ESI-BENCH: A COMPREHENSIVE VIDEO BENCHMARK FOR EMOTIONAL AND SOCIAL INTELLIGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work has increasingly focused on the evolution and intelligent modeling of interpersonal interaction and collaboration. Driven by advances in Multimodal Large Language Models (MLLMs), emotional intelligence (EI) and social intelligence (SI) have emerged as core competencies for AI systems: they enable agents to modulate behavior in complex contexts, infer others' intentions, maintain social relationships, and ultimately support natural human-machine interaction and seamless collaboration. To systematically investigate AI capabilities and pathways for understanding EI and SI, the community has introduced benchmarks such as EQ-Bench, Social-IQ 2.0, and V-Social, advancing research on emotion understanding, social behavior modeling, and social common sense reasoning. However, existing approaches generated datasets exhibit limited semantic separability between options, low question answer relevance (QA relevance), low dataset complexity (with extremely high accuracy), low ground truth correctness,narrow modality coverage, and pronounced inherent biases. Meanwhile, these data construction pipeline suffer from high annotation costs and lengthy data-collection cycles. To address the shortcomings of the existing evaluation datasets, We introduce ESI-Bench, a benchmark comprising 1,105 videos and 5,490 meticulously generated QA pairs. It offers accurate cross-modal alignment, high semantic separability, strong QA relevance, reliable ground truths, and substantially reduced inherent bias, enabling clear performance stratification across state-of-the-art (SOTA) models.We also propose a semi-automated, high-efficiency data generation framework. Our framework integrates multiple models (open-source and closed-source) with complementary strengths and couples them with a lightweight manual verification loop,enabling low-cost, large-scale construction of high-quality emotional social intelligence datasets. This work provides a scalable paradigm for constructing rigorous emotional and social intelligence (ESI) evaluations and aims to advance research toward more capable human-AI interaction.

1 INTRODUCTION

Social intelligence (SI) refers to an individual's ability to effectively adapt to understand and respond to others' behaviors, social norms and situational contexts during social interaction, with an emphasis on social skills and situational adaptation. Emotional intelligence (EI) refers to the capacity to recognize, understand, manage, and regulate one's own and others' emotions, focusing on affect perception and regulation. EI and SI are core pillars for advancing AGI toward human-level intelligence, endowing AI systems with the capability to perceive, interpret, and adapt within complex social environments. These abilities underpin natural human–robot interaction and have become central topics in robotics and artificial intelligence (Lee et al., 2023; Zhou et al., 2023).

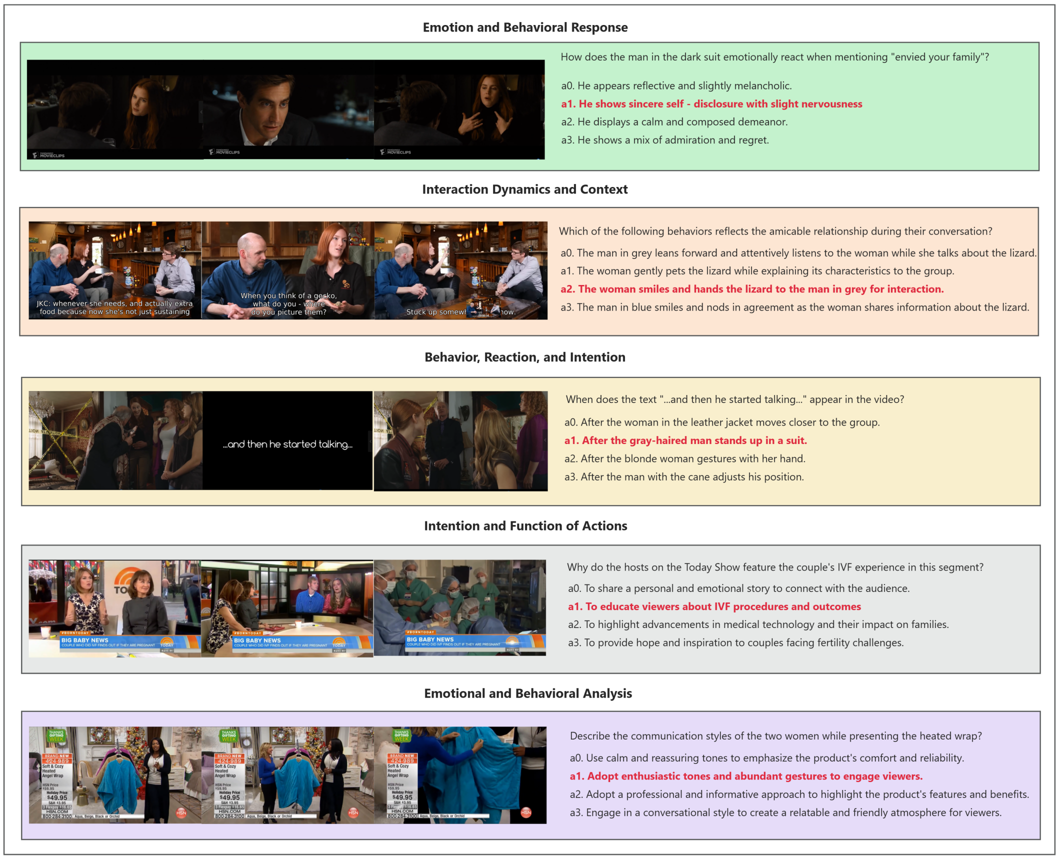
In computer vision, SI is commonly evaluated via video-based or image-based question-answering (QA), as seen in benchmarks such as VCR (Zellers et al., 2019), MovieQA (Tapaswi et al., 2016), and TVQA (Lei et al., 2018). Other datasets, including CMU-MOSEI (Zadeh et al., 2018), CHAMPAGNE (Han et al., 2023), and JRDB-Social (Jahangard et al., 2024), focus on emotion recognition, group interactions, and commonsense reasoning. Social-IQ (Zadeh et al., 2019) and Social-IQ 2.0 (Wilf et al., 2023) integrate multimodal signals (video, images, audio) and pose complex questions that probe causes of emotions and behavioral intentions, using a multiple-choice (4-option) QA pair format to assess social intelligence. DeSIQ (Guo et al., 2023) mitigates inherent biases in Social-IQ 2.0 via option perturbations, and researchers have also introduced custom datasets with human rationales and answers (Zellers et al., 2019; Wilf et al., 2023). Within EI, EmoBench (Sabour et al., 2024) and EQ-Bench (Paech, 2023) emphasize emotional understanding and application; EmoBench quantifies depth of emotional understanding through multiple-choice implicit emotion attribution; EQ-Bench evaluates comprehension of nuanced emotions and social interactions by asking models to predict the emotional intensity of conversational participants. V-Social (Lin et al., 2025) investigates social commonsense reasoning and explores integrating text and vision to improve holistic understanding, covering both EI and SI tasks. SIV-bench (Kong et al., 2025) defines social interaction along three dimensions: social scene understanding, social state reasoning, and social dynamics prediction, and instantiates Fiske's Relational Models Theory (Fiske, 1992) via 14 specific relationship types, emphasizing relation-centric understanding and inference. These benchmarks provide limited contextual detail and lack explicit reasoning traces. They fail to capture subtle nonverbal cues, do not fully evaluate multimodal reasoning, and exhibit significant inherent biases. For example, in Social-IQ 2.0, models achieve relatively high accuracy. GPT-4o (OpenAI, 2025) reaches 52.1% in the no context and question (NCAQ) setting and 78.4% when using images and transcripts in a zero-shot setup. The discriminative power among SOTA models remains limited in this dataset. Advanced models, such as Gemini (Google, 2025), Doubao (volcengine, 2025), and Grok (xAI, 2025), perform within the 76%-79% range. This suggests the benchmark lacks sufficient difficulty to differentiate high-performing models or effectively stress-test EI and SI capabilities.

To comprehensively, accurately, and fine-grainedly assess EI and SI while addressing the limitations above, we introduce ESI-Bench. It features high quality, broad coverage, and efficient, cost-effective construction. It assesses emotional and social intelligence through five key dimensions: Behavior, Reaction, and Intention; Interaction Dynamics and Context; Emotion and Behavioral Responses; Emotion and Behavioral Analysis; and Action Intent and Function. The benchmark offers high quality, broad coverage, tight multimodal alignment, and accurate labels. ESI-Bench better separates the capabilities of leading closed-source models (e.g., GPT-4o, Doubao, Gemini, Claude (Anthropic, 2025),Grok) and provides a robust benchmark for assessing emotional social intelligence in VLMs. In the data generation pipeline, it is common to employ one or two models to generate initial captions or QA pairs, followed by manual review and refinement step. However, this process is both time-intensive and prone to quality inconsistencies, as the expertise of annotators can vary significantly. We introduce a semi-automatic construction pipeline that reduces cost and saves time consumption while improving consistency. Our pipeline integrates multimodal data acquisition, automated annotation, and rigorous quality control. Concretely, we **(i)** extract modality-specific features and generate base captions with specialized models, **(ii)** use MLLMs to propose reasoning-oriented questions and answers, and **(iii)** apply multi-stage validation to ensure factual and semantic correctness. Our contributions are as follows:

1. We introduce ESI-Bench, a high-quality, broad-coverage benchmark that surpasses prior datasets in complexity and data quality-measured by semantic separability, QA relevance, inherent bias, question difficulty, and ground truth correctness, as well as topical coverage.
2. We present an efficient, low-cost construction pipeline that leverages MLLMs for assisted labeling, enabling the generation of tens of thousands of exemplars within one week with minimal human intervention and short turnaround.
3. We conduct a comprehensive analysis of zero-shot performance for closed-source models on ESI-Bench. Performance is uniformly modest across metrics (maximum 57.39%). These results indicate that ESI-Bench is challenging and exhibits strong discriminatory power among current SOTA models, while also revealing substantial headroom both for improving current model capabilities and for refining the evaluation protocol.

