CogniRoute: Learning to Route Social Evidence in Omni-Modal Models

Anonymous authors
Paper under double-blind review

Abstract

Omni-modal models can ingest video, audio, and text, but unified access to multiple modalities does not guarantee that a model uses the right evidence. This gap is especially pronounced in social video question answering, where the answer may hinge on a gesture, vocal tone, temporal cue, or mismatch between what is said and what is visually expressed. We introduce **CogniRoute**, a schema-guided Mixture-of-Experts framework for social omni reasoning. CogniRoute uses a training-only cognitive schema that factorizes each example by cross-modal relation, reasoning demand, and temporal scope, and aligns global routing signatures with this structure during supervised fine-tuning. We further introduce route-aware reinforcement learning, which jointly optimizes token generation and expert allocation using rewards for answer correctness, modality-consistent reasoning, and cognitive temporal grounding. To support training and evaluation, we construct OmniSocialBench, a diagnostic social video QA resource with 118K structured training examples, grounded reasoning traces, schema labels, temporal evidence spans, and a manually verified evaluation split. CogniRoute achieves 59.38% average accuracy on OmniSocialBench, improving over the strongest proprietary baseline by 14.63 percentage points and the strongest open-source omni baseline by 26.75 points, with the largest gains on questions requiring audio-visual coordination, conflict resolution, and temporally grounded social inference.

1 Introduction

Recent omni models can process video, audio, and text within a single autoregressive system, creating a common interface for multimodal perception and reasoning (Xu et al., 2025b;a; Comanici et al., 2025). However, progress in omni modeling has exposed a persistent gap between having access to multiple modalities and using the right evidence for a particular question (Park et al., 2025; Sung-Bin et al., 2024). This gap is especially pronounced in social video question answering. Social meaning is often carried by the relation between expression, action, speech content, tone, laughter, silence, and timing. A question may depend on a glance or hand motion, on what is said and how it is said, on a combination of both, or on a conflict between the two. In such settings, the central challenge extends beyond representing multiple modalities in a shared model to determining which evidence is reliable and where the decisive cue occurs in time.

Existing video QA and audio-visual benchmarks usually specify what the correct answer is (Zadeh et al., 2019; Liu et al., 2024; Li et al., 2024c; Wu et al., 2024), but not why a particular modality should be trusted, whether audio and visual evidence are complementary or conflicting, what reasoning operation is required, or where the relevant evidence occurs in time. As a result, a model may produce the correct answer while relying on the wrong modality, attending to an irrelevant temporal segment, or exploiting dataset priors. This limitation is difficult to diagnose from answer accuracy alone and is particularly harmful for social reasoning, where plausible but incorrect interpretations often arise from missing a subtle cue.

Sparse Mixture-of-Experts architectures offer a natural mechanism for conditional computation (Riquelme et al., 2021; Chi et al., 2022; Pan et al., 2025), but current MoE routing does not solve this problem by itself. In standard multimodal MoE models, routing is typically driven by local token representations and optimized indirectly through next-token prediction and load balancing. Recent MoE and routing methods improve efficiency (Lin et al., 2026; Wu et al., 2025), modality specialization

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.

(Li et al., 2025c; Xia et al., 2025), or adaptive expert usage (Hwang et al., 2023; Yan et al., 2025), but they do not provide explicit sample-level supervision that tells the router what kind of multimodal reasoning the input requires. As a result, the router is not explicitly trained to produce routing patterns that distinguish whether a question requires visual evidence, speech content or prosody, audio-visual coordination, cross-modal conflict resolution, temporal linking, or higher-level social inference.

In this work, we use social video question answering as a strong testbed for omnimodal understanding and introduce **CogniRoute**, a schema-guided MoE training framework for joint visual and audio reasoning. The key idea is to expose the latent evidence structure of each training example through a compact, training-only *Cognitive Schema* that summarizes three sample-level properties: the relation between audio and visual evidence, the reasoning operation required by the question, and the temporal scope of the relevant cue. The schema is used only during training to shape routing behavior; at inference, CogniRoute receives only the original video, audio, and question, and generates the answer directly. During supervised fine-tuning, our *Schema-Aligned Predictive Routing* objective encourages routing patterns to become predictive of the sample's evidence structure, including whether the question depends on visual evidence, audio evidence, audio-visual coordination, cross-modal conflict, temporal reasoning, or higher-level social inference.

Answer likelihood alone provides limited guidance for optimizing expert allocation, because the same answer can be produced through different internal routes, including routes that ignore the relevant modality or temporal region. CogniRoute therefore introduces *Route-Aware MoE Reinforcement Learning*, which optimizes both token generation and expert routing using rollout-level rewards that combine answer correctness with modality-consistent reasoning and cognitive temporal grounding, so that successful generations are those that not only predict the right answer, but also ground the decision in the appropriate modalities and time span.

Because standard QA datasets do not indicate which visual or audio cues support each answer, how those cues relate, or when they occur, we construct **OMNISOCIALBENCH**, a social video QA resource with a structured training set and a manually verified benchmark split, enabling controlled evaluation of whether models use the appropriate modality, resolve audio-visual conflict, and ground their answers in the correct temporal region. The training set augments original video-question-answer pairs with evidence summaries, schema labels, grounded reasoning, routing-related annotations, and temporal evidence spans, providing the supervision required to learn evidence-aware routing. The held-out evaluation set is manually verified and supports diagnostic analysis across social inference types, including mental-state inference, pragmatic meaning, action goals, and social norms. To reduce annotation drift, both splits are built with a staged pipeline that first extracts observation-level visual and audio evidence, then assigns schema labels and generates reasoning from this shared evidence while preserving the original answers.

On OMNISOCIALBENCH, CogniRoute obtains 59.38% average accuracy, improving over the strongest proprietary baseline by 15.33 percentage points and over the strongest open-source omni baseline by 26.77 points. The improvements are especially large on socially grounded categories such as action goal and social-norm inference, where the model must coordinate visual and audio cues, resolve ambiguity or conflict, and ground the interpretation in time. Our contributions are:

- We introduce **CogniRoute**, a schema-guided MoE training framework for omni social video reasoning that aligns expert routing with the evidence structure of each question, including modality relation, reasoning demand, and temporal scope. This enables the model to distinguish when to rely on visual cues, audio cues, audio-visual coordination, cross-modal conflict resolution, or temporally grounded social inference.
- We propose Route-Aware MoE Reinforcement Learning for multimodal MoE reasoning, which jointly optimizes answer generation and expert allocation with rollout-level rewards for answer correctness, modality-consistent reasoning, and cognitive temporal grounding. This encourages the model to produce correct answers through the appropriate modalities and temporal evidence.
- We introduce **OMNISOCIALBENCH**, a social video QA resource designed to evaluate evidence-aware omni reasoning. Unlike standard QA benchmarks that primarily supervise answers, OMNISOCIALBENCH provides structured training supervision over modality evidence, audio-visual relations, reasoning demands, and temporal evidence spans, together with a manually verified evaluation split for fine-grained diagnostic analysis.

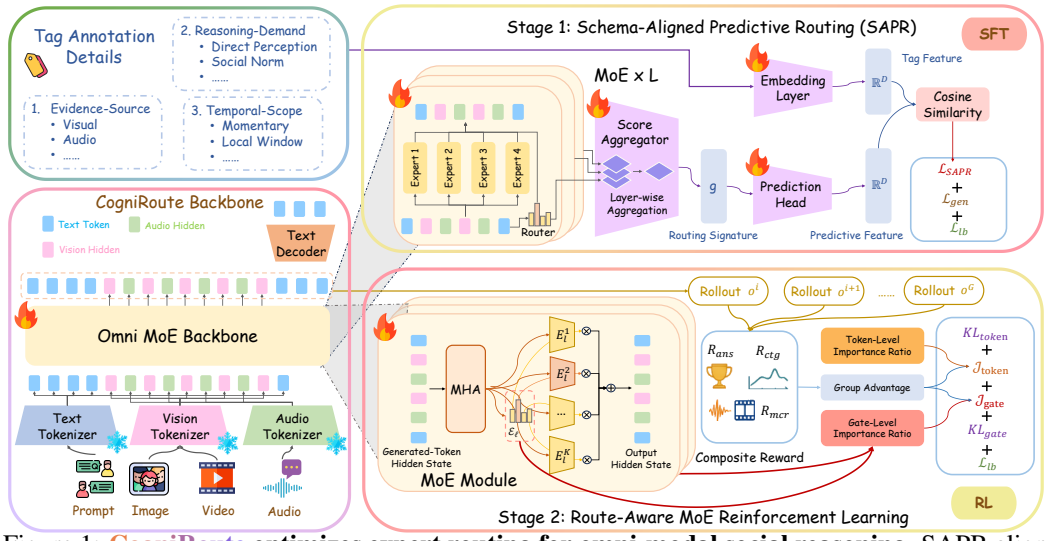

Figure 1: **CogniRoute optimizes expert routing for omni-modal social reasoning.** SAPR aligns routing signatures with semantic annotations of *evidence source* (which modalities provide the relevant cues and how those cues relate), *reasoning demand* (the type of inference needed to interpret the evidence), and *temporal scope* (the minimum temporal context needed to answer the question). RMRL then jointly optimizes answer generation and expert allocation using rewards for answer correctness, modality-consistent reasoning, and temporal grounding.

2 METHOD

MoE architectures can selectively activate experts for different inputs, offering a mechanism for adapting computation for social reasoning task with various evidence requirements across modality, temporal scope, etc. Therefore, we design CogniRoute through two-stage training (Figure 1): First, Schema-Aligned Predictive Routing (SAPR) uses cognitive schema annotation during supervised fine-tuning to encourage routing patterns that reflect evidence sources, reasoning demands, and temporal scope. Then, Route-Aware MoE Reinforcement Learning (RMRL) jointly optimizes answer generation and expert selection with rewards for answer correctness, modality-consistent reasoning, and temporal grounding. Because existing social video QA datasets generally lack the explicit evidence annotations needed to support this training, we curate OMNISOCIALBENCH, including a training set with structured evidences and a manually verified benchmark focused on social scenarios with diverse evidence requirements.

2.1 OMNISOCIALBENCH DATA CURATION

We curate OMNISOCIALBENCH with 118K training set and a 0.8k diagnostic benchmark for the social understanding task. Figure 2 outlines the curation pipeline, where Gemini 3.1 Pro is applied as the core model used for annotation and reasoning generation.

Data Source. We collect examples from OmniMMSI-YouTube (Li et al., 2026), SocialIQ (Zadeh et al., 2019), MTMEUR (Hu et al., 2025), VideoChat-Conversation (Li et al., 2023a), NeXTQA (Xiao et al., 2021), and LLaVA-Video (Zhang et al., 2024), covering multi-person interactions, affective and pragmatic reasoning, egocentric scenarios, and general video understanding with social interpretation.

Training Data Curation. We first extract structured evidence without answering the question, including scene context, participants, cross-modal relations, quality flags, visual cues, and audio cues, providing a shared basis for training and evaluation (Appendix A.1). For SAPR design, each question, ground-truth answer, and evidence record is annotated with cognitive schema tags for evidence source, reasoning demand, and temporal scope (Appendix A.2). Grounded CoT and answers are then generated in the format `<Reasoning>...</Reasoning><Answer>...</Answer>` and filtered to retain only samples with correct generated answers (Appendix A.3).

