THG-OMNI: TRACEABLE HUMAN-CENTRIC GRAPH SUPERVISION FOR VIDEO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni-modal models can answer questions about video, but their predictions rarely show which participant, event, or interval supports an answer. Moreover, answer-level supervision alone does not teach a model to interpret structured video evidence. To systematically mitigate this issue, we present THG-Omni, a human-centric graph-supervision framework that links Segment Anything Model 3 participant tracks with temporally grounded Qwen3-VL annotations. Candidate semantic claims are checked against the source video and available audio before they are used to construct training questions, while track geometry provides explicit spatial and temporal supervision. The resulting questions teach an omni-modal reasoning model to interpret participant attributes, actions, affect, interactions, and their timing from a structured graph. At inference, THG-Omni uses a predicted graph constructed from the evaluation video. Experiments across social reasoning, emotion understanding, and deception perception show consistent improvements over the unadapted backbone. A complementary diagnostic demonstrates that the predicted graph preserves evidence useful for answering questions even without the source video and audio. Together, these findings show that THG-Omni improves human-centric video reasoning while grounding its evidence in identifiable participants and events. Our codes are available at `https://anonymous.4open.science/r/review-n94hhfcwp5666/`

1 INTRODUCTION

Recent advances in multimodal large language models (MLLMs) have improved temporal localization, long-context modeling, and question answering for complex videos (Ren et al., 2024; Wu et al., 2024; Zeng et al., 2025). By sampling more frames and retaining more visual tokens, these models can process longer videos and capture a wider range of events. However, increased visual coverage does not ensure that relevant evidence is correctly localized, associated with the proper participant, or used during inference (Wu et al., 2024; Chen et al., 2025). These limitations are particularly important for human-centric video reasoning, where a prediction may depend on brief and distributed evidence from facial expressions, body motion, speech, vocal delivery, and interactions among participants (Cheng et al., 2024; Yang et al., 2025).

Human-centric video reasoning requires persistent participant identity and explicit temporal association. Visible attributes, actions, and observable affective states must be assigned to the correct person and restricted to the interval in which they occur. Occlusion, camera motion, viewpoint changes, and shot transitions make such associations difficult (Ravi et al., 2025; Gao et al., 2025). Promptable video segmentation provides a useful basis for addressing this problem. Segment Anything Model 3 (SAM 3) grounds concept prompts and propagates instance masks over time, producing participant trajectories with explicit temporal and spatial information (Carion et al., 2026). Video scene graphs further organize trajectories into participant nodes, temporal states, and relations (Kim et al., 2025; Huang et al., 2025). Related studies on multimodal knowledge graphs and structured representations also show that explicit entities and relations can support multimodal reasoning (Lee et al., 2024; Zhang et al., 2024a; Huang et al., 2024b).

Nevertheless, constructing a human-centric video graph does not guarantee improved video reasoning. First, tracking or annotation errors can attach a valid observation to the wrong participant or extend it beyond its supported interval. Such errors may remain linguistically plausible and cannot

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

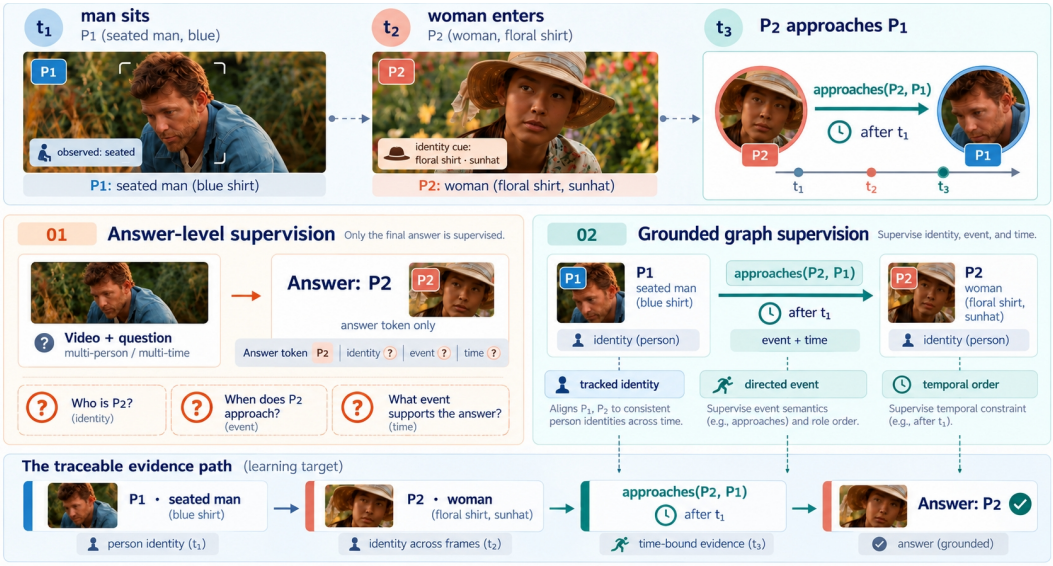


Figure 1: Motivation for traceable graph supervision. This conceptual example contrasts answer-level supervision with identity- and time-grounded graph supervision. The depicted participants and event illustrate the learning objective rather than a benchmark prediction.

be detected through schema validation alone. Second, a pretrained MLLM may not understand a serialized graph without targeted supervision. Direct graph injection can duplicate visual information, propagate uncertain associations, or receive little attention from a model that has not learned the meaning of the graph fields. Modality priors and visual attention biases may further prevent MLLMs from using supplied evidence (Zhou et al., 2025; Kang et al., 2025). Therefore, graph construction and graph utilization must be treated as separate problems (Fig. 1).