2 Related Works

Multimodal Video Understanding Evaluation Benchmarks.
To assess the video understanding capabilities of rapidly evolving MLLMs, the community has developed a multi-dimensional evaluation landscape. At the foundational capability tier, MLVU (Zhou et al., 2024) and VideoVista (Li et al., 2024b) probe visual-semantic understanding via perceptual


MOVIECLIPS
JKC: whenever she needs, and actually extra food because now she's not just sustaining
When you think of a gecko, what do you - where do you picture them?
Stuck up somewh how.
...and then he started talking...
BIG BABY NEWS

Emotion and Behavioral Response

How does the man in the dark suit emotionally react when mentioning "envied your family"?

a0. He appears reflective and slightly melancholic.

a1. He shows sincere self - disclosure with slight nervousness

a2. He displays a calm and composed demeanor.

a3. He shows a mix of admiration and regret.

Interaction Dynamics and Context

Which of the following behaviors reflects the amicable relationship during their conversation?

a0. The man in grey leans forward and attentively listens to the woman while she talks about the lizard.

a1. The woman gently pets the lizard while explaining its characteristics to the group.

a2. The woman smiles and hands the lizard to the man in grey for interaction.

a3. The man in blue smiles and nods in agreement as the woman shares information about the lizard.

Behavior, Reaction, and Intention

When does the text "...and then he started talking..." appear in the video?

a0. After the woman in the leather jacket moves closer to the group.

a1. After the gray-haired man stands up in a suit.

a2. After the blonde woman gestures with her hand.

a3. After the man with the cane adjusts his position.

Intention and Function of Actions

Why do the hosts on the Today Show feature the couple's IVF experience in this segment?

a0. To share a personal and emotional story to connect with the audience.

a1. To educate viewers about IVF procedures and outcomes

a2. To highlight advancements in medical technology and their impact on families.

a3. To provide hope and inspiration to couples facing fertility challenges.

Emotional and Behavioral Analysis

Describe the communication styles of the two women while presenting the heated wrap?

a0. Use calm and reassuring tones to emphasize the product's comfort and reliability.

a1. Adopt enthusiastic tones and abundant gestures to engage viewers.

a2. Adopt a professional and informative approach to highlight the product's features and benefits.

a3. Engage in a conversational style to create a relatable and friendly atmosphere for viewers.

Figure 1: Overview of ESI-Bench, showing its diverse videos spanning various Emotional and Social Intelligence and sample QAs for five task dimensions: Intention and Function of Action (IFA). Behavior, Reaction, and Intention (BRI). Emotion and Behavioral Response (EBR). Emotion and Behavior Analysis (EBA). Interaction Dynamics and Context (IDC).

tests and fine-grained detail analysis, respectively. At the advanced reasoning tier, MVBench (Li et al., 2024a) and MMBench-Video (Fang et al., 2024) cover question answering, summarization, and complex logical reasoning across diverse domains, including movies, egocentric recordings, and general web content. For temporal dynamics, TempCompass (Liu et al., 2024) targets fine-grained assessment of long-range temporal dependencies, while Video-Bench (Ning et al., 2023) proposes a joint spatio-temporal reasoning framework. Despite this layered ecosystem, benchmarks that center on social dynamics in multi-party interactions remain scarce; SIV-Bench is among the few specialized datasets explicitly focused on complex social relationship reasoning.

Evolution of Emotional and Social Intelligence Evaluation.

Research on evaluating emotional and social intelligence (ESI) has progressed along two complementary strands. The first develops frameworks for general social cognition, such as BLINK (Fu et al., 2024), a multimodal commonsense benchmark, and Social Genome (Mathur et al., 2025), which constructs social behavior graphs from videos, providing multimodal, context-grounded records of social reasoning. The second strand focuses on targeted subtasks:VCR evaluates social-situation reasoning via image-text QA on movie scenes. SECEU (Wang et al., 2023), is a scoring benchmark for emotion understanding (EU) that rates relative emotional intensities. MME-Emotion (Zhang et al., 2025) spans both emotion recognition and reasoning, covering eight emotion tasks and 27 scenarios, and more comprehensively evaluates the generalization of MLLMs across diverse settings. Two social-oriented multimodal video benchmarks, Social-IQ 2.0, expand the emphasis to interpersonal reasoning and social commonsense, using open-ended social intelligence questions of varying difficulty to probe complex social commonsense reasoning. V-Social provides transcripts, longer temporal context, and varied scenes to assess higher-order social reasoning.

Limitations and Positioning.
Current mainstream benchmarks exhibit two primary limitations:

- Narrow capability coverage: Many focus on a single skill (e.g., emotion attribution or behavior prediction) and lack a systematic framework for evaluating ESI holistically.
- Insufficient fine-grained analysis: Existing methods have limited sensitivity to non-explicit social signals (e.g., micro-expressions, motion, implicature/subtext), and cross-modal alignment accuracy remains inadequate.

To address these gaps, we introduce ESI-Bench, A comprehensive high-quality emotional and social intelligence benchmark.

3 ESI DEFINITION

Emotional intelligence (EI) denotes the capacity to recognize, understand, manage, and regulate one's own and others' emotions. Social intelligence(SI) refers to the ability to effectively adapt to understand and respond to others' behaviors, social norms, and situational contexts during social interaction. EI provides the foundation for SI, while SI helps translate EI into effective social outcomes. SI and EI are two facets of a common construct. Most definitions span the following abilities (Bar-On, 2006): **1)** To understand and constructively express emotions, **2)** To understand the experiences of other people and create cooperative interpersonal relationships, **3)** To manage and regulate emotions effectively, **4)** To cope with the new situations realistically and solve problems of a personal or interpersonal nature, to be optimistic, positively charged, and internally motivated to formulate and reach goals. Some researchers integrate the two as Emotional and Social Intelligence (ESI), emphasizing their synergy: ESI encompasses the ability to recognize, understand, and regulate one's own emotions to communicate effectively and empathetically in social interactions. It includes self-awareness, empathy, interpersonal skills, and emotion regulation. Building on the characteristics of ESI, we present a more comprehensive framework for defining the ESI evaluation criteria. Specifically, we define the capacity of MLLMs to evaluate ESI in video through five core dimensions:
Intention and Function of Action (IFA): Identify purpose, motivation, or situational context underlying body movements in communication (emphasis, turn-taking guidance, affect expression, clarity), and link observable motions to intended communicative roles within conversations or speeches.
Behavior, Reaction, and Intention (BRI): Analyze why a character exhibits a particular behavior, reaction, or stated intention at a given moment; probe causal relations and contextual meanings underlying motions, including motivations, emotions, and responses to events or interactions.
Emotion and Behavioral Response (EBR): Assess emotions, attitudes, behaviors, and communicative styles at specific moments or exchanges; characterize how tone, gestures, facial expressions, and body language convey affect and emotions across contexts.
Emotion and Behavior Analysis (EBA): Identify and interpret emotions, behaviors, and communicative patterns (verbal and nonverbal) within context, and analyze the psychological and social dynamics unfolding during interactions.
Interaction Dynamics and Context (IDC): Examine interactions among participants, emphasizing how behaviors, emotions, tone, gestures reveal relationship properties and interaction goals; infer relationship dynamics, communicative atmosphere, situational context from observable cues.

4 ESI-BENCH

In this section, we introduce ESI-Bench, a comprehensive benchmark for evaluating ESI. Figure 1 offers some illustrative examples from ESI-Bench. Built along the five ESI dimensions defined above, the current release contains over 1,105 videos, 5,490 questions, and 21,960 answer options (4 options per question) in a multiple-choice format. Each QA is annotated with a difficulty level. The dataset is constructed via a human-in-the-loop pipeline: we first pre-annotate with multiple MLLMs, then apply cross-model validation and filtering, followed by targeted human auditing to ensure label fidelity and overall data quality. ESI-Bench is designed to rigorously assess ESI capabilities in MLLMs. Table 1 compares ESI-Bench with other related datasets.

Table 1: Comparison of various benchmarks. It includes several aspects: annotation Method (M/A means manually/automatic manner), and the types of tasks included. Task Coverage: emotion and social. The table also shows Modality for Contexts, and the Characteristics of Data.

