in which a behavior is produced.

[†]Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

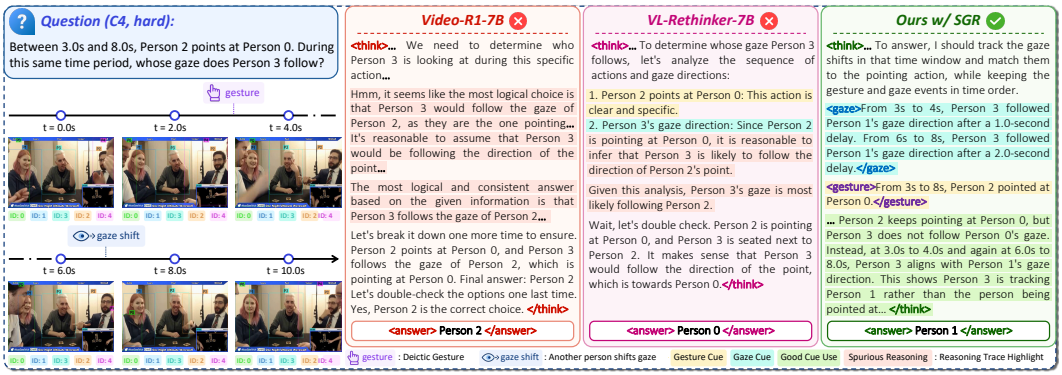


Figure 1: Example from *GRASP*. Multi-person social reasoning requires grounding subtle non-verbal cues in the correct participants over time. Existing MLLMs [19, 77] often take spurious scene-level shortcuts, whereas ours leverage evidence-aware supervision to reason from the relevant social event.

The key hypothesis underlying this work is that modern MLLMs [51, 14, 50], which integrate visual perception with strong reasoning capabilities, can be utilized to integrate nonverbal cues and infer their meaning in the context of social understanding. Several works have introduced new social understanding tasks such as multi-party interaction reasoning [33, 36], speaker target identification [39, 44, 63], and social video QA [55, 61]. However, current MLLMs have only limited ability to capture social interactions as grounded visual events: (i) while they can recognize people and actions in the visual scene, they often fail to organize them into temporally grounded social events among participants, making it difficult to identify the relevant cue, participants, and timing of an interaction, *e.g.,* who is looking at, gesturing toward, or responding to whom. (ii) Their reasoning often does not attend to the fine-grained visual evidence that determines these relations, relying instead on linguistic or scene-level priors as shown in Fig. 1. Effective social interaction understanding requires grounding reasoning in non-verbal cues such as gaze direction, deictic gestures, and their joint coordination across participants, but large-scale supervision that jointly connects these fine-grained cues with high-level social reasoning remains scarce.

In this paper, we close this gap by introducing *G*rounded *R*easoning *A*nd *S*ocial *P*erception (*GRASP*), the first large-scale dataset for grounded social understanding that jointly supports both high-level social reasoning and fine-grained gaze–gesture grounding in multi-person videos. Existing social datasets provide fragmented supervision, focusing on either a single fine-grained cue such as gaze [67, 13, 30, 66] or gesture [7], or high-level social QA tasks [87], rather than jointly supervising how multiple non-verbal cues ground social reasoning. Building such concurrent supervision manually is prohibitively expensive, requiring temporally precise, identity-consistent annotations of multi-person gaze–gesture dynamics. For scalability, we develop a bottom-up pipeline that extracts structured social interactions from video by detecting gaze and deictic gestures across individuals, organizing them into social events (*e.g.,* mutual gaze, joint attention, gaze following), and modeling their joint compositions where gaze and gesture together convey meaning. *GRASP* contains 290K QA instances organized under a 16-category taxonomy, grounded in 245K interaction events across 46K video clips totaling 749 hours. In addition, we provide *GRASP-Bench*, a curated benchmark of 1K test items.

An additional key question that we address in this work is how such structured social events can provide effective learning signals for MLLMs. Existing RL-based policy optimization approaches (*e.g.,* GRPO [69, 52]) encourage models to generate reasoning before answering, but they do not explicitly tie this reasoning to the relevant social cues. In social scenarios, as exemplified in Fig. 1, this can lead models either to produce spurious reasoning traces, consistent with overthinking behavior [72, 84, 11], or to arrive at answers through linguistic and scene-level shortcuts. Accordingly, we introduce *S*ocial *G*rounding *R*eward, a learning signal that verifies whether the model's reasoning references the correct participants involved in the underlying social events. Rather than relying on complex reward functions, we provide a direct signal *"Did the model attend to the right person?"* that shifts supervision toward evidence-aware reasoning.

To demonstrate the utility of *GRASP*, we post-train MLLMs with our data and policy, achieving competitive performance on *GRASP-Bench*. Beyond in-domain evaluation, our model also generalizes to three social reasoning benchmarks [39, 44, 42], indicating that social grounding learned from *GRASP* transfers across task formulations. Collectively, these findings demonstrate that large-scale

participant-level social data enables MLLMs to move beyond primitive perception and reason about social interactions from non-verbal cues.

Our contribution is three-fold:

- We construct *GRASP* and *GRASP-Bench* for training and evaluating MLLMs on grounded social reasoning over multi-person gaze and gesture interactions, using a bottom-up pipeline that scales participant-level social event annotation.
- We propose *Social Grounding Reward*, a simple yet effective RL signal that directly verifies whether a model's reasoning attends to the correct participants involved in the underlying social events, teaching MLLMs to ground social reasoning in the relevant non-verbal evidence.
- We demonstrate that our framework achieves strong performance on *GRASP-Bench*, generalizes to other social reasoning benchmarks, and provides detailed analyses of where and why social grounding is most beneficial.

## 2 Related Works

**Gaze and Gesture in Human Social Behavior.** Non-verbal cues, particularly gaze and gesture, are fundamental to human social interaction, often indicating intention and the targets of utterances [40, 46]. Interpreting these behaviors has driven significant research in computer vision across two parallel tracks. In gaze understanding, several works have addressed the task of gaze target estimation, leveraging benchmarks such as GazeFollow [67] and VideoAttentionTarget [13]. Approaches have evolved from fusion-based methods that process depth, pose, and head crops separately [67, 31, 18, 5, 26], to decoding methods leveraging visual foundation model representations [71, 68]. A related task area is identifying social gaze dynamics such as mutual gaze and joint attention [53, 16, 17, 25]. While some recent methods for gaze target estimation aim to integrate VLM priors into gaze pipelines [27, 54, 64], specialized models for gaze target estimation remain the dominant approach. In contrast to these gaze-specific models, our work demonstrates how to equip MLLMs with the ability to not only estimate gaze, but jointly leverage it as a cue for higher level social understanding.

Concurrently, gesture recognition has been extensively explored through diverse datasets encompassing RGB-D videos (LD-ConGR [49]), high-resolution images (HaGRID [34]), egocentric views (EgoGesture [89]), and multi-person social contexts (SocialGesture [7]). Modeling efforts in this domain have similarly advanced toward leveraging foundation models, including MLLMs and VLMs [47, 6]. Despite extensive progress in these individual domains, a critical gap remains in jointly modeling gaze and gesture to infer social intention [8], an integration essential for a holistic understanding of social behavior.

**Multi-modal Large Language Models in Social AI.** Recent advances in MLLMs have demonstrated remarkable video understanding across diverse domains, including multimedia analysis, and human action recognition [88, 19, 86, 90]. Leading proprietary models, such as the Gemini 3 [24] and GPT-5 [62] series, alongside open-source alternatives like Qwen-VL series [3], leverage extensive context windows to process complex video inputs. While these models have set new records on challenging video question-answering benchmarks [21, 22, 29], a critical limitation remains: many rely disproportionately on their base language models. This reliance often leads to commonsense biases in social contexts, where models may provide accurate answers based on linguistic patterns even without accessing the visual data [43, 40]. While recent datasets [87, 38, 39, 55, 36] have attempted to address this, most existing benchmarks focus on broad social reasoning, overlooking fine-grained interpersonal cues such as gaze shifts and subtle social gestures [44, 45, 73, 85, 75]. In this paper, we bridge this gap by introducing a novel benchmark and a novel post-training strategy designed to enhance the fine-grained social intelligence of MLLMs.

## 3 *GRASP* Dataset: Grounded Reasoning And Social Perception

To enable social reasoning grounded in non-verbal cues, we first curate video sources spanning a broad range of multi-person social interaction types, including multi-party TV dialogue [42, 80], social intelligence reasoning [87], multi-speaker conversation [59], and embodied multi-person behavior [56]. We further include social deduction gameplay [38, 7], which we augment with

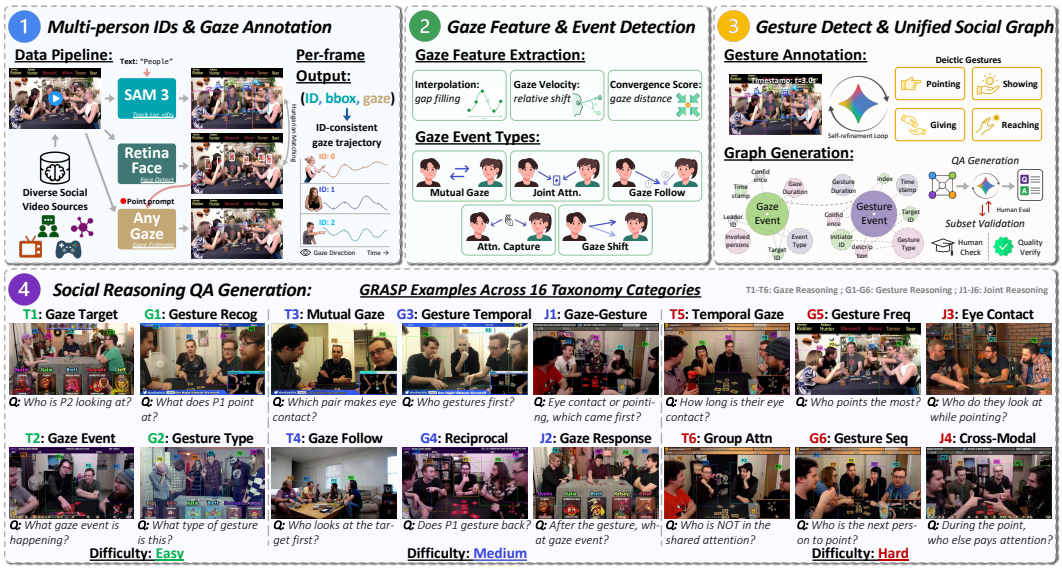


Figure 2: Overview of the *GRASP* construction pipeline and QA examples. We convert multi-person videos into person-consistent gaze and gesture events, compose them into structured social QA pairs, and apply subset validation with human feedback for quality control.

additional online video sources, as such contents contain dense multi-person interactions with rich social signals. Our dataset comprises 46K videos, from which we construct structured representations of social interactions by extracting person-consistent gaze trajectories and deictic gestures and integrating them into higher-order social events. All extracted events are anchored to track person identities, providing structured supervision for both QA generation and our grounding reward. We illustrate the overall collection process in Fig. 2, and provide details in Appendix A.

### 3.1 Data Collection Pipeline

**Multi-person Identification & Gaze Annotation.** Given a raw video $V$, we first split it into shorter segments $\{V_s\}_{s=1}^{S}$ at scene cuts using PySceneDetect to preserve coherent social interactions. Within each segment, our goal is to obtain *person-consistent* gaze trajectories in multi-party scenes, so that every gaze observation can be retained with a stable identity across time. To do so, we design a multi-stage process that jointly tracks person identities and aligns gaze observations across frames.

We leverage an off-the-shelf detection model, SAM 3 [10] with the text prompt *"people"* to obtain temporally consistent bounding boxes $\mathcal{B}^t_{\text{person}} = \{b^t_i\}_{i\in\mathcal{P}}$, where $\mathcal{P}$ denotes the set of tracked person IDs. While the model provides stable identity tracking, it cannot predict precise facial localization for the subsequent gaze estimation process. Accordingly, we detect face location $\mathcal{B}^t_{\text{face}} = \{f^t_j\}_{j=1}^{M_t}$ using RetinaFace [15], where $M_t$ denotes the number of detected faces at time $t$, and associate each face with a tracked identity using Hungarian matching [37] between face boxes and the person boxes.

For each person $i \in \mathcal{P}$ at time $t$, we then use the matched face center $c^t_i$ as a point prompt together with the video frame, and feed them to GazeAnywhere [9] to estimate a normalized gaze point $g^t_i = (x^t_i, y^t_i) \in [0,1]^2$. This yields person-consistent gaze trajectories $\{g^t_i\}_{t=1}^{T}$. The resulting triplet $\{(i, b^t_i, g^t_i)\}_{t=1}^{T}$, where $i \in \mathcal{P}$, preserves person identity, spatial location, and gaze target over time, which are then used to define higher-order social events.

**Gaze Feature Extraction & Event Detection.** Raw per-frame gaze observations $\{g^t_i\}$ are inherently noisy due to missing detections and camera motion. To obtain robust temporal features, we first apply a simple interpolation scheme within each segment to handle short gaps. We then compute a relative gaze velocity by measuring gaze motion relative to the face center. For each person $i$ at time $t$, we define their gaze direction as $d^t_i = g^t_i - c^t_i$, and estimate velocity as $\|\partial_t d^t_i\|$. To capture group-level attention dynamics, we compute a *convergence score* that measures how closely gaze points cluster at each time step. Specifically, we compute the median distance between individual gaze points $g^t_i$ and their centroid, and map it to a normalized score $s^t = \exp(-\alpha \cdot \text{median}_i \|g^t_i - \bar{g}^t\|)$, where $s^t \in [0,1]$.

From these features we define five social gaze events, grounded in established taxonomies of gaze behavior [58, 32]: *mutual gaze* (bidirectional gaze between two participants), *joint attention* (multiple participants attending to a common target), *gaze following* (one participant shifting gaze to another's prior target), *attention capture* (simultaneous gaze shifts across multiple participants), and *sudden gaze shift* (per-person saccade), resulting in 169K gaze events across the corpus (details in Appendix. A).

**Gesture Detection & Unified Social Graph.** Robust detection of communicative hand gestures in untrimmed, multi-person video is challenging due to temporal ambiguity, identity confusion, and the lack of precise grounding signals. Following prior gesture taxonomies in social interaction [57, 7], we focus on deictic gestures and prompt the model to identify four types, namely *pointing*, *showing*, *giving*, and *reaching*. We leverage Gemini [24] as an annotator, augmented with structured visual cues. For each video, we render person-tracked bounding boxes $\mathcal{B}^t_{\text{person}}$ together with their corresponding IDs $i \in \mathcal{P}$ and per-frame timestamps directly onto each frame. This provides explicit spatio-temporal grounding, allowing the model to directly read timestamp from the frame and casting temporal localization into an OCR-like problem, thereby improving the reliability of temporal grounding.

We further employ a self-refinement loop in which low-confidence predictions are iteratively regenerated until convergence, yielding 76K reliable gesture instances with person-level initiator and target identities. We then merge gaze events and gesture annotations into a unified social graph for each video. Each node represents a detected social event with structured metadata, including event type, participating identities, and temporal boundaries, along with modality-specific attributes such as gaze convergence or gesture targets. Joint-modal interactions are captured by linking gaze and gesture events that co-occur within a temporal window, enabling reasoning over joint gaze–gesture patterns. The resulting graph contains 245K events across 46K videos and serves as the foundation for both our QA taxonomy and the grounding supervision used during training.

### 3.2 Social Reasoning QA Taxonomy

Our dataset is designed to probe not only the *perception* of individual cues, but also the ability to *reason* about social dynamics arising from multi-person interactions. Accordingly, we carefully define a 16-category QA taxonomy with increasing levels of difficulty (examples in Appendix. D.1).

- **Gaze reasoning (T1–T6)** spans tasks from perception to inference, including gaze target identification (T1), event classification (T2), mutual gaze recognition (T3), gaze following (T4), temporal reasoning (T5), and group attention dynamics (T6). While T1–T2 focus on identifying *who looks where*, T5–T6 require reasoning over temporal ordering and multi-person attention dynamics.
- **Gesture reasoning (G1–G6)** covers gesture recognition (G1), type classification (G2), temporal reasoning (G3), reciprocal patterns (G4), frequency (G5), and sequence chains (G6), with higher-level categories requiring multi-step interaction reasoning.
- **Gaze–gesture joint reasoning (J1–J4)** requires integrating gaze and gesture signals to reason over co-occurring non-verbal cues, including temporal alignment (J1), gaze response to gesture (J2), eye contact during interaction (J3), and cross-modal person dynamics (J4).

**Social Reasoning QA Generation.** QA pairs are generated from structured event metadata derived from gaze and gesture interactions using a closed-source model [24]. Each question is constructed by querying key attributes such as participant identities, temporal intervals, and interaction types, ensuring that answers are directly verifiable without exhaustive manual annotation. To provide sufficient temporal context, each QA is anchored around a target event and augmented with a short temporal window from its surrounding frames. We design this process to produce diverse questions spanning both perception and higher-order reasoning over multi-person interactions. To ensure data reliability, we further conduct human validation on a subset of generated QA pairs, where sampled QA instances are checked for event consistency, answer verifiability, and ambiguity. The assessment is used to refine the generation and filtering criteria, providing an additional

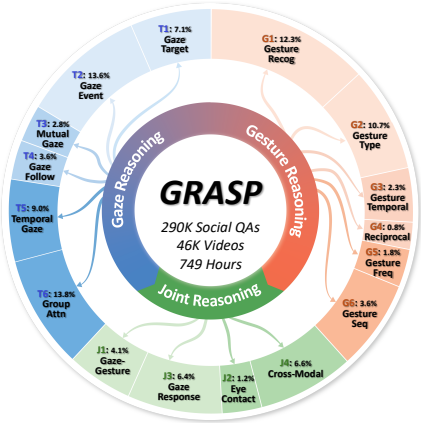

Figure 3: *GRASP* taxonomy and statistics.

quality-control loop before producing the final dataset of 290K QA pairs grounded in structured gaze and gesture events. We provide examples in Fig. 2 and statistics in Fig. 3 and Appendix. D.1. Please see additional details in Appendix B.

