

Cross-Modal Coreference Alignment: Enabling Reliable Information Transfer in Omni-LLMs

Hongcheng Liu*, Yuhao Wang*, Zhe Chen, Pingjie Wang, Zhiyuan Zhu
Yixuan Hou, Yanfeng Wang, Yu Wang†

Shanghai Jiao Tong University

{hongcheng_liu, colane, yuwangshtu}@sjtu.edu.cn

* Equal contribution. † Corresponding author.

Abstract

Omni Large Language Models (Omni-LLMs) have demonstrated impressive capabilities in holistic multi-modal perception, yet they consistently falter in complex scenarios requiring synergistic omni-modal reasoning. Beyond understanding global multimodal context, effective reasoning also hinges on fine-grained cross-modal alignment, especially identifying shared referents across modalities, yet this aspect has been largely overlooked. To bridge this gap, we formalize the challenge as a cross-modal coreference problem, where a model must localize a referent in a source modality and re-identify it in a target modality. Building on this paradigm, we introduce CROSSOMNI, a dataset comprising nine tasks equipped with human-designed reasoning rationales to evaluate and enhance this capability. Experiments on 13 Omni-LLMs reveal systematic weaknesses in cross-modal coreference, which we attribute to the absence of coreference-aware thinking patterns. To address this, we enhance cross-modal alignment via two strategies: a training-free In-Context Learning method and a training-based SFT+GRPO framework designed to induce such thinking patterns. Both approaches yield substantial performance gains and generalize effectively to collaborative reasoning tasks. Overall, our findings highlight cross-modal coreference as a crucial missing piece for advancing robust omni-modal reasoning.

1 Introduction

Omni Large Language Models (Omni-LLMs) have demonstrated impressive capabilities in holistic multi-modal perception, yet they consistently falter in complex scenarios requiring synergistic omni-modal reasoning (Liu et al., 2025b; Yao et al., 2024; Liu et al., 2024b). Most prior work attempts to improve it by aggregating information from multiple modalities, such as expanding data coverage (Ye et al., 2025; Gong et al., 2025) or refining training objectives (Yang et al., 2025b; Zhong et al.,

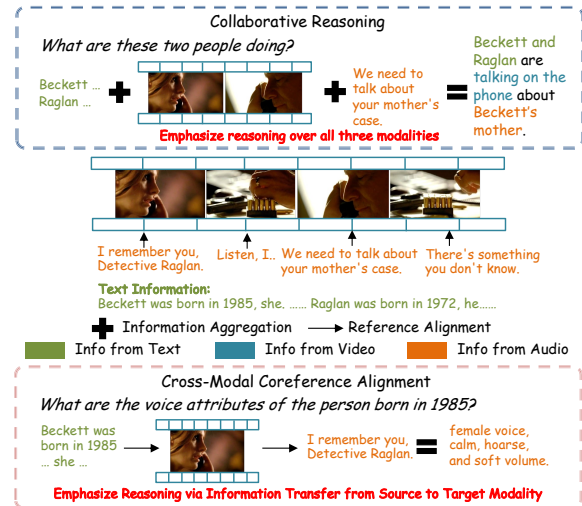


Figure 1: The comparison between collaborative reasoning and cross-modal coreference alignment. The former focuses on coarse aggregation, whereas the latter targets fine-grained alignment of corresponding referents.

2025). However, effective reasoning requires not only modeling global multimodal context but also performing fine-grained cross-modal alignment (Li et al., 2025). Models must bind the same referent across modalities and align its evolving state to compose evidence into coherent reasoning. While aggregation-oriented methods can enhance overall perception, they often underemphasize such alignment and fail to reliably link corresponding content across modalities. Consequently, current Omni-LLMs struggle to maintain referential consistency (Alonso et al., 2025), which severely constrains their reasoning performance.