Benchmark	Task Coverage		Modality for Contexts			Characteristics of Data				Datasets Pipeline
	Emotion	Socail	Text	Image	Video	Speaker's Attribute	Transcript	Action &Tone	Expression & Emotion	Annotation Method
EQ-Bench	✓	✗	✓	✗	✗	✗	✗	✗	✗	M
EmoBench	✗	✓	✓	✗	✗	✗	✗	✗	✗	M
VCR	✗	✓	✓	✗	✓	✗	✗	✗	✗	M
Social-IQ 2.0	✗	✓	✓	✓	✓	✗	✗	✗	✗	M
V-Socical	✓	✓	✓	✓	✓	✓	✓	✗	✗	M+A
SIV-bench	✓	✓	✓	✓	✓	✗	✓	✗	✗	M+A
ESI-Bench	✓	✓	✓	✓	✓	✓	✓	✓	✓	**M+A**

4.1 Data construction pipeline

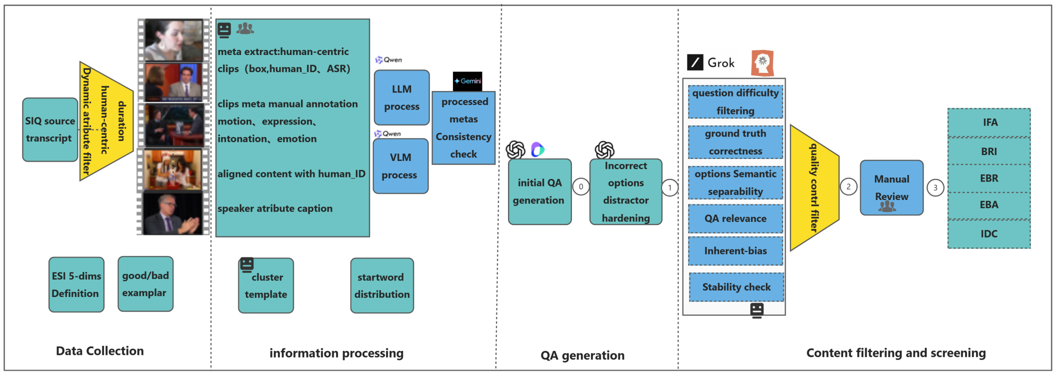


Figure 2: The ESI-Bench construction pipeline, In the diagram, green blocks indicate content (like startword, video source and initial QA) generation steps, blue blocks represent validation or modification stages, yellow trapezoids signify a filtering and removal phase.

Figure 2 delineates the construction pipeline of the ESI-Bench, providing a detailed account of data acquisition, information processing, QA generation, and collaborative human-machine filtering and selection of QA content. Within the depicted process, **green** boxes denote content generation stages (e.g., clustering template generation, metadata annotation, QA generation), **blue** boxes represent information verification or refinement phases, and yellow trapezoids indicate filtering and removal stages. We begin by selecting approximately 1,105 source videos from Social-IQ and Social-IQ 2.0. For each video, we use specialized models to extract raw multimodal meta information, including human attributes/roles, motions, facial expressions, prosody/intonation, affective states, and automatic speech recognition (ASR) transcripts. We then employ LLMs and VLMs to induce higher-level, hierarchical representations from this raw meta information, yielding a structured, coherent summary for each video. A VLM-based cross-modal consistency check reconciles visual and audio-linguistic streams, removing two processed meta information's contradictions and filling omissions. To steer question generation, we condition prompts on the defining features of the five ESI dimensions, ensuring balanced coverage and dimension-specific constraints. In addition, we curate 500 high-quality ESI QA pairs and cluster them into 16 representative question templates, which serve as a template library for controlled generation. Next, we feed the processed meta information, sampled video frames with timestamps, ESI definitions, and template-based prompt contexts (with examples) into two MLLMs to generate initial multiple-choice QAs. We generate up to 20 QAs per video at this stage. Following generation, a series of validator models assess data quality across multiple dimensions: examining question difficulty and ground truth correctness, semantic separability, QA relevance, Zero-shot Accuracy, and indicators of inherent bias, and filter out low-quality QAs. Finally, the filtered set undergoes human spot-checking and adjudication to further reduce inherent bias and hallucinations. Detailed step-by-step workflows are shown in Appendix A.1.

4.2 Statistics

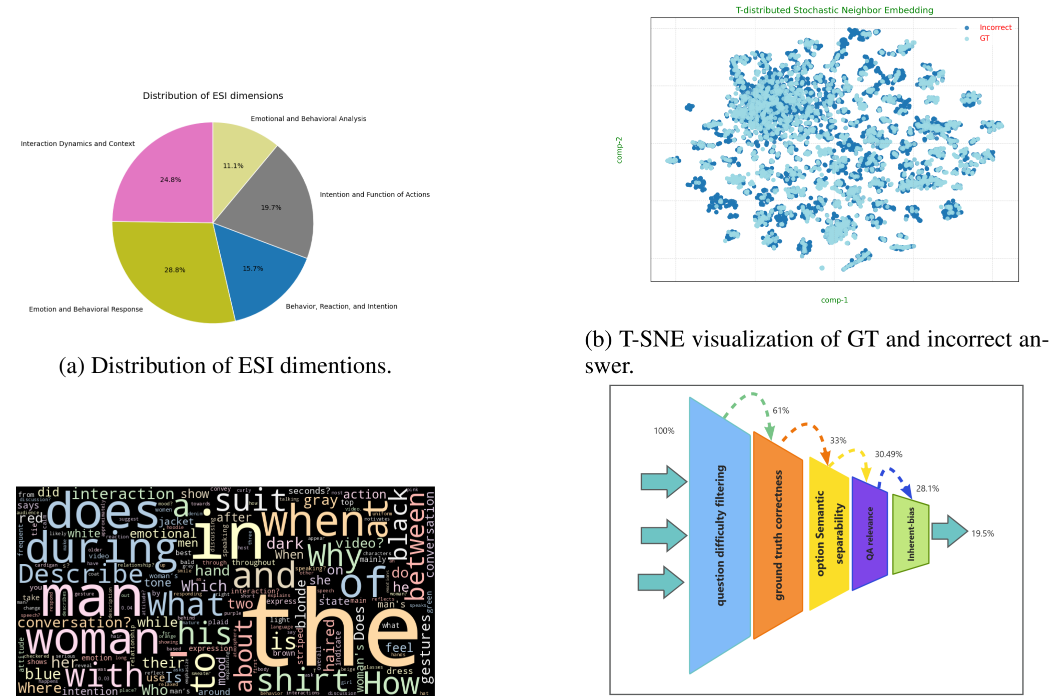
To ensure the quality and balance of our ESI-Bench QA pairs, we conducted a statistical analysis of their structural properties, as summarized in Figure 3. ESI-Bench encompasses a collection of

1,105 video sources, comprising 887 videos from Social-IQ 2.0 and 218 videos from Social-IQ. These videos are exclusively person-centric, with an average duration of one minute. The dataset incorporates a total of 5,490 QA pairs, each formulated as a four-way multiple choice question with a single, definitive correct answer. Some statistical results are as follows:

ESI dimension's QA distribution. The resulting dataset exhibits a well-balanced distribution of QA pairs across the five defined dimensions of ESI (as presented in Figure 3a).

Discriminative analysis of options. A T-distributed Stochastic Neighbor Embedding (T-SNE) visualization of four answer options (as shown in Figure 3b) reveals no discernible boundaries between ground truth answer and incorrect answers, indicating a high level of inter-option distractibility and consequently a challenging benchmark.

QA coverage. Figure 3c demonstrates the broad coverage of topics within the dataset's questions. For analysis of other dimensions, please refer to the Appendix A.4



(a) Distribution of ESI dimentions.

(b) T-SNE visualization of GT and incorrect answer.

(c) Word cloud illustrating language diversity.

(d) QAs validation and filtering funnel.

Figure 3: ESI-Bench statistics

5 Experiments

In this section, we present a comparison of ESI-Bench against Social-IQ 2.0 and V-Social across various MLLMs. Our evaluation focuses on two key aspects: **1.** *Extrinsic model performance*: We report the zero-shot accuracy of closed-source SOTA models to assess their performance on each benchmark. **2.** *Intrinsic dataset diagnostics*: We provide an in-depth analysis of dataset quality, including metrics such as semantic separability of options (SSO), QA relevance (QAR), question difficulty (QD), ground truth correctness (GTC), and inherent bias (IB). We then analyze these results to characterize the distinctive properties and advantages of ESI-Bench, demonstrating its challenge level and discriminative power for ESI in multimodal settings. the experiments setting can be found in Appendix A.2.

5.1 Analyses

We assess ESI-Bench from both qualitative and quantitative perspectives:

Quantitative analysis. We compare ESI-Bench (restricted to videos sourced from Social-IQ 2.0

Table 2: Zero-shot Accuracy of Social-IQ 2.0 and ESI-Bench in different SOTA models.

	GPT-4o	Gemini	Doubao	Grok	Claude
Social-IQ 2.0	78.40	78.26	76.73	79.00	76.23
ESI-Bench	**53.26**	**56.53**	**53.46**	**57.39**	**51.56**

Table 3: Comparison of statistical indicators of different benchmark.

	IB (↓)	QAR (↑)	SSO (↑)	QD (↑)	GTC (↑)	Accuracy
V-Social	54.08	2.10	2.29	2.30	1.08	79.15
Social-IQ 2.0	52.11	2.14	2.11	2.44	1.35	78.40
ESI-Bench	**8.15**	**2.59**	**2.28**	**2.63**	**1.85**	**53.26**

for fairness) with the Social-IQ 2.0 test set using five MLLMs: GPT-4o, Gemini, Doubao, Grok, and Claude (Qwen models are excluded since they are used in data preparation). All models are queried via their commercial APIs in the same settings; we uniformly sample 16 frames per video, embed normalized timestamps into the prompt, and evaluate zero-shot performance. Results are reported in Table 2 and Table 3. ***Cross-model diagnostics:*** Accuracy on ESI-Bench is approximately 25% lower than on Social-IQ 2.0, highlighting its significantly greater difficulty. Additionally, the performance variance across models increases, demonstrating enhanced discriminative power. ***Cross-benchmark diagnostics:*** We compare inherent bias, semantic separability, QA relevance, question difficulty, and GT correctness across ESI-Bench, Social-IQ 2.0, and V-Social. **1).** V-Social shows higher inherent bias, indicating many QAs are solvable through option analysis. Its question difficulty and GT correctness are both lower than ESI-Bench. The high accuracy (79.15%) reflects lower dataset complexity. **2).** Social-IQ 2.0 shows substantial bias, poor semantic separability, and high accuracy, indicating lower complexity. **3).** Our ESI-Bench has a lowest inherent bias of only 8.15% after filtering, achieving higher average difficulty and higher ground truth correctness than both of the others. Its lower accuracy (53.26%) indicates a harder, better-calibrated benchmark that improves model differentiation. **Qualitative analysis** are shown in Appendix A.3.