To address these limitations, we propose THG-Omni, a traceable human-centric graph-supervision framework for omni-modal video reasoning. As illustrated in Fig. 2, SAM 3 first grounds and tracks question-relevant participants, providing visibility intervals, trajectories, and spatial relations. Qwen3-VL then generates whole-video semantic claims describing visible attributes, actions, observable affective states, interactions, and intention hypotheses. These geometric and semantic elements are merged into a human-centric video graph in which every claim is associated with a participant and a temporal interval. Unlike methods that only append a graph to the input prompt, our framework converts individual graph fields into explicit supervision examples.

Because model-generated semantic claims cannot be regarded as ground truth, we introduce cross-model verification before graph supervision. HumanOmniV2 evaluates each semantic claim using the source video and available audio, while human quality control examines participant identity, temporal extent, and visual support (Chen et al., 2024; Yang et al., 2025). Only supported positive claims are retained. This protocol avoids synthetic negative statements that may introduce lexical shortcuts or semantically ambiguous counterfactuals. The verified semantic claims are combined with deterministic questions derived from participant trajectories to supervise graph interpretation. Using low-rank adaptation (Hu et al., 2021), we evaluate graph-specific supervision against the frozen model and compare video-matched, shuffled, and omitted graph input within checkpoint-aligned runs. These experiments examine information retained in the graph and its use during inference.

Our main contributions are summarized as follows:

- We distinguish the information preserved in a human-centric video graph from a reasoning model's ability to use that information, providing a clearer basis for evaluating graph-supervised video reasoning.
- We introduce a traceable video representation that combines question-guided SAM 3 participant tracks with temporally localized annotations of attributes, actions, emotions, interactions, and intentions.

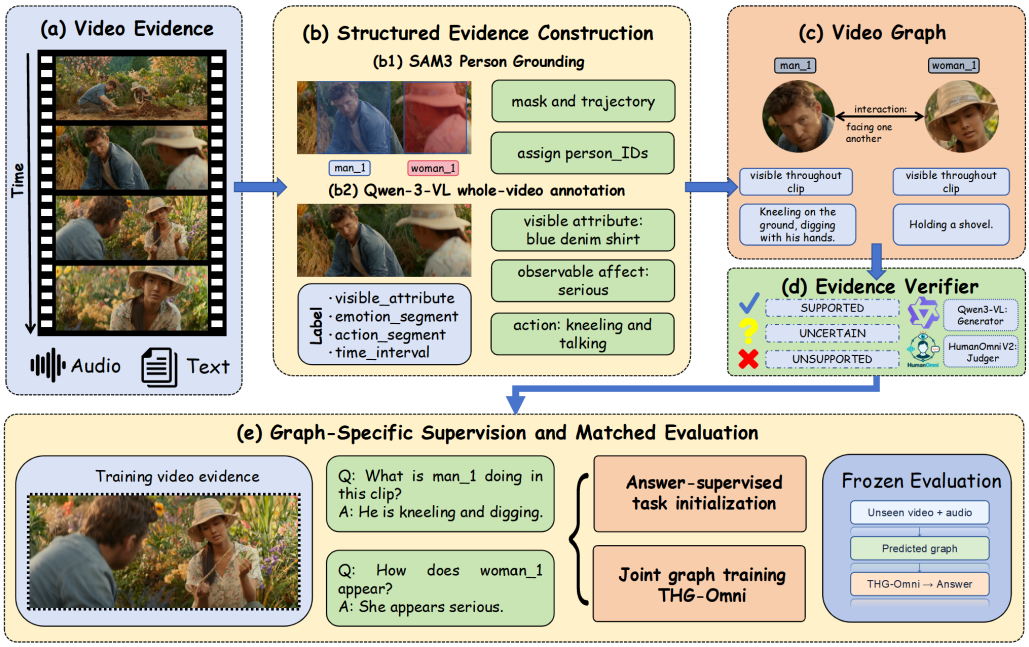


Figure 2: Overview of the proposed traceable human-centric graph-supervision framework. (a) Temporally ordered video evidence. (b) SAM3 grounds and tracks participants, while Qwen3-VL annotates visible attributes, emotions, actions, interactions, and their temporal intervals. (c) The grounded evidence forms a video graph with identity-bound nodes, temporal states, and participant relations. (d) HumanOmniV2 verifies each candidate claim using the source video and available audio; only supported claims are retained. (e) The verified graph supports training of THG-Omni, which is evaluated with video-matched, shuffled, and omitted graph input.

- We develop a positive-only annotation protocol in which Qwen3-VL proposes semantic claims, HumanOmniV2 checks their support against video and available audio, and human quality control reviews identity and temporal grounding.
- THG-Omni outperforms the evaluated baselines across all three IntentBench tasks, achieving strong results in social reasoning, emotion understanding, and deception perception. Additional graph-input experiments confirm that the proposed representation retains answer-relevant evidence and that graph-specific training helps the model use it.

2 Related Work

Human-centric Video Understanding. Human-centric video understanding includes social reasoning, emotion recognition, intention inference, and deception-related judgments (Zadeh et al., 2019; Cai et al., 2025). Answers may depend on facial expressions, motion, speech, vocal delivery, and interactions among participants. Because these cues can occur at different times and across modalities, recent models combine visual, acoustic, and textual evidence (Cheng et al., 2024; Jin et al., 2024). Long-video studies further highlight the difficulty of retaining and retrieving relevant evidence across extended contexts (Wu et al., 2024; Fu et al., 2025). HumanOmniV2's IntentBench evaluates such reasoning through Social-IQ, EMER, and MDPE, which have different prediction formats and evidence requirements (Yang et al., 2025). These tasks motivate representations that preserve participant identity and the time at which each observation holds.