Benchmark Annotation. For the evaluation split, 800 examples are additionally annotated along four social reasoning dimensions (Appendix A.4), providing fine-grained social scenarios with different

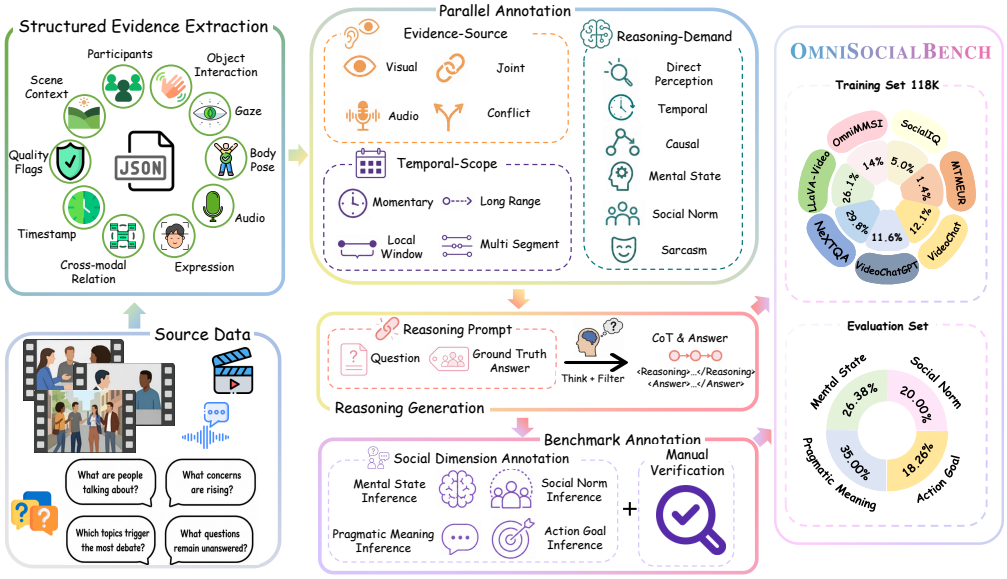

Figure 2: **OMNISOCIALBENCH data curation pipeline.** Structured evidence extraction records observable audio-visual cues. Subsequent stages assign semantic schema tags, generate grounded reasoning, and identify supporting temporal spans. The evaluation split undergoes final manual verification.

social understanding requirements. All benchmark samples are manually verified for quality, with verification details and examples provided in Appendices A.5 and B.

2.2 SCHEMA-ALIGNED PREDICTIVE ROUTING (SAPR)

Standard MoE routers select experts from local token representations and receive supervision primarily through answer generation and load balancing. These objectives do not explicitly specify whether a question requires visual cues, vocal information, cross-modal comparison, or reasoning over events separated in time. SAPR introduces sample-level supervision that encourages routing signatures to reflect these evidence-level requirements.

Cognitive Schema Tag as Semantic Annotation. We construct three types of cognitive schema tags assigned based on the question, the original ground-truth answer, and the structured evidence record. These tags characterize the evidence required to support the answer. We define 3 types of such schema: Evidence source ($\mathcal{C}_{\text{src}}$) identifies how the supportive evidence is distributed across visual and audio modalities and how these modalities relate. Reasoning demand ($\mathcal{C}_{\text{reason}}$) identifies the principal inference operation required to interpret this evidence and support the answer. Temporal scope ($\mathcal{C}_{\text{time}}$) identifies the smallest temporal context sufficient to answer the question, capturing the extent and continuity of the supporting evidence over time. Together, these schema tags provide a compact semantic description of which evidence matters, how it must be interpreted, and how far the reasoning must extend in time. They supervise global routing behavior without prescribing a fixed assignment between individual experts and semantic categories. Tag annotation details and protocols are provided in Appendix A.2.

Predictive Alignment between Routing Signature and Tag Embedding. MoE models use input-dependent routing to selectively activate experts, providing a natural mechanism for adapting computation to different reasoning requirements. However, routing learned through answer generation alone may not reflect the evidence structure needed for social understanding. We therefore design a mechanism to align global routing signatures with semantic tag embeddings, encouraging expert allocation to reflect each question's evidence source, reasoning demand, and temporal scope.

For input token i at MoE layer l, the router ρ_θ^l maps its hidden state $\mathbf{h}_{i,l}$ to a distribution over N experts. Let $\mathcal{I}_v$, $\mathcal{I}_a$, and $\mathcal{I}_q$ denote the input-token indices for video, audio, and question tokens. We construct global routing signature by pooling routing distributions separately for each modality, then projecting them into a shared space with concatenation across modality and averaging across layers:

$$\mathbf{g} = \frac{1}{L}\sum_{l=1}^{L} P_l \ \text{Concat}_{m\in\{v,a,q\}}\left(\frac{1}{|\mathcal{I}_m|}\sum_{i\in\mathcal{I}_m}\rho_\theta^l(\mathbf{h}_{i,l})\right) \tag{1}$$

where L is the number of MoE layers and $P_l \in \mathbb{R}^{D\times 3N}$ is a learnable layer-specific projection that accounts for distinct expert identities across layers. Separate modality pooling prevents longer audio or video streams from dominating the signature. The signature only uses input tokens, excluding generated reasoning and answer tokens. It encourages the routing signature to capture evidence requirements across different modality without confounder revealed by the textual reasoning chain.

Let $\mathcal{V}_{\text{tag}}$ be the schema vocabulary and $W_{\text{tag}} \in \mathbb{R}^{|\mathcal{V}_{\text{tag}}|\times D}$ a learnable embedding matrix, we represent schema by summing tag embeddings with $\mathbf{s} = \sum_{c\in\mathcal{C}} W_{\text{tag}}[c], \mathcal{C} \in \{\mathcal{C}_{\text{src}}, \mathcal{C}_{\text{reason}}, \mathcal{C}_{\text{time}}\}$. With a prediction head H_{pred} maps the global routing signature $\mathbf{g}$ into the schema embedding space, we minimize the cosine distance between the predicted and target embeddings as part of SFT loss:

$$\mathcal{L}_{\text{SFT}} = \mathcal{L}_{\text{nll}} + \lambda_{\text{lb}}\mathcal{L}_{\text{lb}} + \lambda_{\text{SAPR}}\mathcal{L}_{\text{SAPR}}, \quad \mathcal{L}_{\text{SAPR}} = \left(1 - \frac{H_{\text{pred}}(\mathbf{g})^\top \mathbf{s}}{\|H_{\text{pred}}(\mathbf{g})\|_2 \|\mathbf{s}\|_2}\right) \tag{2}$$

where $\mathcal{L}_{\text{nll}}$ is the autoregressive negative log-likelihood of the reasoning CoT and answer, and $\mathcal{L}_{\text{lb}}$ is the standard MoE load-balancing loss. The coefficients λ_{lb} and λ_{SAPR} control the contributions of load balancing and schema alignment.

In inference stage, schema tags are neither appended to the model input nor required as generated outputs because they are auxiliary training annotations. The model only receives ordinary visual, audio, textual input and uses its learned router, while the tag embedding matrix, routing-summary projections, and prediction head are only available during training.

2.3 Route-Aware MoE Reinforcement Learning (RMRL)

SAPR encourages routing to reflect the evidence support in the input, but supervised generation does not directly distinguish successful expert allocations from routes that produce a correct answer using inappropriate evidence. Based on this target, we propose an RL method called RMRL, including routing guidance design along with multi-modal specific reward design to customize GRPO. Therefore, RMRL treats expert selection as an explicit policy action and uses rollout-level rewards to optimize both generated tokens and routing decisions.

Routing Gate Aware GRPO in MoE. In classical GRPO setting for MoE model, the old policy generates a group of G rollouts $\{(o^{(i)}, \mathcal{E}^{(i)})\}_{i=1}^{G}$ for each input x, where $o^{(i)}$ is the output sequence and $\mathcal{E}^{(i)}$ records expert selected over generated tokens and MoE layers. Each rollout receives a reward $\mathcal{R}^{(i)} = \mathcal{R}(x, o^{(i)})$. Then we can get the token-level GRPO loss function:

$$\begin{aligned} \mathcal{J}_{\text{token}} &= -\frac{1}{G}\sum_{i=1}^{G}\sum_{t} \min\left(u_t^{(i)}(\theta)\hat{A}^{(i)},\ \text{clip}(u_t^{(i)}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}^{(i)}\right) \\ u_t^{(i)}(\theta) &= \frac{\pi_\theta\left(o_t^{(i)} \mid x, o_{<t}^{(i)}, \mathcal{E}_{\le t}^{(i)}\right)}{\pi_{\theta_{\text{old}}}\left(o_t^{(i)} \mid x, o_{<t}^{(i)}, \mathcal{E}_{\le t}^{(i)}\right)}, \quad \hat{A}^{(i)} = \frac{\mathcal{R}^{(i)} - \text{mean}_j\, \mathcal{R}^{(j)}}{\text{std}_j\, \mathcal{R}^{(j)} + \epsilon_A} \end{aligned} \tag{3}$$

To apply the same feedback to expert allocation, let $\mathcal{E}_{t,l}^{(i)}$ be the selected expert set for token t at layer l. We score this set using the router distribution before top-k selection:

$$\ell_{\theta,t,l}^{(i)} = \prod_{e\in\mathcal{E}_{t,l}^{(i)}} \rho_\theta^l\left(\mathbf{h}_{t,l}^{(i)}(\theta)\right)[e] \tag{4}$$

This order-agnostic product is a likelihood surrogate; it is not the exact probability of sampling the set without replacement from the restricted candidate pool. Therefore, the gate branch has a surrogate importance ratio $v_{t,l}^{(i)}(\theta)$ and then we can apply gate-level GRPO on MoE:

$$\mathcal{J}_{\text{gate}} = -\frac{1}{G}\sum_{i=1}^{G}\sum_{t,l} \min\left(v_{t,l}^{(i)}(\theta)\hat{A}^{(i)},\ \text{clip}(v_{t,l}^{(i)}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}^{(i)}\right), \quad v_{t,l}^{(i)}(\theta) = \frac{\ell_{\theta,t,l}^{(i)}}{\ell_{\theta_{\text{old}},t,l}^{(i)}} \tag{5}$$

During policy updates, the recorded expert selections are based on old policy model. The current and old policies model use the same selected experts, with hidden states are computed under their respective parameters. Both branches share the rollout-level advantage, allowing rewards to reinforce the generated sequence and the expert allocations that produced it.

The full reinforcement learning objective is

$$\mathcal{J}_{\mathrm{RL}} = \mathcal{J}_{\mathrm{token}} + \eta \mathcal{J}_{\mathrm{gate}} + \beta_{\mathrm{tok}} \mathrm{KL}_{\mathrm{tok}} + \beta_{\mathrm{gate}} \mathrm{KL}_{\mathrm{gate}} + \lambda_{\mathrm{lb}} \mathcal{L}_{\mathrm{lb}} \tag{6}$$

which is minimized over θ. The KL terms regularize the language policy and router against a frozen reference model, and the load-balancing term discourages concentration on a small subset of experts.

As for reward design, we combine three complementary rewards:

$$\mathcal{R}(x, o) = \lambda_{\mathrm{ans}} \mathcal{R}_{\mathrm{ans}}(x, o) + \lambda_{\mathrm{mcr}} \mathcal{R}_{\mathrm{mcr}}(x, o) + \lambda_{\mathrm{ctg}} \mathcal{R}_{\mathrm{ctg}}(x, o) \tag{7}$$

They evaluate the final answer correctness, the modality evidence consistency in the reasoning, and the temporal focus of the model's reasoning state, where weights of all three reward sums to 1. More details about the reward design are provided in Appendix C.

Answer Correctness Reward. We extract the final answer y and compare it with the ground-truth answer $y^\star$ under after normalization, which gives reward $\mathcal{R}_{\mathrm{ans}}(x, o) = \mathbb{I}\left[(y = y^\star)\right]$.

Modality-Consistent Reward. To encourage reasoning grounded in the appropriate modality evidence, we use a LLM judge to assess the reasoning trace against structured annotations of the required modalities and their relation. The judge assigns $\mathcal{R}_{\mathrm{mcr}}(x, o) \in \{0, 0.25, 0.5, 0.75, 1.0\}$, with higher scores indicating stronger consistency between the reasoning and the annotated evidence requirements. Our prompt is designed to assign high score to reasoning which cites concrete visual or audio observations and correctly interprets their supportive or conflicting relation.