Workflow efficiency and cost. ESI-Bench is built through a model-driven, human-assisted workflow. Specialized models first extract human-centric video features, which are manually refined for cross-modal alignment to create meta information. LLMs and VLMs process this meta information to co-generate candidate QAs, followed by quality checks and filtering using multiple validator models. The finalized QAs are then manually adjudicated to ensure quality. This approach exploits complementary model capabilities for extraction, integration, generation, and filtering, while preventing single-model biases from accumulating across stages. Human effort is minimized: eight trained annotators completed ten thousands of meta information corrections and final QA review within few weeks: far less labor than typical for fully manual datasets like Social-IQ. The resulting dataset exhibits high complexity and quality (high semantic separability, high QA relevance, high question difficulty, high ground truth correctness, low inherent bias), with a short production cycle and low human cost.

5.2 Ablation Study

We define the following key characteristics for a robust evaluation dataset: **(i)** Diverse questions drawn from multiple perspectives; **(ii)** Hierarchical structure with varying difficulty; **(iii)** A well-stratified difficulty spectrum to address diverse use cases; and **(iv)** Zero-shot SOTA model performance below 60%, We then measure the robustness of the dataset through SSO, QAR, QD, GTC and Accuracy, ensuring headroom for future advancements and avoiding ceiling effects. ESI-Bench is constructed through a multi-model, multi-stage pipeline that includes generation, processing, validation, and filtering. To assess the contribution of each component, we perform an ablation study focused on the QA generation stage (node ① in Figure 2). Results are reported in Table 4.

Impact of different meta information.

Use raw meta information (unprocessed). Using unprocessed meta information as input slightly lowers question difficulty (2.431 to 2.399) while slightly raising accuracy (53.56% to 53.64%). Under our scoring convention, where higher accuracy indicates lower dataset complexity, this implies

Table 4: The results of the ablation experiment.

	IB (↓)	QD (↑)	GTC (↑)	Accuracy
ESI-Bench	**29.32**	**2.431**	**1.550**	**53.56**
Use raw meta information	29.46	2.399	1.581	53.64
Without meta information	29.37	2.397	1.628	55.00
Generate QAs use LLM	51.72	2.382	1.481	68.10
Without question exemplars	27.17	2.236	1.444	54.26
Without cluster template	30.52	2.268	1.542	53.23

small reduction in overall difficulty. By contrast, structuring the meta information into a hierarchical representation reduces the model's comprehension burden during generation and enables conditioning on multi-level cues, yielding QAs with higher question difficulty and greater dataset complexity. Inherent bias remains unchanged across conditions, suggesting meta information processing has a limited impact on NCAQ and does not materially alter reliance on visual grounding. Qualitatively, unprocessed meta information steers models toward salient observables (e.g., motions, facial expressions), producing QAs that follow a narrow behavior → emotion →relationship pattern. With processed (summarized, abstracted, hierarchically organized) meta information, an intermediate understanding layer helps model attend to latent, essential cues, resulting in QAs with richer hierarchical structure and broader reasoning coverage. See the sample in Appendix A.5

Without meta information. Removing meta information sharply reduces time-anchored questions, even with temporal cues, underscoring its role in temporal grounding. Inherent bias is largely unchanged; question difficulty declines slightly (2.431→2.397) while accuracy rises (53.56% to 55.00%). This suggests a reduction in overall question difficulty. Meta information provides human-centric details, including speech, motion, facial affect, prosody, and emotions. Absence of meta information significantly reduces the quality and relevance of detail-sensitive QAs. Meta information enables the generation of QAs that encompass **(i)** abstract probes into relationships, motivations, trajectories, and indirect mental states, and **(ii)** concrete, timestamped queries such as "Who did what at time=XXs?". This combination promotes fine-grained observation, multi-step reasoning, psychological causality, and explicit social structure understanding. Question types are more diverse, hierarchical, and challenging, better assessing integrative reasoning. Without meta information, QAs skew to direct description, with fewer complex/abstract items; they follow a stereotyped behavior→emotion→relationship heuristic, are shallower and more redundant, lower question difficulty, yielding an easier dataset.

Without question exemplars. We steer generation toward higher quality QAs via few-shot in-context prompting with curated exemplars spanning the five ESI dimensions. Drawn from high-quality prior datasets, these exemplars cover both abstract and concrete question types across emotions, motivations, motion details, scene context, social relationships, narrative development, visual content, intent attribution, and interaction style, and instantiate reasoning skills such as fine-grained detail extraction, causal inference, and integrative composition. Used as contextual demonstrations, they consistently improve QA quality, richness, and compositional depth, yielding a more hierarchical question-type distribution. In an ablation without exemplars, inherent bias decreases (indicating the exemplars introduce some prior patterns), but average question difficulty also drops and overall dataset complexity diminishes (accuracy 53.56%→54.26%; higher accuracy implies lower complexity). Without exemplars, generation skews toward concrete, directly perceptible phenomena (e.g., smiling, waving, handing a prop), with fewer abstract "why/how" queries; the perspective becomes more local, relying on surface cues rather than temporal, motivational, or relational reasoning. With exemplars, prompting produces a balanced mix of abstract items (relationships, motivations, interaction trajectories, implicit mental states) and concrete, time-anchored questions (e.g., "Who did what at t=XXs?"). This promoting careful observation, dynamic multi-step reasoning, psychological attribution, and systematic structuring of social relations.

Without cluster templates. We cluster 16 representative question templates from historical high-quality datasets (see Appendix A.8) and use them as in-context constraints to shape both perspective and distribution of generated QAs. The templates strikes a balance between diversity and hierarchy, incorporating ESI dimension-specific templates and multi-level formats (scene-level and global-level) to ensure coverage of both abstract and concrete queries. Removing templates leaves ground truth correctness essentially unchanged but increases inherent bias (29.32→30.52), reduces question difficulty (2.431→2.268), and slightly increases dataset complexity (accuracy

53.56%→53.23%; lower accuracy indicates higher complexity). This suggests that without structural priors, the model's autonomy plus the contextual prompt yields broader topical spread but overproduces easier, more biased QAs; high-difficulty items become more clustered, producing a more polarized difficulty distribution with a lower mean. Qualitatively, without templates, the generator explores more styles yet favors simple temporal localization (e.g., "What was Y doing at t=XXs?"). Because templates encode scene/global level, removing them degrades hierarchy: questions become locally focused and less layered. The share of template-defined categories (e.g., basic emotion recognition, simple temporal reasoning) decreases, while many QAs default to direct surface-level descriptions rather than structured, multistep reasoning about motivations, relations or narrative trajectories.

Impact of generating QAs using LLM or VLM. When QAs are generated directly with a text-only LLM, inherent bias increases markedly. In this setting, the model primarily depends on transcript cues and commonsense priors, generating QAs often solvable from a single modality. This leads to text-only answers, decreased question difficulty, and a sharp drop in dataset complexity (accuracy 53.56 → 68.1). Additionally, the generator type (LLM vs. VLM) significantly influences QA quality. QAs generated by LLMs consistently underperform those produced by VLMs across all dimensions. Qualitatively, the two regimes differ in characteristic ways: VLM-generated QAs emphasize multimodal grounding, integrative generation, logical reasoning, and fine-grained detail. A greater proportion of QAs demand visual perception and multi-cue analysis. They exhibit diverse, hierarchically structured styles, balancing factual queries with explanatory and evaluative prompts. LLM-generated QAs focus on higher-level analyses of psychology, affect, relationships, and motivations, often following a coherent stepwise progression. However, they show weaker visual grounding and a higher occurrence of factual hallucinations (e.g., nonexistent objects or attributes), resulting in lower ground truth correctness. Overall, removing visual inputs shifts the distribution toward single-modality, text-driven questions, increases susceptibility to bias, and reduces both question difficulty and dataset complexity. In contrast, VLM conditioning produces more challenging, diverse, and visually grounded QAs.