To bridge this gap, we formalize the challenge as a cross-modal coreference problem, which we view as a fundamental capability for Omni-LLMs. Moving beyond the collaborative reasoning focused on coarse aggregation in prior work, our approach targets a critical intermediary step: precisely locating a referent in a source modality and re-identifying it within a target modality (Figure 1). We introduce CROSSOMNI, a comprehensive dataset designed

Datasets	# Videos	# QA Pairs	Modality	w. Cross-Modality	Cross-Modal Coreference	w. Training Data
Pano-AVQA (2021)	5,400	51,700	A, V	✗	-	✓
AVQA (2022)	57,015	57,335	A, V	✗	-	✓
Music-AVQA (2022)	9,288	45,867	A, V	✗	-	✓
OmniBench (2025)	-	4,555	A, I	✗	-	✗
WorldSense (2025)	1,662	3,172	A, V	✗	-	✗
DailyOmni (2025)	684	1,197	A, V	✗	-	✗
AVUT (2025c)	2,662	11,609	A, V	✓	A→V	✗
CROSSOMNI	4,147	39,726	A, V, T	✓	A↔V, A↔T, V↔T	✓

Table 1: Comparison of CROSSOMNI with representative omni-modal datasets. Cross-modality denotes questions that require information transfer across modalities. A/I/V/T denote audio, image, video, and text.

to evaluate and elicit this capability. CROSSOMNI encompasses three single-modality and six cross-modality coreference tasks spanning text, audio, and video. Crucially, it provides human-annotated reasoning rationales that supervise step-wise alignment, facilitating both granular evaluation and effective model tuning.

Benchmarking 13 leading Omni-LLMs reveals a persistent performance gap between single- and cross-modal coreference, highlighting a systemic deficiency. Through Chain-of-Thought (CoT) prompting, we observe that models often lack the necessary “thinking patterns” to transfer information reliably between modalities. Motivated by this, we explore two strategies to induce coreference-aware thinking: a training-free In-Context Learning approach and a training-based method combining SFT with Group Relative Policy Optimization (GRPO). Both methods yield significant gains across all coreference tasks. Furthermore, our ablation studies demonstrate that these improvements generalize to broader collaborative reasoning benchmarks, suggesting that cross-modal coreference is a foundational “missing piece” in the evolution of Omni-LLM intelligence.

The main contributions can be summarized as:

- **Cross-Modal Coreference Alignment.** We formulate cross-modal coreference alignment, where the model must identify information in a target modality by first locating it in a source modality. This paradigm shifts the focus for improving Omni-LLMs from coarse modality aggregation to fine-grained cross-modal coreference alignment.
- **Rationale-Augmented Multi-Task Dataset.** This dataset comprises 9 diverse task types enriched with human-designed rationales, specifically developed to facilitate the training and evaluation of Omni-LLMs in complex cross-modal coreference alignment.
- **Coreference-aware Thinking Patterns.** We

introduce both novel training-free and training-based methods to elicit structured thinking patterns. We show that these patterns significantly boost performance on both coreference-specific tasks and general collaborative reasoning benchmarks.

2 CROSSOMNI Dataset

2.1 Dataset Overview

2.1.1 Data Principles

The CROSSOMNI dataset is designed to evaluate and support training for cross-modal coreference alignment over three modalities in Omni-LLMs. Unlike existing omni benchmarks that simply fuse information from different modalities and aggregate various cues to answer a question, CROSSOMNI emphasizes cross-modality coreference among audio, visual, and textual data, as illustrated in Figure 2. Specifically, CROSSOMNI pairs single-modality coreference tasks with cross-modality coreference tasks, allowing unimodal performance to serve as a baseline for directly assessing and highlighting cross-modal evidence transfer within a unified framework.

Concretely, we construct a diverse set of videos with comprehensive annotations, providing training sets and test splits for cross-modality alignment, consisting of 4,147 videos and 39,726 question-answer pairs (Table 2). All videos are sampled from TVQA¹ and come from six long-running TV shows with multi-person conversations and complex visual scenes, which pose significant challenges for cross-modality coreference alignment. For each modality, we perform detailed annotation and automatic verification, followed by human checking to ensure data quality. We also provide explicit step-by-step rationales for each instance, offering effective thinking patterns that enable more