Cognitive Temporal Grounding Reward. Correct answers and plausible reasoning do not ensure that the model attends to the temporal evidence supporting its conclusion. We introduce Cognitive Temporal Grounding (CTG), which rewards alignment between the temporal attention of the final reasoning state and annotated evidence regions. Attention serves as a proxy for temporal focus, requiring no separate grounding model.

We aggregate causal attention from a short reasoning suffix $\mathcal{Q}_{\mathrm{suffix}}$, including the closing reasoning token, to audio-visual input tokens. Averaging across deep layers $\mathcal{L}_{\mathrm{ctg}}$ and attention heads reduces dependence on individual queries or heads. Tokens at the same position on the backbone's shared temporal grid are pooled to obtain

$$\alpha(\tau) = \frac{1}{Z_\alpha} \sum_{\substack{j \in \mathcal{I}_{\mathrm{av}} \\ \tau_j = \tau}} \frac{1}{|\mathcal{Q}_{\mathrm{suffix}}||\mathcal{L}_{\mathrm{ctg}}|} \sum_{u \in \mathcal{Q}_{\mathrm{suffix}}} \sum_{l \in \mathcal{L}_{\mathrm{ctg}}} \frac{1}{H_l} \sum_{h=1}^{H_l} A_{u \to j}^{l,h} \tag{8}$$

where $A_{u \to j}^{l,h}$ is the attention weight from query u to input token j, H_l is the number of heads, and Z_α normalizes the distribution over time. $\mathcal{I}_{\mathrm{av}}$ contains only audio-visual tokens; question tokens are excluded because they have no physical timestamp.

To model uncertainty nearby boundary, we convert the evidence spans $\mathcal{S}_{\mathrm{gt}}$ to a smooth distribution:

$$P_{\mathrm{gt}}(\tau) = \frac{1}{Z_{\mathrm{gt}}} \max_{[s,e] \in \mathcal{S}_{\mathrm{gt}}} \exp\left(-\frac{d(\tau, [s, e])^2}{2\sigma^2}\right) \tag{9}$$

where $d(\tau, [s, e])$ is zero inside the interval and otherwise equals the distance to its nearest boundary. Z_{gt} is a normalization constant, and σ controls temporal tolerance.

We apply additive smoothing and renormalization, $\tilde{\alpha}(\tau) \propto \alpha(\tau) + \epsilon_{\mathrm{ctg}}$, to ensure finite divergence. The reward is then

$$\mathcal{R}_{\mathrm{ctg}}(x, o) = \begin{cases} \exp\left(-\gamma D_{\mathrm{KL}}\left(P_{\mathrm{gt}} \| \tilde{\alpha}\right)\right), & \text{if a valid reasoning suffix } \mathcal{Q}_{\mathrm{suffix}} \text{ exists} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where $\gamma > 0$ controls sensitivity to temporal mismatch. Higher rewards indicate closer alignment with the annotated evidence distribution, complementing answer correctness and modality consistency with feedback on temporal focus.

3 Experiments

We train **CogniRoute** to assess the effectiveness of the proposed training framework for cognitive social understanding and its impact on general audio-visual, audio-only, and video-only capabilities. Our experiments compare the model with competing baselines and standard in-domain fine-tuning,

Table 1: **Model performance in cognitive social understanding task across models and social dimensions.** : VLMs without omni inputs; : open-source omni models; : VLMs with omni inputs; : finetuned open-source omni models. Accuracy is the metric across models for evaluation.

	Mental State	Pragmatic Meaning	Action Goal	Social Norm	Avg
GPT 5.4 (Singh et al., 2025)	18.96	18.57	20.13	14.37	18.12
Qwen 3.6 Plus (Yang et al., 2025a)	14.69	19.29	26.17	15.00	18.50
GLM 5V Turbo (Hong et al., 2025b)	13.27	17.50	18.12	14.37	15.88
Baichuan-Omni-1.5 (Li et al., 2025b)	12.80	12.86	12.75	14.37	13.12
OmniVinci (Ye et al., 2025)	26.07	21.07	20.81	18.75	21.88
Qwen3 Omni (Xu et al., 2025b)	38.86	31.43	32.89	26.25	32.63
MiMo V2 Omni (Xiao et al., 2026)	31.28	30.36	32.89	30.63	31.13
Qwen 3.5 Omni (Team, 2026)	36.02	28.57	30.87	31.25	31.50
Gemini 3.1 Flash (Comanici et al., 2025)	43.13	41.79	47.65	27.50	40.38
Gemini 3.1 Pro (Comanici et al., 2025)	**60.19**	43.93	36.91	33.13	44.75
OmniVinci + Standard SFT	46.92	43.21	47.65	39.38	44.25
Qwen3 Omni + Standard SFT	40.28	38.93	44.97	37.50	40.13
Qwen3 Omni + SAPR	51.66	49.29	53.69	46.25	50.13
Qwen3 Omni + SAPR + RMRL (CogniRoute)	**60.19**	**55.71**	**66.44**	**58.13**	**59.38**

followed by controlled ablations of schema-aligned routing, joint token and gate optimization, and the individual reward components.

Dataset. We train CogniRoute on the 118K structured training data. We allocate 10.6K examples to training the routing-related prediction components, 96.8K for supervised fine-tuning (SFT), and 10.6K for reinforcement learning (RL).

Benchmarks. We evaluate CogniRoute on OMNISOCIALBENCH test set. Comparisons include models with different modality capabilities, as indicated in Table 1. We report accuracy across dataset and models. We also evaluate on 10 general public benchmarks. Audio-visual benchmarks comprise OMNIBENCH (Li et al., 2024d), DAILY-OMNI (Zhou et al., 2025b), AV-SPEAKERBENCH (Nguyen et al., 2025), AV-ODYSSEY (Gong et al., 2024), and OMNIVIDEOBENCH (Li et al., 2025a). We additionally evaluate audio-only understanding on MMAU (Sakshi et al., 2024) and video-only understanding on VIDEO-MME-V2 (Fu et al., 2026), MMVU (Zhao et al., 2025c), and MVBENCH (Li et al., 2024c). These benchmarks are not used for training.

Training Details. We use Qwen3-Omni-30B-A3B (Xu et al., 2025b) as the backbone, with 128 experts and 8 active experts per routed token. SFT uses low-rank adaptation with rank 64 and a learning rate of 1×10^{-4}. Training runs on eight NVIDIA H200 GPUs for approximately one day, with a per-device batch size of 4 and one gradient accumulation step. We use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay of 0.1, together with cosine learning rate scheduling, warmup, bfloat16 precision, and gradient clipping. Additional implementation details appear in Appendix D.

3.1 MAIN RESULTS

Performance on Cognitive Social Understanding Tasks. We evaluate our method and various baselines in cognitive social understanding ability of model via OMNISOCIALBENCH. In Table 1, CogniRoute achieves the strongest overall performance across all baselines, especially powerful proprietary omni models. Its consistent strength across multiple social dimensions demonstrates the effectiveness of our approach for general cognitive social understanding. Compared with fine-tuned open-source baselines using standard SFT, our method delivers stronger performance. This comparison highlights the benefit of our method design, showing the effectiveness of our evidence-guided strategy like SAPR. We further observe a stage-wise improvement: SAPR provides a substantial boost while the subsequent RMRL stage further improves performance across all evaluated dimensions. This progressive gain shows that the two stages contribute incremental benefits to the final model. Qualitative comparisons with Gemini 3.1 Pro are shown in Figure 3, with more examples provided in Appendix E.

Generalizability across Public Audio-Visual Benchmarks. Table 2 shows that CogniRoute improves over its backbone on most public benchmarks, including all video-only benchmarks and nearly all audio-visual benchmarks. Performance remains similar on WORLDSENSE and MMAU. These results suggest that improved social understanding is achieved while broadly preserving general capabilities, without signs of catastrophic forgetting.

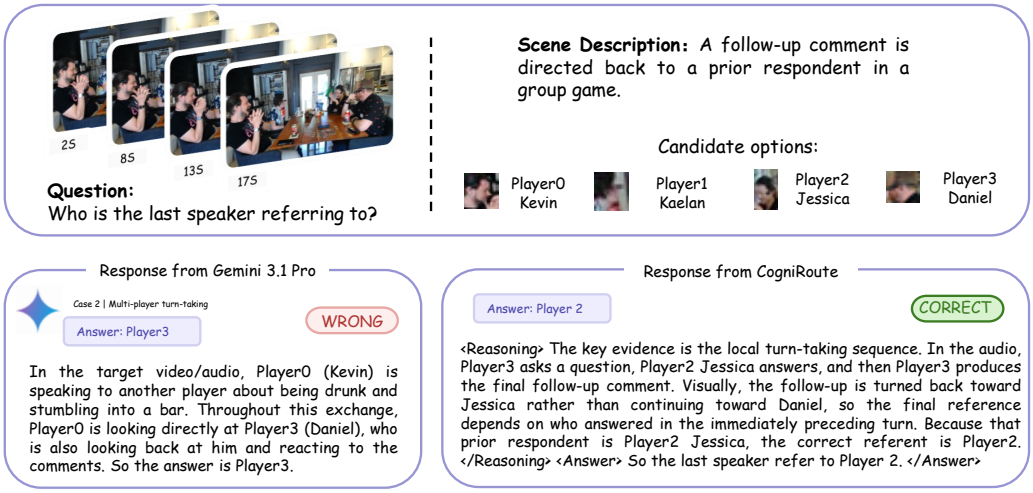

Figure 3: **Qualitative comparison on multi-party turn-taking.** Gemini 3.1 Pro over-relies on the visually salient gaze direction toward Player3 (Daniel) and predicts Player3 as the referent. CogniRoute follows the local dialogue sequence: Player3 asks a question, Player2 (Jessica) answers, and the subsequent follow-up is directed back to the prior respondent. This turn-taking evidence identifies Player2 as the correct referent.

Table 2: **Model performance (accuracy) in public benchmarks.** We compare CogniRoute with its Qwen3-Omni backbone on ten audio-visual, audio-only, and video-only benchmarks.

	Audio-Visual						Audio	Video	
	OMNIBENCH	DAILY-OMNI	AV-SPEAKERBENCH	AV-ODYSSEY	OMNIVIDEOBENCH	MMAU	VIDEO-MME-V2	MMVU	MVBENCH
Qwen3 Omni 30B	55.5	73.6	54.1	34.7	38.4	**75.4**	36.5	59.8	69.5
CogniRoute **(Ours)**	**56.7**	**74.3**	**56.2**	**36.8**	**38.9**	73.2	**36.8**	**60.7**	**70.2**

Table 3: **Ablation of SAPR in SFT stage.** Full SAPR uses correct schema tags and layer-specific projections, comparing with 4 SFT variants: 1) Standard SFT, 2) SAPR with shuffled tags (shuffle the item-tag mapping while maintaining proportion of each tag category), 3) SAPR with random tags (assign a random tag from uniform random distribution), 4) SAPR with shared projection across layers (correct tag, but use same project for each layer).

Model Variant	Mental State	Pragmatic Meaning	Action Goal	Social Norm	Avg.
Standard SFT (w/o SAPR)	40.28	38.93	44.97	37.50	40.13
SAPR with shuffled tags	41.71	39.64	46.31	39.38	41.38
SAPR with random tags	40.76	38.93	45.64	38.75	40.63
SAPR with shared projection across layers	45.02	43.21	48.99	41.88	44.50
Full SAPR	**51.66**	**49.29**	**53.69**	**46.25**	**50.13**

3.2 ABLATION STUDY

Effectiveness of SAPR in SFT stage. As Table 3, full SAPR consistently outperforms standard SFT across all other settings. Shuffled and random tags provide only small improvements, indicating that SAPR benefits primarily from meaningful input, which means cognitive schema correspondence rather than the auxiliary objective alone. Sharing the projection across layers also reduces performance, supporting the use of layer-specific mappings to account for distinct expert identities when aligning routing signatures with schema representations.