Video Grounding and Scene Graphs. Promptable segmentation provides spatial grounding: Segment Anything introduced image-level prompting, SAM 2 propagated masks through video, and SAM 3 added concept-prompted detection and tracking (Kirillov et al., 2023; Ravi et al., 2025; Carion et al., 2026). Persistent identity across shots remains a separate challenge from mask localization (Gao et al., 2025). Video scene graphs organize entities and changing relations over time (Huang

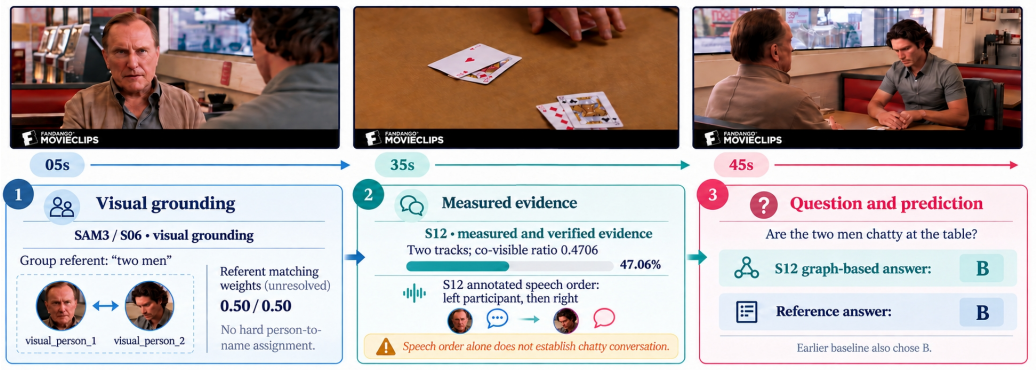


Figure 3: Social-IQ example of grounded evidence and model decision. Video frames at 5, 35, and 45 seconds show the two participants and the card game. SAM3/S06 ground the group referent "two men" to two visual tracks, while the person-to-name assignment remains unresolved (matching weights 0.50/0.50). S12 records a co-visible ratio of 0.4706 and an annotated speech order; speech order alone does not establish a chatty conversation. The S12 graph-based answer and the reference answer are both B. An earlier baseline also chose B, so this example illustrates evidence traceability rather than an accuracy gain.

et al., 2024a). Existing approaches use temporal caption alignment or neuro-symbolic learning to obtain such structure (Kim et al., 2025; Huang et al., 2025). However, a well-formed graph can still assign an action or attribute to the wrong person. Our representation therefore links claims to tracked participants and temporal evidence, and evaluates semantic support separately from structural validity.

Structured Reasoning and Annotation Verification. Multimodal knowledge graphs and structure-aware vision-language models expose entities and relations for reasoning (Lee et al., 2024; Huang et al., 2024b). Video-language methods also use localized supervision or question-conditioned evidence acquisition to improve temporal grounding (Zeng et al., 2025; Wang et al., 2024). Unlike knowledge-graph completion, which predicts missing relations from entity representations (Zhang et al., 2024a), our task uses observations grounded in video trajectories and trains an omni-modal model through graph-specific supervision with low-rank adaptation (Hu et al., 2021; Wang et al., 2025). For semantic fields, Qwen3-VL generates candidate claims and HumanOmniV2 checks their support against the source video and available audio. Prior work on multimodal critics and model-generated feedback motivates this verification step, but model agreement does not guarantee correctness (Xiong et al., 2025; Yu et al., 2025). Human review therefore remains necessary for identity binding, temporal extent, and interaction arguments.

3 Method

3.1 Overview

Our framework constructs a human-centric graph from video evidence and uses the graph to supervise omni-modal video reasoning. SAM3 grounds and tracks people to obtain temporal and spatial evidence. Qwen3-VL then describes visible attributes, actions, affective states, interactions, and possible intentions using grounded person identifiers. The two sources are assembled into a graph while retaining the provenance of each claim. Human quality control and HumanOmniV2 verification filter semantic claims before training. The accepted claims, together with deterministic questions derived from person tracks, form graph-specific supervision. At evaluation time, the model receives a predicted graph from the evaluation video; these evaluation annotations are not manually corrected or claim-verified. Figure 2 summarizes these stages.

3.2 Question-Guided Human Grounding

We first parse the question to identify explicit human references, grammatical number, and the requested semantic category. The resulting prompt set may include *person*, *man*, *woman*, *boy*, and *girl*. Answer candidates are not used to construct the grounding prompts. When several questions share the same video and prompt set, the corresponding grounding record is computed once and reused. Given a video V and a question Q, geometric grounding produces

$$T = g_{\mathrm{SAM3}}\big(V, p(Q)\big), \tag{1}$$

where $p(Q)$ denotes the human references extracted from the question, and T contains person tracks and geometric measurements.

SAM3 performs promptable person segmentation and temporal mask propagation (Carion et al., 2026). Each video is sampled at up to 20 frames for construction of the geometric graph. Candidate masks are filtered using fixed constraints on mask area, bounding-box area, side length, and aspect ratio. Intersection-over-union and containment tests remove redundant detections. Track association uses mask overlap, normalized center displacement, and a fixed maximum temporal gap.