# 4 Learning to Ground Social Reasoning

**Preliminaries.** GRPO [69] provides an effective framework for training MLLMs with reasoning capabilities [19, 12, 86]. Given a video-question pair query $q$, the policy $\pi_\theta$ generates $K$ reasoning trajectories $\{y_1, \ldots, y_K\}$. The policy is optimized by maximizing the following objective:

$$\max_{\pi_\theta} \mathbb{E}_{y \sim \pi_{\theta_{\text{old}}}} \left[ \sum_{i=1}^{K} \frac{\pi_\theta(y_i \mid q)}{\pi_{\theta_{\text{old}}}(y_i \mid q)} \cdot \hat{A}_i \; - \; \beta \, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) \right], \tag{1}$$

where $\pi_{\theta_{\text{old}}}$ is the policy from the previous iteration, $\pi_{\text{ref}}$ is a fixed reference (*e.g.,* SFT model), and $\beta$ controls the KL regularization. The advantage for each trajectory is computed relative to the group:

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_K\})}{\text{std}(\{r_1, \ldots, r_K\})}, \tag{2}$$

where $r_i$ denotes the reward assigned to trajectory $y_i$. In practice, rewards are typically defined using simple verifiable signals such as answer correctness and output format consistency. Each trajectory is formatted with special tokens, where the model generates intermediate reasoning within `<think>` and outputs the final answer within `<answer>`.

## 4.1 Training Strategy for Social Reasoning

We adopt a two-stage training strategy: (i) warm-up the model with structured SFT responses that explicitly separate gaze and gesture evidence and (ii) RL post-training with rewards for correctness, format consistency, structural grounding, and social grounding.

**Stage 1: Structured Format Supervised Fine-Tuning.** Applying RL directly to the base models induces a cold-start problem, as reasoning about fine-grained non-verbal cues is largely absent from existing multimodal corpora. Thus, stochastic rollouts rarely produce trajectories that reference relevant gaze or gesture events, resulting in a reward landscape that is too sparse to effectively bootstrap learning. Following prior works [19, 65] that introduces a warm-up before policy optimization, we first utilize the 52K *GRASP* subset for SFT, where each target response is formatted as:

`<think>` `<gaze>` $\mathcal{E}_{\text{gaze}}$ `</gaze>` `<gesture>` $\mathcal{E}_{\text{ges}}$ `</gesture>` `</think>` `<answer>` $a$ `</answer>`,

where $\mathcal{E}_{\text{gaze}}$, $\mathcal{E}_{\text{ges}}$ enumerate the temporally grounded gaze, gesture events drawn from the social graph, and $a$ is an answer for those events. This warm-up teaches models to decompose social reasoning into its non-verbal components before arriving at an answer, yielding a reference model $\pi_{\text{ref}}$ that can generate the dual-tag format and provides a stable starting point for the subsequent RL stage.

**Stage 2: Grounded Reinforcement Post-Training.** Starting from $\pi_{\text{ref}}$, we continue training on the *GRASP* dataset with policy optimization. Given a video-question pair $q$ and a group of $K$ rollouts $\{y_i\}_{i=1}^{K}$, we define the scalar reward for trajectory $y_i$ as a weighted combination of four signals:

$$r_i \;=\; \lambda_{\text{acc}}\, r_i^{\text{acc}} \;+\; \lambda_{\text{fmt}}\, r_i^{\text{fmt}} \;+\; \lambda_{\text{str}}\, r_i^{\text{str}} \;+\; \lambda_{\text{gnd}}\, r_i^{\text{gnd}}, \tag{3}$$

where $r_i^{\text{acc}} = \mathbb{1}[\hat{a}_i = a_i]$ enforces answer correctness, and $r_i^{\text{fmt}} = \mathbb{1}[y_i \in \mathcal{F}]$ ensures that the output adheres to the required structure, where $\mathcal{F}$ denotes the set of valid outputs. We further include a structural term $r_i^{\text{str}}$ that preserves the use of `<gaze>` and `<gesture>` tags, and a social grounding term $r_i^{\text{gnd}}$ that encourages the model to identify the relevant participants in the interaction.

## 4.2 Social Grounding Reward for Social Learning Signal

To make the reasoning trace reflect non-verbal cues, we introduce a structural grounding term that encourages the model to express its observations through the dual-tags. This enforces a consistent structure where gaze and gesture information is explicitly articulated rather than implicitly inferred:

$$r_i^{\text{str}} \;=\; \mathbb{1}[\, y_i \text{ contains } \langle \texttt{gaze} \rangle \text{ or/and } \langle \texttt{gesture} \rangle \text{ blocks} \,]\,. \tag{4}$$

Table 1: Evaluation results on *GRASP-Bench*. We report accuracy for all 16 categories. Columns are grouped by reasoning type and difficulty level.

| Method | Gaze Reasoning | | | | | | | Gesture Reasoning | | | | | | | Joint Reasoning | | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Easy | | Medium | | Hard | | Avg. | Easy | | Medium | | Hard | | Avg. | Medium | | Hard | | Avg. | |
| | T1 | T2 | T3 | T4 | T5 | T6 | | G1 | G2 | G3 | G4 | G5 | G6 | | J1 | J2 | J3 | J4 | | |
| *Proprietary Models* | | | | | | | | | | | | | | | | | | | | |
| Claude Sonnet 4.6 [2] | 17.1 | 56.1 | 18.2 | 11.9 | 30.7 | 35.7 | 30.2 | 42.6 | 64.0 | 30.6 | 39.0 | 29.0 | 54.5 | 46.5 | 38.3 | 18.2 | 42.1 | 36.5 | 37.3 | 37.0 |
| GPT-5.4 [62] | 38.6 | 50.9 | 26.4 | 23.9 | 27.3 | 38.1 | 34.9 | 63.1 | 69.4 | 55.1 | 51.2 | 38.7 | 59.1 | 60.4 | 39.1 | 31.8 | 40.8 | 39.2 | 39.0 | 43.9 |
| Gemini 3.1 Pro [24] | 42.9 | 54.4 | 49.1 | 40.3 | 28.4 | 29.8 | 41.8 | 63.1 | 84.7 | 61.2 | 56.1 | 48.4 | 68.2 | 67.6 | 42.6 | 50.0 | 44.7 | 44.6 | 44.3 | 50.5 |
| *Supervised Fine-Tuned Models* | | | | | | | | | | | | | | | | | | | | |
| Qwen2.5-VL-7B-Instruct [4] | 30.0 | 55.3 | 27.3 | 28.4 | 27.3 | 31.0 | 34.3 | 41.0 | 48.6 | 36.7 | 34.1 | 22.6 | 45.5 | 40.7 | 33.9 | 31.8 | 31.6 | 36.5 | 33.8 | 36.2 |
| InternVL3.5-8B-Instruct [79] | 30.0 | 31.6 | 30.0 | 28.4 | 36.4 | **34.5** | 31.9 | 46.7 | 53.2 | 26.5 | 34.1 | 19.4 | 40.9 | 42.0 | 43.5 | **45.5** | 31.6 | 35.1 | 38.3 | 36.6 |
| LLaVA-OV-1.5-8B-Instruct [1] | 24.3 | 36.8 | 33.6 | 17.9 | 30.7 | 29.8 | 30.0 | 35.2 | 82.9 | 32.7 | 22.0 | 35.5 | 27.3 | 47.1 | 42.6 | 31.8 | 35.5 | 35.1 | 38.0 | 37.3 |
| Qwen3-VL-8B-Instruct [3] | 24.3 | 35.1 | 25.5 | 14.9 | 31.8 | 33.3 | 28.3 | 45.1 | 86.5 | 28.6 | 41.5 | 22.6 | 40.9 | 52.7 | 33.9 | 31.8 | 35.5 | 48.6 | 38.0 | 38.3 |
| Qwen3.5-9B (Instruct Mode) [74] | 22.9 | 52.6 | 30.0 | 22.4 | 34.1 | 25.0 | 32.8 | 51.6 | 87.4 | 40.8 | **46.3** | 25.8 | **68.2** | 59.0 | 40.0 | 31.8 | 39.5 | 45.9 | 40.8 | 43.0 |
| *RL Post-Training Models* | | | | | | | | | | | | | | | | | | | | |
| VL-Rethinker-7B [77] | 24.3 | 36.0 | 22.7 | 25.4 | 39.8 | 27.4 | 29.6 | 36.1 | 49.5 | 38.8 | 29.3 | 25.8 | 22.7 | 38.0 | 32.2 | 27.3 | 36.8 | **50.0** | 37.6 | 34.2 |
| VideoChat-R1.5-7B [86] | 32.9 | 39.5 | 25.5 | 35.8 | 30.7 | 32.1 | 32.6 | 44.3 | 29.7 | 36.7 | 43.9 | 12.9 | 27.3 | 35.4 | 33.9 | 27.3 | 31.6 | 36.5 | 33.4 | 33.7 |
| Video-R1-7B [19] | 30.0 | 51.8 | 38.2 | 28.4 | 35.2 | 31.0 | 37.1 | 42.6 | 29.7 | 36.7 | 26.8 | 25.8 | 40.9 | 34.8 | 36.5 | 22.7 | 30.3 | 47.3 | 36.6 | 36.3 |
| LongVILA-R1-7B [12] | 28.6 | 29.8 | 32.7 | 20.9 | 30.7 | 28.6 | 29.1 | 29.5 | 26.1 | 28.6 | 24.4 | 32.3 | 36.4 | 28.5 | 37.4 | 36.4 | 38.2 | 24.3 | 34.1 | 30.1 |
| InternVL3.5-8B [79] | 20.0 | 27.2 | 31.8 | 19.4 | 28.4 | 23.8 | 25.9 | 46.7 | 55.9 | 32.7 | 31.7 | 16.1 | 36.4 | 42.8 | 36.5 | 31.8 | 38.2 | 37.8 | 36.9 | 33.9 |
| LLaVA-OV-1.5-8B-RL [1] | 22.9 | 30.7 | 30.0 | 26.9 | 27.3 | 29.8 | 28.3 | 31.1 | 76.6 | 32.7 | 19.5 | 12.9 | 22.7 | 41.5 | 42.6 | 40.9 | 30.3 | 28.4 | 35.5 | 34.2 |
| Qwen3-VL-8B-Thinking [3] | 31.4 | 36.0 | 23.6 | 16.4 | 30.7 | 27.4 | 28.1 | 41.0 | 74.8 | 40.8 | 34.1 | 19.4 | 40.9 | 48.4 | 40.9 | 36.4 | 39.5 | 32.4 | 38.0 | 36.9 |
| Qwen3.5-9B (Thinking Mode) [74] | 32.9 | 21.9 | 14.5 | 19.4 | 22.7 | **34.5** | 23.6 | 45.9 | 66.7 | 32.7 | 29.3 | 25.8 | 27.3 | 45.7 | 24.3 | 31.8 | 30.3 | 31.1 | 28.2 | 31.7 |
| *Socially-Grounded RL (Ours)* | | | | | | | | | | | | | | | | | | | | |
| Qwen3-VL-8B + *SGR* | **37.1** | 65.8 | 47.3 | 41.8 | 43.2 | 33.3 | 46.3 | 50.8 | **90.1** | 38.8 | 39.0 | 25.8 | 59.1 | 58.0 | **53.9** | 22.7 | **51.3** | 43.2 | 48.1 | 50.4 |
| Qwen3.5-9B + *SGR* | **37.1** | **70.2** | **50.0** | **46.3** | **46.6** | 27.4 | **48.0** | **54.9** | **90.1** | **59.2** | **46.3** | **41.9** | 63.6 | **64.4** | 48.7 | 40.9 | 48.7 | 39.2 | **45.6** | **52.6** |

We then define a social grounding reward that verifies whether the reasoning refers to the correct participants of the interaction. Unlike perception tasks that rely on categorical labels, social interactions are inherently *relational*, requiring the model to identify involved participants and their interactions. Let $\mathcal{P}_i^{\text{pred}}$ denote the set of person identities mentioned within the `<gaze>` and `<gesture>` blocks of trajectory $y_i$, and let $\mathcal{P}^{\text{gt}}(q)$ denote the set of participants involved in the target social event specified in $q$. The grounding reward is defined using a weighted precision–recall formulation:

$$r_i^{\text{gnd}} = w_i \cdot P(\mathcal{P}_i^{\text{pred}}, \mathcal{P}^{\text{gt}}(q)), \qquad w_i = 1 + R(\mathcal{P}_i^{\text{pred}}, \mathcal{P}^{\text{gt}}(q)), \tag{5}$$

where $P(\cdot)$, $R(\cdot)$ measure precision, recall over participant IDs by comparing predicted and target sets based on their overlap. By balancing coverage and selectivity, our reward penalizes both noisy mentions of all visible individuals and overly narrow predictions that miss key participants, which encourages models to identify relevant interacting individuals without introducing irrelevant ones.

# 5 Experiments

**Implementation Details.** We build our framework on two baselines [3, 74]. The input videos are sampled at 2 fps and rendered with person-tracked bounding boxes, where each individual is marked with IDs (*e.g., P0*, *P1*, ...) to support grounded references. Training proceeds in two stages: we first fine-tune the models on the 52K small subset of *GRASP*, followed by our policy optimization on the 238K MCQ subset with $K = 8$ rollouts with KL regularization. The composite reward is set as $(\lambda_{\text{acc}}, \lambda_{\text{fmt}}, \lambda_{\text{str}}, \lambda_{\text{gnd}}) = (1.0, 0.1, 0.05, 0.2)$. All models are trained in bfloat16 on 8×NVIDIA H200 GPUs with DeepSpeed ZeRO-2, and full training details are provided in the Appendix C.

**Experimental Setting.** We evaluate on *GRASP-Bench*, a curated test set covering 16 categories, as our in-domain benchmark, and on three public social-reasoning benchmarks to assess cross-benchmark generalization: MMSI [39], Online-MMSI [44], and TVQA+ [42]. MMSI probes multi-party social reasoning through three tasks, including speaker target identification (STI), pronoun coreference resolution (PCR), and mentioned-player prediction (MPP), while Online-MMSI evaluates the same tasks under a causally truncated setting that admits only past frames. TVQA+ measures spatio-temporally grounded video QA over multi-character dialogues. We compare against three groups of MLLMs: (i) proprietary models [2, 62, 24], (ii) instruction-tuned models [4, 79, 1, 3, 74], and (iii) RL-based reasoning models [77, 86, 19, 12, 79, 1, 3, 74].

## 5.1 Main Results on *GRASP-Bench*

We report the main results on *GRASP-Bench* in Tab. 1. Overall, our socially-grounded RL consistently outperforms both supervised fine-tuned models and existing RL post-training baselines across all reasoning types. The gains are particularly evident in gaze and gesture reasoning, where accurate identification of interacting individuals is critical, as well as in cross-modal reasoning that requires

Table 2: Cross-benchmark generalization results on MMSI, Online-MMSI, and TVQA+. MMSI and Online-MMSI are evaluated on three social reasoning tasks: STI, PCR, and MPP.

| Method | MMSI | | | | Online-MMSI | | | | TVQA+ |
|---|---|---|---|---|---|---|---|---|---|
| | STI | PCR | MPP | Avg. | STI | PCR | MPP | Avg. | Acc. |
| *Supervised Fine-Tuned Models* | | | | | | | | | |
| Qwen2.5-VL-7B-Instruct [4] | 57.4 | 64.1 | **52.3** | 57.4 | 55.2 | 65.5 | 42.8 | 53.7 | 70.4 |
| InternVL3.5-8B-Instruct [79] | 51.5 | 58.3 | 50.3 | 53.0 | 47.4 | 60.6 | 44.0 | 49.9 | 67.1 |
| LLaVA-OV-1.5-8B-Instruct [1] | 64.9 | 65.8 | 50.9 | 60.2 | 55.8 | 63.3 | 41.1 | 52.7 | 72.5 |
| Qwen3-VL-8B-Instruct [3] | 67.5 | 63.6 | 48.7 | 59.7 | 59.3 | 66.6 | 43.9 | 55.9 | 71.9 |
| Qwen3.5-9B (Instruct Mode) [74] | 67.9 | **68.2** | 50.1 | 61.6 | 56.6 | 68.0 | 43.6 | 55.2 | 70.6 |
| *RL Post-Training Models* | | | | | | | | | |
| InternVL3.5-8B [79] | 49.2 | 42.7 | 37.5 | 43.2 | 43.8 | 58.1 | 43.3 | 47.6 | 70.9 |
| VideoChat-R1.5-7B [86] | 60.4 | 65.8 | 46.9 | 57.1 | 56.9 | 68.3 | 38.5 | 53.5 | 69.9 |
| LongVILA-R1-7B [12] | 50.5 | 44.6 | 30.5 | 41.7 | 40.9 | 45.7 | 28.4 | 37.8 | 59.4 |
| Video-R1-7B [19] | 58.6 | 60.8 | 48.8 | 55.7 | 55.0 | 63.6 | 40.3 | 52.2 | 71.2 |
| LLaVA-OV-1.5-8B-RL [1] | 63.3 | 59.8 | 44.1 | 55.5 | 52.8 | 57.6 | 38.1 | 48.9 | 68.4 |
| VL-Rethinker-7B [77] | 62.1 | 62.7 | 47.1 | 56.9 | 56.3 | 64.1 | 41.8 | 53.3 | 66.8 |
| Qwen3-VL-8B-Thinking [3] | 70.4 | 57.0 | 48.8 | 59.0 | 62.0 | 60.8 | 42.5 | 54.7 | 70.0 |
| Qwen3.5-9B (Thinking Mode) [74] | 70.2 | 61.9 | 48.2 | 60.0 | **63.1** | 61.4 | 39.7 | 54.3 | 65.6 |
| *Socially-Grounded RL (Ours)* | | | | | | | | | |
| Qwen3-VL-8B + *SGR* | 67.4 | 63.6 | 49.8 | 60.1 | 60.6 | 63.9 | **45.1** | 56.0 | **73.2** |
| Qwen3.5-9B + *SGR* | **71.2** | 66.9 | 49.2 | **62.1** | 60.3 | **70.4** | 42.8 | **56.9** | 70.9 |

Table 3: Ablation on reward components for social grounding and their impact on social reasoning performance.

| Method | Reward Comp. | | | | *GRASP-Bench* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $r_i^{\text{acc}}$ | $r_i^{\text{fmt}}$ | $r_i^{\text{str}}$ | $r_i^{\text{gnd}}$ | Gaze R. | Gest. R. | Joint R. | Easy | Med | Hard | Avg. |
| Baseline: *Qwen3-VL-8B* | | | | | | | | | | | |
| Instruct (SFT) | – | – | – | – | 28.3 | 52.7 | 38.0 | 49.9 | 28.5 | 36.0 | 38.3 |
| *GRPO* (RFT) | ✓ | ✓ | – | – | 34.3 | 50.8 | 41.1 | 56.1 | 32.4 | 33.9 | 41.1 |
| *SGR* (RFT) | ✓ | ✓ | ✓ | ✓ | **46.3** | **58.0** | **48.1** | **63.1** | **45.0** | **42.1** | **50.4** |
| Baseline: *Qwen3.5-9B* | | | | | | | | | | | |
| Instruct (SFT) | – | – | – | – | 32.8 | 59.0 | 40.8 | 56.6 | 34.7 | 36.8 | 43.0 |
| *GRPO* (RFT) | ✓ | ✓ | – | – | 34.1 | 56.6 | 48.4 | 57.7 | 36.9 | 39.5 | 44.6 |
| *SGR* (RFT) | ✓ | ✓ | ✓ | ✓ | **48.0** | **64.4** | **45.6** | **65.5** | **49.3** | **41.9** | **52.6** |

Table 4: Model scaling effects with *SGR* on *GRASP-Bench* evaluation.

| Model | #Params | *GRASP-Bench* | | | **Avg.** |
|---|---|---|---|---|---|
| | | Gaze R. | Gest. R. | Joint R. | |
| *Baseline: Qwen3-VL* | | | | | |
| Q3VL + *SGR* | 2B | 43.0 | 56.9 | 38.0 | 46.2 |
| Q3VL + *SGR* | 4B | 43.2 | 58.5 | 44.6 | 48.3 |
| Q3VL + *SGR* | 8B | 46.3 | 58.0 | 48.1 | 50.4 |
| *Baseline: Qwen3.5* | | | | | |
| Q3.5 + *SGR* | 2B | 41.4 | 57.2 | 36.8 | 45.3 |
| Q3.5 + *SGR* | 4B | 46.2 | 62.0 | 41.4 | 48.5 |
| Q3.5 + *SGR* | 9B | 48.0 | 64.4 | 45.6 | 52.6 |

integrating multiple social cues. These results indicate that explicitly grounding reasoning in non-verbal evidence leads to more reliable and coherent social understanding, rather than relying on coarse visual perception or language priors.