¹https://nlp.cs.unc.edu/data/jielei/tvqa/tvqa_public_html/index.html

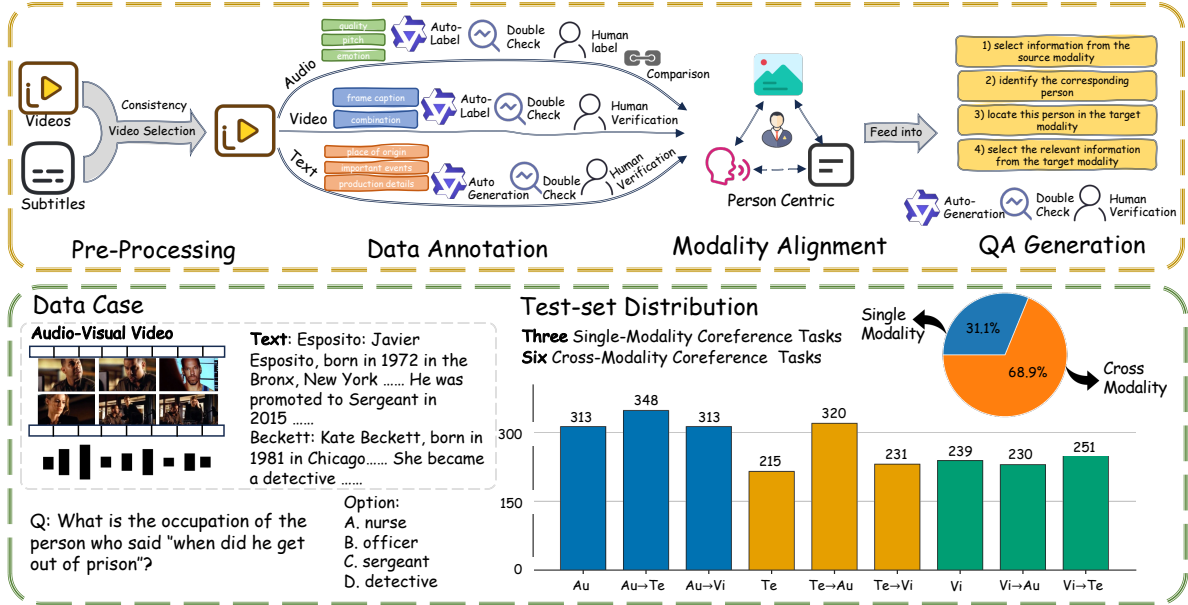


Figure 2: Overview of the CROSSOMNI dataset. The top section illustrates the annotation pipeline, including pre-process, data annotation, modality alignment, and QA generation. The bottom section presents example instances and the distribution of the test set.

Type	# Train	# Test	# Total
Videos	2,725	1,422	4,147
QA pairs	37,266	2,460	39,726

Table 2: Statistics of the CROSSOMNI dataset.

precise analysis and training of cross-modal coreference alignment.

2.1.2 Task Definition

We model each instance as a source-to-target coreference alignment problem. Given a question q , a source modality $m_s \in \{Au, Vi, Te\}$, and a target modality $m_t \in \{Au, Vi, Te\}$, the model is required to: (1) identify the referent in the source modality that is needed to answer q , and (2) retrieve the aligned information in the target modality to produce the final answer a . Single-modality coreference corresponds to $m_s = m_t$, while cross-modality coreference corresponds to $m_s \neq m_t$.

2.1.3 Task Classification

Under task definition, we define nine task types: three single-modality coreference tasks and six cross-modality coreference tasks over audio, video, and text. Table 3 presents representative examples for each type. The information available in each modality is summarized below, and more details are provided in Appendix B.1.

- **Audio:** Speech content and audio characteristics.
- **Vision:** Visual understanding, including actions, attributes, spatial reasoning, and counting.

- **Text:** Factual information from text, such as birthplace, career, and related attributes.

2.2 Data Construction

Our data construction pipeline consists of four stages: pre-processing, data annotation, modality alignment, and question-answer pair generation, as summarized in Figure 2. We describe it below.

2.2.1 Pre-Processing

TVQA provides rich video and subtitle annotations, but their coverage and quality are uneven. Since our goal is cross-modal alignment that links each utterance to a speaker identity and synchronizes it with the corresponding visual evidence and text information, reliable human-labeled subtitles with accurate timestamps are essential. Therefore, we restrict our data to clips with manually annotated transcripts and discard segments that lack human subtitles or contain temporally discontinuous timestamps. Finally, we split each video into subtitle-aligned clips, which serve as the basic unit for the subsequent modality-specific annotation and question-answer pair construction.