Effectiveness of Routing Gate Aware GRPO in RL stage. We compare token-only, gate-only, and joint optimization in Table 4. All variants start from the same full-SAPR checkpoint and use the same reward formulation. Token-only RL improves over the SAPR checkpoint, while gate-only optimization yields a smaller gain. Full RMRL outperforms both variants across all four categories, showing that explicit routing optimization provides benefits beyond token-level reinforcement learning. These results support jointly optimizing generation and expert allocation to improve cognitive social understanding.

Reward Components Ablation. Table 5 examines the contribution of each reward by individually removing one of them from the full RL reward. Removing any component reduces accuracy across all four categories, supporting the contribution of each reward within the combined objective. The

Table 4: **Ablation of optimization branch level in RMRL.** The SAPR checkpoint is included as the pre-RL reference. Token-level and gate-level branch optimization is applied separately and simultaneously to verify effectiveness of routing gate aware GRPO method.

Model Variant	Mental State	Pragmatic Meaning	Action Goal	Social Norm	Avg.
SAPR	51.66	49.29	53.69	46.25	50.13
SAPR + Token branch	57.36	52.86	62.42	54.38	56.13
SAPR + Gate branch	53.55	50.35	55.70	48.75	51.88
SAPR + both branches (RMRL)	**60.19**	**55.71**	**66.44**	**58.13**	**59.38**

Table 5: **Ablation of RMRL reward components.** We individually remove the answer correctness ($\mathcal{R}_{ans}$), modality-consistent reasoning ($\mathcal{R}_{mcr}$), and cognitive temporal grounding ($\mathcal{R}_{ctg}$) rewards. SFT denotes the SAPR checkpoint before RL.

Model Variant	Mental State	Pragmatic Meaning	Action Goal	Social Norm	Avg.
SFT	51.66	49.29	53.69	46.25	50.13
SFT + RL w/o $\mathcal{R}_{ctg}$	54.98	51.43	57.72	51.25	53.50
SFT + RL w/o $\mathcal{R}_{mcr}$	56.40	53.21	60.40	53.13	55.38
SFT + RL w/o $\mathcal{R}_{ans}$	57.82	54.64	62.42	55.63	57.13
CogniRoute (SFT + Full RL)	**60.19**	**55.71**	**66.44**	**58.13**	**59.38**

largest reduction occurs without temporal grounding, followed by modality consistency, indicating that feedback on when and from which modality to obtain evidence is particularly valuable in this setting. Removing answer correctness also lowers performance, showing that evidence-oriented rewards benefit from direct supervision of the final answer. The full reward combination performs best throughout, supporting the joint use of answer-level and evidence-based feedback.

More Ablation Studies. Appendix F provides further analyses. Appendix F.1 shows that the gains of our method transfer to another backbone, OmniVinci. Appendix F.2 shows that SAPR learns well-separated schema tag embeddings that are predictable from routing signatures. Appendix F.3 shows that full RMRL yields more schema-aware, balanced, and modality-sensitive expert routing. Appendix F.4 visualizes that routing signatures form clearer schema-aligned clusters after SAPR. Appendix F.5 shows that the smooth CTG reward gives the most precise temporal grounding. Appendix F.6 and Appendix F.7 show that CogniRoute causally relies on the annotated temporal evidence and the question-relevant modality, respectively. We further deploy CogniRoute on VR smart glasses for real-world social assistance in Appendix G.

4 Related Work

Multimodal large language models have expanded from image–text understanding (Bai et al., 2025b;a; Lin et al., 2024; Liu et al., 2023; Chen et al., 2024b; Li et al., 2024a) to joint processing of text, vision, and audio (Hurst et al., 2024; Fu et al., 2025). Systems such as Qwen-Omni (Xu et al., 2025a;b) and Gemini (Team et al., 2023; 2024; Comanici et al., 2025) support audio-visual understanding and real-time interaction. However, processing multiple modalities does not ensure that a model uses the evidence relevant to each question. Existing omni-modal objectives (Ye et al., 2025; Yu et al., 2024) supervise generated responses rather than guiding how computation is allocated to the evidence and reasoning each question requires. CogniRoute addresses this gap by using sample-level schema supervision to guide MoE routing to reflect which modalitiesnprovide the supporting cues, how those cues must be interpreted, and how much temporal context is needed. Appendix H provides a broader review.

5 Conclusion

We introduce CogniRoute, a schema-guided MoE framework for omni-modal social reasoning that aligns expert routing with the evidence structure of each question. CogniRoute improves omni-social understanding accuracy while encouraging more evidence-aware computation across modalities and time. Across our OmniSocialBench benchmark and public audio-visual tasks, CogniRoute shows consistent gains over strong omni baselines, especially on questions requiring cross-modal coordination, conflict resolution, and temporally grounded social inference. Results show that evidence-structured routing is an effective training signal for improving social omni-modal reasoning.

ETHICS STATEMENT

This work aims to improve omni-modal social video understanding by encouraging models to use appropriate visual, audio, and temporal evidence rather than relying only on final-answer supervision. The proposed dataset annotations and routing-aware training framework may benefit applications such as assistive video understanding, educational analysis, human-computer interaction, and socially aware dialogue systems. CogniRoute and OmniSocialBench are intended to support research on evidence-grounded social video understanding. Potential applications include assistive video understanding and human–computer interaction. Because social interpretations depend on context and may vary across individuals and cultures, model outputs should be treated as uncertain interpretations. Deployment would require appropriate consent, privacy protections, and evaluation of performance across demographic and cultural groups. The benchmark is not designed for high-stakes decisions about individuals.

AI USE STATEMENT

We used generative AI tools to assist with data synthesis and writing refinement. Specifically, these tools were used to help synthesize annotations while engaging human reviewers to verify data generation. Meantime, we also use AI tools to improve expression, rectify grammar, and polish wording. All AI-assisted content was carefully reviewed, verified, and revised by the authors before being included in the paper.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the SAPR and RMRL training objectives and the evaluation setup in the main text. Appendix A documents the data sources, annotation procedures, prompt templates, consistency filtering, and human verification used to construct OMNISOCIALBENCH. Appendix C provides additional reward details, including the prompt and scoring rubric used by the modality-consistent reasoning judge. Appendix D reports the model configuration, optimization settings, training hardware, and expert-selection procedures used during training and evaluation. CogniRoute model checkpoints and the OMNISOCIALBENCH benchmark will be open-sourced.

REFERENCES

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

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, 2025b.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in neural information processing systems*, 35:32897–32912, 2022.

Meng Cao, Haoze Zhao, Can Zhang, Xiaojun Chang, Ian Reid, and Xiaodan Liang. Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning, 2026.

Junyi Chen, Longteng Guo, Jia Sun, Shuai Shao, Zehuan Yuan, Liang Lin, and Dongyu Zhang. Eve: Efficient vision-language pre-training with masked prediction and modality-aware moe. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1110–1119, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, et al. On the representation collapse of sparse mixture of experts. *Advances in Neural Information Processing Systems*, 35:34600–34613, 2022.

Xu Chu, Xinrong Chen, Guanyu Wang, Zhijie Tan, Kui Huang, Wenyu Lv, Tong Mo, and Weiping Li. Qwen look again: Guiding vision-language reasoning models to re-attention visual information, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Chengqi Duan, Rongyao Fang, Yuqing Wang, Kun Wang, Linjiang Huang, Xingyu Zeng, Hongsheng Li, and Xihui Liu. Got-r1: Unleashing reasoning capability of mllm for visual generation with reinforcement learning, 2025.

David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images, 2025.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025.

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

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, et al. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? *arXiv preprint arXiv:2412.02611*, 2024.

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

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025a.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025b.

Jinpeng Hu, Hongchang Shi, Chongyuan Dai, Zhuo Li, Peipei Song, and Meng Wang. Beyond emotion recognition: A multi-turn multimodal emotion understanding and reasoning benchmark. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5814–5823, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Changho Hwang, Wei Cui, Yifan Xiong, Ziyue Yang, Ze Liu, Han Hu, Zilong Wang, Rafael Salas, Jithin Jose, Prabhat Ram, et al. Tutel: Adaptive mixture-of-experts at scale. *Proceedings of Machine Learning and Systems*, 5:269–287, 2023.

Haonan Jiang, Yuji Wang, Yongjie Zhu, Xin Lu, Wenyu Qin, Meng Wang, Pengfei Wan, and Yansong Tang. Embed-rl: Reinforcement learning for reasoning-driven multimodal embeddings. *arXiv preprint arXiv:2602.13823*, 2026.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Wentao Wang, Zhenghao Song, Dingling Zhang, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. *arXiv preprint arXiv:2510.10689*, 2025a.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, et al. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024b.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023a.

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, 2024c.

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

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025b.

Yizhi Li, Yinghao Ma, Ge Zhang, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*, 2024d.

Yunshui Li, Binyuan Hui, ZhiChao Yin, Min Yang, Fei Huang, and Yongbin Li. Pace: Unified multi-modal dialogue pre-training with progressive and compositional experts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13402–13416, 2023b.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025c.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2023.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. Moe-llava: Mixture of experts for large vision-language models. *IEEE Transactions on Multimedia*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Xiulong Liu, Zhikang Dong, and Peng Zhang. Tackling data bias in music-avqa: Crafting a balanced dataset for unbiased question-answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4478–4487, 2024.

Zijun Long, George Killick, Richard McCreadie, and Gerardo Aragon Camarasa. Multiway-adapater: Adapting large-scale multi-modal models for scalable image-text retrieval, 2024.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022.

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

Brandon Ong, Tej Deep Pala, Vernon Toh, William Chandra Tjhi, and Soujanya Poria. Training vision-language process reward models for test-time scaling in multimodal reasoning: Key insights and lessons learned, 2025.

Xinglin Pan, Wenxiang Lin, Lin Zhang, Shaohuai Shi, Zhenheng Tang, Rui Wang, Bo Li, and Xiaowen Chu. Fsmoe: A flexible and scalable training system for sparse mixture-of-experts models. In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*, pp. 524–539, 2025.

Jean Park, Kuk Jin Jang, Basam Alasaly, Sriharsha Mopidevi, Andrew Zolensky, Eric Eaton, Insup Lee, and Kevin Johnson. Assessing modality bias in video question answering benchmarks with multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19821–19829, 2025.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024.

Burak Satar, Hongyuan Zhu, Hanwang Zhang, and Joo Hwee Lim. Rome: Role-aware mixture-of-expert transformer for text-to-video retrieval, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model, 2025.

Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. Mome: Mixture of multimodal experts for generalist multimodal large language models. *Advances in neural information processing systems*, 37:42048–42070, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, and Yu Cheng. Openthinkimg: Learning to think with images via visual tool reinforcement learning, 2025.

Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. *arXiv preprint arXiv:2410.18325*, 2024.

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

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

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

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, and Zhaoxiang Zhang. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology, 2026.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, Lewei Lu, Haodong Duan, Yu Qiao, Jifeng Dai, and Wenhai Wang. Visualprm: An effective process reward model for multimodal reasoning, 2025a.

Yuji Wang, Wenlong Liu, Jingxuan Niu, Haoji Zhang, and Yansong Tang. Vg-refiner: Towards tool-refined referring grounded reasoning via agentic reinforcement learning, 2025b.

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.

Qiong Wu, Zhaoxi Ke, Yiyi Zhou, Xiaoshuai Sun, and Rongrong Ji. Routing experts: Learning to route dynamic experts in existing multi-modal large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Guoyang Xia, Yifeng Ding, Fengfa Li, Lei Ren, Wei Chen, Fangxiang Feng, and Xiaojie Wang. Smar: Soft modality-aware routing strategy for moe-based multimodal large language models preserving language capabilities. *arXiv preprint arXiv:2506.06406*, 2025.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9777–9786, June 2021.