A notable trend is that SFT models are generally more competitive than RL baselines, suggesting that generic RL post-training for longer reasoning does not necessarily improve social understanding. Without an explicit grounding signal, additional reasoning can amplify spurious assumptions and distract models from the non-verbal cues and participant relations that determine the answer. In contrast, *SGR* rewards reasoning that refers to the correct participants in the underlying event, retaining useful deliberation while suppressing unsupported shortcuts. We provide qualitative comparisons in Appendix D, where RL baselines often produce plausible but ungrounded rationales, whereas our method reasons from the relevant people and non-verbal cues.

### 5.2 Generalization to Other Social Understanding Datasets

We evaluate cross-benchmark generalization on other social understanding datasets [39, 44, 42] in a zero-shot setting, without further task-specific fine-tuning. As shown in Tab. 2, *SGR*-trained models outperform or remain competitive with corresponding instruction-tuned backbones and RL post-training baselines. Given that our model's post-training is performed only on *GRASP*, these results indicate that (i) the models do not over-specialize to the *GRASP-Bench* format and (ii) *SGR* training is beneficial for broader social reasoning tasks beyond *GRASP*'s gaze and gesture focused questions. Furthermore, we provide qualitative reasoning-trace comparisons for these benchmarks in Appendix D.2, demonstrating that *SGR*-trained models effectively leverage gaze and gesture as intermediate cues on these benchmarks.

### 5.3 Ablation and Diagnostic Analysis

**Detailed Comparison with GRPO.** We conduct a detailed comparison with GRPO [69] to understand whether improvements arise from stronger optimization alone or from explicitly grounding

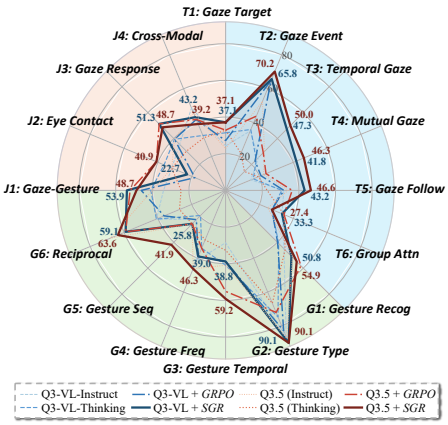


Figure 4: Per-category comparison on *GRASP-Bench* across baselines including SFT, reasoning, GRPO-tuned, and ours.

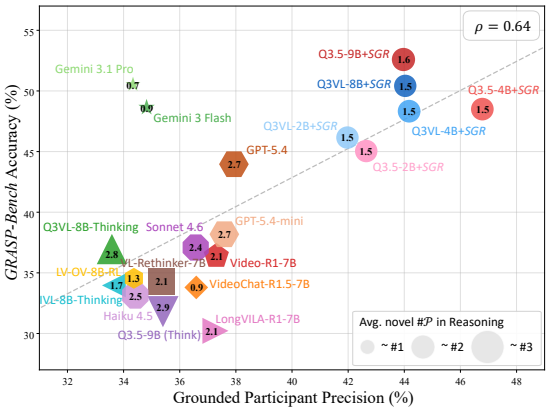


Figure 5: Grounded participant precision—accuracy on *GRASP-Bench* across various reasoning baselines. Marker size reflects the average number of novel participants mentioned in the reasoning trace.

social reasoning. While the RL method encourages models to generate reasoning before answering, it does not enforce that such reasoning is aligned with the correct social evidence. As a result, models often produce verbose but ungrounded reasoning traces that fail to identify the relevant individuals involved in social interactions. As shown in Tab. 3, incorporating GRPO on top of the baseline yields moderate improvements, indicating that reasoning supervision alone provides some benefit. However, adding our *SGR* leads to substantially larger gains across all metrics, highlighting the importance of explicitly grounding reasoning in non-verbal cues.

Fig. 4 further provides a detailed breakdown across all categories. While instruction model and zero-shot reasoning models exhibit inconsistent performance, and GRPO improves certain categories, it still struggles to reliably capture fine-grained social cues. In contrast, ours consistently improves performance, demonstrating that our grounding reasoning leads to coherent social understanding.

**Scaling with Model Size.** We analyze model-size scaling with *SGR* in Tab. 4 and compare the corresponding grounding behavior in Fig. 5. Across the baseline backbones [3, 74], average accuracy improves with model size, increasing from 46.2 to 50.4 and from 45.3 to 52.6, respectively. In Fig. 5, the scaled *SGR*-trained models remain in the high grounded-participant-precision region while achieving increasingly strong *GRASP-Bench* accuracy. This shows that the scaling gains come together with participant-grounded reasoning, rather than only stronger answer prediction.

**Grounding Precision Analysis.** To verify whether our reward indeed encourages models to identify the correct participants, we analyze grounding precision across reasoning models. We measure the proportion of correctly mentioned participants in the reasoning trace, alongside overall accuracy. As shown in Fig. 5, the precision across baselines exhibits positive correlation with accuracy ($\rho = 0.64$), indicating that correctly identifying relevant individuals is closely tied to successful social reasoning. We further observe distinct differences in grounding behavior across models. Many RL baselines achieve moderate accuracy despite low grounding precision, often mentioning multiple irrelevant participants, which leads to noisy reasoning traces and higher false positives. In contrast, our method achieves higher grounding precision with fewer participant mentions, reflecting more selective and targeted reasoning. This behavior aligns with the objective of *SGR*, which explicitly encourages the model to focus on the correct interacting individuals rather than enumerating all visible candidates.

## 6 Discussion and Conclusion

**Discussion.** We introduced *GRASP* as a large-scale social reasoning dataset grounded in fine-grained non-verbal events. Constructing such supervision is inherently challenging: despite our quality-control pipeline and human validation (Appendix A.7), the resulting annotations can still be noisy in ambiguous scenes. Moreover, we propose *SGR*, a simple but effective reward signal that encourages models to ground reasoning in the participants involved in the target interaction. We also examine a possible ID-copying shortcut through ID corruption analysis (Appendix C.3); the observed degradation indicates that copying question IDs alone does not explain the learned behavior. Still,

*SGR* mainly verifies participant-level grounding, and incorporating stricter temporal alignment and richer social cues remains important future work. Broader limitations are discussed in Appendix E.

**Conclusion.** We introduced *GRASP*, a large-scale dataset for grounded social reasoning that connects high-level social QA with fine-grained gaze and gesture events in multi-person videos. *GRASP* provides 290K QA pairs over 46K videos, organized by a 16-category taxonomy, together with *GRASP-Bench* for social reasoning evaluation. Moreover, we proposed *SGR*, a reward signal that uses structured social events to encourage models to reason about the participants involved in each interaction. Experiments show that participant-grounded supervision improves performance on *GRASP-Bench* while preserving zero-shot performance on related social video QA tasks.

## Acknowledgements

We acknowledge the support of the UIUC Investment for Growth Program and the Health Care Engineering Systems Center within the Coordinated Science Lab at UIUC. This work was also supported in part by MSIT through NRF Korea (RS-2025-00563942), IITP (IITP-2026-RS-2020-II201819, 10%; IITP-2026-RS-2026-25616664), and the IITP AI Computing Support Project for R&D (RS-2026-25505492, High-Performance Research AI Computing Infrastructure Support at the 2 PFLOPS Scale, 10%).

## References

[1] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

[2] Anthropic. System card: Claude sonnet 4.6. `https://www.anthropic.com/claude-sonnet-4-6-system-card`, feb 2026. Official system card.

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[5] Jun Bao, Buyu Liu, and Jun Yu. Escnet: Gaze target detection with the understanding of 3d scenes. In *CVPR*, pages 14126–14135, 2022.

[6] Tonko EW Bossen, Andreas Møgelmose, and Ross Greer. Can vision-language models understand and interpret dynamic gestures from pedestrians? pilot datasets and exploration towards instructive nonverbal commands for cooperative autonomous vehicles. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4779–4788, 2025.

[7] Xu Cao, Pranav Virupaksha, Wenqi Jia, Bolin Lai, Fiona Ryan, Sangmin Lee, and James M Rehg. Socialgesture: Delving into multi-person gesture understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19509–19519, 2025.

[8] Xu Cao, Pranav Virupaksha, Sangmin Lee, Bolin Lai, Wenqi Jia, Jintai Chen, and James Matthew Rehg. Toward human deictic gesture target estimation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[9] Xu Cao, Houze Yang, Vipin Gunda, Zhongyi Zhou, Tianyu Xu, Adarsh Kowdle, Inki Kim, and James M Rehg. Gaze target estimation anywhere with concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[10] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[11] Wei-Lin Chen, Liqian Peng, Tian Tan, Chao Zhao, Blake JianHang Chen, Ziqian Lin, Alec Go, and Yu Meng. Think deep, not just long: Measuring llm reasoning effort via deep-thinking tokens. *arXiv preprint arXiv:2602.13517*, 2026.

[12] Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, et al. Scaling rl to long videos. *arXiv preprint arXiv:2507.07966*, 2025.

[13] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *CVPR*, pages 5396–5406, 2020.

[14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

[15] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020.

[16] Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. Inferring shared attention in social scene videos. In *CVPR*, pages 6460–6468, 2018.

[17] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5724–5733, 2019.

[18] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. Dual attention guided gaze target detection in the wild. In *CVPR*, pages 11390–11399, 2021.

[19] Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

[20] Chris D Frith and Uta Frith. Mechanisms of social cognition. *Annual review of psychology*, 63:287–313, 2012.

[21] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118, 2025.

[22] Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, et al. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026.

[23] Émilie Gagnon-St-Pierre, Marina M Doucerain, and Henry Markovits. Reasoning strategies explain individual differences in social reasoning. *Journal of Experimental Psychology: General*, 150(2):340, 2021.

[24] Google Deepmind. Gemini 3.1 pro model card. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`, feb 2026. Official system card.

[25] Anshul Gupta, Samy Tafasca, Arya Farkhondeh, Pierre Vuillecard, and Jean-marc Odobez. Mtgs: A novel framework for multi-person temporal gaze following and social gaze prediction. *Advances in Neural Information Processing Systems*, 37:15646–15673, 2024.

[26] Anshul Gupta, Samy Tafasca, and Jean-Marc Odobez. A modular multimodal architecture for gaze target prediction: Application to privacy-sensitive settings. In *CVPRW*, pages 5041–5050, 2022.

[27] Anshul Gupta, Pierre Vuillecard, Arya Farkhondeh, and Jean-Marc Odobez. Exploring the zero-shot capabilities of vision-language models for improving gaze following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 615–624, 2024.

[28] Judith A Hall, Terrence G Horgan, and Nora A Murphy. Nonverbal communication. *Annual review of psychology*, 70(2019):271–294, 2019.

[29] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

[30] Muhammet Ilaslan, Chenan Song, Joya Chen, Difei Gao, Weixian Lei, Qianli Xu, Joo Lim, and Mike Shou. Gazevqa: A video question answering dataset for multiview eye-gaze task-oriented collaborations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10462–10479, 2023.

[31] Tianlei Jin, Qizhi Yu, Shiqiang Zhu, Zheyuan Lin, Jie Ren, Yuanhai Zhou, and Wei Song. Depth-aware gaze-following via auxiliary networks for robotics. *Engineering Applications of Artificial Intelligence*, 113:104924, 2022.

[32] Mathis Jording, Arne Hartz, Gary Bente, Martin Schulte-Rüther, and Kai Vogeley. The "social gaze space": A taxonomy for gaze-based communication in triadic interactions. *Frontiers in psychology*, 9:226, 2018.

[33] Caixin Kang, Yifei Huang, Liangyang Ouyang, Mingfang Zhang, Ruicong Liu, and Yoichi Sato. Can mllms read the room? a multimodal benchmark for assessing deception in multi-party social interactions. *arXiv preprint arXiv:2511.16221*, 2025.

[34] Alexander Kapitanov, Karina Kvanchiani, Alexander Nagaev, Roman Kraynov, and Andrei Makhliarchuk. Hagrid–hand gesture recognition image dataset. In *WACV*, pages 4572–4581, 2024.

[35] Kobin H Kendrick, Judith Holler, and Stephen C Levinson. Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical transactions of the royal society B*, 378(1875):20210473, 2023.

[36] Fanqi Kong, Weiqin Zu, Xinyu Chen, Yaodong Yang, Song-Chun Zhu, and Xue Feng. Siv-bench: A video benchmark for social interaction understanding and reasoning. *arXiv preprint arXiv:2506.05425*, 2025.

[37] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[38] Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James Rehg, and Diyi Yang. Werewolf among us: Multimodal resources for modeling persuasion behaviors in social deduction games. In *Findings of ACL*, pages 6570–6588, 2023.

[39] Sangmin Lee, Bolin Lai, Fiona Ryan, Bikram Boote, and James M Rehg. Modeling multimodal social interactions: new challenges and baselines with densely aligned representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14585–14595, 2024.

[40] Sangmin Lee, Minzhi Li, Bolin Lai, Wenqi Jia, Fiona Ryan, Xu Cao, Ozgur Kara, Bikram Boote, Weiyan Shi, Diyi Yang, et al. Towards social ai: A survey on understanding social interactions. *arXiv preprint arXiv:2409.15316*, 2024.

[41] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018.

[42] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8211–8225, 2020.

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

[44] Xinpeng Li, Shijian Deng, Bolin Lai, Weiguo Pian, James M Rehg, and Yapeng Tian. Towards online multi-modal social interaction understanding. *arXiv preprint arXiv:2503.19851*, 2025.

[45] Xinpeng Li, Bolin Lai, Hardy Chen, Shijian Deng, Cihang Xie, Yuyin Zhou, James Matthew Rehg, and Yapeng Tian. Omni-mmsi: Toward identity-attributed social interaction understanding. *arXiv preprint arXiv:2604.00267*, 2026.

[46] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *ECCV*, pages 619–635, 2018.

[47] Zhuoming Li, Aitong Liu, Mengxi Jia, Yubo Lu, Tengxiang Zhang, Changzhi Sun, Dell Zhang, and Xuelong Li. Gestura: A lvlm-powered system bridging motion and semantics for real-time free-form gesture understanding. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(4):1–29, 2025.

[48] Zongyu Lin, Zhikun Xu, Xiaohan Song, Yixin Wan, Xingcheng Yao, Tsung-Han Lin, Selina Song, Pranav Subbaraman, Ben Zhou, Kai-Wei Chang, et al. V-alphasocial: Benchmark and self-reflective chain-of-thought generation for visual social commonsense reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19025–19047, 2025.

[49] Dan Liu, Libo Zhang, and Yanjun Wu. Ld-congr: A large rgb-d video dataset for long-distance continuous gesture recognition. In *CVPR*, pages 3304–3312, 2022.

[50] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[51] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

[52] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

[53] Manuel Marin-Jimenez, Andrew Zisserman, and Vittorio Ferrari. " here's looking at you, kid": Detecting people looking at each other in videos. In *BMVC*. British Machine Vision Association and Society for Pattern Recognition, 2011.

[54] Athul M Mathew, Haithem Hermassi, Thariq Khalid, and Arshad Ali Khan. Gazevlm: A vision-language model for multi-task gaze understanding. *arXiv preprint arXiv:2511.06348*, 2025.

[55] Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. Social genome: Grounded social reasoning abilities of multimodal models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24879–24902, 2025.

[56] Claire McLean, Makenzie Meendering, Tristan Swartz, Orri Gabbay, Alexandra Olsen, Rachel Jacobs, Nicholas Rosen, Philippe de Bree, Tony Garcia, Gadsden Merrill, et al. Embody 3d: A large-scale multimodal motion and behavior dataset. *arXiv preprint arXiv:2510.16258*, 2025.

[57] David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[58] Chris Moore, Philip J Dunham, and Phil Dunham. *Joint attention: Its origins and role in development*. Psychology Press, 2014.

[59] Le Thien Phuc Nguyen, Zhuoran Yu, Samuel Low Yu Hang, Subin An, Jeongik Lee, Yohan Ban, SeungEun Chung, Thanh-Huy Nguyen, JuWan Maeng, Soochahn Lee, et al. See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models. *arXiv preprint arXiv:2512.02231*, 2025.

[60] Lixing Niu, Jiapeng Li, Xingping Yu, Xinyi Dong, Shu Wang, Ruining Feng, Bo Wu, Ping Wei, Yisen Wang, and Lifeng Fan. Read the room: Video social reasoning with mental-physical causal chains. In *The Fourteenth International Conference on Learning Representations*, 2026.

[61] Lixing Niu, Jiapeng Li, Xingping Yu, Shu Wang, Ruining Feng, Bo Wu, Ping Wei, Yisen Wang, and Lifeng Fan. Rˆ 3-vqa:" read the room" by video social reasoning. *arXiv preprint arXiv:2505.04147*, 2025.

[62] OpenAI. Gpt-5.4 thinking system card. `https://openai.com/index/gpt-5-4-thinking-system-card/`, mar 2026. Official system card.

[63] Liangyang Ouyang, Yifei Huang, Mingfang Zhang, Caixin Kang, Ryosuke Furuta, and Yoichi Sato. Multi-speaker attention alignment for multimodal social interaction. *arXiv preprint arXiv:2511.17952*, 2025.

[64] Anupam Pani and Yanchao Yang. Gaze-vlm: Bridging gaze and vlms through attention regularization for egocentric understanding. *arXiv preprint arXiv:2510.21356*, 2025.