For track i at sampled frame t, the observation is:

$$o_i^t = (t, b_i^t, c_i^t, a_i^t), \tag{2}$$

where b_i^t is the bounding box, c_i^t is its center, and a_i^t is the mask-area ratio relative to the frame. The trajectory:

$$\mathcal{T}_i = \{o_i^t : t \in F_i\} \tag{3}$$

records the visible extent and spatial motion of the participant. At this stage, the system does not assign emotions, intentions, or social roles.

3.3 Deterministic Spatio-Temporal Graph

The geometric component is represented as:

$$G_T = (\mathcal{V}_T, \mathcal{E}_T). \tag{4}$$

Each valid human track defines a node containing its prompt category, temporal extent, visibility coverage, and compact trajectory statistics. Pairwise relations are calculated only when two tracks are visible in a sufficient number of common sampled frames.

For tracks i and j, their center distance at frame t is normalized by the frame diagonal:

$$d_{ij}^t = \frac{\|c_i^t - c_j^t\|_2}{\sqrt{W^2 + H^2}}, \tag{5}$$

where W and H are the frame width and height. The graph records co-visibility, relative horizontal and vertical position, mask overlap, normalized distance, and distance change. A video-level relation is retained only when its temporal support exceeds a fixed threshold. The supporting frames and source tracks remain attached to each relation, allowing every deterministic graph field to be traced back to the video.

3.4 Whole-Video Semantic Annotation and Graph Assembly

Geometric tracks do not identify actions, affect, or interactions. Qwen3-VL-8B therefore annotates the complete video and an automatically generated contact sheet, using grounded SAM3 identifiers as identity anchors. Let $H(T)$ be the grounded identifier set and L_C a task-specific candidate vocabulary, when one exists. The semantic annotations are generated as

$$S = g_{\mathrm{ann}}\big(V, \mathrm{contact}(V), H(T), L_C\big). \tag{6}$$

The annotator receives neither the reference answer nor a graph corrected using that answer. For MER24, L_C contains the six candidate emotion labels associated with the video; Social-IQ uses free-form descriptions because it has no corresponding closed inventory.

We represent a semantic claim as

$$s_k = (u_k, \ell_k, I_k, E_k, \gamma_k), \qquad I_k = [r_k^{\mathrm{s}}, r_k^{\mathrm{e}}], \quad 0 \le r_k^{\mathrm{s}} \le r_k^{\mathrm{e}} \le 1, \tag{7}$$

where u_k is a person identifier or an ordered pair of people for an interaction, ℓ_k is the claimed attribute, affect, action, interaction, or intent, I_k is its normalized video interval, E_k contains one to three supporting evidence ratios in $[0, 1]$, and γ_k is a confidence category. The schema also records the people present and annotation limits. Every argument of a claim must name a declared person, and self-interactions are excluded. Appearance claims crossing a camera transition require renewed visual evidence.

An identity map π attaches an annotation identifier to its SAM3 track whenever the track is available. The annotator may introduce a neutral identifier for an additional visible person, but the resulting node remains semantic-only until geometric support exists. Invalid or ambiguous bindings are excluded rather than assigned to a convenient track. The assembled graph is

$$\begin{aligned} G &= g_{\text{merge}}(G_T, S; \pi) \\ &= \big(\mathcal{V}_T \cup \mathcal{V}_S,\, \mathcal{E}_T \cup \mathcal{E}_S\big), \end{aligned} \tag{8}$$

where geometric fields in $(\mathcal{V}_T, \mathcal{E}_T)$ retain their track and frame provenance, while semantic fields in $(\mathcal{V}_S, \mathcal{E}_S)$ retain their claim, interval, evidence ratios, and confidence. A semantic-only person has no deterministic geometric fields. Whole-video annotation is used in the formal pipeline; a fixed-window pilot produced fragmented identities when participants overlapped.

3.5 Training-Only Quality Control and Claim Verification

Human inspection of development outputs checks person binding, temporal extent, label validity, and visual support before large-scale construction. Schema normalization is conservative: it may standardize a supported format, set an inapplicable optional label to `null`, or retain existing evidence points from an overlong list. It does not create claims. Malformed person bindings, invalid intervals, non-human nodes, and unsupported interaction arguments are excluded; changes and exclusions are logged. Appearance changes across camera transitions receive interval-level review.

HumanOmniV2 independently checks each remaining semantic claim against the source video and available audio (Yang et al., 2025). It sees neither the benchmark answer nor a synthetic negation. For a candidate s_k, define

$$z_k = \text{Verify}_{\text{HumanOmniV2}}(s_k, V, A) \in \{\texttt{supported}, \texttt{unsupported}, \texttt{uncertain}\}. \tag{9}$$

The retained semantic collection is

$$S^+ = \{s_k \in S : \text{schema_valid}(s_k) \wedge \text{identity_valid}(s_k) \wedge z_k = \texttt{supported}\}. \tag{10}$$

Manually rejected, unsupported, uncertain, and malformed claims are excluded. When a multi-claim response cannot be parsed, only its affected verification unit is retried; original outputs and parser repairs are preserved. This positive-only protocol avoids the lexical shortcuts and equivalent counterfactuals observed in an earlier original-versus-negated discrimination setting. Both human review and claim verification apply to training and development construction, not evaluation graphs.

3.6 Graph-Specific Supervision

The final training collection combines deterministic graph questions with verified semantic questions. Deterministic targets are computed from frozen track statistics and cover temporal duration, first appearance, relative presence, co-visibility, overlap, distance change, and directional relations. Semantic targets are derived only from accepted positive claims and cover visible attributes, actions, emotions, interactions, and intentions.