2.2.2 Data Annotation

We annotate each video with visual, audio, and textual information. For the visual modality, we use Qwen2.5-VL-7B (Bai et al., 2025) to caption representative frames and then use Qwen3-8B (Yang et al., 2025a) to summarize them into clip-level

Single-modality Coreference Tasks	
Audio-only (Au)	What is the emotion in the sentence “why would she do that”? A. calm B. excited C. angry D. sad
Vision-only (Vi)	Where does the third woman enter from? A. front side B. back side C. left side D. right side
Text-only (Te)	What was beckett’s first screenplay? A. the weight of silence B. silent echoes C. the weight of words D. echoes of the past
Cross-modality Coreference Tasks	
Audio→Vision (Au→Vi)	What facial expression does the person who spoke with aversion emotion exhibit? A. disoriented look B. calm expression C. serious or concerned expression D. attentive look
Audio→Text (Au→Te)	What is the occupation of the person who said “when did he get out of prison”? A. nurse B. officer C. sergeant D. detective
Vision→Audio (Vi→Au)	What does the person’s emotion in voice who has wide eyes and open mouth in the video? A. sadness B. excitement C. happiness D. fear
Vision→Text (Vi→Te)	What is the birthdate of the person wearing brown pants and a black t-shirt? A. april 5, 1985 B. january 1, 1985 C. march 12, 1985 D. march 12, 1984
Text→Audio (Te→Au)	What does the person say who joined the princeton-plainsboro teaching hospital? A. I’d be too nervous. Couldn’t be objective. B. I need to know everything about you. C. here’s what happens when doctors care too much. D. She’s mine. Stay away.
Text→Vision (Te→Vi)	Where is the Vision space of the person who led a drug trafficking investigation? A. in the vicinity of the second man B. in front of the woman C. in the left of the woman D. in the background

Table 3: Examples of single-modality and cross-modality coreference tasks. **In all cases, the model receives an audio-visual video and a textual biography as input.**

descriptions, as well as to produce person-centric descriptions with external knowledge from subtitles (e.g., role cues explicitly mentioned in the subtitle). For audio, we treat the human-provided subtitles with speaker identities and timestamps as the ground truth for speech content. We additionally label basic voice characteristics (e.g., quality, pitch, emotion) using Qwen3-Omni-30B-A3B (Jin et al., 2025) through structured binary questions. Finally, we use Qwen3-8B (Yang et al., 2025a) to generate text biographies for people in the videos, including key attributes such as place of origin, important events, and production details. All descriptions are verified by cross-comparing annotations across different samples to ensure consistency and correctness. Prompts are provided in Table 13 and Table 14 in Appendix D.

2.2.3 Modality Alignment

To align information across different modalities, we use visual features as the anchor. In particular, we treat each person as a bridge that links visual, audio, and textual information, thereby enabling tri-modal alignment. For visual–audio alignment, we use speaker names as anchors. The same name appearing in person-centric visual descriptions and human-labeled transcripts is used to associate the corresponding visual and audio segments. For vi-

Model	Base model	Procedure
Baichuan-Omni (2024)	Qwen2.5-7B	P+SFT
HumanomniV2 (2025b)	Qwen2.5-Omni-7B	SFT+RL
M2-Omni (2025)	Llama3.1-8B	P+SFT
Ming-Omni (2025)	Ling-moe	P+SFT
Minicpm-o (2024)	Qwen2.5-7B	P+SFT+RL
Ola (2025b)	Qwen2.5-7B	P+SFT
Omni-r1 (2025)	Qwen2.5-Omni-7B	SFT+RL
OmniVinci (2025)	Qwen2.5-7B	SFT
Qwen2.5-Omni (2025)	Qwen2.5-3B/7B	P+SFT
Qwen3-Omni-Cap (2025)	Qwen3-30B-A3B	P+SFT+RL
Qwen3-Omni-Ins (2025)	Qwen3-30B-A3B	P+SFT+RL
Qwen3-Omni-Thk (2025)	Qwen3-30B-A3B	P+SFT+RL

Table 4: Model statistics of different omni models. The P, SFT, and RL denote the pre-training, supervised fine-tuning, and reinforcement learning stages, respectively.

sual–textual alignment, we integrate person-centric universal descriptions into textual biographies so that the biographies can be grounded in visual cues.