[65] Sungjune Park, Hyunjun Kim, Junho Kim, Seongho Kim, and Yong Man Ro. Dip-r1: Deep inspection and perception with rl looking through and understanding complex scenes. *arXiv preprint arXiv:2505.23179*, 2025.

[66] Taiying Peng, Jiacheng Hua, Miao Liu, and Feng Lu. In the eye of mllm: Benchmarking egocentric video intent understanding with gaze-guided prompting. *arXiv preprint arXiv:2509.07447*, 2025.

[67] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? *NeurIPS*, 28, 2015.

[68] Fiona Ryan, Ajay Bati, Sangmin Lee, Daniel Bolya, Judy Hoffman, and James M Rehg. Gaze-lle: Gaze target estimation via large-scale learned encoders. 2025.

[69] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[70] Heung-Yeung Shum, Xiao-dong He, and Di Li. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19:10–26, 2018.

[71] Yuehao Song, Xinggang Wang, Jingfeng Yao, Wenyu Liu, Jinglin Zhang, and Xiangmin Xu. Vitgaze: gaze following with interaction features in vision transformers. *Visual Intelligence*, 2(1):1–15, 2024.

[72] Jinyan Su, Jennifer Healey, Preslav Nakov, and Claire Cardie. Between underthinking and overthinking: An empirical study of reasoning length and correctness in llms. *arXiv preprint arXiv:2505.00127*, 2025.

[73] Hamza Tahboub, Weiyan Shi, Gang Hua, and Huaizu Jiang. Socialfusion: Addressing social degradation in pre-trained vision-language models. *arXiv preprint arXiv:2512.01148*, 2025.

[74] Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

[75] Bhaavanaa Thumu, Leena Mathur, Youssouf Kebe, and Louis-Philippe Morency. Social caption: Evaluating social understanding in multimodal models. *arXiv preprint arXiv:2601.14569*, 2026.

[76] Michael Tomasello and Michael Jeffrey Farrar. Joint attention and early language. *Child development*, pages 1454–1463, 1986.

[77] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *arXiv preprint arXiv:2504.08837*, 2025.

[78] Shijing Wang, Chaoqun Cui, Yihua Cheng, and Yaping Huang. Gaze following in question answering: A comprehensive benchmark for vision-language models, 2025.

[79] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[80] Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Qun Liu, and Dongyan Zhao. Friends-mmc: A dataset for multi-modal multi-party conversation understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25425–25433, 2025.

[81] Jianan Wei, Tianfei Zhou, Yi Yang, and Wenguan Wang. Nonverbal interaction detection. In *European Conference on Computer Vision*, pages 277–295. Springer, 2024.

[82] Alex Wilf, Leena Mathur, Sheryl Mathew, Claire Ko, Youssouf Kebe, Paul Pu Liang, and Louis-Philippe Morency. Social-iq 2.0 challenge: Benchmarking multimodal social understanding. `https://github.com/abwilf/Social-IQ-2.0-Challenge`, 2023.

[83] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 47(8):1087–1102, 2023.

[84] Yuyang Wu, Yifei Wang, Ziyu Ye, Tianqi Du, Stefanie Jegelka, and Yisen Wang. When more is less: Understanding chain-of-thought length in llms. *arXiv preprint arXiv:2502.07266*, 2025.

[85] Tianyu Xie, Jinfa Huang, Yuexiao Ma, Rongfang Luo, Yan Yang, Wang Chen, Yuhui Zeng, Ruize Fang, Yixuan Zou, Xiawu Zheng, et al. Socialomni: Benchmarking audio-visual social interactivity in omni models. *arXiv preprint arXiv:2603.16859*, 2026.

[86] Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *arXiv preprint arXiv:2509.21100*, 2025.

[87] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817, 2019.

[88] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

[89] Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu. Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. *TMM*, 20(5):1038–1050, 2018.

[90] Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Xihan Wei, Liefeng Bo, et al. Humanomni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*, 2025.

## Appendix Contents


**A Detailed *GRASP* Construction** **15**

A.1 Video Pre-processing and Person-Consistent Gaze Trajectories . . . . . 15

A.2 Gaze Feature Extraction . . . . . 15

A.3 Social Gaze Event Detection . . . . . 16

A.4 Deictic Gesture Annotation . . . . . 16

A.5 Unified Social Graph Construction . . . . . 17

A.6 Social Reasoning QA Generation . . . . . 18

A.7 Quality Control and Human Validation . . . . . 18

**B Detailed *GRASP* Data Statistics** **19**

B.1 Per-Dataset Scale . . . . . 20

B.2 Gaze and Gesture Event Distributions . . . . . 21

B.3 QA Category Distribution . . . . . 22

B.4 Construction Yield . . . . . 22

B.5 Event Density and Training Distribution . . . . . 22

B.6 Comparison with Specialized Social VQA Benchmarks . . . . . 23

**C Additional Analyses and Experimental Details** **23**

C.1 Training Details . . . . . 23

C.2 Other Benchmark Details . . . . . 24

C.3 Participant-ID Corruption Analysis . . . . . 25

C.4 Reasoning Length Breakdown . . . . . 25

C.5 *SGR* Error Profile . . . . . 26

**D Qualitative Comparison** **27**

D.1 *GRASP-Bench* Examples . . . . . 27

D.2 Reasoning Trace Comparison with Baselines . . . . . 27

D.3 Failure Cases . . . . . 27

**E Broader Impacts** **27**

# A Detailed *GRASP* Construction

## A.1 Video Pre-processing and Person-Consistent Gaze Trajectories

To construct temporally grounded social interactions, we first normalize raw videos into short interaction-centered clips. Long-form videos are segmented using scene-cut detection (PySceneDetect[3]). We enforce a minimum segment length of 45 seconds by merging overly short adjacent segments, and a maximum segment length of 120 seconds by splitting long segments at the nearest available scene boundary. All downstream gaze processing samples videos at 2 fps, so the temporal resolution of gaze-derived annotations is 0.5 seconds.

For each processed clip, we estimate person-consistent gaze trajectories through a multi-stage identity-aware pipeline. We first detect and track people using SAM 3 [10] across frames using a text-prompted person tracking model with the prompt "people", yielding temporally consistent person bounding boxes and participant identities. Since these person tracks do not provide sufficiently precise facial localization for gaze estimation, we additionally detect faces using [15] in each sampled frame. Each detected face is associated with a tracked person identity by solving a Hungarian matching problem [37] between face boxes and the upper half of the corresponding person boxes, which approximates the head region. Matches with zero overlap are rejected to avoid assigning a face to an unrelated body track.

After face-person association, we estimate the gaze target with AnyGaze [9] for each matched face. The gaze estimator predicts a normalized gaze point $g_i^t = (x_i^t, y_i^t) \in [0,1]^2$ for person $i$ at time $t$, together with an in-frame indicator. The final per-frame representation therefore aligns each participant identity with a person box, face box, gaze point, and gaze validity flag. This identity-preserving representation is the basis for all subsequent gaze feature computation and social event detection.

Table 5: Key components used to construct person-consistent gaze trajectories.

| Component | Role | Key setting |
|---|---|---|
| Scene segmentation | Split long-form videos into coherent interaction segments. | 45–120s segments |
| Frame sampling | Normalize the temporal resolution for downstream gaze processing. | 2 fps; 0.5s interval |
| Person tracking | Maintain stable participant identities across frames. | Text prompt "people"; confidence threshold 0.45 |
| Face detection | Localize faces for gaze estimation. | Per sampled frame; confidence threshold 0.54 |
| Face-person matching | Assign each detected face to a tracked participant identity. | Hungarian matching on the upper 50% person region; zero-overlap matches rejected |
| Gaze estimation | Predict where each participant is looking. | Normalized frame coordinate $g_i^t \in [0,1]^2$ with in-frame flag |

## A.2 Gaze Feature Extraction

Raw gaze trajectories contain missing observations caused by missed face detections, out-of-frame gaze, occlusion, and abrupt camera changes. We therefore apply a conservative interpolation scheme before computing temporal features. Missing gaps of up to 3 frames are linearly interpolated, as shown in Tab. 6 . Gaps of 4–10 frames are filled by carrying forward the last valid observation with decayed confidence. Longer gaps are left missing. To prevent spurious continuity across cuts or large viewpoint changes, interpolation is blocked whenever the temporal gap exceeds 3 seconds or the face displacement exceeds 30% of the frame width.

To reduce sensitivity to camera motion, gaze velocity is computed relative to the person's face center rather than from raw gaze point displacement. Let $c_i^t$ denote the face center of person $i$ at time $t$. We define the face-centered gaze direction as

$$d_i^t = g_i^t - c_i^t,$$

and estimate gaze velocity as:

$$v_i^t = \frac{\|d_i^t - d_i^{t-1}\|}{\Delta t}.$$

[3]We use AdaptiveDetector with default setup in `https://github.com/Breakthrough/PySceneDetect`.

This formulation captures changes in where a person looks relative to their own head position, making it more robust to camera pans and global frame motion.

We also compute a group-level convergence score to identify shared attention. For all valid in-frame gaze points at time $t$, we compute their centroid $\bar{g}^t$ and measure the median distance from individual gaze points to this centroid. The convergence score is

$$s^t = \exp\left(-3.0 \cdot \text{median}_i \|g_i^t - \bar{g}^t\|\right).$$

Using the median distance makes the score robust to outliers, such as one participant looking away while several others attend to a common region.

Table 6: Missing-gaze handling before temporal feature computation.

| Missing gap | Strategy | Assigned confidence |
|---|---|---|
| $\leq 3$ frames | Linear interpolation | $1.0 - 0.1 \times \text{gap}$ |
| 4–10 frames | Carry forward | $0.5 \times \exp(-0.2 \times \text{gap})$ |
| $> 10$ frames | Keep missing | 0 |

### A.3 Social Gaze Event Detection

From the per-person velocities and group convergence features, we detect five classes of social gaze events: sudden gaze shift, joint attention, gaze following, attention capture, and mutual gaze, as summarized in Tab. 7. Each event type is defined through an operational detector over identity-consistent gaze trajectories.

Table 7: Operational definitions for the five social gaze event types used in *GRASP*.

| Event type | Event definition | Main criterion |
|---|---|---|
| Sudden gaze shift | One participant rapidly changes gaze direction. | $v_i^t > 0.7$ |
| Joint attention | At least two participants attend to a common region. | $s^t \geq 0.6$, duration $\geq 0.5$s |
| Gaze following | One participant looks toward another participant's previous gaze target. | distance $< 0.03$, lag 1.0–2.0s |
| Attention capture | At least three participants shift gaze within a short temporal window. | $v_i^t > 0.4$ for $\geq 3$ persons |
| Mutual gaze | Two participants look at each other's face regions. | bidirectional face-box hit, duration $\geq 1.0$s |

For sudden gaze shifts, we cluster consecutive high-velocity frames using a maximum temporal gap of 0.6 seconds and retain events whose durations fall between 0.5 and 1.5 seconds. Joint attention is detected by identifying frames with high convergence score and at least two visible faces, then clustering consecutive frames whose participant sets have at least 70% overlap. Participants far from the convergence center are filtered out to avoid including late joiners or peripheral observers.

Gaze following is detected by comparing a participant's current gaze target against another participant's prior high-confidence measured gaze target. We only use measured gaze histories for this detector and require a temporal lag between 1.0 and 2.0 seconds. Attention capture is detected when at least three visible participants exhibit high gaze velocity within a short temporal window, indicating a simultaneous shift of attention. Mutual gaze is detected through a bidirectional hit test: person $A$'s gaze must fall inside person $B$'s face box while person $B$'s gaze falls inside person $A$'s face box. To reduce false positives, mutual gaze uses measured gaze only, not interpolated gaze, and applies a 2% bounding-box margin. Only high-confidence gaze events are passed to the unified graph construction stage. In the final construction pipeline, downstream processing keeps gaze events with confidence at least 0.9.

### A.4 Deictic Gesture Annotation

In addition to gaze, *GRASP* models communicative hand gestures that direct attention or indicate interaction targets. We use a `Gemini 3.1 Pro` [24] to detect four deictic gesture types: pointing, showing, giving, and reaching (Tab. 8). To provide explicit identity and temporal grounding, each input video is rendered with tracked person boxes, visible participant labels $P0, P1, \ldots$, and timestamp overlays $t = \text{X.XXs}$. The model is instructed to scan the entire video, identify all visible deictic gestures, assign initiator and target IDs from the overlay, and use the displayed timestamp for start and end times. The full gesture classification prompt is shown in Fig. 26.

We use a strict and conservative gesture definition. A gesture is annotated only when its type and intent are clearly identifiable from hand, finger, or arm motion alone. The prompt explicitly excludes inference from gaze direction, head pose, speech, facial expression, scene context, or gaze visualization artifacts. Ambiguous hand movements, beat gestures, casual arm motion, and resting poses are skipped. We require a minimum duration of 1.0 second, allow simultaneous gestures by different participants, and return an empty set when no high-confidence deictic gesture is visible.

We further apply an iterative verification-and-regeneration step. After the initial Gemini prediction, a separate verification pass scores the predicted gesture set out of 7 points: 2 points for timeline alignment and 5 points for visual-context accuracy, including gesture type, initiator ID, target person ID, reliance on deictic hand/arm motion, and coverage of visible gesture events. Each subcriterion is scored as 1.0, 0.5, or 0.0. If the total score is at most 5/7, we regenerate the gesture prediction by providing the previous prediction and verification report as feedback. This process reduces timestamp errors, swapped participant identities, false positives from non-deictic motion, and missed salient gestures.

Each retained gesture is stored with structured fields including gesture type, initiator identity, target type, target person identity when applicable, start time, end time, and confidence. The final construction keeps gestures with confidence at least 0.85. Free-form target descriptions and reasoning text produced by Gemini are used only as contextual language for graph and QA generation. Ground-truth answers are derived only from structured, verifiable fields such as gesture type, initiator identity, target identity, and temporal boundaries.

Table 8: Gesture types annotated in *GRASP* and their conservative annotation criteria.

| Gesture type | Annotation criterion |
|---|---|
| Pointing | Clear arm, hand, or finger extension toward a specific person or target. |
| Showing | Intentionally orienting or presenting an object for another participant to inspect. |
| Giving | Offering an object to another participant with transfer intent. |
| Reaching | Directed hand extension toward an object or target with acquisition intent. |

### A.5 Unified Social Graph Construction

After extracting gaze events and gesture annotations, we merge them into a unified social graph for each video. The graph represents social interactions as temporally ordered event nodes. Each node stores the event source, event type, involved participant identities, start and end times, confidence, and modality-specific attributes such as gaze convergence statistics, leader-follower identities, gesture initiators, or gesture targets.

Before graph construction, gaze events and gestures are filtered using their respective confidence thresholds. Gaze events are retained at confidence $\geq 0.9$, and gestures are retained at confidence $\geq 0.85$. All timestamps are snapped to 0.5-second intervals to match the 2 fps processing resolution. Events whose temporal boundaries exceed the video duration are removed. Duplicate events are removed when they share the same type, the same participants, and an overlapping temporal window.

To support joint reasoning over coordinated non-verbal cues, we identify gaze–gesture pairs whose temporal distance is at most 3.0 seconds. These pairs become the basis for Joint Reasoning questions in categories J1–J4. During graph pruning, gaze–gesture pairs are protected before single-modality events are removed, because randomly dropping one side of a pair would destroy the evidence needed for joint reasoning. When a video contains many events, we cap the graph at 25 events, prioritizing protected gaze–gesture pairs, event-type diversity, and high-confidence events.

Table 9: Key filtering and merging parameters for unified social graph construction.

| Parameter | Value | Purpose |
|---|---|---|
| Gaze confidence threshold | $\geq 0.9$ | Retain high-precision gaze events. |
| Gesture confidence threshold | $\geq 0.85$ | Retain high-confidence gesture annotations. |
| Timestamp snapping | 0.5s | Match the 2 fps processing resolution. |
| Joint-pair proximity | $\leq 3.0$s | Link temporally adjacent gaze and gesture events. |
| Maximum events per video | 25 | Control graph size while preserving key social evidence. |

| Modality | Subcategory | Example |
|---|---|---|
| **Gaze** | T1: Gaze Target ID | *Q: At around 2.7 seconds, who is Person 0 looking at?*<br>A: Person 2 |
| | T2: Gaze Event Classification | *Q: What best describes the interaction between Person 0 and Person 1 around 2.0 seconds?*<br>A: Looking at the same thing |
| | T3: Mutual Gaze Recognition | *Q: How long do Person 1 and Person 2 maintain eye contact starting at 1.5 seconds?*<br>A: 3.0 seconds |
| | T4: Gaze Following Detection | *Q: Which pair of people makes eye contact between 2.5 and 4.5 seconds?*<br>A: Person 3 and Person 5 |
| | T5: Temporal Gaze Reasoning | *Q: In the gaze following event between 1.0s and 3.0s, who looks at the target first?*<br>A: Person 6 |
| | T6: Group Attention Dynamics | *Q: During the shared attention event between 2.0 and 3.0 seconds, which person is NOT part of the group?*<br>A: Person 3 |
| **Gesture** | G1: Gesture Recognition | *Q: Between 3.0 and 5.0 seconds, who is Person 3 pointing at?*<br>A: Person 0 |
| | G2: Gesture Type | *Q: What type of gesture does Person 1 perform at 3.0 seconds?*<br>A: Reaching |
| | G3: Gesture Temporal | *Q: Between 2.5 and 5.0 seconds, Person 10 performs two gestures. Which happens first?*<br>A: Showing |
| | G4: Reciprocal Patterns | *Q: What is the most common gesture type performed throughout the video clip?*<br>A: Pointing |
| | G5: Gesture Frequency | *Q: Person 2 gives to Person 1 between 1.5–6.5s. Who does Person 1 give to next?*<br>A: Person 0 |
| | G6: Sequence Chains | *Q: At 2.5s, Person 3 reaches toward Person 2. Who is Person 2 simultaneously reaching toward?*<br>A: Person 1 |
| **Joint** | J1: Gaze-Gesture Alignment | *Q: Which happens first for Person 0: their sudden gaze shift, or their pointing gesture towards Person 2?*<br>A: The sudden gaze shift happens first |
| | J2: Gaze Response to Gesture | *Q: Just as Person 0 starts pointing at a card at 2.0s, who are they making eye contact with?*<br>A: Person 1 |
| | J3: Eye Contact / Interaction | *Q: Right before Person 1 starts a giving gesture at 3.5s, which two people share attention?*<br>A: Person 0 and Person 2 |
| | J4: Cross-Modal Dynamics | *Q: Across 1.5–30.0s, who initiates both showing gestures and joins every gaze event?*<br>A: Person 0 |

Table 10: Task examples in our 16-category social-reasoning benchmark. Each row shows one representative MCQ-style question–answer pair (only the correct answer text is shown for brevity). Difficulty progresses Easy → Hard along the rows within each modality (T1–T2 Easy, T3–T4 Medium, T5–T6 Hard; analogous for G/J).