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

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, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Jiaming Yan, Jianchun Liu, Hongli Xu, and Liusheng Huang. Accelerating mixture-of-expert inference with adaptive expert split mechanism. *arXiv preprint arXiv:2509.08342*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

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*, 2025b.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization, 2025c.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*, 2025.

Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. Salmonn-omni: A codec-free llm for full-duplex speech understanding and generation. *arXiv preprint arXiv:2411.18138*, 2024.

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.

Jianghangfan Zhang, Yibo Yan, Kening Zheng, Xin Zou, Song Dai, and Xuming Hu. Gm-prm: A generative multimodal process reward model for multimodal mathematical reasoning. *arXiv preprint arXiv:2508.04088*, 2025.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning. *arXiv preprint arXiv:2503.05379*, 2025a.

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

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8475–8489, 2025c.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning, 2026.

Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's "aha moment" in visual reasoning on a 2b non-sft model, 2025a.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025b.

APPENDIX

A Detailed Annotation Protocol **17**
- A.1 Structured Evidence Extraction 17
- A.2 Cognitive Schema Tags Annotation 18
- A.3 CoT Curation and Consistency Filter 19
- A.4 Social Dimension Taxonomy for Benchmark 19
- A.5 Human Review of Annotations 20
- A.6 Comparison with Public Benchmarks 20

B OMNISOCIALBENCH Examples **20**

C Reward Details **21**
- C.1 Modality-Consistent Reasoning Reward 21
- C.2 Human Validation of the Modality Verifier 22

D Additional Experimental Details **23**

E Qualitative Analysis **23**

F Additional Studies **24**
- F.1 Method Generalization Across Backbone 24
- F.2 Learnable Tag Embedding Collapse Diagnostics 25
- F.3 Routing Behavior Under Gate Optimization 26
- F.4 Routing Signature Visualization 27
- F.5 Effect of CTG Reward Design on Temporal Grounding 27
- F.6 Causal Evidence Deletion 27
- F.7 Modality Evidence Use 28

G Real-world VR Smart Glasses Deployment **28**

H Related Work **28**

A Detailed Annotation Protocol

Data Collection and Curation. The source pool is drawn from OmniMMSI-YouTube (Li et al., 2026), SocialIQ (Zadeh et al., 2019), MTMEUR (Hu et al., 2025), VideoChat-Conversation (Li et al., 2023a), NeXTQA (Xiao et al., 2021), and LLaVA-Video (Zhang et al., 2024). Together, these sources cover multi-person interaction, affect-centered scenes, egocentric videos, and general videos that still require social interpretation. A single prompt that tries to observe the clip, solve the question, assign labels, and generate reasoning at the same time is not reliable enough for dataset construction, because it often mixes observation with interpretation and gives an unstable output format. We therefore use a staged pipeline. The model first extracts structured evidence from the clip, and later prompts read the same evidence to assign task labels and generate grounded reasoning. We keep a sample only if the generated answer matches the original ground truth answer after normalization. The evaluation set follows the same pipeline and adds final manual verification.

Example Structure and Evidence Coverage. Each example preserves the original video clip, audio track, question, and ground-truth answer, and augments them with structured audio-visual evidence, schema labels, grounded reasoning, and temporal evidence spans. The evidence records observable cues such as participants, scene context, gaze, expression, body pose, gesture, object interaction, speech content, speaker changes, laughter, silence, and tone-relevant acoustic events. The reasoning trace explains the answer using this recorded evidence, while the temporal spans identify the supporting time region. These fields capture cues needed for social interpretation, including temporal ordering and mismatches between spoken and visually expressed meaning.

A.1 Structured Evidence Extraction

In the first stage, the model output JSON recording various evidence observed in video clip input with question and ground truth answer. The structured JSON has seven top level parts: `scene_context`, `participants`, `visual_nodes`, `audio_nodes`, `dialogue_nodes`, `crossmodal_nodes`, and `quality_flags`. Each node should stay at the observation level. For example, the extractor may record that a speaker looks away, pauses, and speaks with flat prosody, but it should not assign sarcasm or hidden frustration. This shared evidence source makes later labels easier to check.

Structured Evidence Extraction Prompt

System. You are given a pre-segmented video clip with audio, its original question, and its original ground truth answer. Convert the clip into structured evidence JSON. Record only observable cues. Do not answer the question. Do not infer hidden mental states.

Task. Use the video and audio to extract scene context, participants, timestamped visual events, timestamped audio events, dialogue structure, simple cross modal relations, and quality flags.

Required JSON fields. `scene_context`, `participants`, `visual_nodes`, `audio_nodes`, `dialogue_nodes`, `crossmodal_nodes`, `quality_flags`.

Field guidance. `scene_context`: place, social setting, salient objects, camera distance, notable background cues. `participants`: participant id, role cues, position, visibility, face visibility. `visual_nodes`: start time, end time, actor, event type, description. `audio_nodes`: start time, end time, speaker, transcript, prosody, non-speech audio. `dialogue_nodes`: start time, end time, speaker, addressee, dialogue act, utterance summary. `crossmodal_nodes`: start time, end time, relation type, linked visual nodes, linked audio nodes, description. `quality_flags`: low light, occlusion, off-screen speaker, weak audio, background noise, uncertainty note.

Rules. Every event should have timestamps when possible. Use short factual descriptions. If a field is absent, return an empty list or `null`. Do not output markdown. Output valid JSON only.

Figure 4: **Structured evidence extraction prompt.** Gemini-3.1-Pro first converts each clip into observation-level JSON.

A.2 COGNITIVE SCHEMA TAGS ANNOTATION

The three types of cognitive schema tags are assigned from the original question, the original ground truth answer, and the structured evidence JSON. They are sample-level labels, so the same clip may receive different labels under different questions.

Evidence Source. The field `evidence_source` records how the minimum answer evidence is distributed across modalities. VISUAL is used when the main evidence lies in visual cues such as face, gaze, posture, action, object use, or scene context, and audio adds little useful support. AUDIO is used when the main evidence lies in transcript, prosody, speaker turn, or other audio cues, while the image is weak or static. JOINT is used when the answer needs support from both modalities. CONFLICT is used when visual and audio cues point to different meanings and the answer depends on resolving the mismatch. This label gives a modality prior to the router.

Cross Modal Relation Annotation Prompt

System. You are given the original question, the original ground truth answer, and the structured evidence JSON for one sample. Assign the single best `evidence_source` label. Use the question as the main reference.

Label definitions. VISUAL: the main evidence is in `visual_nodes`, `participants`, or `scene_context`. AUDIO: the main evidence is in `audio_nodes` or speaker cues. JOINT: the answer needs support from both modalities. CONFLICT: visual and audio cues point to different meanings, and the answer comes from resolving the mismatch.

Decision rule. Choose the label that best matches the minimum evidence needed to support the original answer.

Output format. Return valid JSON only in the form `{"evidence_source": "LABEL"}`.

Figure 5: **Evidence source prompt.** Gemini-3.1-Pro assigns one modality label from the structured evidence JSON.

Reasoning-Demand. The field `reasoning_demand` records the main reasoning operation required by the question. DIRECT PERCEPTION is used when the answer follows from directly visible or audible cues. TEMPORAL is used when the answer depends on event order, duration, or relation across time. CAUSAL is used when the question asks about cause, effect, or trigger. MENTAL STATE is used for hidden emotion, belief, desire, or intention. SOCIAL NORM is used when the answer depends on role expectation, local rule, or appropriateness in context. SARCASM is used when the surface utterance differs from the intended meaning. This label guides the choice of reasoning experts.

Reasoning Demand Annotation Prompt

System. You are given the original question, the original ground truth answer, and the structured evidence JSON for one sample. Assign the single best `reasoning_demand` label. Use the question as the main reference.

Label definitions. DIRECT PERCEPTION: the answer follows from directly observable cues. TEMPORAL: the answer needs event order, duration, or relation across time. CAUSAL: the answer needs cause, effect, or trigger. MENTAL-STATE: the answer needs inference about hidden emotion, belief, desire, or intention. SOCIAL-NORM: the answer needs a rule, role expectation, or judgment of appropriateness in context. SARCASM: the intended meaning differs from the literal utterance.

Decision rule. Choose the label that best matches the main reasoning step needed to support the original answer.

Output format. Return valid JSON only in the form `{"reasoning_demand": "LABEL"}`.

Figure 6: **Reasoning demand prompt.** Gemini-3.1-Pro assigns one reasoning label from the structured evidence JSON.

Temporal-Scope. The field `temporal_scope` records the smallest temporal field needed to answer the question correctly. MOMENTARY is used when a short instant is enough. LOCAL WINDOW is used when the answer depends on a short continuous span around the key event. LONG RANGE is

used when evidence is spread across a large part of the clip. MULTI SEGMENT is used when several separated spans must be combined. This label guides the temporal field used by the router.

Temporal Scope Annotation Prompt

System. You are given the original question, the original ground truth answer, and the structured evidence JSON for one sample. Assign the single best `temporal_scope` label. Use the question as the main reference.

Label definitions. MOMENTARY: a short instant is enough. LOCAL WINDOW: a short continuous span around the key event is enough. LONG RANGE: evidence is spread across a large part of the clip. MULTI SEGMENT: several separated spans must be combined.

Decision rule. Choose the smallest temporal field that still supports the original answer.

Output format. Return valid JSON only in the form `{"temporal_scope": "LABEL"}`.

Figure 7: **Temporal scope prompt.** Gemini-3.1-Pro assigns one temporal label from the structured evidence JSON.

A.3 CoT Curation and Consistency Filter

After the three task labels are predicted, Gemini-3.1-Pro generates the reasoning CoT based on the clip, the original question, the original answer, the structured evidence JSON, and the cognitive schema tags. The prompt asks for grounded explanation only and does not allow facts that are not already present in the JSON. Specifically, the cognitive schema tags are used to keep the reasoning aligned with the intended answer path. For example, VISUAL should focus on visual evidence, TEMPORAL should explain event order when needed, and LONG RANGE or MULTI SEGMENT should connect cues across the required spans.

Besides, this stage also includes a consistency filtering step. After curating CoT, we use it as additional input to original question answering task to generate answer. Then, we normalize both the generated answer and the original answer to compare them. If the generated answer does not match the original answer, we discard the sample to maintain high quality for training examples with reasoning CoT.

Reasoning CoT Generation Prompt

System. You are given a video clip with audio, the original question, the original ground truth answer, the structured evidence JSON, and the predicted task labels. Generate the final tagged response.

Input labels. `evidence_source`, `reasoning_demand`, `temporal_scope`.

Output format. Return exactly this format and nothing else: `<Reasoning>...</Reasoning><Answer>...</Answer>`.

Reasoning rules. Use only evidence that appears in the JSON. Keep the reasoning short, factual, and grounded. Mention only the cues that directly support the answer. Do not add new people, actions, emotions, or events that are not present in the JSON.

Label-aware rules. Follow the predicted labels when selecting cues. VISUAL should focus on visual evidence, AUDIO should focus on audio evidence, JOINT should explain what comes from each modality, and CONFLICT should explain the mismatch. TEMPORAL should explain event order, CAUSAL should explain cause or effect, and LONG RANGE or MULTI SEGMENT should connect evidence across the needed spans.

Answer rule. Keep `<Answer>` concise and semantically aligned with the original ground truth answer.

Figure 8: **Reasoning CoT generation prompt.** Gemini-3.1-Pro generates the final tagged response conditioned on the structured evidence and the predicted task labels.

A.4 Social Dimension Taxonomy for Benchmark

The field `social_dimension` is used only in OMNISOCIALBENCH. It is not part of the training target and is not used for expert routing. This field describes the main social phenomenon tested by

Table 6: **Comparison with public benchmarks.** OMNISOCIALBENCH provides rich evidences for social video understanding task with human verification.