For video V, optional audio A, question Q, answer candidates C, and graph G, the downstream prediction is

$$\hat{Y} = f_{\theta,\phi}(V, A, Q, C, G), \tag{11}$$

where θ denotes the frozen omni-modal backbone and ϕ denotes the trainable LoRA parameters.

Training first fits an answer-supervised adapter, merges it into the frozen backbone, and then performs one epoch of joint adaptation. The joint collection includes answer questions with and without graph input, verified graph questions, and matched or mismatched graph identification. With the

training examples denoted by $\mathcal{D}_{\text{joint}}$, the joint-stage objective is

$$\mathcal{L}_{\text{joint}} = - \sum_{(V,A,G,Q,Y)\in\mathcal{D}_{\text{joint}}} \log P_{\theta,\phi}\big(Y \mid V, A, G, Q\big), \tag{12}$$

where G may be omitted for graph-dropout examples. For multiple-choice questions, the candidate set C is included in Q in the training objective.

LoRA modules are attached to the language-model attention projections (Hu et al., 2021). The visual tower, audio tower, and base language-model parameters remain frozen. Human review and cross-model verification are restricted to training and development data construction. During evaluation, the model receives only a graph predicted from the evaluation video. No evaluation annotation is manually corrected, and no verifier verdict is used as an inference input.

4 Experiments

4.1 Experimental Setup

We evaluate THG-Omni on human-centric video reasoning tasks. MER24 contains 150 multi-label emotion samples and is used to test the information retained in the scene graph. IntentBench covers social reasoning, emotion understanding, and deception perception. Its original manifest contains 2,689 questions from 633 videos. One MDPE video is excluded by a stable task identifier because it repeatedly exhausts GPU memory even when processed alone on an NVIDIA L20. The resulting evaluation set contains 2,688 aligned question–graph pairs from 632 videos: 2,356 Social-IQ questions, 133 EMER questions, and 199 MDPE questions. The exclusion is fixed before prediction and scoring.

Graph-specific supervision is constructed from person tracks and accepted semantic claims. The graph-question collection contains 7,339 training examples, comprising 1,555 deterministic temporal and spatial questions and 5,784 verified positive semantic questions; its development collection contains 1,146 examples. The final joint-training stage presents 2,887 examples in one epoch, combining answer questions, graph questions, and matched or mismatched graph examples. The graph-question training and development partitions do not share videos. Qwen2.5-Omni-7B serves as the reasoning backbone. SAM3 provides person tracks and geometric relations, Qwen3-VL-8B generates whole-video semantic annotations, and HumanOmniV2 assesses the support for candidate semantic claims. In the video reasoning pipeline, frames are sampled at 0.25 FPS, generation is deterministic, and available audio is provided to the reasoning model. The annotation model receives video frames and a contact sheet without audio. The verifier may use audio when assessing a claim. LoRA modules are applied to the language attention projections; the visual tower, audio tower, and base language-model weights remain frozen. Additional training and checkpoint details are given in the appendix.

Social reasoning and deception perception are measured by accuracy. Emotion understanding is measured by sample-level multi-label F1. F1 is computed from each question's predicted and reference label sets and then averaged over questions. Exact set match requires equality of the two sets. For MER24, Table 2 reports this sample-averaged F1 as macro F1; micro F1 instead pools true positives, false positives, and false negatives across all samples before computing F1. Its set precision and recall are averaged separately over samples. The overall IntentBench score is a sample-weighted aggregation across the 2,688 questions. Every condition in an IntentBench comparison uses the same questions and evaluation videos.

4.2 Main Experiments

Table 1 illustrates the IntentBench running on the same 2,688 questions. The matched THG-Omni achieves 71.19% Social-IQ accuracy and 68.30% MDPE accuracy, compared with 53.99% and 61.81% for Qwen2.5-Omni-7B. EMER sample F1 is 86.26% versus 68.09% for Qwen2.5-0mni-7B, while EMER exact match is 66.17% versus 41.35%. Our method has a sample-weighted overall score of 71.72%. Figure 4 visualizes the baseline and THG-Omni scores alongside the separate MER24 graph-information diagnostic. Graph-SFT denotes a separate LoRA adapter trained for graph-field question answering and evaluated with either no graph or a compact graph; these two

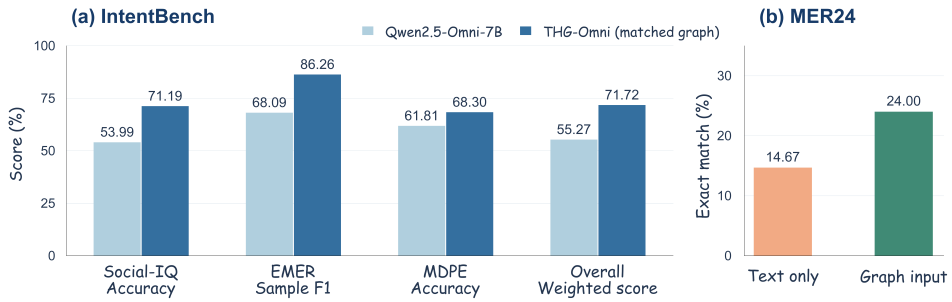


Figure 4: IntentBench results and MER24 graph diagnostic. (a) Qwen2.5-Omni-7B and video-matched THG-Omni are compared on 2,688 questions. (b) On 150 MER24 samples without source video or audio, graph input improves exact set match over text-only input.

Table 1: IntentBench results on 2,688 questions. Overall is the sample-weighted aggregate of Social-IQ accuracy, EMER sample F1, and MDPE accuracy. LLaVA-NeXT-Video uses video without audio. Graph-SFT uses a distinct adapter; the THG-Omni rows report different graph-input conditions for the same trained model.