### A.6 Social Reasoning QA Generation

*GRASP* question–answer pairs are generated from the unified social graphs using Gemini 3.1 Pro [24]. The goal is to convert structured gaze, gesture, and joint gaze–gesture events into natural-language social reasoning questions whose answers remain directly verifiable from event metadata. Each QA instance is anchored to one or more source events and includes a time range that identifies the relevant visual evidence.

We use the 16-category taxonomy defined in § 3.2 and Tab. 10, covering gaze reasoning (T1–T6), gesture reasoning (G1–G6), and joint reasoning (J1–J4). QA generation is adaptive to the event density of each graph. Clips with only a few gaze or gesture events generate simple perception-oriented questions, such as gaze target identification, gaze event classification, gesture recognition, and gesture type classification. Clips with richer event timelines generate harder temporal, reciprocal, frequency, sequence, group-attention, and joint reasoning questions. Joint reasoning questions are generated only when the graph contains temporally linked gaze–gesture pairs. The full QA generation prompt is shown in Fig. 27.

All generated QA pairs are validated before inclusion. We discard malformed multiple-choice questions, questions with missing or invalid time ranges, answers containing speculative or affective language, invalid person references, and category–modality mismatches. The correct answer is derived from structured graph fields such as event type, participant IDs, initiator and target IDs, timestamps, and gaze–gesture pair links, rather than from free-form generated descriptions. Metadata such as source event IDs, time ranges, and internal reasoning fields is used for verification and filtering.

### A.7 Quality Control and Human Validation

*GRASP* uses multiple levels of quality control to reduce noise from automatic annotation. First, all gaze and gesture events are filtered by confidence before graph construction. Second, gesture annotation follows a conservative protocol that skips ambiguous gestures and avoids inferring intent from non-gesture cues. Third, graph construction removes invalid temporal boundaries, duplicate events, and malformed event records while preserving gaze–gesture pairs required for joint reasoning.

Table 11: Human validation results on *GRASP-Bench*. Four evaluators audit approximately 25% of the benchmark in a category-balanced manner. We report the percentage of QA instances judged visually supported, temporally grounded, and unambiguous.

| Gaze Reasoning | | | Gesture Reasoning | | | Joint Reasoning | | |
|---|---|---|---|---|---|---|---|---|
| **Cat.** | **Difficulty** | **Validity (%)** | **Cat.** | **Difficulty** | **Validity (%)** | **Cat.** | **Difficulty** | **Validity (%)** |
| T1 Gaze Target ID | Easy | 64.3 | G1 Gesture Recog. | Easy | 73.3 | J1 Gaze-Gest. Align. | Medium | 86.7 |
| T2 Gaze Event Class. | Easy | 57.1 | G2 Gesture Type | Easy | 86.7 | J2 Gaze Resp. to Gest. | Medium | 81.2 |
| T3 Mutual Gaze | Medium | 73.3 | G3 Gesture Temporal | Medium | 78.6 | J3 Eye Contact / Inter. | Hard | 71.4 |
| T4 Gaze Following | Medium | 88.2 | G4 Reciprocal Gesture | Medium | 93.8 | J4 Joint Person Dynamics | Hard | 80.0 |
| T5 Temporal Gaze | Hard | 42.9 | G5 Gesture Frequency | Hard | 80.0 | – | – | – |
| T6 Group Attention | Hard | 62.5 | G6 Sequence Chains | Hard | 60.0 | – | – | – |
| **Gaze Avg.** | – | **65.6** | **Gesture Avg.** | – | **78.9** | **Joint Avg.** | – | **80.0** |

A graph-level health check verifies that required structured fields such as event type, participants, start time, end time, and confidence are present across the corpus.

We additionally conduct human validation as an independent reliability audit. We randomly sample approximately 25% of *GRASP-Bench* in a category-balanced manner. Four human evaluators participate in the study. As shown in Fig. 6, evaluators inspect each QA instance with the corresponding video clip, participant IDs, temporal context, category label, question, and answer options. They judge whether the provided answer is visually supported, temporally grounded, and unambiguous.

Tab. 11 summarizes the validation results. The audited subset obtains an overall validity of 74.2%, with gesture and joint reasoning showing higher validity rates of 78.9% and 80.0%, respectively. Gaze reasoning is comparatively more challenging, reflecting the difficulty of verifying fine-grained gaze direction, temporal ordering, and group attention in dense multi-person scenes. This validation serves as an independent sanity check on dataset reliability. The final dataset relies on structured event fields, confidence filtering, and deterministic validation rules, with human validation providing an additional estimate of QA quality.

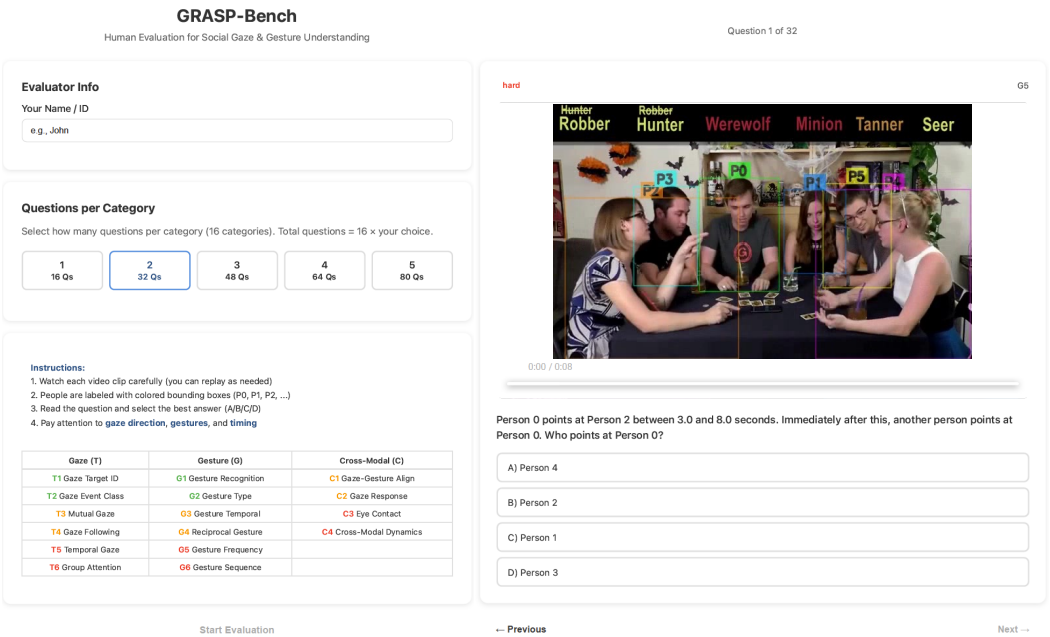


Figure 6: Human validation interface. Evaluators inspect each QA instance with the corresponding video clip, participant IDs, temporal context, category label, question, and answer options, then verify whether the QA pair is visually grounded and unambiguous.

## B Detailed *GRASP* Data Statistics

We summarizes the scale, composition, and construction yield of *GRASP*. Our video sources span diverse multi-person social settings, including multi-party TV dialogue [42, 80], social intelligence

reasoning [87], multi-speaker conversation [59], embodied multi-person behavior[4] [56], and social deduction gameplay [38, 7]. We additionally augment the social deduction domain with online video sources, since these videos contain dense multi-person interactions with rich gaze and gesture signals.

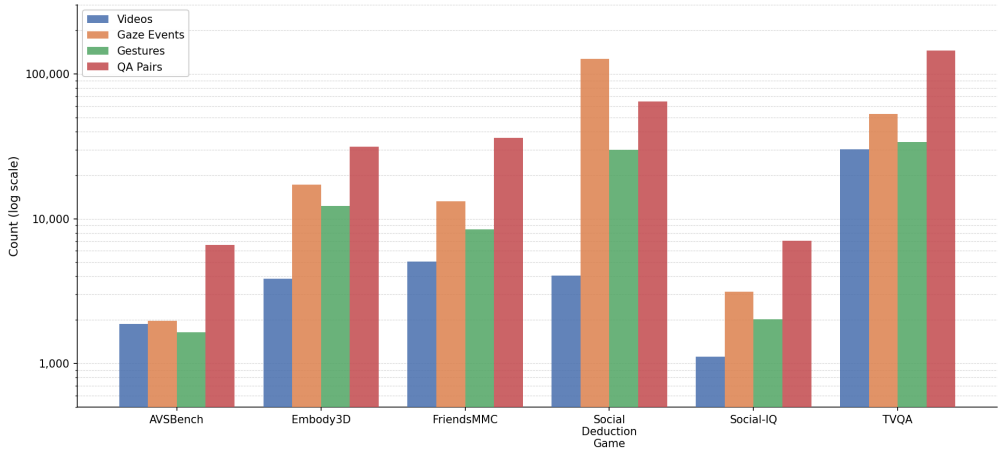


Figure 7: Dataset-level scale across the six source domains. We report the number of source videos, retained gaze events, detected gestures, and generated QA pairs for each dataset. The y-axis is shown in log scale to make datasets of different sizes comparable.

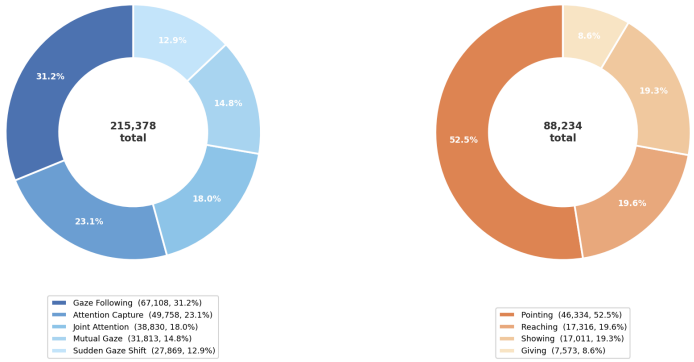


Figure 8: Distribution of retained gaze event types and gesture types. Gaze events are filtered at confidence $\geq 0.9$, and gestures are filtered at confidence $\geq 0.85$.

Across these sources, *GRASP* contains 46,158 processed videos. From these videos, we construct structured social representations by extracting person-consistent gaze trajectories and deictic gestures, then integrating them into temporally ordered social graphs anchored to tracked participant identities. The pipeline detects 429,663 raw gaze events, retains 215,378 high-confidence gaze events using a confidence threshold of 0.9, and retains 88,234 gesture events using a confidence threshold of 0.85. After temporal merging and event deduplication, the pipeline produces 290,148 QA pairs across 16 categories. The final training data contains 238,296 multiple-choice QA pairs for RL training and 51,852 open-ended QA pairs for SFT training.

### B.1 Per-Dataset Scale

Fig. 7 shows the per-dataset scale. In particular, Social Deduction Game contains 4,044 videos but contributes 127,033 gaze events and 30,006 gestures, reflecting the dense multi-person interaction patterns in this domain.

[4]We use only the multi-person subsets of Embody3D: *multi-person conversation*, *day-in-the-life*, and *scenarios*.

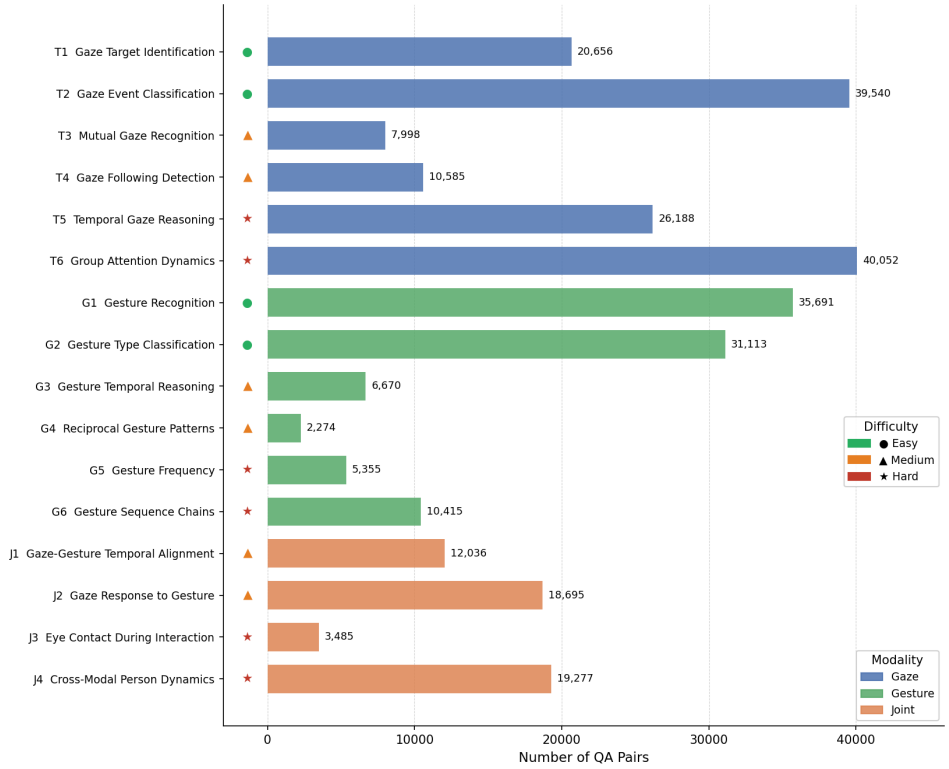


Figure 9: QA category distribution. The benchmark contains 16 categories: six gaze categories (T1–T6), six gesture categories (G1–G6), and four joint reasoning categories (J1–J4). Bar colors indicate modality, and markers indicate difficulty.

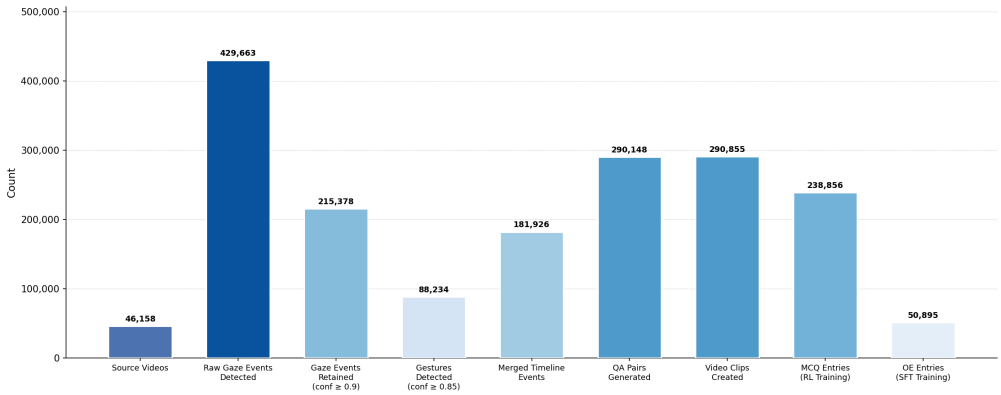


Figure 10: Data construction yield. The pipeline starts from source videos, detects raw gaze and gesture events, filters high-confidence gaze events, merges timeline events, and generates QA pairs and per-QA video clips for training.

## B.2 Gaze and Gesture Event Distributions

As shown in Fig. 8, the retained gaze events cover five event types: gaze following, attention capture, joint attention, mutual gaze, and sudden gaze shift. Gaze following is the most frequent category, accounting for 31.2% of retained gaze events, followed by attention capture at 23.1%. Gesture annotations cover four types: pointing, reaching, showing, and giving. Pointing is the dominant gesture type, accounting for 52.5% of all detected gestures, which is consistent with its role as a frequent cue in social and referential interaction.

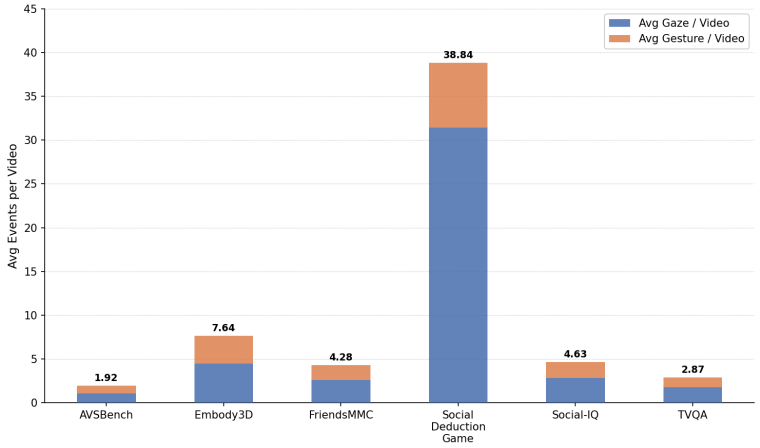


Figure 11: Average gaze and gesture event density per video. Social Deduction Game has the highest event density due to continuous multi-person social interaction.

### B.3 QA Category Distribution

The QA taxonomy is organized into three groups: gaze reasoning, gesture reasoning, and joint reasoning. Gaze categories include target identification, event classification, mutual gaze recognition, gaze following detection, temporal gaze reasoning, and group attention dynamics. Gesture categories cover recognition, type classification, temporal reasoning, reciprocal patterns, frequency, and sequence chains. Joint reasoning categories require relating gaze and gesture events over time, including gaze-gesture temporal alignment, gaze response to gesture, eye contact during interaction, and cross-modal person dynamics. We show their distribution in Fig. 9

### B.4 Construction Yield

Fig. 10 summarizes the full construction pipeline. The confidence filter retains 215,378 of 429,663 raw gaze detections, corresponding to 50.1% retention. The retained gaze events and detected gestures are then merged into 181,926 timeline events. Since each timeline event can support different question forms, including single-event recognition, temporal comparison, sequence reasoning, and cross-modal reasoning, the final number of QA pairs is larger than the number of merged timeline events.

### B.5 Event Density and Training Distribution

Fig. 11 reports the average number of retained gaze events and detected gestures per video across source domains. Event density varies substantially across domains. Social Deduction Game has the highest average density, with 31.42 gaze events and 7.42 gestures per video. Other datasets are sparser, typically containing fewer than eight total events per video. This variation is expected because the source datasets differ in interaction structure, video duration, and the frequency of visible social cues.

Fig. 12 summarizes the distribution of MCQ training examples by modality and difficulty. Easy examples are mostly direct gaze or gesture recognition questions, while medium and hard examples require temporal or cross-modal reasoning. Joint reasoning does not include easy questions by design, since these examples require relating gaze and gesture events or reasoning about person-level interaction dynamics.