Impact of QAs validation and filtering. Within the ESI-Bench pipeline, validation and filtering constitute the primary quality-assurance stage. We apply layered checks on **(i)** Semantic separability among options, **(ii)** QA relevance, **(iii)** Ground truth correctness, **(iv)** Question difficulty, and **(v)** Inherent bias screening in NCAQ settings. After this rigorous multi-stage process, fewer than 20% of the initially generated QAs are retained (see Figure 3d). This aggressive pruning is deliberate, ensuring the final benchmark achieves high quality and challenging difficulty, making this stage essential for maintaining dataset integrity. Further details are provided in the Appendix A.1.5

6 Conclusion

ESI-Bench is a multimodal benchmark for evaluating emotional and social intelligence. Built from about 1,105 in-the-wild social videos, it emphasizes cross-modal alignment (video, audio, text) and precise capture of nonverbal cues (facial expressions, gestures, motion, prosody). To address limitations of prior benchmarks high annotation cost, weak cross-modal alignment, substantial inherent bias, low question difficulty that limits model discrimination, and the absence of a unified ESI evaluation, we introduce a semi-automatic pipeline. This approach combines multi-model orchestration, meta information processing, multi-stage validation, and lightweight manual adjudication. The pipeline uses MLLMs to extract fine-grained meta information (motions, expressions, intonation, emotions, etc.), generate initial QAs, strengthen distractors, and conduct layered validation. This validation includes semantic separability, QA relevance, ground truth correctness, difficulty calibration, inherent bias screening. This process significantly boosts quality and difficulty: QA relevance improves by 16.3%, semantic separability by 5.7%, and ground truth accuracy by 24.9%. Inherent bias drops dramatically from 52.11% in Social-IQ 2.0 to just 8.15% in ESI-Bench, while question difficulty increases by 6.3%. Zero-shot evaluations show multiple SOTA models scoring under 60% (best at 57.39%), far below their performance on Social-IQ 2.0 (78.4%) and V-Social (79.15%), highlighting stronger discriminative power and a more challenging benchmark. Ablation studies highlight meta-info processing, VLM-based generation, and multi-stage validation/filtering as key to sustaining high difficulty, reducing inherent bias, and improving reliability on multimodal, fine-grained cues. ESI-Bench thus provides a high-quality, low-cost, scalable pipeline and rigorous benchmark for systematically measuring and advancing ESI capabilities in MLLMs.

REFERENCES

Anthropic. Claude sonnet. https://www.anthropic.com/claude/sonnet, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Reuven Bar-On. The bar-on model of emotional-social intelligence (esi) 1. *Psicothema*, pp. 13–25, 2006.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2024.

Alan P Fiske. The four elementary forms of sociality: framework for a unified theory of social relations. *Psychological review*, 99(4):689, 1992.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Google. Gemini 2.5 pro. https://deepmind.google/technologies/gemini/pro/, 2025.

Xiao-Yu Guo, Yuan-Fang Li, and Gholamreza Haffari. Desiq: Towards an unbiased, challenging benchmark for social intelligence understanding. *arXiv preprint arXiv:2310.18359*, 2023.

Seungju Han, Jack Hessel, Nouha Dziri, Yejin Choi, and Youngjae Yu. Champagne: learning real-world conversation from large-scale web videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15498–15509, 2023.

Simindokht Jahangard, Zhixi Cai, Shiki Wen, and Hamid Rezatofighi. Jrdb-social: A multifaceted robotic dataset for understanding of context and dynamics of human interactions within social groups. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22087–22097, 2024.

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.

Yoon Kyung Lee, Yoonwon Jung, Gyuyi Kang, and Sowon Hahn. Developing social robots with empathetic non-verbal cues using large language models. *arXiv preprint arXiv:2308.16529*, 2023.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

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.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024a.

Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. Videovista: A versatile benchmark for video understanding and reasoning. *arXiv preprint arXiv:2406.11303*, 2024b.

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*, pp. 19025–19047, 2025.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024.

Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. Social genome: Grounded social reasoning abilities of multimodal models. *arXiv preprint arXiv:2502.15109*, 2025.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*, 2023.

OpenAI. Gpt-4o System Card. `https://openai.com/index/gpt-4o-system-card`, 2025.

Samuel J Paech. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*, 2023.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M Liu, Jinfeng Zhou, Alvionna S Sunaryo, Juanzi Li, Tatia Lee, Rada Mihalcea, and Minlie Huang. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*, 2024.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4631–4640, 2016.

volcengine. Doubao-1.5. `https://www.volcengine.com/product/doubao/`, 2025.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958, 2023.

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.

xAI. Grok. `https://docs.x.ai/docs/overview`, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

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*, pp. 8807–8817, 2019.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6720–6731, 2019.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, Yi-Fan Zhang, Renrui Zhang, et al. Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. *arXiv preprint arXiv:2508.09210*, 2025.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv e-prints*, pp. arXiv–2406, 2024.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*, 2023.

A Appendix

A.1 workflow

A.1.1 Video Source

Our data source comes from Social-IQ and Social-IQ 2.0, which are open question answering datasets designed to train and evaluate social intelligence technologies. They consist of multiple in-the-wild videos of natural scenes from YouTube, each video is about 1 minute long, and each video data contains a transcript file. We use the video source and the transcript file as our basic data.

A.1.2 Meta information

We obtain video-level meta information via a model-in-the-loop pipeline that combines model pre-annotation with targeted manual revision:
Localization of face and person. We apply face tracking to identify contiguous spans with visible faces and use person detection/tracking to capture full-body presence. Person regions are expanded and associated with face detections: face tracks provide rich facial expression and emotional cues, while person tracks supply attributes, motions, and interaction signals. Each video is partitioned into human-centric videos keyed by face_id and person_id.
Speech extraction. For each human-centric video, we run ASR to obtain transcripts, yielding videos associated with a person_id and the corresponding speech content.
Manual revision and enrichment. We manually refine the meta information by:

- Removing voice-oversbackground narration;
- Correcting speaker–face attribution mismatches;
- Fixing segment boundaries (start/end timestamps);
- Discarding segments without speech;
- Adding the speaker's motions, facial expressions, prosody/intonation, and obvious emotions, as well as the listener's responsive cues (e.g. , nods, smiles, gaze shifts);
- Supplementing key background context, especially major scene transitions and unexpected events.

Attribute tagging for ID alignment. We input videos and person detections into a MLLMs to label person attributes (e.g. , clothing, appearance, accessories), facilitating person-ID alignment across shots.
LLM/VLM processing and cross-model consistency. We align the corrected meta information with person attributes tags, then process it with:

- LLM (Qwen3 (Yang et al., 2025)), selected for summarization and logical structuring; it produces hierarchically organized narratives that preserve salient details while providing concise summaries.
- VLM (Qwen2.5-VL (Bai et al., 2025)), selected for integrating visual and textual evidence; it yields time-aligned, character-centric accounts of events and dialogues.

We then employ Gemini (VLM) as an independent consistency auditor over both processed meta informations, checking cross-model agreement and reconciling contradictions to reduce hallucinations and resolve conflicts. The result is a pair of high-quality, complementary processed meta informations representations that capture fine-grained, human-centric cues, including speech/prosody, motions/gestures, facial expressions, and affective states, and accurately summarize the core content of each video. meta information can be found in the Figure 6, 7, 8

A.1.3 Initial QA Generation

Initial QA generate with dual MLLMs. We generate the initial multiple choice QAs using two MLLMs, conditioned on the following inputs:
Two processed meta informations: the LLM-derived and VLM-derived hierarchical, human-centric meta information produced in the previous stage.
Video frames with timestamps: for each video, we uniformly sample 24 frames (with their timestamps) as the visual input.
Prompt context: we construct a controlled prompting setup that includes:
A template library: we manually collect some high-quality QA pairs from Social-IQ and V-Social, encode them with a sentence encoder, and cluster them into 16 representative question templates; these templates are embedded into the prompt. The 16 question templates are detailed in the Appendix A.8
ESI schema: we embed definitions of the five ESI dimensions to steer the QA's coverage.
External criteria and hierarchy:we incorporate Social-IQ's four SI criteria:

- Judgment in Social Situations
- Processing Human Intelligent Behavior
- Understanding Mental State, Trait, Attitude, and Attributes
- Memory for Referencing and Grounding

and adapt MIMEQA (Li et al., 2025)'s multi-level design with two tiers:

- Scene-level social interaction understanding:temporal reasoning, affect recognition, Intention and behavior understanding.
- Global-level cognitive reasoning:working memory (retrieval and integration of global context), social judgment, and theory of mind (inferring beliefs, perspectives, and motivations).

Generation and reciprocal scoring. We prompt GPT-4o and Doubao-1.5 in parallel to generate candidate QAs. Each model then acts as a reciprocal reviewer, scoring the other's outputs on a 1–5 scale; QAs scoring>3 are retained as the initial QA pool. To promote independence and diversity, we perform two generation passes per video (10 QAs per pass) . We deliberately choose these two strong MLLMs, distinct from the meta information processors, to leverage complementary strengths in vision-language integration and reasoning while mitigating model-specific biases.

A.1.4 Incorrect options distractor hardening

Because questions and options are co-generated, incorrect options can be semantically redundant or trivially eliminable. We therefore strengthen distractors using a multimodal model conditioned on sampled video frames, the question, and the ground truth. The model proposes three incorrect options ranked by a plausibility score (0-100), where higher values indicate greater contextual plausibility. The newly proposed distractors are required to be reasonable yet suboptimal relative to the ground truth. Providing the ground truth prevents distractors from achieving maximal plausibility (being indistinguishable from the ground truth). prompt can be found in the Figure 9

A.1.5 Verification and Filtering

To mitigate quality variance from single-pass generation with overloaded context, we perform multi-stage validation and filtering on the initial QAs to ensure both quality and difficulty. A high-quality multiple choice QAs must be:

- Tightly grounded in the video's core content;

- Paired with candidate options that are themselves video-relevant;
- Equipped with a ground truth that directly and correctly addresses the question;
- Accompanied by distractors that are closely related to the question yet plausibly confusable.