2.2.4 Question–Answer Pair Construction

We construct QA pairs in two steps: generation and verification. We use Qwen3-8B (Yang et al., 2025a) to generate questions based on the visual descriptions, subtitles, audio characteristics, and biographies, following the unified source-to-target format. For each question type, we enforce a fixed coreference transfer structure: locating the referent in a source modality and then selecting the corresponding information in a target modality. We then verify each candidate instance by checking (i) source-side

Model	Audio-centric			Text-centric			Visual-centric			Overall		
	Au	Au→Te	Au→Vi	Te	Te→Au	Te→Vi	Vi	Vi→Au	Vi→Te	Single	Cross	Overall
Baichuan-Omni	34.19	29.31	21.09	51.63	20.94	24.24	30.13	15.65	22.78	37.81	22.69	27.39
HumanomniV2	28.75	39.08	<u>45.37</u>	44.65	24.06	38.53	46.44	29.13	37.84	38.72	35.79	36.71
M2-Omni	45.69	35.92	29.39	75.35	26.56	45.89	<u>56.49</u>	33.04	51.35	57.37	36.20	42.83
Ming-Omni	45.05	45.98	55.91	73.49	41.88	56.71	<u>64.44</u>	37.39	49.81	59.06	47.91	51.38
Minicpm	36.74	38.51	29.07	73.02	23.12	36.80	45.19	16.52	<u>47.49</u>	49.54	31.97	37.48
Ola	46.33	46.26	42.49	<u>49.30</u>	39.38	46.75	54.81	29.57	<u>37.07</u>	49.80	40.70	43.52
Omni-r1	49.20	46.84	45.69	83.26	38.44	54.11	<u>66.11</u>	33.91	61.00	64.02	46.37	51.90
OmniVinci	51.76	55.17	<u>56.23</u>	55.35	39.69	47.62	59.83	32.17	44.79	55.28	46.75	49.39
Qwen2.5-Omni-3B	45.05	44.54	39.62	<u>57.21</u>	39.06	50.65	63.18	33.04	44.02	54.11	41.79	45.62
Qwen2.5-Omni-7B	48.88	45.11	40.26	83.26	37.19	52.81	<u>64.44</u>	33.91	59.85	63.37	44.43	50.36
Qwen3-Omni-Cap	55.14	50.75	36.93	52.09	40.00	49.44	<u>53.47</u>	38.26	50.12	53.76	44.19	47.19
Qwen3-Omni-Ins	<u>60.58</u>	53.05	44.92	64.65	44.69	55.06	58.91	43.48	56.29	61.20	49.42	53.10
Qwen3-Omni-Thk	62.17	67.13	56.10	73.49	52.19	63.29	62.26	47.39	<u>69.42</u>	65.37	59.40	61.29
Average	46.89	45.97	41.77	64.36	35.94	47.84	<u>55.82</u>	32.58	48.60	54.57	42.12	46.01

Table 5: Performance of different models across modalities. Au, Te, and Vi denote Audio, Text, and Vision, respectively; X→Y denotes source to target modality (e.g., Au→Te: Audio to Text). Single and Cross denote the average performance on single-modal and cross-modal tasks, respectively. In each row, the best is in **bold** and the second-best is underlined.

uniqueness of the referent and (ii) target-side answerability from the annotations. Prompts are provided in Table 15 in Appendix D.

2.3 Chain-of-Thought Rationales for Training

For the training split, we provide structured rationales to support both analysis and training. Each rationale follows a four-step template aligned with our task format: (1) identify the source and target modalities from the question; (2) locate the referent in the source modality; (3) find the aligned referent in the target modality; and (4) extract the answer from the target modality. We generate these rationales with Qwen3-8B (Yang et al., 2025a) conditioned on the annotations and ground-truth answers, and verify them for consistency with both the evidence and the intended reasoning path. Prompts are provided in Table 16 in Appendix D.

2.4 Quality Control and Human Verification

We perform human verification to assess annotation quality and QA validity. For audio characteristics, we randomly sample 1,000 clips and re-label quality, pitch, and emotion, comparing human labels with those from the automatic pipeline and obtaining 93.1% agreement. For video descriptions, we inspect 1,000 descriptions and obtain a correctness rate of 96.8%. For textual biographies, we sample 1,000 cases and verify that there is no inappropriate or illegal content. For questions, we sample 1,000 instances and evaluate whether each question is answerable given the provided information, obtaining an answerability rate of 98.3%. These results suggest that CROSSOMNI meets the quality

requirements for reliable training and evaluation. More details are provided in Appendix B.4.