Overall, the constructed data provides 145,049 gaze QA pairs, 91,536 gesture QA pairs, and 53,504 joint reasoning QA pairs. In terms of format, the final data is primarily multiple-choice, with 238,296 MCQ examples and 51,852 open-ended examples. The MCQ split is used for RL training, while the open-ended split is used for SFT.

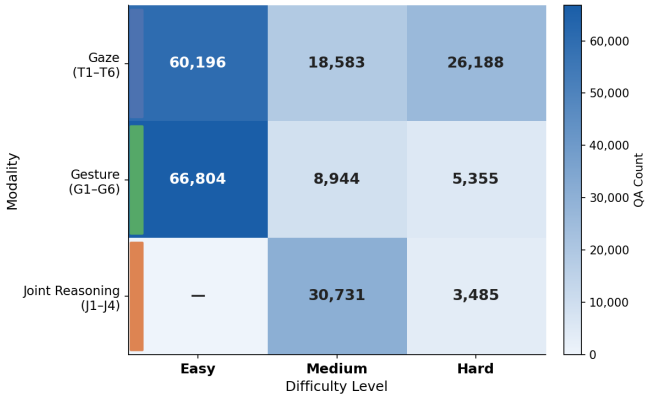


Figure 12: Distribution of MCQ training examples by modality and difficulty. Open-ended examples are excluded from this heatmap because they are used for SFT and are categorized as hard reasoning examples.

Table 12: Comparison of *GRASP* with existing multimodal social understanding benchmarks. Our dataset scales significantly beyond prior works while distinctly focusing on fine-grained, participant-level non-verbal grounding.

| **Benchmark** | **Input** | **Grounded Cues** | **#Samples** | **#QAs** | **Core Focus** |
|---|---|---|---|---|---|
| GazeVQA [78] | Image | Gaze | 122K | 410K | Re-label GazeFollow [67]. Static gaze target understanding in single frames |
| Social-IQ 2.0 [82] | Video | - | 1.2K | 7.5K | Unconstrained social situations, explicit human reasoning |
| SocialGesture [7] | Video | Gesture | 10K | 10K | Multi-person deictic gesture understanding |
| SIV-Bench [36] | Video | - | 2.7K | 5.4K | Social relation theory, state reasoning, dynamics prediction |
| Social Genome [55] | Video | - | 272 | 1.5K | Grounded multi-step reasoning traces and external knowledge |
| SocialOmni [85] | Video | - | 2K | 2.2K | Omni-modal conversational interactivity (who, when, how) |
| Omni-MMSI [45] | Video | - | 7K | 5K | Re-label MMSI [39]. Identity-attributed interaction and multi-party speaker reference |
| V-Social [48] | Video | - | 214 | 956 | Social relation recognition and grounding in long-form videos |
| $R^3$-VQA [60] | Video | - | 316 | 4.8K | Mental state estimation and social causal reasoning |
| MimeQA [43] | Video | - | 221 | 806 | Social reasoning for silent theatrical storytelling |
| **GRASP (Ours)** | **Video** | **Gaze, Gesture** | **46K** | **290K** | **Grounded participant-level events and dynamic social tracking** |

### B.6 Comparison with Specialized Social VQA Benchmarks

As summarized in Table 12, *GRASP* occupies a unique position in the social AI landscape. Unlike single grounded cues datasets such as GazeVQA [78] and SocialGesture [7], which excel in static spatial gaze target or gesture understanding and localization, *GRASP* emphasizes the temporal dynamics of social events (*e.g.,* joint attention and turn-taking). Furthermore, while reasoning-centric benchmarks like $R^3$-VQA [60] and Social Genome [55] offer deep cognitive insights, their limited scale often restricts their use to evaluation rather than post-training. *GRASP* provides the necessary scale (46K videos) and multi-modal grounding (gaze and gesture) to facilitate the alignment of MLLMs with nuanced human social behaviors.

## C Additional Analyses and Experimental Details

### C.1 Training Details

We train our models with the two-stage procedure described in § 4. First, we perform supervised fine-tuning on the open-ended *GRASP* subset to warm up the model with the structured reasoning format. We then apply GRPO-based reinforcement learning on the multiple-choice training split,

using answer correctness, output-format consistency, structural grounding, and social grounding as the reward components. The detailed hyperparameters are summarized in Tab. 13.

For both SFT and RL, input videos are sampled at 0.5 FPS and training is performed in *bfloat16* with AdamW, cosine learning-rate scheduling, gradient clipping, and DeepSpeed ZeRO-2. The SFT stage is trained for two epochs with a learning rate of $2 \times 10^{-6}$, while the RL stage is trained for one epoch with a learning rate of $3 \times 10^{-6}$. For GRPO, we use $K = 8$ rollouts per query, a KL penalty of $0.05$, mean-centered advantage normalization, and advantage clipping to $\pm 5.0$. The reward weights are set to $1.0/0.1/0.05/0.2$ for accuracy, format, structural grounding, and social grounding, respectively.

Table 13: Detailed training hyperparameters.

| Config | SFT | RL (GRPO) |
|---|---|---|
| Input frames | 2 FPS | 2 FPS |
| LR scheduler | Cosine with warm-up | Cosine with warm-up |
| Warmup steps | 200 | 256 |
| Optimizer | AdamW ($\beta_1$=0.9, $\beta_2$=0.999) | AdamW |
| Global batch size | 32 | 32 |
| Learning rate | $2 \times 10^{-6}$ | $3 \times 10^{-6}$ |
| Weight decay | $1 \times 10^{-3}$ | $1 \times 10^{-3}$ |
| Gradient clipping | 1.0 | 1.0 |
| Training precision | *bfloat16* | *bfloat16* |
| DeepSpeed | ZeRO-2 | ZeRO-2 |
| Epochs | 2 | 1 |
| Max generation length | – | 1024 |
| Number of rollouts $K$ | – | 8 |
| KL penalty $\beta$ | – | 0.05 |
| Advantage normalization | – | Mean-center |
| Advantage clipping | – | $\pm 5.0$ |
| Reward weights | – | 1.0 / 0.1 / 0.05 / 0.2 |

### C.2 Other Benchmark Details

We use MMSI [39], Online-MMSI [44], and TVQA+ [42] to test whether GRASP-Bench transfers beyond in-domain gaze/gesture reasoning to related forms of human-centered video understanding, including person-reference tracking in multi-party conversation and spatio-temporally grounded video QA.

MMSI is built from real multi-party social deduction game videos with 151 YouTube games totaling 14.8 hours and 48 Ego4D games totaling 7.3 hours. It defines three dialogue-based tasks: Speaking Target Identification (STI), Pronoun Coreference Resolution (PCR), and Mentioned Player Prediction (MPP). These tasks ask models to identify who is being addressed, referred to, or mentioned, with 4,087 STI, 3,182 PCR, and 3,832 MPP instances across the YouTube and Ego4D subsets. In this game setting, references occur in conversations shaped by accusation, defense, hidden-role claims, persuasion, and shifting group attention. Although gaze and gesture are not the labels being predicted, they are part of the visible evidence humans use to track who a speaker is addressing or discussing. Online-MMSI uses the same source data and tasks but restricts the input to context before the target utterance, giving us a related test of person-reference tracking in a setup more realistic to active dialogue.

TVQA+ comes from TVQA [41], which contains 152K QA pairs from 21.8K clips across 460 hours of six TV shows. TVQA+ augments the Big Bang Theory subset with spatial grounding annotations, yielding 29,383 multiple-choice questions from 4,198 clips, 148,468 annotated images, 310,826 bounding boxes, and 2,527 visual concept categories. The benchmark asks models to answer questions from subtitles and video while localizing relevant moments and grounding visual concepts such as people and objects. The questions are not strictly social reasoning, but they are human-centered, asking who did something, what someone was holding or wearing, where someone was, what direction someone turned, or what object was involved. These questions often require interpreting people's actions and relations to objects, space, and time, where gaze, gesture, body orientation, and movement can provide useful visual evidence.

Table 14: Participant-ID corruption test on *GRASP-Bench*. We consistently remap Person IDs in the question, answer options, and metadata answer text while keeping the video unchanged and preserving the correct answer letter. Normal and Corrupt denote accuracy (%), and Drop denotes Corrupt minus Normal in percentage points.

| Cat. | Task | Qwen3-VL-8B + *SGR* | | | Qwen3.5-9B + *SGR* | | |
|---|---|---|---|---|---|---|---|
| | | **Normal** | **Corrupt** | **Drop** | **Normal** | **Corrupt** | **Drop** |
| *Gaze Reasoning* | | | | | | | |
| T1 | Gaze Target ID | 37.1 | 27.1 | −10.0 | 37.1 | 21.4 | −15.7 |
| T2 | Gaze Event Class. | 65.8 | 61.4 | −4.4 | 70.2 | 59.1 | −11.1 |
| T3 | Mutual Gaze Recog. | 47.3 | 49.1 | +1.8 | 50.0 | 48.6 | −1.4 |
| T4 | Gaze Following | 41.8 | 19.4 | −22.4 | 46.3 | 16.4 | −29.9 |
| T5 | Temporal Gaze | 43.2 | 23.9 | −19.3 | 46.6 | 21.6 | −25.0 |
| T6 | Group Attention | 33.3 | 25.0 | −8.3 | 27.4 | 24.7 | −2.7 |
| → Gaze Avg. | | 46.3 | 35.9 | −10.4 | 48.0 | 33.7 | −14.3 |
| *Gesture Reasoning* | | | | | | | |
| G1 | Gesture Recog. | 50.8 | 26.4 | −24.4 | 54.9 | 29.8 | −25.1 |
| G2 | Gesture Type | 90.1 | 100.0 | +9.9 | 90.1 | 50.2 | −39.9 |
| G3 | Gesture Temporal | 38.8 | 36.6 | −2.2 | 59.2 | 22.0 | −37.2 |
| G4 | Reciprocal Gesture | 39.0 | 11.8 | −27.2 | 46.3 | 10.8 | −35.5 |
| G5 | Gesture Frequency | 25.8 | 23.3 | −2.5 | 41.9 | 22.6 | −19.3 |
| G6 | Sequence Chains | 59.1 | 63.6 | +4.5 | 63.6 | 50.0 | −13.6 |
| → Gesture Avg. | | 58.0 | 30.2 | −27.8 | 64.4 | 27.7 | −36.7 |
| *Joint Reasoning* | | | | | | | |
| J1 | Gaze-Gest. Align. | 53.9 | 34.9 | −19.0 | 48.7 | 37.6 | −11.1 |
| J2 | Gaze Resp. to Gest. | 22.7 | 31.8 | +9.1 | 40.9 | 18.2 | −22.7 |
| J3 | Eye Contact / Inter. | 51.3 | 34.2 | −17.1 | 48.7 | 35.5 | −13.2 |
| J4 | Joint Person Dynamics | 43.2 | 13.9 | −29.3 | 39.2 | 27.8 | −11.4 |
| → Joint Avg. | | 48.1 | 29.0 | −19.1 | 45.6 | 33.0 | −12.6 |
| **Overall** | | **50.4** | **32.6** | −17.8 | **52.6** | **32.0** | −20.6 |

### C.3 Participant-ID Corruption Analysis

A potential concern is that *SGR* could be exploited by simply repeating participant IDs that already appear in the question, rather than grounding the reasoning trace in the video. To diagnose this shortcut, we conduct a participant-ID corruption test on *GRASP-Bench*. For each MCQ example, we consistently remap all Person IDs in the question, answer options, and metadata answer text, while leaving the video unchanged and preserving the correct answer letter. This keeps the textual format and answer-position prior intact, but breaks the alignment between textual Person IDs and the visual Person IDs overlaid in the video. If a model mainly relied on text-only ID shortcuts, its accuracy should remain stable under this corruption. In contrast, a visually grounded model should degrade when the answer requires mapping textual participant references to the correct people in the video.

As shown in Tab. 14, both SGR-trained models drop substantially under participant-ID corruption. Qwen3-VL-8B+*SGR* drops from 50.4% to 32.6%, and Qwen3.5-9B+*SGR* drops from 52.6% to 32.0%. The degradation is strongest in categories that require resolving specific participants, such as gaze following (T4), temporal gaze reasoning (T5), gesture recognition (G1), reciprocal gestures (G4), and joint person dynamics (J4). For example, T4 drops by 22.4pp and 29.9pp, G4 drops by 27.2pp and 35.5pp, and J4 drops by 29.3pp and 11.4pp for the two models, respectively.

The category-level pattern also clarifies when corruption should or should not affect performance. Categories with weaker direct dependence on participant identity, such as mutual gaze duration or some group-attention questions, show smaller or mixed changes. In contrast, categories that require tracking who looks at whom, who gestures to whom, or who participates in both gaze and gesture events consistently show large drops. This diagnostic does not rule out every possible form of reward gaming, but it directly addresses the ID-repetition shortcut: the final SGR-trained policies depend on the consistency between textual participant references and visually grounded participant identities, rather than merely reusing Person IDs from the question.

### C.4 Reasoning Length Breakdown

To understand how accuracy and reasoning length relate, we plot the accuracy/length in Fig. 13 and perform Pearson correlation analysis. We find no correlation (Pearson r = + 0.039) between accuracy and average reasoning length for all baseline models for the collection of social reasoning tasks in

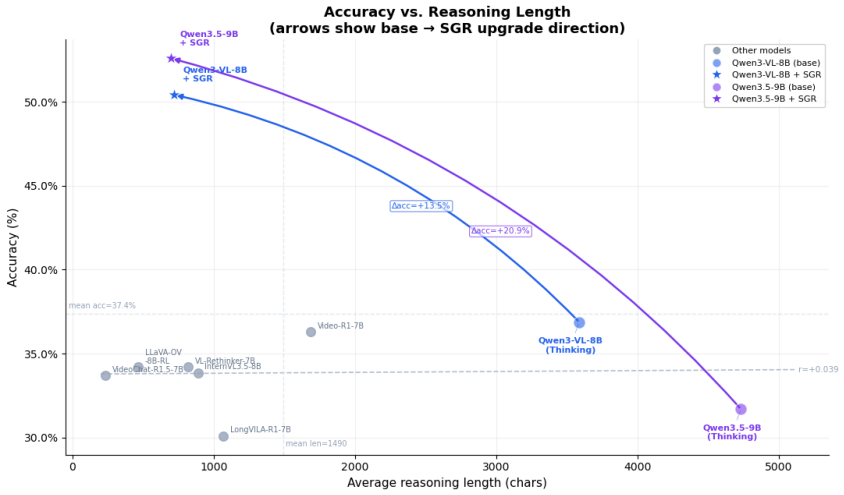


Figure 13: Accuracy compared against average reasoning length for all baselines.

*GRASP-Bench*. This runs counter to findings for other VQA tasks where longer reasoning helps performance. Here longer reasoning is not necessarily better reasoning. Meanwhile, our method simultaneously shortens the reasoning length for the our models and improves the accuracy.

### C.5 *SGR* Error Profile

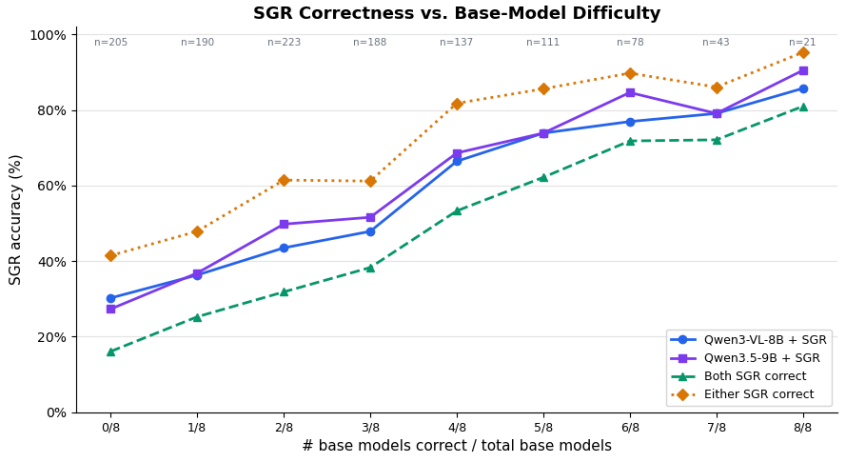


Figure 14: Error profile for *SGR* compared against difficulty of the questions for the base models.

To understand the difficulty of the *GRASP-Bench* tasks, we group cases into buckets defined by how many base models were correct, and the accuracy of the our models per bucket was computed as shown in Fig. 14. This gives an estimation of how hard the *GRASP* questions are for our comparison models off the shelf and allows us to evaluate the *SGR* error profile. There are three primary takeaways:

- The *SGR* performance tracks along the scale of difficulty described by the base model performance. This suggests there are other things contributing to the complexity of the benchmark tasks which are not being captured by our reward.
- There exist "hard" cases that all other models fail on where we recover using the *SGR* approach (like Fig. 1), but there are also "easier" cases where all other models are correct but our approach fails (examples in Appendix. D.3).

- Our models separately follow similar error profiles but our two models are making different errors, creating an opening for future work to investigate the characteristics of the errors being made by each model when supported by *SGR*.

# D Qualitative Comparison

We provide qualitative examples to complement the main results in the paper. The examples illustrate how *GRASP-Bench* questions are grounded in fine-grained gaze and gesture events, how different models reason over the same multi-person interaction, and where our method still fails.

## D.1 *GRASP-Bench* Examples

We present representative *GRASP-Bench* examples in Fig. 23, 24, and 25. These examples cover gaze reasoning (T1–T6), gesture reasoning (G1–G6), and joint gaze–gesture reasoning (J1–J4), respectively. Each example shows the video context with tracked participant IDs, the relevant non-verbal cue, and the corresponding question–answer pair. Together, they illustrate the range of *GRASP-Bench*, from direct perception questions such as gaze target or gesture type recognition to harder questions requiring temporal ordering, reciprocal interaction, and gaze–gesture alignment.

## D.2 Reasoning Trace Comparison with Baselines

We compare reasoning traces from our SGR-trained models against representative supervised, reasoning, and RL post-training baselines. Fig. 15, 16, and 17 show *GRASP-Bench* examples across easy, medium, and hard difficulty levels. In addition to *GRASP-Bench*, we include examples from MMSI [39] and TVQA+ [42]. Fig. 18, 19, and 20 show the three MMSI tasks: speaker target identification (STI), pronoun coreference resolution (PCR), and mentioned-player prediction (MPP). Fig. 21 shows a TVQA+ example. Across these cases, baselines often rely on salient participants or dialogue priors, while our models more consistently ground their responses in the relevant participants and visual/social cues.

## D.3 Failure Cases

Fig. 22 shows representative failure cases of our method. In the first example, the gesture is visually ambiguous: the queried interval contains a point-like phase of a reaching motion, while later frames make the reaching/recoil pattern clearer. Such cases are challenging because the distinction between pointing and reaching can be subtle even under human inspection. In the second example, the target interval contains multiple gaze events from the same participant, with several people involved in nearby social activity. The model misses one relevant gaze cue and therefore grounds its answer in incomplete evidence. These examples suggest that remaining errors often arise from ambiguous gesture semantics and limited resolution of dense multi-person gaze dynamics.