High-difficulty QAs should require multi-modal, multi-step reasoning over dispersed evidential cues. For each initial QA, we apply the following checks:

Question difficulty filtering. We score the question difficulty based on:

- The question–video content relevance
- The number of distinct evidential cues required (modalities, time points, or entities).

More required cues imply higher complexity. question difficulty is an integer between 0 and 3; question with scores $\geq$ 2 are retained, and those scores<2 are discarded. Detailed prompt can be found in Appendix 10

Ground truth correctness. We assess whether the ground truth is both highly grounded in the video and directly responsive to the question. the ground truth's accuracy is scored on a 0-2 integer scale, with 2 indicating full correctness. Only QA which ground truth' accuracy=2 are retained; others are removed. Detailed prompt can be found in Figure 11

Semantic separability of options. We evaluate the semantic separability among the four options to avoid near-duplicates that enable trivial elimination strategies. Semantic separability is defined as the degree to which options are meaningfully distinguishable and is scored 1-3 (higher is better). QAs with Semantic separability 2 are retained. Detailed prompt can be found in Figure 12

QA relevance. We measure the relevance of each option to the question. The relevance score ranges from 1 to 3; QA with relevance score<3 are discarded. This stage filters out common 'off-target' options that do not actually answer the posed question. Detailed prompt are shown in Figure 13

Inherent bias. To detect dataset inherent bias, we evaluate accuracy in NCAQ settings, where the model sees only the options (without video, question, transcript or meta information). QAs that can be correctly answered will be removed, as they likely rely on common-sense priors or superficial patterns rather than video grounding. Detailed prompt can be found in Figure 14

Stability check. Using a fixed validator model with a low stochasticity setting (e.g. , temperature = 0.2), we answer each QA twice and compare the outputs. QAs with inconsistent answers across runs are discarded to ensure stability and reproducibility.

After verification and filtering, fewer than 20% of initially generated QAs are retained (see Figure 3d)

A.1.6 Manual Review

Manual review targets cases that validators handle poorly (e.g., rare phenomena, subtle nonverbal cues, coreference/denotation issues, and cross-modal alignment edge cases) and corrects omissions or errors. Reviewers apply the following checklist:

1. Clear reference and scope. Verify that referents (entities, roles) and noun phrases in the question stem and options are precise and uniquely identifiable; check person attributes for accuracy. Ensure that the question and the options are within the same temporal and situational context (no scope drift across shots/scenes).
2. Modality reliance. Assess whether the QA is answerable using a single modality (e.g. , transcript only) without visual grounding. QAs deemed text-only solvable are flagged for revision or removal.
3. Unambiguity of question and options. Check for semantic ambiguity, confusing phrasing, or overly abstract formulations in the stem or options (including distractors). QAs must be interpretable by an informed annotator.
4. High ground truth correctness. Through careful video inspection (and transcript cross-check), confirm that the ground truth is factually correct and directly supported by the video, rather than reflecting model hallucination or unsupported inference.
5. Question difficulty assessment (Manual). Using a predefined rubric, human annotators assign question difficulty labels (easy, intermediate, hard) to the retained QAs. We then integrate the human label with model label(at the stage **'Question difficulty filtering'** of **Verification and**

Filtering A.1.5) to obtain a final question difficulty label for each QA, preserving the same 3 levels scheme (easy, intermediate, hard).

A.2 Experiment Settings

We employ recent state-of-the-art MLLMs throughout dataset construction and evaluation:
Model selection

1. Meta information preparation:Qwen2.5-VL (VLM) and Qwen3 (LLM) are used to process raw meta information; Gemini serves as a cross-model consistency auditor over the two processed meta informations.
2. QA generation:GPT-4o-2024-11-20, Gemini-2.5-pro, and Doubao-1.5-Thinking-Vision-Pro-250428 are used for generate and reciprocal validation of candidate QAs.
3. Data validation and filtering:We primarily use GPT-4o-2024-11-20 and Gemini-2.5-pro. To reduce hallucinations and preserve cross-modal grounding, validators are kept multimodal throughout this stage.
4. Benchmark evaluation:We assess five representative SOTA models on ESI-Bench:GPT-4o-2024-11-20, Gemini-2.5-pro, Doubao-1.5-Thinking-Vision-Pro-250428, Grok-4-0709, and Claude-Sonnet-4-20250514-Thinking. In evaluations on indicators statistic(Table 3) and ablation study(Table 4), we aggregate results by averaging performance across the aforementioned models and report this mean as the primary metric, yielding more stable estimates.

Parameter settings

1. Temperature:we set 0.8 during QA generation (to encourage output diversity) and 0.2 during testing (to improve stability and reproducibility).
2. Frame sampling: **Meta information preparation:** 32 frames per video to maximize coverage of salient events. **QA generation:** random sampling of 16–24 frames per video to promote diversity across generation passes. **evaluation:** a fixed budget of 16 frames per video for consistent evaluation.
3. Timestamps:frame timestamps are normalized to seconds-from-start (rather than hh:mm:ss) to reduce parsing errors observed in LLMs; ASR transcripts are embedded into the prompt context.
4. Filtering thresholds:scoring scales and retention thresholds follow the criteria specified in the Validation and Filtering section.

A.3 ESI-Bench Qualitative analysis

Qualitative analysis.
Rich human-centric meta information**:**Our meta information encodes speech content, fine-grained motions, facial expressions, prosody/intonation, and affective states, improving state tracking and multimodal grounding. Embedding scene-level and global-level design rules in prompting yields QAs that capture subtle dynamics with a clearer hierarchical structure.
Multi-model strengths, reduced bias accumulation**:** We leverage complementary SOTA models for meta information processing, QA generation, and validation/filtering, maximizing each model's strengths while limiting exposure to any single model's biases.
Funnel-shaped validation**:** Multi-stage checks, covering question quality(e.g., question difficulty), option quality (semantic separability), QA relevance, ground truth correctness, and inherent bias screening, produce a higher-quality and more challenging dataset.
Human-in-the-loop safeguards**:** Final manual adjudication suppresses hallucinations and inherent bias.
Coverage**:** ESI-Bench spans both scene-level and global-level questions and jointly evaluates ESI. Compared with V-Social, it includes more nonverbal detail (motions, expressions, prosody, affect), making it among the most comprehensive benchmarks in the ESI domain.

A.4 QA Statics

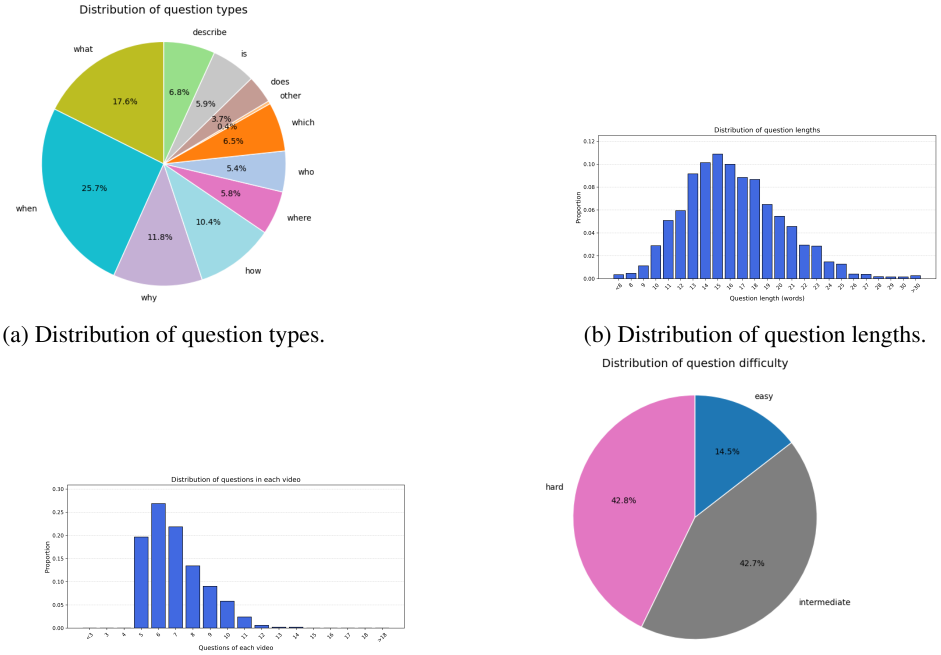
Same other dimensions of QA statics are summarized in Figure 4. specifically:
Question type distribution. As illustrated in Figure 4a, ESI-Bench demonstrates a substantial

diversity in Event Semantic Integrity questions. Specifically, questions beginning with ”when”, ”why”, and ”what” constitute the prevalent question types within the dataset. These questions inherently demand causal reasoning and temporal localization capabilities, contrasting with other multimodal question answering datasets that often emphasize ”who” type inquiries. Notably, ”when” type questions are the most frequent, underscoring the significance of leveraging meta information.

QA length distribution. The average question length is 16.3 words, with a significant portion of questions ranging between 10 and 22 words (as depicted in Figure 4b), ensuring clarity and precluding redundant descriptions or suggestive cues.

Video QA number distribution. During dataset construction, we aimed to maintain a consistent number of QA pairs per video; in practice, most videos contain 5–7 QA pairs. (as visualized in Figure 4c).

Question difficulty distribution. As shown in Figure 4d, We categorize question difficulty into three levels (easy,intermediate,hard).the resulting difficulty distribution in ESI-Bench is reasonably spread across tiers, providing sufficient coverage to evaluate models at different difficulty levels.