	Social-IQ	EMER		MDPE	Overall
Model / input	Accuracy	Sample F1	Exact	Accuracy	Score
LLaVA-NeXT-Video Zhang et al. (2024b)	31.58	29.68	12.03	39.20	32.05
HumanOmniV2 (Yang et al., 2025)	51.57	80.03	54.89	53.77	53.14
Qwen2.5-Omni-7B Xu et al. (2025)	53.99	68.09	41.35	61.81	55.27
Graph-SFT (w/o graph)	49.66	70.52	40.60	60.30	51.48
Graph-SFT (compact graph)	56.66	71.32	46.62	61.31	57.73
THG-Omni (w/o graph)	67.28	84.63	66.17	62.31	67.77
THG-Omni (shuffled graph)	67.11	81.29	61.65	61.31	67.38
THG-Omni (matched graph; Ours)	**71.19**	**86.26**	**66.17**	**68.30**	**71.72**

rows share weights. THG-Omni denotes the staged, joint-training model described in Section 3. The THG-Omni rows form a matched graph-input ablation of the same trained model on the same 2,688 questions: they differ only in whether the input graph is matched to the video, shuffled, or omitted. Graph-SFT uses a separate adapter.

LLaVA-NeXT-Video-7B-32K provides an additional video-only baseline on the same 2,688 questions. It receives eight uniformly sampled frames and the question with its answer options, without audio or graph input, and obtains 31.58% Social-IQ accuracy, 29.68% EMER sample F1, and 39.20% MDPE accuracy. These scores use less input evidence than the audio-video runs and should be interpreted with that modality difference in mind. Of the 2,688 responses, 94 Social-IQ predictions were recovered from an unambiguous option letter at the start of the saved raw response; the eight remaining unparsed responses are counted as incorrect. HumanOmniV2 (Yang et al., 2025) is also evaluated without graph input on the same 2,688 questions. It obtains 51.57% Social-IQ accuracy, 80.03% EMER sample F1, and 53.77% MDPE accuracy; its 21 unparsed responses are counted as incorrect. THG-Omni improves EMER sample F1 to 86.26% on the same questions.

Table 2 reports a MER24 graph-information diagnostic on 150 samples. Both conditions receive the same question and answer options, without source video or audio; only the graph input differs. Adding the predicted graph increases exact match from 14.67% to 24.00% and macro F1 from 42.18% to 57.18%. There are 21 rescued and 7 harmed examples (exact McNemar $p = 0.0125$), and the bootstrap 95% interval for the exact-match difference is $[2.67\%, 16.00\%]$. This diagnostic tests information recoverable from the graph, rather than improvement to video-conditioned reasoning.

Table 2: MER24 graph-information diagnostic on 150 samples. Both conditions exclude source video and audio; the only input difference is the predicted scene graph. All values are percentages. Set precision and recall are sample-level averages.

Input	Exact	Set precision	Set recall	Macro F1	Micro F1
Question and options	14.67	37.90	56.67	42.18	44.23
Question, options, and graph	**24.00**	**54.40**	**69.33**	**57.18**	**56.16**

4.3 Graph Evidence and Input Analysis

A separate, fixed Graph-SFT adapter tests whether graph-specific training helps the model use a compact graph. With the same adapter, adding graph input raises the overall IntentBench score from 51.4805% to 57.7330%, a gain of 6.2525%. Social-IQ accuracy rises from 49.6604% to 56.6638%, and EMER exact match from 40.6015% to 46.6165%. Its no-graph score is below the unadapted backbone's 55.27%; the within-adapter benefit is therefore specific to the compact-graph condition, rather than an ablation of the final THG-Omni model. Figure 3 illustrates how one Social-IQ answer can be traced to grounded evidence. The earlier baseline also chose the reference answer in this case, so the example does not represent a corrected baseline error.

The THG-Omni rows in Table 1 further examine the role of graph input. The updated matched-graph run scores 71.72% overall, compared with 67.77% without graph input and 67.38% with a shuffled graph. On Social-IQ, the corresponding accuracies are 71.19%, 67.28%, and 67.11%; on MDPE, they are 68.30%, 62.31%, and 61.31%. EMER sample F1 is 86.26% with the matched graph, versus 84.63% without a graph and 81.29% with a shuffled graph. These observations are consistent with the model benefiting from video-matched structured evidence, particularly for social reasoning and emotion understanding. The no-graph and shuffled-graph controls were evaluated in an earlier run, however, so the differences from the updated matched-graph row should be read as comparisons of reported conditions rather than paired estimates from one evaluation run.

The three input conditions also address different questions. Removing the graph tests whether the trained model can answer from the video and available audio alone. Shuffling supplies graph text whose participant and event descriptions no longer correspond to the evaluated video; it therefore probes the importance of graph–video correspondence rather than simply the presence of additional tokens. The lower EMER score with a shuffled graph is consistent with mismatched evidence being unhelpful for emotion questions, where the relevant person and interval matter. At the same time, the no-graph condition remains substantially stronger than the unadapted backbone on Social-IQ and EMER. Part of THG-Omni's performance is therefore retained without graph input, while THG-Omni with updated matched-graph run gives the highest reported overall score. The MER24 experiment tests a separate property: when source video and audio are withheld, the predicted graph raises exact match from 14.67% to 24.00%. It shows that the representation contains answer-relevant information, but does not by itself measure its contribution in video-conditioned inference. Together, the controlled Graph-SFT comparison, THG-Omni input conditions, and MER24 diagnostic provide distinct evidence about learning to use a graph, preserving video correspondence, and retaining information in the graph.