Benchmark	Modalities	Evidence Annot.	Temporal Span	Schema Labels	Reasoning Trace	Human-Verified Test
SocialIQ (Zadeh et al., 2019)	V+A+T	✗	✗	✗	✗	✓ (partial)
Omni-MMSI (Li et al., 2026)	V+A+T	✗	✗	✗	✗	✓
SocialOmni (Xie et al., 2026)	V+A+T	✗	✗	✗	✗	✓
Daily-Omni (Zhou et al., 2025b)	V+A	✗	✓	✗	✗	✓
WorldSense (Hong et al., 2025a)	V+A+T	✗	✗	✗	✗	✓
OMNISOCIALBENCH (Ours)	V+A+T	✓	✓	✓	✓	✓

the question, which can provide insight about fine-grained level evaluation for model performance on different social scenarios. We have 4 social dimensions: MENTAL STATE INFERENCE is used for hidden emotion, belief, desire, uncertainty, or masked affect. PRAGMATIC MEANING INFERENCE is used for indirect meaning, sarcasm, politeness, teasing, or relation signals in dialogue. ACTION GOAL INFERENCE is used for goal, plan, next step, or intention behind an action sequence. SOCIAL NORM INFERENCE is used for role expectation, setting rule, social convention, or appropriateness in context. The same clip may receive different `social_dimension` labels under different questions.

Social Dimension Annotation Prompt

System. You are given the original question, the original ground truth answer, and the structured evidence JSON for one benchmark sample. Assign the single best `social_dimension` label. Use the question as the main reference.

Label definitions. MENTAL STATE INFERENCE: hidden emotion, belief, desire, uncertainty, or other internal state. PRAGMATIC MEANING INFERENCE: indirect meaning, sarcasm, teasing, politeness, power relation, or relation signal in dialogue. ACTION GOAL INFERENCE: goal, plan, next step, purpose, or intention behind an action sequence. SOCIAL NORM INFERENCE: norm, role expectation, setting rule, appropriateness, or social convention.

Decision rule. Choose the single label that best captures the main kind of social reasoning tested by the question.

Output format. Return valid JSON only in the form `{"social_dimension": "LABEL"}`.

Figure 9: **Social dimension prompt.** Gemini-3.1-Pro assigns one benchmark label from the structured evidence JSON.

A.5 HUMAN REVIEW OF ANNOTATIONS

We apply manual human review across training set and benchmark in OMNISOCIALBENCH to evaluate the data quality to ensure reliability of our annotation policy and outcome datasets. For the training set, 20 annotators check the quality of structured evidence, cognitive schema tags and CoT of a random sample with size 500, and the overall annotation accuracy is 99.1%. For the evaluation split, 5 annotators check all 800 examples. In this step, we find 6 problematic cases where more than one human reviewer mark it as incorrect, so we manually rectify these cases to ensure it aligns with human review fully.

A.6 COMPARISON WITH PUBLIC BENCHMARKS

Table 6 compares the annotation fields between OMNISOCIALBENCH and public general / social-specific audio-visual video QA benchmarks. OMNISOCIALBENCH includes richest human-verified information, including structured audio-visual evidence, temporal evidence span, cognitive schema tag, and high-quality reasoning trace. These fields are not only supporting our method design, but also can serve as cornerstone for other task requiring rich evidences to solve question in social scenario.

B OMNISOCIALBENCH EXAMPLES

We provide qualitative examples from OMNISOCIALBENCH to illustrate how each question is paired with structured audio-visual evidence, schema labels, grounded reasoning, and temporal evidence spans. As shown in Figures 10 and 11, the benchmark covers diverse social reasoning cases where the answer may rely on visual cues, audio cues, their joint interpretation, or conflicts between what is

said and what is visually expressed. These examples also demonstrate how OMNISOCIALBENCH makes the supporting evidence explicit, enabling diagnostic evaluation beyond final-answer accuracy.



Figure 10: **Visualization of OMNISOCIALBENCH examples,** showing representative social video QA instances with corresponding questions, answers, schema tags, grounded evidence, reasoning traces, and temporal evidence spans.

C REWARD DETAILS

We provide a visualization in Fig 12 to explain details of 3 reward designs used in RMRL stage.

C.1 MODALITY-CONSISTENT REASONING REWARD

While the Cognitive Temporal Grounding reward encourages the model to focus on the correct temporal region, it does not directly assess whether the generated reasoning grounds the answer in the appropriate modality evidence. To complement it, we introduce a Modality-Consistent Reasoning reward computed by a frozen LLM judge (*e.g.*, Gemini-3.1-Pro). This reward is built upon the annotations produced by our Structured Evidence Extraction pipeline. For each sample, the annotation specifies whether the reasoning is expected to rely on visual evidence, audio evidence, or both, and, for dual-modality cases, whether the relation between the two modalities is supportive or conflicting. During reinforcement learning, we extract the text span between `<Reasoning>` and `</Reasoning>` and ask the frozen judge to score how well the reasoning matches the target modality requirement. The judge outputs a scalar score in $[0, 1]$ using a five-level rubric: $\{0, 0.25, 0.5, 0.75, 1.0\}$. A score of 1.0 is assigned when the reasoning explicitly uses the required modality evidence with concrete observations and clearly links them to the answer. When both modalities are required, the reasoning must also correctly characterize their relation, *e.g.*, whether they are mutually



Figure 11: **Additional visualization of OMNISOCIALBENCH examples.** These examples further illustrate the diversity of modality requirements, reasoning demands, and temporal scopes in OMNISOCIALBENCH, including cases requiring audio-visual integration and socially grounded inference.

supportive or in conflict. A score of 0.75 indicates that the reasoning is largely modality-consistent but remains incomplete, weakly connected to the answer, or only partially captures the cross-modal relation. A score of 0.5 is used when the reasoning only partially satisfies the requirement, such as relying on only one required modality or mentioning the correct modality without concrete evidence. A score of 0.25 is assigned when the reasoning contains weak, vague, or ambiguous modality cues or mixes correct and incorrect evidence sources. A score of 0 is assigned when the reasoning fails to ground the answer in the required modality, relies on the wrong evidence source, or does not contain a valid reasoning span. Importantly, the judge is instructed to reward concrete modality-specific observations rather than superficial keyword matching. For example, statements such as "the speaker sounds hesitant" or "the person smiles while stepping back" are treated as grounded evidence, whereas generic phrases such as "based on the video" or "from the audio" without specific observations are not sufficient for a high score. We keep the judge frozen and use the same prompt throughout all experiments. The full prompt is shown in Fig. 13. The judge evaluates only the modality consistency of the reasoning, while answer correctness is handled separately by the binary answer reward.

C.2 HUMAN VALIDATION OF THE MODALITY VERIFIER

To evaluate the realiablity of our modality consistent reasoning reward, we engage human review to validate it. We sample 500 rollouts from the reinforcement learning pool and give 4 independent human reviewers the same rubric and the same structured evidence annotation that the judge receives. Each reviewer will give score for each rollout independently, then we measure whether this score

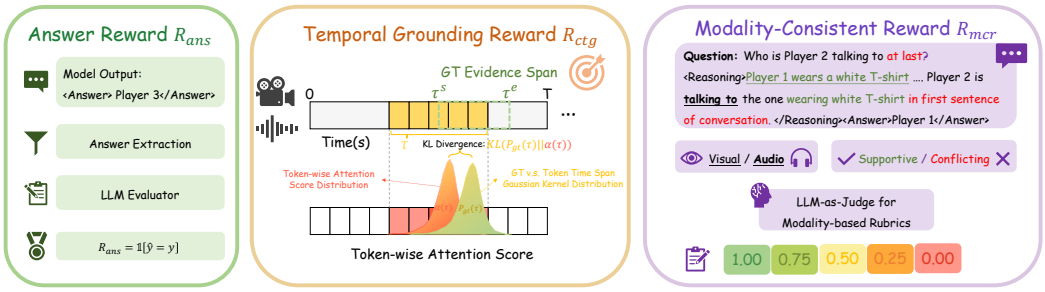

Figure 12: **Visualization of the reward design.** The overall reward consists of three complementary components: the answer reward $\mathcal{R}_{\text{ans}}$ encourages correct final answers, the Cognitive Temporal Grounding reward $\mathcal{R}_{\text{ctg}}$ guides the model to attend to the annotated temporal evidence span, and the Modality-Consistent Reasoning reward $\mathcal{R}_{\text{mcr}}$ promotes reasoning grounded in the required visual and/or audio evidence.

Table 7: **Statistics for agreement between human reviewer and LM judge.** Pairwise Cohen's κ experiment for agreement evaluation. Gemini-3.1-Pro is used as LM judge.

Statistics	Mean Pairwise Agreement Rate	Agreement Rate 95% CI	Mean Pairwise κ
Outcome	0.893	0.874–0.911	0.785

matches the score given by LM judge. Further, we did statistical analysis by calculating mean of Cohen's κ of each reviwer pair to evaluate agreement between LM judge and human reviewer.

As shown in Table 7, a mean pairwise κ of 0.785 indicates a high level of agreement between human reviewer and LM judge. It supports the use of the modality consistent reasoning judge score as a reliable reward.

D Additional Experimental Details

For supervised fine-tuning (SFT), we train the base model, Qwen3 Omni 30B, on 8 NVIDIA H200 GPUs for one day, using a per-device batch size of 4 with a gradient accumulation step of 1. The learning rate is set to 1×10^{-4}, and we apply low-rank adaptation with rank 64. For reinforcement learning (RL), the loss coefficients are set to $\lambda_{\text{ans}} = \lambda_{\text{ctg}} = \lambda_{\text{mcr}} = 1$. We use AdamW optimization with standard $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay of 0.1, along with a cosine learning rate schedule and a short warmup phase. Mixed-precision training (bfloat16) is enabled for efficiency, and gradient clipping is applied to stabilize training. Additionally, we adopt a mixture-of-experts configuration with $N = 128$ total experts, from which $k = 8$ are activated per routed token. As defined in §2.3, expert selection is stochastic during RL rollout collection: we restrict the router to its $M = \lceil 1.5k \rceil = 12$ highest-scoring experts and sample the 8 active experts without replacement from the renormalized distribution over this candidate pool. Evaluation and all reported results use deterministic top-k routing.

E Qualitative Analysis

We provide qualitative response comparisons between CogniRoute and Gemini 3.1 Pro on manually verified examples from OmniSocialBench. These examples are intended as diagnostic visualizations rather than an additional quantitative metric. Each case shows the input frames, question, candidate options, Gemini's response, and the correct response produced by CogniRoute. The examples below cover discourse-grounded reference resolution and long-range affect continuity, complementing the multi-party turn-taking case in Figure 3 of the main paper.

Figure 14, the main-paper Figure 3, and Figure 15 show a consistent pattern: Gemini 3.1 Pro often identifies a locally plausible cue, but maps it to the wrong evidence structure. In the first two cases, the answer depends on discourse-level reference tracking rather than a single utterance or a single gaze direction. In the third case, the answer depends on temporal continuity rather than a local facial expression. We find CogniRoute has better performance in these scenario, which indicates the effectiveness of our evidence-guiding routing strategy in method design. The qualitative results

Modality-Consistent Reasoning Judge Prompt

System Prompt:
You are a strict evaluator for modality-consistent reasoning in multimodal question answering. Given a question, a final answer, a reasoning trace, and a structured evidence annotation, your task is to score whether the reasoning grounds the answer in the required modality evidence.
Evaluate only the reasoning trace. Do **not** judge whether the final answer itself is correct.
Scoring Rubric:

- **1.00**: The reasoning explicitly uses the required modality evidence with concrete observations and clearly links them to the answer. If both modalities are required, it also correctly describes their relation (supportive or conflicting).
- **0.75**: The reasoning is mostly modality-consistent, but the evidence grounding is incomplete, weakly connected to the answer, or only partially captures the cross-modal relation.
- **0.50**: The reasoning only partially satisfies the modality requirement, *e.g.*, it relies on only one required modality, or mentions the correct modality without concrete evidence.
- **0.25**: The reasoning provides weak, vague, or ambiguous modality evidence, or mixes correct and incorrect evidence sources.
- **0.00**: The reasoning fails to ground the answer in the required modality, relies on the wrong evidence source, or does not contain a usable reasoning span.