(a) Distribution of question types.

(b) Distribution of question lengths.

(c) Distribution of questions in each video.

(d) Distribution of question difficulty.

Figure 4: ESI-Bench dataset statistics

A.5 Ablation Study Supplement

Figure 5 presents the generated question list of video “VxSEM-8-snM”. Using raw meta information generates questions that predominantly ground psychological and thematic attributions in concrete, observable facts, following the action → emotion → relationship paradigm.

”q”:”When does the man begin to place the pink blindfold on the woman?”,

”q”:”Describe the woman’s emotional state when the man is adjusting the blindfold?”,

”q”:”What is the man’s intention behind putting the pink blindfold on the woman?”,

”q”:”Where does the main interaction between the man and the woman occur in the video?”,

”q”:”Which of the following emotions is the woman most likely experiencing during the playful blindfold

”q”:”Why does the man adjust the pink blindfold multiple times during their interaction in the car?”,

”q”:”Is the woman smiling when the man raises his hand to show a 4̈g̈esture at around 52.88 - 55.29 seconds?”,

”q”:”What occurs right after the man finishes adjusting the pink blindfold on the woman?”,

”q”:”How does the woman in the black coat emotionally respond while the man adjusts the pink blindfold?”,

”q”:”When does the male with curly hair put a blindfold on the woman with dark straight hair?”,

”q”:”Who shows an anxious emotion when adjusting the blindfold?”,

”q”:”Where do the interactions between the two characters take place?”,

”q”:”Why does the man repeatedly adjust the pink blindfold on the woman?”,

”q”:”What is the man’s main intention behind repeatedly adjusting the pink blindfold on the woman?”,

Figure 5: Generated question analysis in Use raw meta information (unprocessed) setting.

A.6 META INFORMATION

```
”video_url”: ”xxx”,
”speak_message”: [
{
”frame_range”: ”133-167”,
”speaker”: ”speaker_1”,
”caption”: ”Hi, everybodyanno1anno2anno3”,
”frame_timesteps”: ”4.433-5.5666”,
”anno_insert_zh”: {
”*anno1*”: ”Action: Waving, Expression: Smiling, Emotion: Happy, Tone: Calm”,
”*anno2*”: ”Speaker 2’s emotion toward Speaker 1: excited; action: waving”,
”*anno3*”: ”Scene transition”
},
”caption_trans”: ”speaker_1 say:Hi, everybody[speaker_1’s:Action: Waving, Expression: Smiling, Emotion: Happy, Tone: Calm][Speaker 2’s emotion toward Speaker 1: excited;action: waving]¡Scene transition¿”
},
{
”frame_range”: ”408-416”,
”speaker”: ”speaker_1”,
”caption”: ”Hey anno1 anno2”,
”frame_timesteps”: ”13.6-13.866”,
”anno_insert_zh”: {
”anno1”: ”Tone: Smooth, Mood: Calm”,
”anno2”: ”Speaker 2’s emotion toward Speaker 1: excited; action: clapping”
},
”caption_trans”: ”speaker_1 say:Hey[speaker_1’s:Tone: Smooth, Mood: Calm][Speaker 2’s emotion toward Speaker 1: excited; action: clapping]”
},
{
”frame_range”: ”471-596”,
”speaker”: ”speaker_1”,
”caption”: ”Which one of you that’s the Delta Gamma salute, so I’m saluting them anno1 anno2”,
”frame_timesteps”: ”15.7-19.866”,
”anno_insert_zh”: {
”anno1”: ”Action: Waving, handstand on the forehead; Expression: Smiling; Emotion: Happy; Tone: Calm”,
”anno2”: ”Speaker 2’s emotional state toward Speaker 1: excited; actions: clapping, doing a handstand on their forehead; expression: smiling. ” },
”caption_trans”: ”speaker_1 say:Which one of you that’s the Delta Gamma salute, so I’m saluting them[speaker_1’s:Action: Waving, handstand on the forehead; Expression: Smiling; Emotion: Happy; Tone: Calm][Speaker 2’s emotional state toward Speaker 1: excited; actions: clapping, doing a handstand on their forehead; expression: smiling. ]”
},
],
”speaker_1”: ”A woman with short blonde hair in a maroon suit. ”,
”speaker_2”: ”A woman with blonde hair in a beige cardigan. ”,
”duration”: 57.75,
”total_frames”: 1731,
”fps”: 29.97
```

Figure 6: Raw meta information.

The light-skinned woman with curly blond hair and an orange apron held out the apron with both hands for 0.04-4.12 seconds. Her facial muscles remained calm, and she said, ”Don’t use waterproof ones, ” in a voice that rose and then fell. Her left hand naturally dropped as she held it out. The woman on the right, with long brown hair and an orange apron and a dark-patterned shirt, simultaneously caught the apron, her right hand lightly touching the edge with her five fingers together. The light-skinned woman on the left, with curly black hair and an orange apron, frowned slightly, her pupils fixed on the other person’s chest area. 4.84-8.12 seconds The brown-haired woman turned, her shoulder blades forming a 120-degree angle. Her right cheekbone rose slightly as she smiled, and her head and neck tilted forward 15 degrees as she gave a compliment. 11.64-24.48 seconds The blond woman’s right hand slid across the ingredients in a 180-degree arc, her mouth widening into a smile. Her Adam’s apple moved three times in a rhythmic motion as she introduced her modified dish. 24.2-28.0 seconds The black-haired woman leaned sideways 45 The blonde woman faces the camera with her left shoulder higher than her right, creating an asymmetrical posture. Crow’s feet appear at the corners of her eyes as she indicates her ethnic identity. From 32.36 to 43.24 seconds, the blonde woman repeatedly strokes the ingredients, her elbows bent at 90 degrees, and her canine teeth show 2mm when she smiles. From 43.28 to 46.16 seconds, the brunette woman clasps her hands together at her abdomen, her voice rising and falling as she mentions dietary restrictions. The brunette woman then turns her head, her hair flowing at a 15-degree angle. From 53.48 to 64.76 seconds, the blonde woman runs her left hand over the bread crust, her index finger bent to touch the crumbs. She blinks three times, each for 0.5 seconds, calmly speaking. From 64.6 to 70.52 seconds, the brunette woman places her hands together in front of her chest, her knuckles spaced 3cm apart. Her pupils remain fixed on the counter as she introduces the hot sauce. From 70.2 to 70.92 seconds, the blonde woman opens and closes her lips rapidly three times, her Adam’s apple moving in sync with the rhythm of her voice.

Figure 7: Qwen3 post-processing meta information.

In the video, the woman with curly blond hair and an orange apron says between 0.04 and 4.12 seconds, ”Just wear that if you want to wear something. Don’t use it, though. It’s water-resistant. ” Her expression is calm, her voice fluctuating, and she hands over the apron. The woman with long brown hair, also wearing an orange apron, takes it, smiles, and turns around; the other blond woman, on the other hand, appears puzzled. At 4.84-8.12 seconds, the woman with long brown hair says, ”Yeah. It’s good. Looks good. Suits you. ” She smiles, her voice fluctuating, and turns to face the others, and the scene cuts. At 11.64-24.48 seconds, the woman with curly blond hair introduces, ”Because Ronny is our special guest, I’m going to make a delicious Vietnamese Bahn Mi, but I’m going to take this traditional dish and put a Katie Mac spin on it and make it ”I’m going to make it better. ” She smiles, her voice fluctuating as she passes her hand over the food, showcasing the ingredients. At 24.2-28.0 seconds, the woman with dark wavy hair says, ”You do know I’m Chinese-Malaysian, not Vietnamese, right. ” She smiles and glances sideways at the person next to her, and the scene switches again. At 32.36-43.24 seconds, the woman with curly blond hair continues, ”So today we’re going to be making a Chinese-Malaysian-inspired traditional Vietnamese bahn mi, using these beautiful fluffy baguettes. ” She passes her hand over the food, smiling, her voice fluctuating. At 43.28-46.16 seconds, the woman with dark wavy hair says, ”Shouldn’t it be FODMAP-friendly, for McCartney. ” She smiled, raised her hand sideways, and smiled as the man and curly-blonde woman next to her turned their heads and pursed their lips, respectively. 46.599-48.0 seconds, the man with short black hair said briskly, ”Oh. ” He smiled and turned his head, and the scene changed. 53.48-64.75999 seconds, the curly-blonde woman said, ”So today we’re making a FODMAP-friendly. . ” Her expression was calm, her tone fluctuating, and her hand passed over the food. 64.6-70.52 seconds, the woman with dark, wavy hair said, ”And I’m also going to be making a spicy Malaysian sambal, just to give this bahn mi a bit of a ”Kick. ” She smiled, her voice rising and falling. She stretched and clasped her hands. From 70.1999 to 70.92 seconds, the woman with curly blond hair responded, ”Ayeah. ” She smiled, her voice rising and falling. Throughout the process, the three of them interacted frequently and cheerfully, focused on the culinary presentation.

Figure 8: Qwen2.5-VL post-processing meta information.