5 Conclusion

In this work, we present THG-Omni, a framework for human-centric video reasoning with traceable graph supervision. It links SAM 3 participant tracks with temporally localized semantic annotations, verifies candidate claims against the source evidence, and uses the resulting graph to train an omni-modal model to interpret people, actions, affect, and interactions. Experiments on IntentBench show consistent improvements over the unadapted backbone across social reasoning, emotion understanding, and deception perception. Additional graph-input experiments show that the structured representation retains answer-relevant evidence and can improve reasoning when the model is trained to use it. THG-Omni thus offers a way to improve video question answering while making the supporting people and events easier to identify.

AI USE STATEMENT

Generative AI tools assisted with manuscript writing, photo beautifying and language polishing, as well as aspects of code and figure preparation. The authors take responsibility for the final work.

REPRODUCIBILITY STATEMENT

The method, graph construction procedure, training setup, and evaluation metrics are described in Sections 3 and 4. The appendix provides additional details on annotation checks, training examples, and evaluation alignment. An anonymous code snapshot is available at `https://anonymous.4open.science/r/review-n94hhfcwp5666/`. Reproducing the reported results also requires access to the underlying video datasets and pretrained model weights.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 62606208 and Grant 62606209 and in part by the Supercomputing Center of Lanzhou University. This work was also supported by the Talent Scientific Fund of Lanzhou University.

REFERENCES

Cong Cai, Shan Liang, Xuefei Liu, Kang Zhu, Zhengqi Wen, Jianhua Tao, Heng Xie, Jizhou Cui, Yiming Ma, Zhenhua Cheng, et al. Mdpe: A multimodal deception dataset with personality and emotional characteristics. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12957–12964, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *arXiv preprint arXiv:2402.04788*, 2024.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 45647–45682, 2025.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander G Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37: 110805–110853, 2024.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Ruopeng Gao, Ji Qi, and Limin Wang. Multiple object tracking as id prediction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27883–27893. IEEE, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14271–14280. IEEE, 2024a.

Jiani Huang, Ziyang Li, Mayur Naik, and Ser-Nam Lim. Laser: A neuro-symbolic framework for learning spatio-temporal scene graphs with weak supervision. In *International Conference on Learning Representations*, volume 2025, pp. 41977–41999, 2025.

Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Zhou Zhao, Tangjie Lv, Zhipeng Hu, et al. Structure-clip: Towards scene graph knowledge to enhance multi-modal structured representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 2417–2425, 2024b.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-UniVi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13700–13710, 2024.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 87676–87703, 2025.

Kibum Kim, Kanghoon Yoon, Yeonjun In, Jaehyeong Jeon, Jinyoung Moon, Donghyun Kim, and Chanyoung Park. Weakly supervised video scene graph generation via natural language supervision. In *International Conference on Learning Representations*, volume 2025, pp. 15480–15502, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. Multimodal reasoning with multimodal knowledge graph. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10767–10782, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323. IEEE, 2024.

Song Wang, Junhong Lin, Xiaojie Guo, Julian Shun, Jundong Li, and Yada Zhu. Reasoning of large language models over knowledge graphs with super-relations. In *International Conference on Learning Representations*, volume 2025, pp. 4208–4226, 2025.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76. Springer, 2024.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Lllava-critic: Learning to evaluate multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13618–13628. IEEE, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. Humanomniv2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, et al. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19985–19995. IEEE, 2025.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8799–8809. IEEE, 2019.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. Timesuite: Improving mllms for long video understanding via grounded tuning. In *International Conference on Learning Representations*, volume 2025, pp. 38057–38081, 2025.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. Native: Multi-modal knowledge graph completion in the wild. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pp. 91–101, 2024a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *International Conference on Learning Representations*, volume 2025, pp. 54415–54439, 2025.

A Additional Experimental Details and Evidence Analysis

A.1 Annotation and Claim Verification

Whole-video Qwen3-VL annotation was piloted on two development videos and then inspected on ten. All ten records passed schema validation. Manual review checked participant identity, temporal extent, label validity, and support in the video. In one case, an appearance attribute persisted across a camera transition even though the person's clothing changed; the affected interval was flagged for review. A subsequent check of 66 development videos produced 66 schema-valid records.

On the development set, HumanOmniV2 assessed 632 candidate claims from 395 valid verification units. It classified 591 as supported, 26 as unsupported, and 15 as uncertain. Only supported claims that also pass schema and participant-identity checks are eligible for semantic question construction. These verdicts are model-based assessments using the source video and available audio; the supported count is not a human-verified accuracy measure. The 632 development claims are distinct from the 632 videos in the IntentBench evaluation set.

The evaluation-video annotation run covers 632 videos. After invalid outputs are retried, all 632 annotation records pass schema validation. Schema validity checks the output format, not the truth of every semantic claim. Evaluation graphs are generated automatically: they receive no manual semantic corrections, and verifier decisions are not provided to the reasoning model at inference.

A.2 Graph Questions and Adapter Training

The graph-question collection contains 7,339 training examples: 1,555 deterministic temporal or geometric questions and 5,784 questions constructed from accepted semantic evidence. Its development collection contains 1,146 examples and has no video overlap with the graph-question training collection. Deterministic answers are computed from tracked-person statistics; semantic questions cover attributes, actions, affect, interactions, and intentions only when their source claims pass verification.