Important Rules:

- Reward **concrete modality-specific observations**, not superficial keywords.
- Generic phrases such as "based on the video" or "from the audio" do not count as strong evidence unless followed by specific observations.
- Visual evidence includes facial expressions, body movements, gaze, scene changes, visible interactions, or other observable events.
- Audio evidence includes spoken content, tone, laughter, crying, silence, or salient background sounds.
- If both modalities are required, check whether the reasoning uses both and whether it correctly describes their relation.

Input:
Question: {question}
Final Answer: {final_answer}
Reasoning: {reasoning}
Structured Evidence Annotation:

- Visual evidence required: {yes/no}
- Audio evidence required: {yes/no}
- Cross-modal relation: {none/support/conflict}

Output Format:
Return a JSON object: `{"score": 0.00/0.25/0.50/0.75/1.00, "brief_rationale": "..." }`

Figure 13: **Prompt used for the frozen LLM judge to compute the Modality-Consistent Reasoning reward.** The judge scores whether the generated reasoning grounds the answer in the modality evidence specified by the structured evidence annotation.

Table 8: **Effectiveness of method across backbones.** Evaluation of an additional open-source omni-modal backbone OmniVinci (Ye et al., 2025) under 3 settings. All fine-tuning settings use the same dataset as the Qwen-Omni-30B-A3B backbone.

	Mental State	Pragmatic Meaning	Action Goal	Social Norm	Avg.
OmniVinci	26.07	21.07	20.81	18.75	21.88
OmniVinci + SFT	46.92	43.21	47.65	39.38	44.25
OmniVinci + SAPR + RMRL	**60.19**	**55.71**	**66.44**	**58.13**	**59.38**

complement the quantitative gains in Table 1 by illustrating how CogniRoute improves evidence selection, cross-modal grounding, and temporal reasoning in social video question answering.

F Additional Studies

F.1 Method Generalization Across Backbone

To evaluate generalization under identical training settings, we compare CogniRoute with a baseline model OmniVinci and its retrained variant. As shown in Table 8, CogniRoute consistently achieves

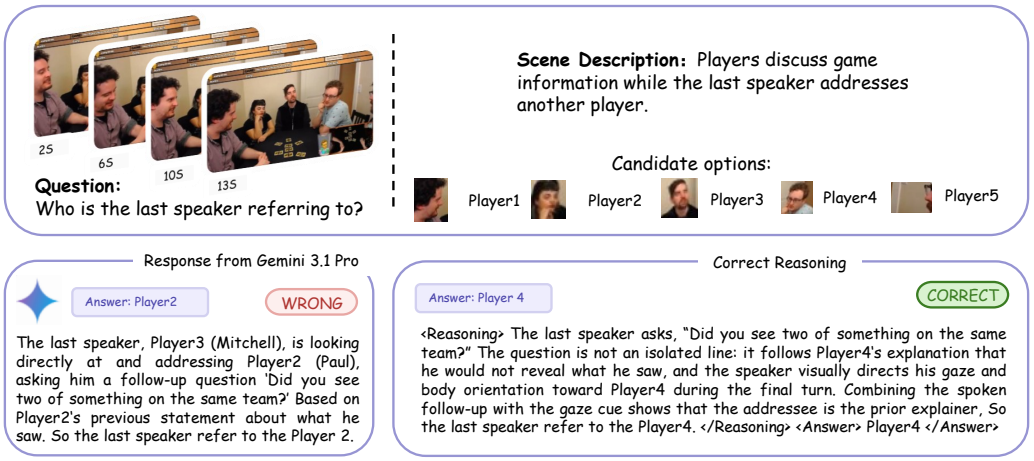

Figure 14: **Qualitative comparison on discourse-grounded reference resolution.** Gemini 3.1 Pro selects Player2 by treating the final utterance as a local follow-up to Player2's previous statement. CogniRoute instead links the final question to Player4's prior explanation and uses the speaker's gaze and body orientation toward Player4, leading to the correct referent.

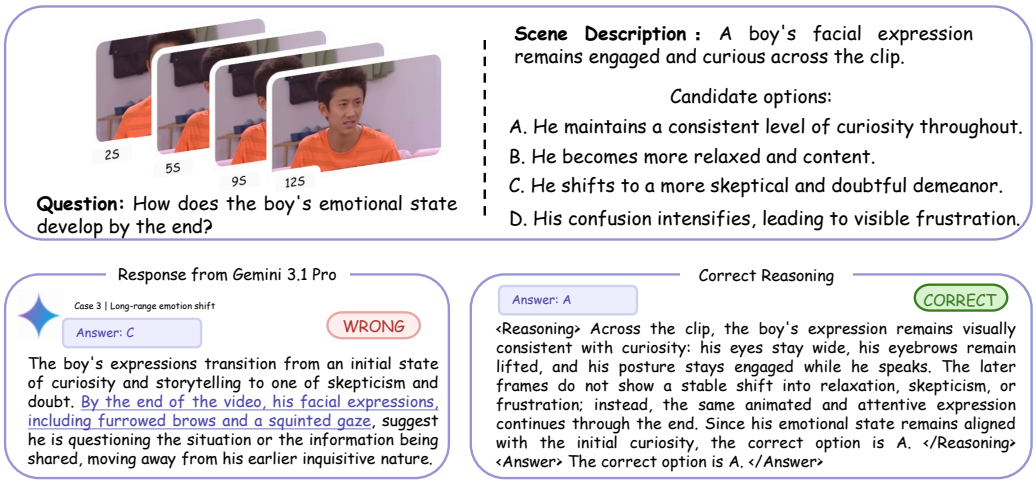

Figure 15: **Qualitative comparison on long-range affect grounding.** Gemini 3.1 Pro predicts an emotional shift toward skepticism or doubt from isolated late-frame cues. CogniRoute aggregates the boy's expression across the clip and observes that his wide-eyed, engaged expression remains consistent, supporting the correct answer that his curiosity is maintained throughout the video.

stronger performance across all four social dimension categories, and OmniVinci also gains a performance increase under the same training setting, indicating better transfer and generalization ability under the same supervision regime.

F.2 Learnable Tag Embedding Collapse Diagnostics

SAPR learns schema tag embeddings jointly with routing alignment, so the experiment examines whether different cognitive schemas form clearly distinguishable representations. First, the tag embedding space is evaluated using mean/max pairwise cosine similarity and normalized effective rank, which measure how well the schema tags are separated and how broadly they span the representation space. Second, the routing signatures g are frozen and used to train a linear probe for predicting the true schema tags. These diagnostics evaluate whether SAPR produces a well-structured schema space and whether that structure is reflected in schema-aware routing patterns. Standard SFT does not use tag embedding, so it is used only as a routing baseline.

Table 9: **Tag embedding collapse diagnostics.** Correct cognitive schema tags yield well-separated, high-rank embeddings and higher routing predictability than shuffled or random tags

Model Variant	Route mAP	Mean Cos.	Max Cos.	Norm. ERank	Avg. Acc.
Standard SFT without SAPR	32.41	N.A.	N.A.	N.A.	40.13
SAPR with shuffled tags	34.18	0.09	0.34	0.87	41.38
SAPR with random tags	33.27	0.11	0.37	0.84	40.63
Full SAPR	**62.85**	**0.06**	**0.28**	**0.92**	**50.13**

Table 10: **Routing behavior under gate optimization.** Full RMRL improves schema- and modality-aware routing while keeping expert use more balanced than any single-branch ablation.

Model Variant	Avg. Acc. ↑	Route mAP ↑	Load CV ↓	Active Exp. ↑	Δ_{rel} ↑	Δ_{irrel} ↓
SFT with Full SAPR	50.13	62.85	0.26	116	0.141	0.090
Token branch only	56.13	64.07	0.27	114	0.154	0.095
Gate branch only	51.88	67.48	0.24	120	0.181	0.087
Full RMRL	**59.38**	**73.92**	**0.22**	**124**	**0.224**	**0.082**

Let σ_r be the singular values of the row normalized tag embedding matrix, and let $p_r = \sigma_r / \sum_s \sigma_s$. The normalized effective rank is defined as

$$\text{ERank} = \frac{\exp\left(-\sum_r p_r \log p_r\right)}{\min(|\mathcal{V}_{\text{tag}}|, D)}.$$

A value close to 1 means that the tag embedding matrix uses many independent directions, while a low value means that the embeddings are close to a low rank solution.

As shown in Table 9, Full SAPR yields well-separated tag embeddings, with low mean and maximum pairwise cosine similarity and a high normalized effective rank. Its substantially higher route mAP further shows that these structured tag representations are reflected in the routing signatures, supporting effective alignment between cognitive schemas and expert routing.

F.3 Routing Behavior Under Gate Optimization

We further analyze how the gate branch changes routing behavior. Route mAP measures whether routing signatures predict the schema tags. Load CV and Active Exp. measure whether the model keeps a balanced use of experts. The two route shift metrics measure whether routing changes more when the required modality is removed than when an irrelevant modality is removed.

For this analysis, routing statistics are computed on generated reasoning tokens, since the gate branch receives RL feedback through generated token routes. Route mAP is obtained by training a linear probe on frozen routing signatures to predict the true schema tags. Load CV is the coefficient of variation of selected expert counts over the 128 experts. Active Exp. is the number of experts that receive at least 0.1% of selected routes. To compute modality route shift, we first generate the output with the full input and then run teacher forcing on the same output under masked inputs. Let g_{gen} denote the routing signature over generated reasoning tokens and let $d(\cdot, \cdot)$ be cosine distance. We define

$$\Delta_{\text{rel}} = d\left(g_{\text{gen}}(x), g_{\text{gen}}(x_{\setminus m_{\text{rel}}})\right), \qquad \Delta_{\text{irrel}} = d\left(g_{\text{gen}}(x), g_{\text{gen}}(x_{\setminus m_{\text{irrel}}})\right).$$

For visual only samples, video is the relevant modality and audio is the irrelevant modality. For audio only samples, audio is the relevant modality and video is the irrelevant modality. For samples that require both audio and video, removing either modality is treated as relevant and these samples are not used for Δ_{irrel}.

As shown in Table 10, Full RMRL achieves the highest route mAP and active-expert count, together with the lowest load CV, indicating more schema-aware routing with broader and more balanced expert utilization. It also produces the largest route shift when the relevant modality is removed (Δ_{rel}) and the smallest shift for irrelevant-modality removal (Δ_{irrel}), showing stronger sensitivity to task-relevant evidence. Compared with the single-branch variants, these results suggest that jointly optimizing token generation and expert routing yields the most effective combination of routing quality and answer accuracy.

Table 11: **Effectiveness of CTG reward design** The full soft-distribution CTG reward gives the most precise temporal grounding because it performs better than other settings without CTG reward or with different design.

Model Variant	Avg. Acc.	Mass@0.5s ↑	Mass@1.0s ↑	Hit@0.5s ↑	Hit@1.0s ↑	PeakErr ↓
SFT with Full SAPR	50.13	15.6	26.8	16.3	28.1	2.94
Full RL without $\mathcal{R}_{\text{ctg}}$	53.50	17.4	30.2	18.5	31.6	2.67
Full RL with binary CTG reward	58.37	27.8	44.0	29.1	45.3	1.98
Full RL with shuffled CTG time labels	53.25	16.9	29.1	17.8	30.4	2.73
Full RMRL	**59.38**	**31.4**	**49.2**	**33.1**	**51.0**	**1.56**

F.4 Routing Signature Visualization

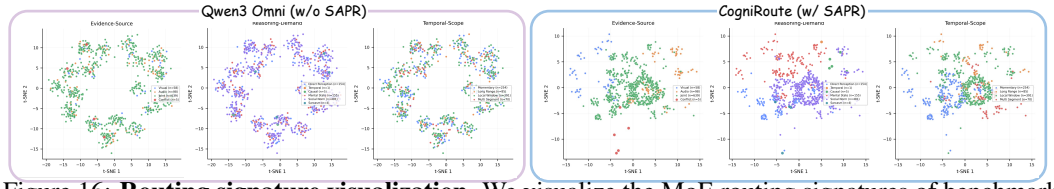

Figure 16: **Routing signature visualization.** We visualize the MoE routing signatures of benchmark samples before and after applying SAPR. For each model, the same two-dimensional coordinates are colored by Cross Modal Relation, Reasoning Demand, and Temporal Scope.