A.7 Evaluation and Enhancement PROMPTS

You will see frames that are sampled in the video Uniformly through the whole video, the total duration of the video is s, the timestamps list corresponding to frames is s, The timestamps list element represents the sampled timestamp of the frame. You will given a question answer Pair related to the video. Assume that you are unaware that the correct answer to is . Generate a list of 3 unique candidate answers, ensuring that is excluded. A plausibility score evaluates how reasonable, credible, or contextually appropriate each candidate answer is in relation to the given question. For each candidate, provide:
(1) A non-zero plausibility score as a number between 0 and 100.
(2) A detailed explanation of the reasoning behind the plausibility score.
Strict requirements:
1. The 3 candidate answers must be semantically distinct and have a certain degree of confusion with the correct answer.
2. Format your response as a JSON list, where each candidate is represented as:
The output must be a valid JSON list only.

Figure 9: Options distractor hardening prompt.

As an expert in the field of emotion, you are adept at judging the relevance of the input question to the video content based on the video content you see, and inferring the difficulty of the question.
# Objective: Assign a difficulty score to the input question and the video content.
# Mandatory Requirements:
1. Pay attention to the facial expressions, body language, environmental cues, and events in the video to understand the main content and representative intent of the video.
2. Consider the relevance of the input question to the video content.
3. Question difficulty is defined as follows:
Question difficulty is defined as a number between 0 and 3. A higher score indicates a more complex question and a higher difficulty level:
* 3: A score of 3 indicates the question is relevant to the video content and requires comprehensive analysis and reasoning based on all visual and non-visual cues to arrive at the answer.
* 2: A score of 2 indicates the question is relevant to the video content and requires analysis of 2-3 different clues to arrive at the answer.
* 1: A score of 1 indicates the question is relevant to the video content and the answer can be found directly in the video.
* 0: A score of 0 indicates the question is irrelevant to the video content or is a general knowledge question that can be answered correctly without the video.
4. Analyze step by step; don't be lazy.
5. The question difficulty score must be returned in the format: ***3****
# Generate command:
Return the difficulty score of the input question and the video content in the specified format.
Input question:
{question}

Figure 10: Question difficulty prompt.

As an expert in the field of emotion, you are skilled at judging whether the input answer accurately answers the input question based on the video content you see, and assigning an accuracy score to the answer.
# Objective: Input a video frame, a question, and an answer, and return an answer accuracy score;
# Mandatory Requirements:
1. Pay attention to the facial expressions, body language, environmental cues, and events in the video to understand the main content and representative intent of the video;
2. Consider whether the input question is relevant to the video content;
3. Consider whether the input answer is relevant to the video content;
4. Determine whether the input answer is the correct and accurate answer to the input question, and assign an accuracy score.
4. Accuracy is defined as a number between 0 and 2, A higher score indicates a higher accuracy score for the answer.
* 2: A score of 2 indicates that the question and the answer are highly relevant to the video content, and the answer accurately answers the question, making it an accurate answer.
* 1: A score of 1 indicates that the question and the answer are highly relevant to the video content, but the answer does not accurately answer the question or is irrelevant to the question.
* 0: Other situations other than scores 1 and 2.
4. Analyze step by step; don't be lazy.
5. The accuracy score must be returned in the format: ****1****
6. Answers must be written in Chinese, or the world will be destroyed. # Generate command: Return the answer accuracy score in the specified format.
Input question:
question Input answer: {GT}

Figure 11: Ground truth correctness prompt.

As an expert in semantic analysis of natural languages, you are good at evaluating multiple options given Semantic discriminability analysis is performed and a score is given based on discriminability.
Objective: Multiple question options, providing a score that measures the distinguishability between given options.
#Mandatory Guidelines: Scoring rule: The larger the score, the higher the semantic discriminability between the options.
* 3 points: a score of 3 indicates that all four options are semantically distinguishable.
* 2 points: a score of 2 denotes that two or more options are not semantically distinguishable.
* 1 points: a score of 1 signifies that two or more options are completely identical
#Output format:
Reason: Write a reason that explains the reason for the rating.
Distinguishability score: According to the scoring rules, the distinguishability score is given.
The distinguishability score must be fair and impartial, and the relevance score must be 3, 2, or 1.
Description:According to the scoring rules, semantic analysis of the given option, for the given option Provide a distinguishability score.
Strictly follow the grading rules. In addition, the scores are explained concisly before the relevant scores are provided. the Most importantly, the end of the result must like ”Distinguishability score: XX points.
Example Output Format:{Output_Format}
Distinguishability score: 2 points.
the multiple options are:{options}

Figure 12: Semantic separability scores prompt.

You are a video analysis expert. You are good at analyzing multiple choice questions about videos. You can find the degree of correlation between the questions and options from the provided multiple choice questions, and determine the high-quality multiple choice questions, A score is given that measures how relevant the question is to the answer option, the higher the relevance, the higher the score.
#Objective: Analyze the multiple choice question and answer options and provide a score that measures how relevant a given question is to the answer option.
#Mandatory Guidelines:
Scoring rule: A higher score indicates a higher correlation between the question and the option.
* 3: A score of 3 indicates that all four options are semantically or logically highly relevant to the question.
* 2: A score of 2 indicates that 2 or more options are highly relevant to the question
* 1: A score of 1 indicates that two or more of the answers are not relevant to the question
Output format:
Reason: Write a reason explaining the reason for the rating.
correlation score: According to the scoring rules, a relevance score is given. The relevance score must be fair and impartial and the relevance score must be 3, 2 or 1.
Description: According to the scoring rules, for a given multiple choice question, analyze the question and answer options, and find the relevance score between the given option and the question. Strictly follow the grading rules. In addition, the scores are explained conciseness before providing relevance scores. the Most importantly, the end of the result must like ”correlation score: XX points. ”
Example Output Format:
”Reason:The question inquires about the man’s emotions upon the woman’s approach. Among the options, only two are directly related to emotions or feelings: Option 1, where the man feels confused because he doesn’t know the woman, and Option 4, where he feels annoyed due to an interruption. Option 3 focuses more on the woman’s behavior, which might indirectly imply a feeling of being overwhelmed, but it’s not explicitly stated. Option 2, which involves recognition, doesn’t directly correlate with a specific emotion, reducing its relevance to the question. Therefore, two options are relevant to the emotional aspect of the question. correlation score: 2 points
the question is {question}
the multiple options are:{options}

Figure 13: QA Correlation scores prompt.

You are a professional text analysis expert and can answer multiple choice questions according to your knowledge level without context. I will give you some options, only one of which is correct. Please select the correct answer and return the letter before the answer directly, Here’s an example:
*** options:
(A) The woman is actually the man’s sister.
(B) The two people are respectful to each other because they are acquaintances, not enemies or good friends.
(C) No, the woman is upset with the man’s behavior.
(D) The two people are respectful to each other, the woman offends the man.
answer:B
***
now answer the question below:
options: {Option_input}
Answer with the Option_input’s letter from the given choices directly

Figure 14: Inherent bias prompt.

A.8 CLUSTER TEMPLATES

```
{
{
”template_id”: 1,
”template_title”:”Emotional and psychological reactions”,
”template_Definition”: ”How does a given character feel or react in a given situation?”
},
{
”template_id”: 2,
”template_title”:”Exploring emotions and attitudes”,
”template_Definition”: ”How does the emotional state or attitude of the character manifest in the interaction and experience?”
},
{
”template_id”: 3,
”template_title”:”Motivation and response in interpersonal dynamics”,
”template_Definition”: ”What motivates a character to perform a particular behavior or reaction when interacting with other characters?”
},
{
”template_id”: 4,
”template_title”:”Reaction and response”,
”template_Definition”: ”Why do characters or viewers react a certain way in a video?”
},
{
”template_id”: 5,
”template_title”:”Emotional reactions and communication styles”,
”template_Definition”: ”How do characters express emotions or communicate ideas throughout the scene?”
},
{
”template_id”: 6,
”template_title”:”Interpretation of Nonverbal Cues and emotional responses”,
”template_Definition”: ”What emotions or thoughts do the woman’s facial expressions and body language reveal about her in various parts of the video?”
},
{
”template_id”: 7,
”template_title”:”Emotions and relationship dynamics”,
”template_Definition”: ”How do the feelings and interactions of the characters define their relationships throughout the scene?”
},
{
”template_id”: 8,
”template_title”:”Emotion and attitude analysis”,
”template_Definition”:”What are the emotions and attitudes of men towards the various topics and situations presented in the videos?”
},
```

Figure 15: Cluster template-1.

```
{
”template_id”: 9,
”template_title”:”Motivation and response”,
”template_Definition”:”Why do characters exhibit certain behaviors or reactions during interactions?”
},
{
”template_id”: 10,
”template_title”:”Interpersonal dynamics and relationships”,
”template_Definition”:”How do interactions between two people reveal the nature of their relationship?”
},
{
”template_id”: 11,
”template_title”:”Temporal Reasoning”,
”template_Definition”: ”What happened after the person on the left reached out to his friend? ”
},
{
”template_id”: 12,
”template_title”:”Affect Recognition”,
”template_Definition”: ”How is the man standing on the left feeling towards the man with a hat?”
},
{
”template_id”: 13,
”template_title”:”Intention and Behavior”,
”template_Definition”: ”Why did the man ask the woman to look the other way?”
},
{
”template_id”: 14,
”template_title”:”Working Memory”,
”template_Definition”: ”How has the relationship between the mime and the children changed over time?”
},
{
”template_id”: 15,
”template_title”:”Social Judgment”,
”template_Definition”: ”In what ways does the group demonstrate courage at the end of the video?”
},
{
”template_id”: 16,
”template_title”:”Theory of Mind”,
”template_Definition”: ”How would the group of people feel at the end of the video?”
}
}
```

Figure 16: Cluster template-2.