The final reasoning model is trained in two stages. First, a task question-answering LoRA adapter is trained and merged into Qwen2.5-Omni-7B. One epoch of joint training then uses the resulting

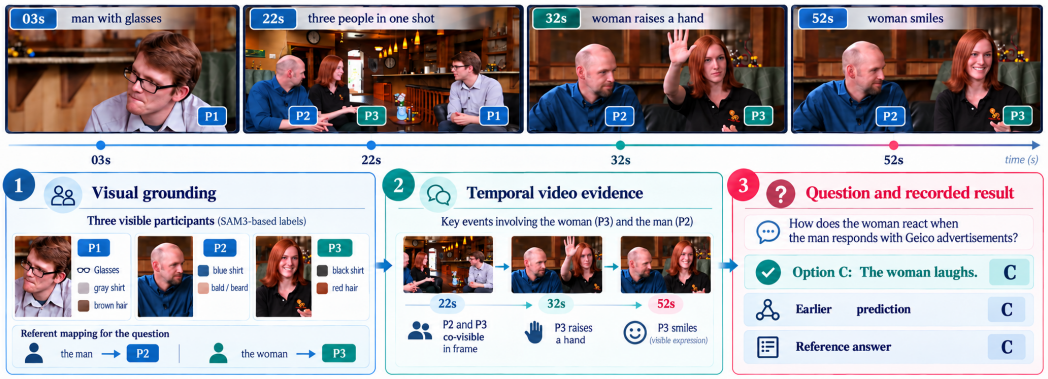


Figure 5: Illustrative Social-IQ video-to-graph example. Original video frames show the visible participants and the woman's later reaction; the graph links in panel C are schematic.

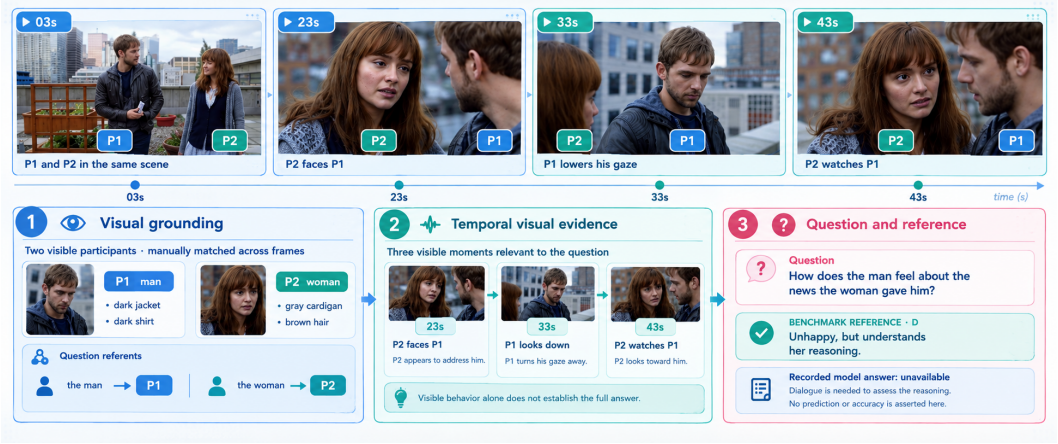


Figure 6: Illustrative Social-IQ video-to-graph example. Original video frames show the visible participants and the woman's later reaction; the graph links in panel C are schematic. The figure further illustrates participant grounding, temporally localized visual evidence, and the alignment between the Social-IQ question and its reference answer.

checkpoint at a learning rate of 10^{-5}. The joint epoch comprises 2,887 presentations and 181 optimizer steps: 1,155 task questions with graph dropout, 866 task questions with matched graphs, 578 verified graph questions, 144 matched-graph identification examples, and 144 mismatched-graph identification examples. These are training presentations rather than 2,887 distinct videos. LoRA updates the language-model attention projections; the visual tower, audio tower, and remaining backbone parameters are frozen during optimization. Inference uses the merged task-adapted checkpoint and the joint-stage adapter.

A.3 Evaluation Alignment and Graph-Input Audit

IntentBench evaluation uses a frozen set of 2,688 questions and aligned predicted graphs from 632 videos: 2,356 Social-IQ, 133 EMER, and 199 MDPE questions. One video from the original 2,689-question manifest was excluded by a fixed task identifier because person tracking repeatedly exhausted GPU memory; the exclusion preceded prediction and scoring. Training-example construction does not read the held-out benchmark answers. Evaluation predictions are saved before a separate scoring step accesses reference labels.

An initial shuffled-graph control changed `structured_evidence_text`, while the inference formatter read `scene_graph`. Consequently, the actual graph-text input and raw output were identical for all 2,688 questions across that control and its matched run. Those pre-

dictions are excluded from the graph-correspondence analysis. The corrected formatter reads `structured_evidence_text`; a pre-inference audit verified different graph-text inputs for every aligned question. Both corrected runs yielded 2,688 parsed predictions. This input audit establishes that the model received different graph text, without itself establishing a performance gain.

A.4 Illustrative Video-to-Graph Example

Figure 5 illustrates how participant identity and a temporally localized reaction can be represented alongside a Social-IQ question. The video frames and the question's reference answer come from sample `social_iq_001983`. The P1–P3 identifiers and the links in the illustrative graph explain the intended representation. Figure 6 further illustrates the same example from the perspective of participant grounding and temporal visual evidence. In particular, the upper timeline organizes the selected video frames chronologically, while the visual grounding panel associates the natural-language referents with the corresponding visible participants. The temporal visual evidence panel highlights the relevant moments in the interaction, and the final panel presents the Social-IQ question together with its benchmark reference answer. Together, the two illustrations show how participant identity, temporally localized evidence, and question–answer information can be organized within the proposed video-to-graph representation.