Figure 16 projects the routing signatures into two dimensions. Without SAPR, routing assignments are diffuse and weakly structured, with substantial overlap across samples. After SAPR, routing signatures form more distinct and coherent clusters that align with `evidence_source`, `reasoning_demand`, and `temporal_scope`, indicating that samples with similar evidence structure become more likely to share expert usage. We report this projection as a qualitative companion to the route mAP probe rather than as independent evidence, since both are computed from the same routing signatures.

F.5 Effect of CTG Reward Design on Temporal Grounding

We evaluate whether CTG improves temporal evidence alignment and answer accuracy beyond increasing its own reward value. The study isolates three factors: the presence of temporal supervision, the benefit of a smooth reward, and the correctness of evidence timestamps. We compare SFT with Full SAPR, full RL without CTG, full RL with a binary CTG reward, full RL with shuffled CTG timestamps, and Full RMRL with the proposed smooth CTG reward. The binary variant rewards an attention peak within 0.5 seconds of the annotated time, whereas smooth CTG provides partial credit for near misses; shuffled timestamps test whether correct temporal supervision drives the gains. For every setting, we extract temporal attention from the reasoning suffix, including the closing reasoning token, using the same layer/head averaging and audio-visual token pooling procedure. Since each example has a single annotated evidence time, we report Mass@0.5s and Mass@1.0s, the attention mass within the corresponding windows ($\pm 0.5/1$ from annotated time). Besides, Hit@0.5s and Hit@1.0s are measured as the proportion of examples whose attention peak falls within evidence windows. Last, PeakErr is calculated as the mean absolute distance between the attention peak and the annotated time.

As shown in Table 11, full RMRL consistently achieves the strongest temporal grounding and answer accuracy. Binary CTG also improves performance over RL without CTG, supporting the value of explicit temporal supervision, while its weaker results relative to smooth CTG suggest a benefit from graded feedback for near misses. Shuffled timestamps offer no comparable gains, indicating that correct evidence timing is essential. Together, these findings support smooth CTG as an effective means of improving attention-based temporal alignment alongside task performance.

F.6 Causal Evidence Deletion

This experiment tests whether the temporal region identified by the annotations is actually important for the model's prediction, rather than merely correlated with its internal attention pattern. The design follows a causal intervention intuition: if CogniRoute truly relies on the annotated evidence, removing that region should hurt performance substantially more than removing an unrelated region

Table 12: **Causal evidence deletion.** GT Mask removes a 1-second window centered at the annotated evidence time, while Random Mask removes an equal-length window elsewhere. The larger performance change under GT Mask indicates stronger causal reliance on the annotated temporal evidence.

Input Variant	Mental State	Pragmatic Meaning	Action Goal	Social Norm	Avg.
GT Mask	47.87	43.21	53.69	44.38	46.63
Random Mask	57.35	52.14	62.42	55.63	56.13
Full Input	**60.19**	**55.71**	**66.44**	**58.13**	**59.38**

Table 13: **Modality Evidence Usage Ablation.** CogniRoute relies on relevant modality more significantly than base model, indicating the importance of correct evidence support.

Model / Subset	Full Input	Remove Required Modality	Remove Irrelevant Modality
Qwen3-Omni-30B, VISUAL	33.0	26.7 (-6.3)	31.4 (-1.6)
Qwen3-Omni-30B, AUDIO	31.0	25.8 (-5.2)	29.6 (-1.4)
CogniRoute, VISUAL	60.0	41.3 (-18.7)	57.9 (-2.1)
CogniRoute, AUDIO	57.0	39.4 (-17.6)	54.6 (-2.4)

of the same duration. In order to study it, we apply different kind of mask in RL stages to evaluate causal reliance of temporal region.

As shown in Table 12, removing the annotated evidence window reduces average accuracy from 59.38 to 46.63, whereas removing a matched random window only reduces it to 56.13. The same trend appears across all four social-reasoning categories, indicating that CogniRoute relies specifically on the annotated temporal evidence when forming its predictions.

F.7 Modality Evidence Use

This experiment explores whether CogniRoute's predictions depend selectively on the modality required by each question. We split the benchmark by the annotated evidence source. For questions whose evidence source is video only, we run one setting that removes the video stream and one that removes the audio stream. For questions whose evidence source is audio only, we do the same in the opposite direction. The removed stream is dropped from the input and the question tokens stay unchanged. The design compares the effect of removing the required modality with removing an irrelevant modality: a model with stronger modality-grounded reasoning should show a substantially larger performance change when task-relevant evidence is removed.

As shown in Table 13, CogniRoute exhibits a clear separation between required and irrelevant modality removal, while the backbone shows a weaker distinction. This indicates that CogniRoute develops stronger modality-specific evidence use and more selectively relies on the information relevant to each question.

G Real-world VR Smart Glasses Deployment

We deploy CogniRoute on VR smart glasses for real-time social interaction, as shown in Figure 17. In the first example (top row), CogniRoute performs synergistic routing by combining hand trajectories with partial speech to provide contextual assistance. In the second example (bottom row), CogniRoute captures subtle social cues, detecting hesitation and conflicting gestures, to infer true intent (*e.g.*, polite refusal).

H Related Work

Omnimodal Large Language Models and audio-visual reasoning. Recent MLLMs have expanded from image-text understanding to omni-modal models that jointly process text, vision, and audio within a unified interaction framework (Bai et al., 2025b;a; Lin et al., 2024; Liu et al., 2023; Chen et al., 2024b; Li et al., 2024a; Hurst et al., 2024; Fu et al., 2025). Systems such as Qwen-Omni and

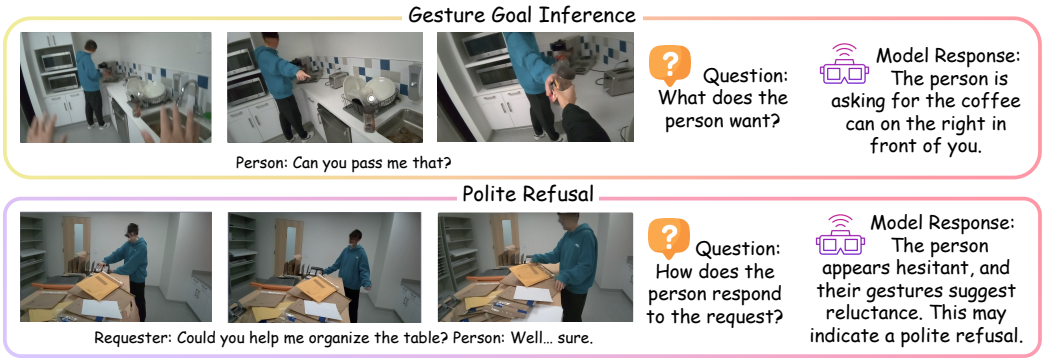

Figure 17: **Real-world VR Application** illustrating CogniRoute's ability to infer human intent.

Gemini (Xu et al., 2025a;b; Team et al., 2023; 2024; Comanici et al., 2025) demonstrate strong real-time and holistic audio-visual understanding capabilities. However, unified multimodal access (Xu et al., 2025a;b; Comanici et al., 2025) does not specify how evidence should be selected or how computation should be allocated for each question. Existing OLM training (Ye et al., 2025; Yu et al., 2024) primarily optimizes generated responses, leaving modality reliability, cross-modal agreement or conflict, temporal grounding, and social evidence use only indirectly supervised. CogniRoute addresses this gap by using sample-level schema supervision to guide MoE routing according to modality relation, reasoning demand, and temporal scope.

MoE-based Multimodal Models. Mixture-of-Experts (MoE) models enable conditional computation by routing inputs to a subset of experts (Eigen et al., 2013). In multimodal learning, MoE has been used to specialize computation across modalities (Bao et al., 2022; Mustafa et al., 2022; Chen et al., 2024a), tasks (Satar et al., 2022; Li et al., 2023b), and domains through modality- or task-conditioned experts, sparse routing (Shen et al., 2024), and parameter-efficient expert adapters (Long et al., 2024). These methods improve capacity and efficiency (Li et al., 2024b; Lin et al., 2026), but their routers are typically optimized through next-token prediction, load balancing, or general efficiency objectives, rather than fine-grained supervision to learn which modality is reliable, how audio and visual cues relate, what reasoning is required, and where the decisive cue occurs. In contrast, CogniRoute aligns routing signatures with sample-level schema labels and uses route-aware reinforcement learning to optimize expert allocation for correct, modality-consistent, and temporally grounded answers.

Reinforced Multimodal Reasoning. Recent advances in Reinforcement Learning (RL), particularly through process supervision and verifiable reward modeling in language models (Lightman et al., 2023; Guo et al., 2025), have begun to extend into Multimodal Large Language Models (MLLMs), enabling improved reasoning capabilities across multimodal tasks (Cao et al., 2026; Duan et al., 2025; Su et al., 2025; Wang et al., 2025b; Yang et al., 2025c; Zhou et al., 2025a; Shen et al., 2025; Fan et al., 2025; Zheng et al., 2026). These approaches introduce reward signals that supervise not only final predictions but also intermediate reasoning steps, often by verifying consistency with visual evidence through process reward models (Wang et al., 2025a; Ong et al., 2025; Zhang et al., 2025). Building on R1-style training paradigms and RL algorithms such as GRPO (Shao et al., 2024), recent works further encourage more structured reasoning trajectories to achieve grounded visual reasoning (Yang et al., 2025c; Shen et al., 2025; Zhou et al., 2025a; Cao et al., 2026; Chu et al., 2025). In parallel, methods like GRIT (Fan et al., 2025), DeepEyes (Zheng et al., 2026), and related frameworks (Wang et al., 2026; Jiang et al., 2026) incorporate explicit visual grounding, *e.g.*, bounding boxes or region-level supervision, into the reasoning process. This alignment between textual inference and localized visual evidence improves interpretability and helps reduce hallucination (Jiang et al., 2026).

Social omni-modal reasoning. Recent work has begun to move omni-modal modeling toward human-centered and socially interactive understanding. HumanOmni (Zhao et al., 2025b) develops a vision-speech-language model for human-centric video understanding, with emphasis on emotion recognition, facial-expression description, and action understanding. R1-Omni (Zhao et al., 2025a) applies reinforcement learning with verifiable rewards to omni-modal emotion recognition, improving reasoning, accuracy, and generalization. HumanOmniV2 (Yang et al., 2025b) further studies context-aware omni-modal reasoning over human intentions and emotions, addressing global-context misunderstanding and shortcut behavior through context and logical rewards. Complementary benchmarks examine social interaction from more targeted perspectives: SocialOmni (Xie et al., 2026)

evaluates conversational interactivity through speaker identification, interruption timing, and interruption generation, while Omni-MMSI (Li et al., 2026) focuses on identity-attributed social interaction understanding from raw audio, vision, and speech cues. These works establish social omni-modal understanding as an important direction, but they primarily focus on improving or evaluating high-level social perception, emotion/intent reasoning, turn-taking, or identity attribution. In contrast, our work treats social video question answering as an evidence-allocation problem. CogniRoute explicitly supervises the model's internal routing behavior according to the evidence structure of each question, including modality relation, reasoning demand, and temporal scope, and further optimizes expert allocation with rewards for answer correctness, modality-consistent reasoning, and temporal grounding.