# E Broader Impacts

*GRASP* and *SGR* are intended to advance research on multimodal social understanding. By encouraging models to reason about who interacts with whom from gaze and deictic gestures, our work may benefit human-centered video understanding, assistive agents, collaborative robotics, and socially aware interfaces where participant-level grounding is important.

At the same time, social reasoning from video requires careful interpretation. *GRASP* is constructed from publicly available or research-oriented video sources, and people are represented through anonymous participant IDs rather than real identities. The benchmark focuses on event-level visual reasoning, such as who looks at whom or who gestures toward whom, and is not designed for identifying individuals, profiling people, or inferring private mental states. Since gaze and gesture can be ambiguous, culturally variable, and sensitive to occlusion or camera viewpoint, models trained or evaluated on *GRASP* should not be used for high-stakes decisions about people.

Our dataset and models may also inherit biases from the source videos and from the foundation models used in the construction pipeline. Social settings, camera styles, gestures, and interaction norms in the data may not represent all populations or contexts. In addition, automatic gaze and

gesture extraction can introduce annotation noise in complex scenes. We mitigate these issues through confidence filtering, structured validation, and subset human feedback, but improving diversity and event reliability remains important future work.

Finally, *GRASP* uses multimodal and language models for gesture annotation, event description, and QA generation. To reduce unsupported inferences, ground-truth answers are derived from structured fields such as participant IDs, event types, timestamps, and gaze–gesture links, rather than free-form generated text. We believe that releasing *GRASP* as a documented research benchmark can help the community study socially grounded MLLMs more transparently, while making the limitations and intended scope of the dataset explicit.

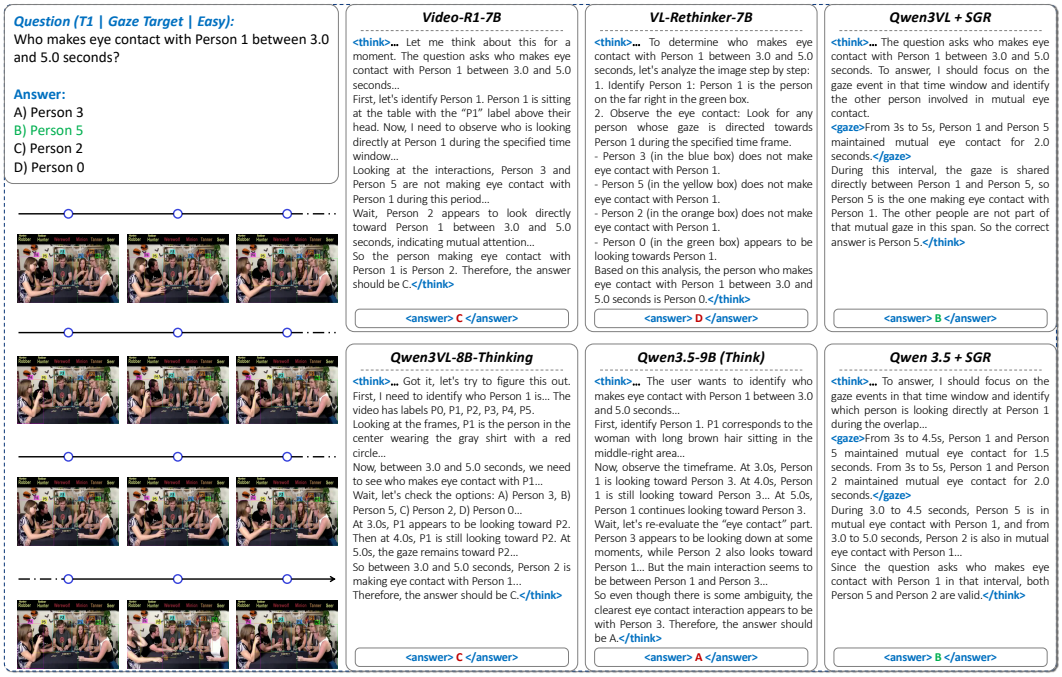


Figure 15: Qualitative comparison on an easy *GRASP-Bench* example.

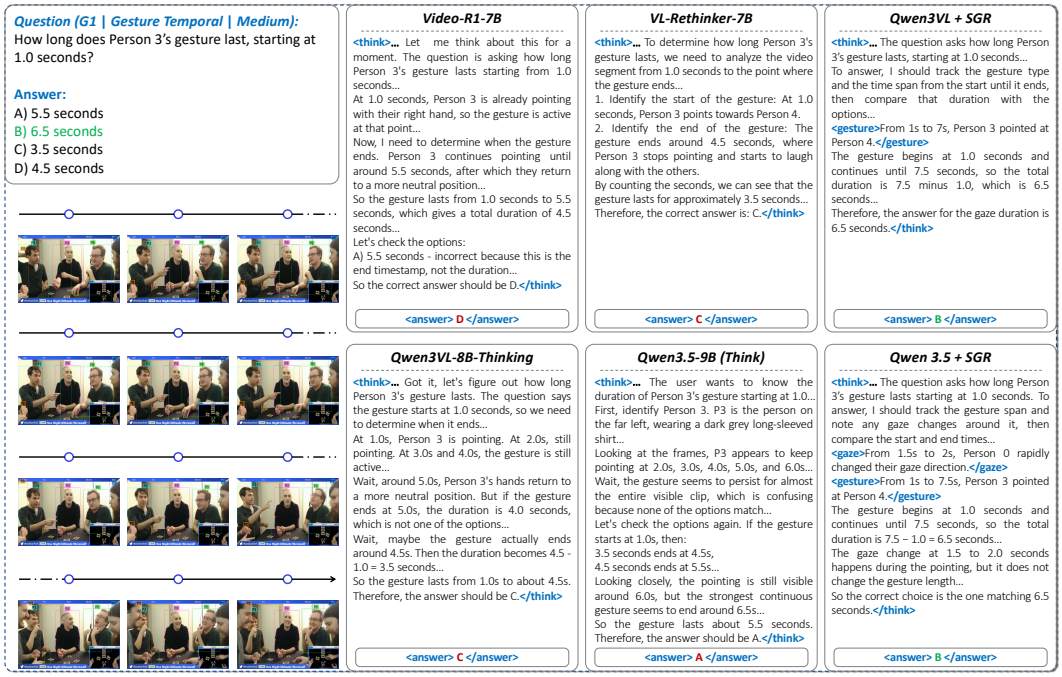


Figure 16: Qualitative comparison on a medium *GRASP-Bench* example.

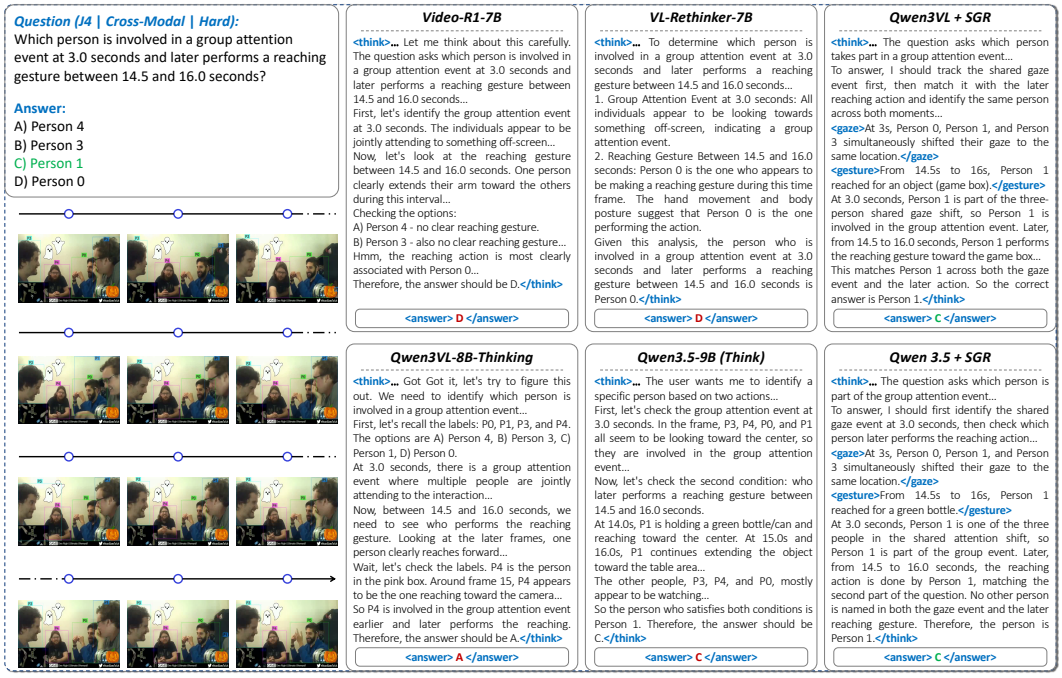


Figure 17: Qualitative comparison on a hard *GRASP-Bench* example.

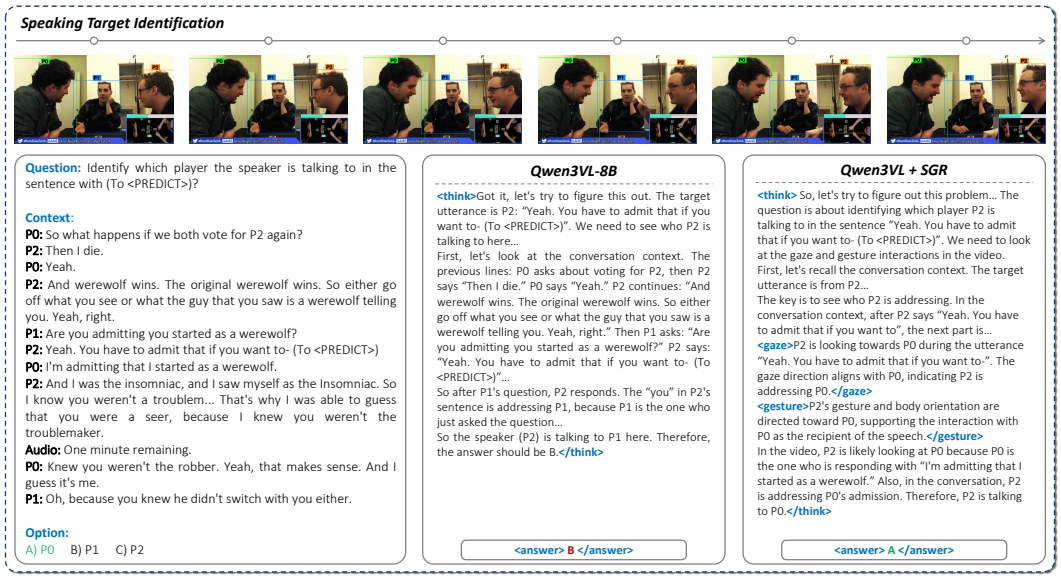


Figure 18: Qualitative comparison on MMSI speaker target identification (STI).

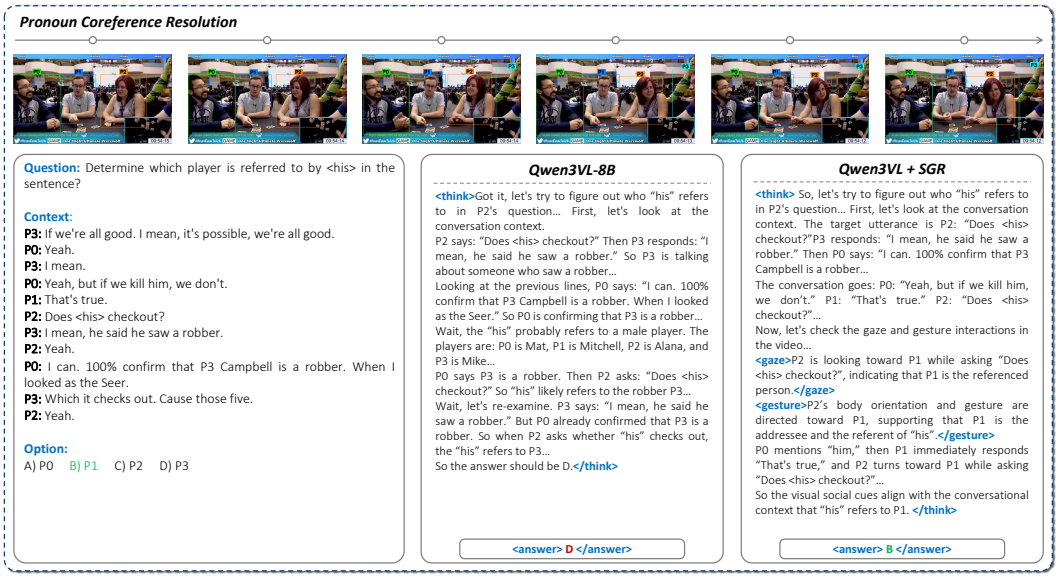


Figure 19: Qualitative comparison on MMSI pronoun coreference resolution (PCR).

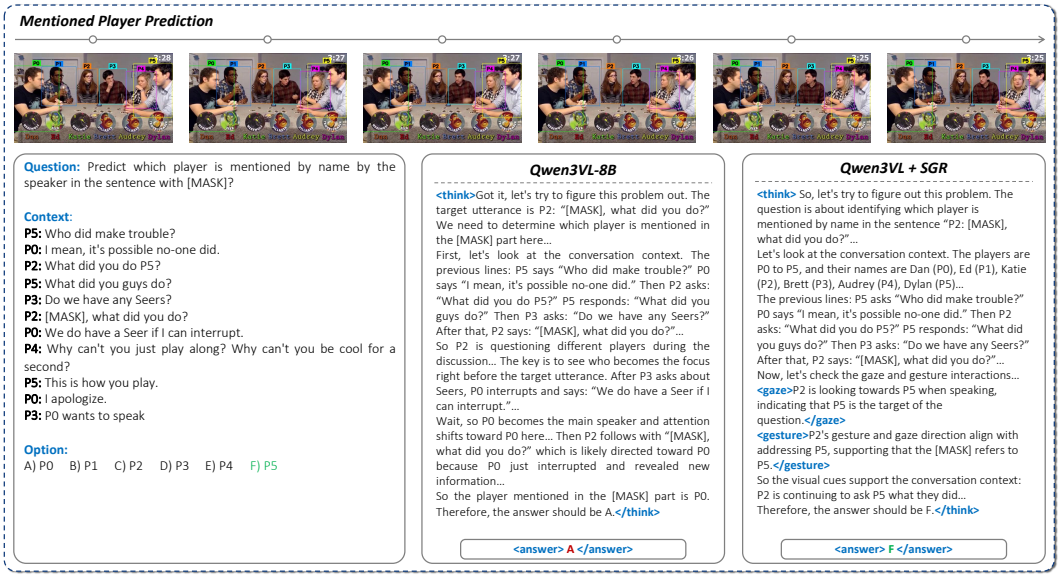


Figure 20: Qualitative comparison on MMSI mentioned-player prediction (MPP).

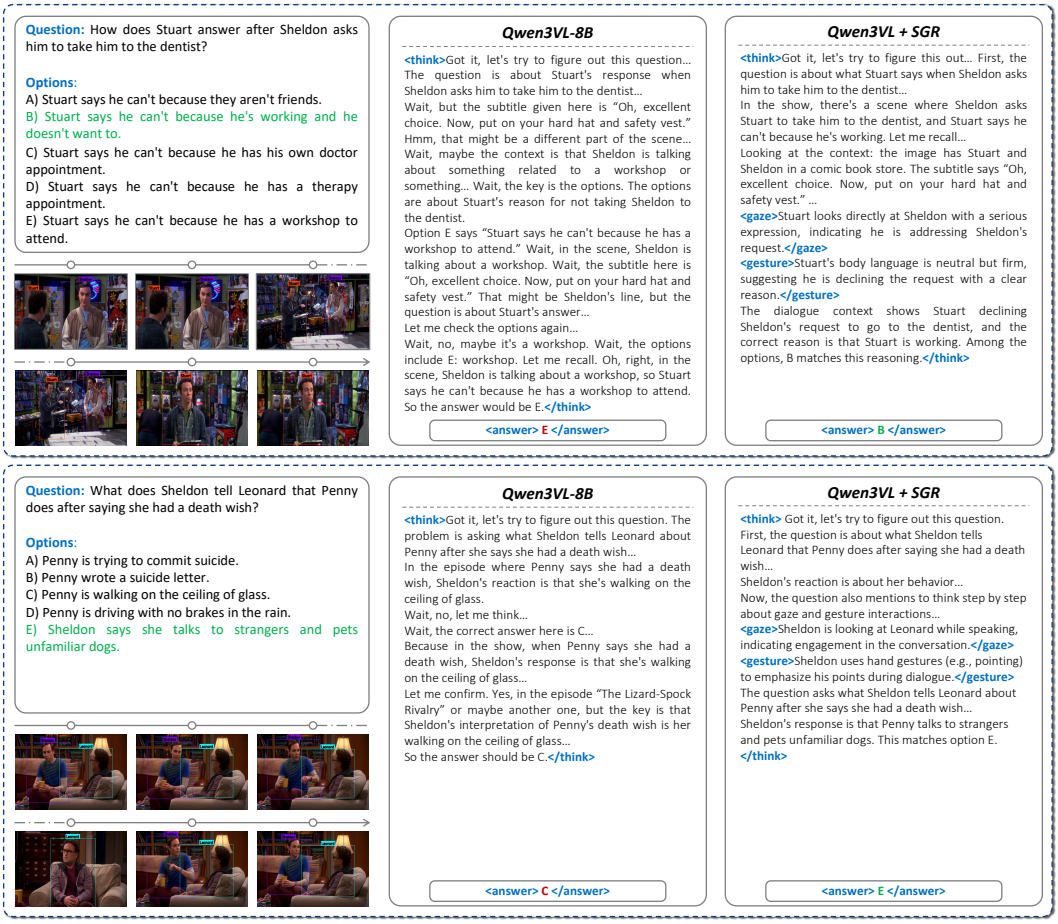


Figure 21: Qualitative comparison on TVQA+.

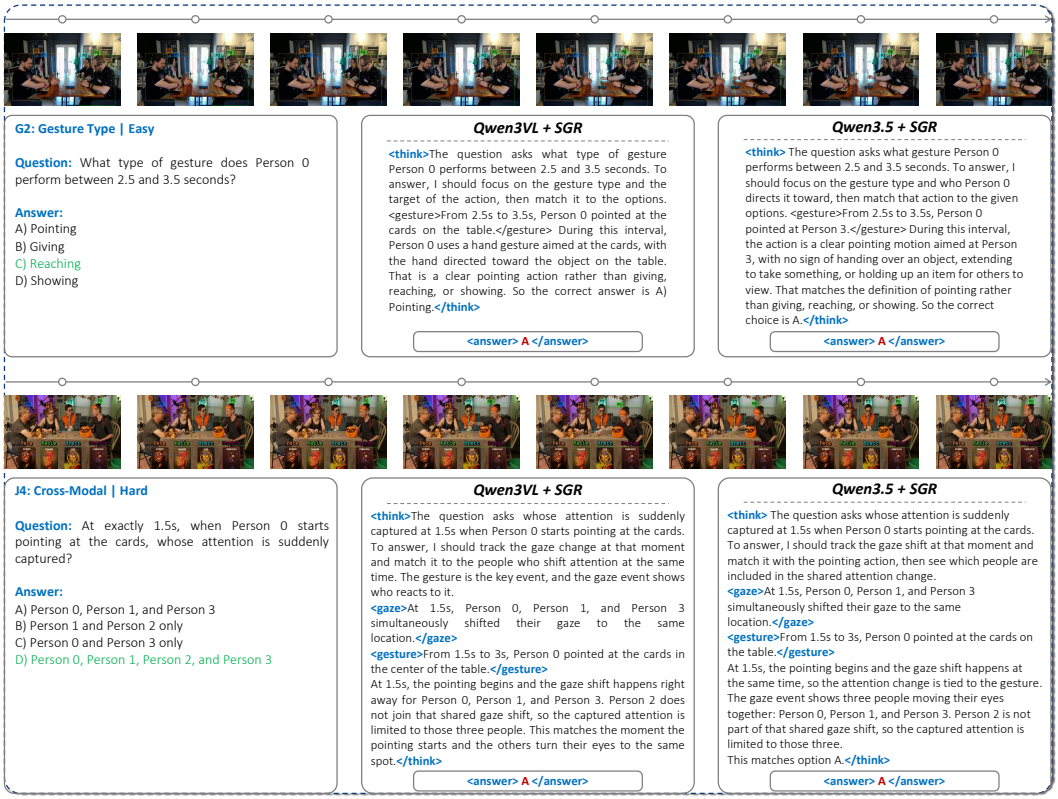


Figure 22: Failure cases on *GRASP-Bench*. We show two representative errors: an ambiguous deictic gesture where reaching and pointing cues are visually close, and a crowded gaze-reasoning case where multiple gaze events occur within the target interval.

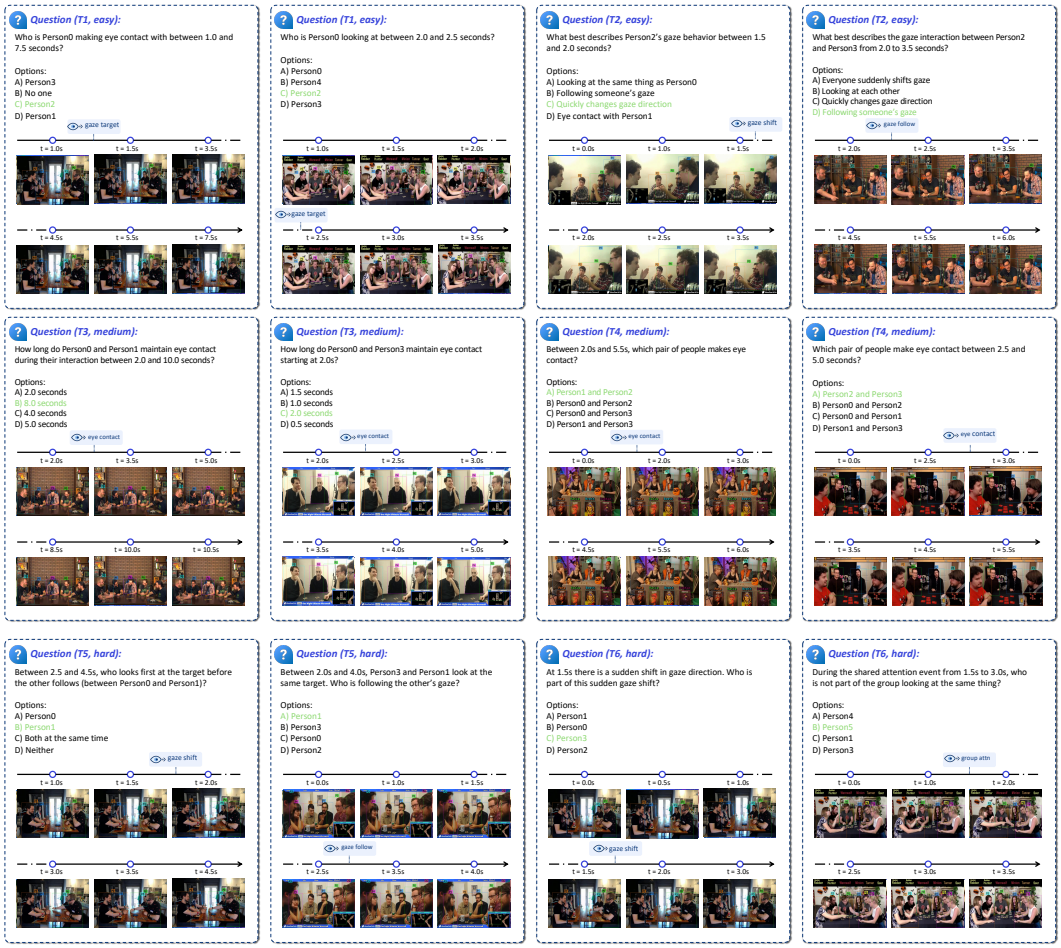


Figure 23: Qualitative *GRASP-Bench* examples for gaze reasoning, covering T1–T6.

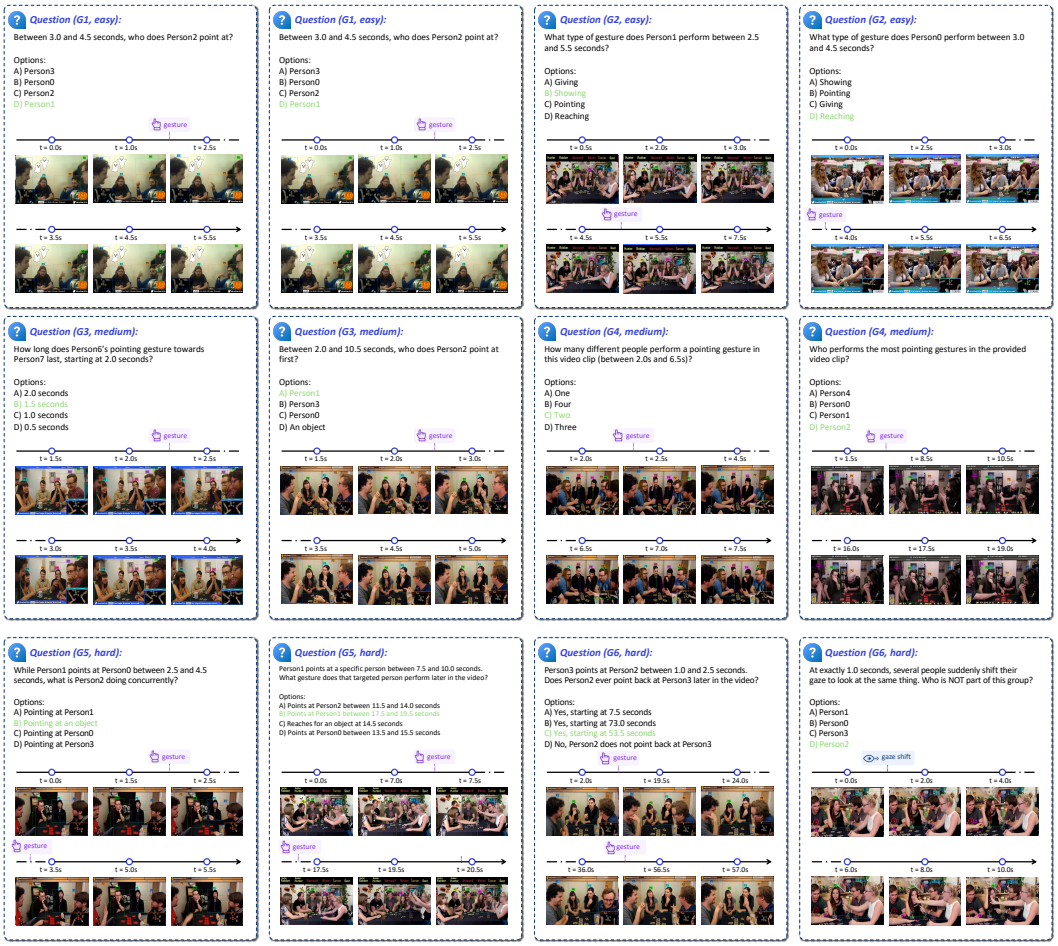


Figure 24: Qualitative *GRASP-Bench* examples for gesture reasoning, covering G1–G6.

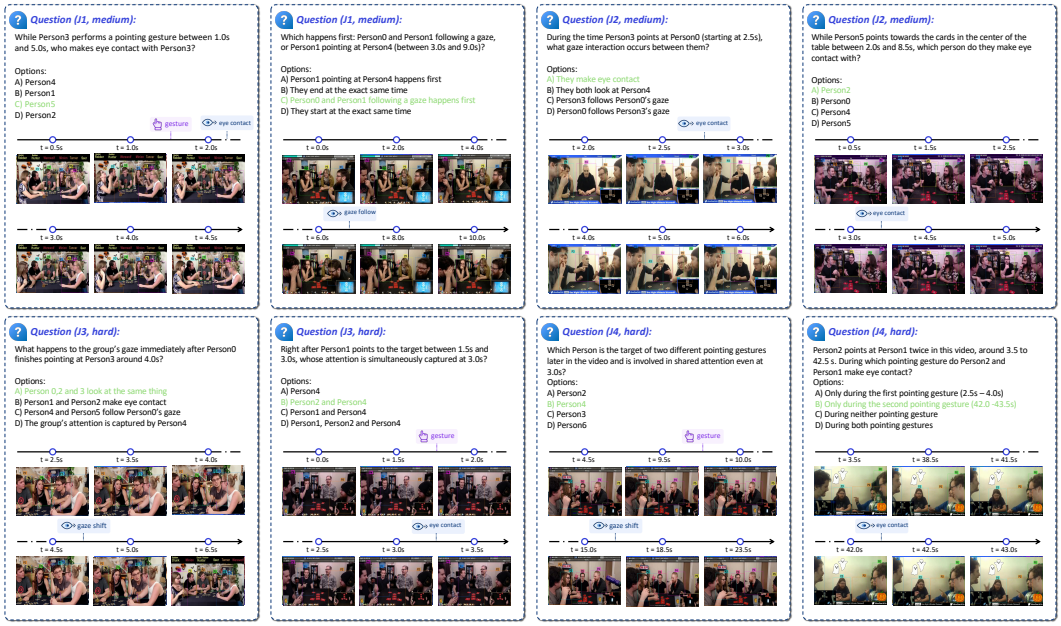


Figure 25: Qualitative *GRASP-Bench* examples for joint gaze–gesture reasoning, covering J1–J4.

## Gesture prompt

```
Analyze this entire video and detect ALL deictic gestures.

## About This Video
- The video shows a social interaction between several people.
- P0, P1, P2... represents person IDs.
- Scan the ENTIRE video from start to end and annotate ALL deictic gestures that are defined below with the correct
↪  timestamps.
- Deictic gestures are related to HAND, FINGER, and ARM movements, and are NOT related to other motions like head pose.

## Timestamp - Visual Reference (CRITICAL)
- Look at the text overlay at the top-left of the video, which represents the current time in seconds: "t=X.XXs".
- ALWAYS use the 't' value (seconds) shown on the screen for your timestamps.

## Deictic Gesture Types

1. **pointing** - Hand gesture intentionally indicating a specific target person
  - YES: Clear arm extension with finger/hand pointing at a specific person
  - NOT the case: Talking with hands (beat/emphasis gestures), casual or rhythmic hand movements, arm resting, or
 ↪  reaching/manipulating gestures where the primary intent is interaction rather than indicating
  - Start: When the arm begins to lift and extend.
  - End: When the hand begins to retract after pointing.

2. **showing** - Presenting an object for others to visually inspect
  - YES: The subject intentionally presents an object by orienting, tilting, sliding, or positioning it so that its
 ↪  visual properties are clearly exposed to others
  - NOT the case: Simply holding or carrying an object without presenting it, object resting on table/lap without
 ↪  intentional presentation, or manipulating an object for use rather than display
  - Start: When the object begins to be intentionally oriented or positioned for visual inspection
  - End: When the object returns to a neutral/use position or is no longer presented

3. **giving** - Intentionally offering an object to transfer possession
  - YES: The subject intentionally offers an object to another person in a way that clearly invites transfer of
 ↪  possession, regardless of whether the recipient actually takes it
  - NOT the case: Simply holding or displaying an object, presenting an object for viewing (showing), or moving an
 ↪  object without intent to hand it over
  - Start: When the object begins to be purposefully moved or offered toward another person for transfer
  - End: When the offering action clearly ends (e.g., the object is released, pulled back, or the offer is withdrawn)

4. **reaching** - Extending the hand toward an object / person with intent to touch it
  - YES: The subject clearly expresses intent to acquire the object or touch the person through a directed hand/arm
 ↪  extension toward an object not in their possession, often accompanied by pre-grasp configuration (open/curved
 ↪  fingers) or body movement (e.g., leaning forward)
  - NOT the case: Pointing or indicating gestures, casual arm extension without acquisition intent, expressive or beat
 ↪  gestures, or any motion where the hand is not plausibly attempting to obtain the object
  - Start: When the hand/arm begins a purposeful reach toward an object with acquisition intent
  - End: When the reach terminates

## Rules
1. Be **strict and conservative**: Annotate a gesture ONLY when the gesture type and intent are **clearly and
↪  unambiguously identifiable** from hand/arm motion alone.
2. If a gesture's **intent or type cannot be clearly determined** from the hand/arm motion alone, do NOT annotate it.
3. Each gesture must have a clear start_time and end_time (in seconds).
4. Person IDs must exactly match the visible Person IDs (P0, P1, P2, ...).
5. If no **high-confidence** deictic gestures are visible, return an EMPTY array.

## IMPORTANT NOTE
- Focus ONLY on hand, finger, and arm movements that clearly express communicative intent.
- Do NOT infer intent from gaze, head orientation, speech, or context outside the visible gesture.
- Only annotate gestures that can be CLEARLY identified as one of the above four types: pointing, showing, giving,
↪  reaching.
- The DURATION (end_time - start_time) of the gesture should be at LEAST 1 SECOND, if not, do NOT annotate it.
- There will be multiple gesture events in the video even simultaneously. You should annotate all simultaneous gestures.
- If a previous prediction or critique report is provided, use it to reduce false positives rather than increase
↪  coverage.

## Response Format (JSON only)

{
  "gestures": [
    {
      "gesture_type": "pointing" | "showing" | "giving" | "reaching",
        "initiator_id": <number>,
        "start_time": <t value from screen>,
        "end_time": <t value from screen>,
        "target_type": "person" | "object",
        "target_person_id": <ID or null>,
        "target_description": "<brief description>",
        "confidence": <0.0 to 1.0>
        "reasoning": <why you categorized this gesture as <gesture_type>>
    }
  ]
}
```

Figure 26: Prompt for deictic gesture annotation.

## QA prompt

```
You are a QA generator for a non-verbal social reasoning dataset.

You will receive structured event annotations detected in a video clip. Your job: generate question-answer pairs that
↪  the model must answer by WATCHING THE VIDEO.

=== STRICT OUTPUT RULES ===

1. ANSWER FORMAT - two types only:
   - MCQ: answer field = single letter (A/B/C/D). answer_text = full text of correct option.
   - Open-ended: answer field = answer_text = 2-4 sentence descriptive answer. Must reference specific Person IDs and
  ↪  timestamps from the event data.

2. ANSWER CONTENT:
   - Factual, no reasoning. NEVER include these words in answers:
     "suggesting", "indicating", "likely", "because", "implies", "seems", "probably", "emotion", "feeling"

3. QUESTION QUALITY:
   - Must require WATCHING THE VIDEO to answer. Not answerable from question text alone.
   - Use natural language: say "eye contact" not "mutual gaze event", "points at" not "performs a pointing gesture",
  ↪  "looking at the same thing" not "joint attention with high convergence".
   - Include specific timestamps from the event data.
   - All person references: "Person {id}" format.

4. MCQ RULES:
   - Exactly 4 options (A/B/C/D).
   - Distractors must be plausible: other person IDs in the video, other event types that didn't happen, swapped roles,
  ↪  nearby but incorrect timestamps.
   - Option text must be concise (<=15 words per option).

5. OUTPUT FORMAT:
   Return ONLY a JSON array. No explanation, no markdown, no commentary.
   Each element:
   {
     "category": "T1",
     "difficulty": "easy",
     "format": "mcq" | "open_ended",
     "question": "...",
     "options": ["A) ...", "B) ...", "C) ...", "D) ..."],
     "answer": "C",
     "answer_text": "Person 3",
     "source_event_ids": [4000],
     "time_range": [13.0, 15.5]
   }
   - "options" field: ONLY for mcq.
   - "source_event_ids": gaze event_id(s) or gesture index(es) used as GT.
   - "time_range": [start, end] of relevant events.
```

Figure 27: Prompt QA generation.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction summarize the dataset, benchmark, and learning method, with claims supported by our experimental evaluations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include the discussion and limitation section in the main manuscript and appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not involve theoretical result to prove.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Implementation section is included with further details in appendix. We will open the source code and data to public.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We included all the details in appendix and will open to public in next step.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, we have baselines and benchmarks explanations in the manuscript.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: Our experiments do not need statistical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include the computational details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We reviewed CoE.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We add it in appendix: broader impact section.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We do not redistribute raw videos from online sources. We release only derived annotations and source identifiers under applicable usage terms, with a removal-request procedure.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We acknowledge the original sources and document the applicable licenses or usage terms for existing datasets, models, APIs, and online video sources.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes, we introduce a new dataset, and all the details are described in the appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Four internal evaluators conducted a benchmark QA quality check. The evaluation procedure and interface are described in the appendix. No external crowdsourcing was conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The human evaluation was conducted as an internal benchmark quality-control activity, rather than research involving the evaluators as study subjects. No personal or sensitive information was collected, and IRB review was not required under the applicable institutional guidelines.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use Gemini 3.1 Pro for constrained gesture annotation, verification, and question generation from structured social-event graphs. Benchmark answers are derived from structured event fields rather than free-form LLM outputs. These procedures are described in the manuscript.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.