3 Evaluation Experiment

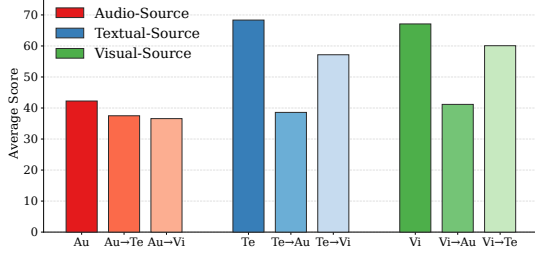
To examine the performance of coreference alignment across modalities, we evaluate Omni-LLMs on CROSSOMNI, which consists of single-modality questions and cross-modality counterparts under a coreference alignment task format.

3.1 Experimental Details

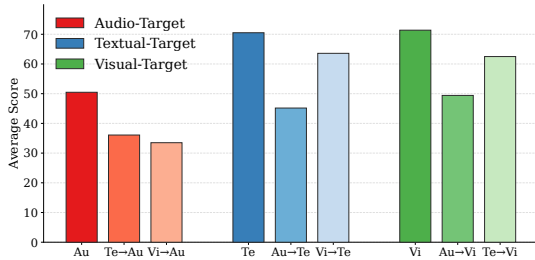
To comprehensively examine performance on cross-modal coreference alignment, we evaluate 13 mainstream Omni-LLMs with varying sizes, architectures, and training strategies. The model details are summarized in Table 4. For all experiments, we use the prompt “Answer this question based on the video and text information in A/B/C/D.” and report accuracy as the evaluation metric.

3.2 Main Evaluation Results

The main results across different tasks are shown in Table 5. On average, cross-modality coreference tasks are about 21% worse than single-modality coreference tasks. For almost all models, the top two results are obtained on single-modality coreference tasks, indicating that current Omni-LLMs have strong unimodal perception but do not consistently carry it over to cross-modal settings. Therefore, improving omni-modal reasoning calls for directly strengthening cross-modal coreference alignment, which appears to be a key missing component in current Omni-LLMs.



(a) Average scores on fixed-source questions.



(b) Average scores on fixed-target questions.

Figure 3: Average performance on fixed-source and fixed-target cross-modal questions.

3.3 Controlled Cross-modal Experiments

To disentangle the effect of information content from that of cross-modal coreference alignment, we construct two controlled settings. For fixed-source questions, we select 150 questions from each single-modality coreference task and convert them into cross-modality coreference questions by changing only the target modality while keeping the source information fixed. For fixed-target questions, we follow a similar process but keep the target modality fixed and vary the source modality. As shown in Figure 3b and Figure 3a, performance consistently drops once the source modality differs from the target modality, even though the information in the fixed modality remains unchanged. This indicates that current models struggle to identify the correct referent in the source modality and transfer it reliably to the target modality, further demonstrating the necessity of improving cross-modal coreference alignment.

4 Analysis and Improvements

To better understand and improve cross-modal coreference alignment in Omni-LLMs, we conduct a systematic analysis of model failures and develop targeted remedies. We first diagnose why models perform well when evidence is contained within a single modality but fail when the answer requires transferring a referent across modalities. Our analyses indicate that a central bottleneck is the lack of explicit coreference-aware thinking pat-

terns for cross-modality reasoning. We then present two complementary approaches to induce such procedures: a training-free In-Context Learning method and a training-based framework via SFT and GRPO. Experimental results show that enhancing it effectively improves overall reasoning performance.

4.1 Diagnosing Failures Cause

To pinpoint the sources of cross-modal coreference failures, we adopt Chain-of-Thought (CoT) prompting to make the model explicitly reveal its coreference alignment process by prepending “Let’s think step by step.” to the default prompt. This setting allows us to inspect the resulting reasoning rationales of the whole process for systematic errors.

Importantly, we find that models rarely establish explicit and correct links between entities across modalities, as illustrated in Figure 4. This suggests that current models lack coreference-aware reasoning patterns, namely, the ability to explicitly and step-by-step establish and coordinate referential correspondences across modalities. Furthermore, as shown in Figure 5 and Table 6, Chain-of-Thought prompting yields only modest gains, with an overall relative improvement of 1.7%. These limited improvements further suggest that current models do not possess sufficiently strong cross-modal coreference reasoning capabilities on their own, and that simply eliciting more detailed reasoning is far from sufficient to close the gap. Consequently, improving cross-modal coreference alignment likely requires explicitly fostering coreference-aware reasoning patterns for reference alignment.

4.2 Improving Cross-Modal Coreference

Building on the above experiments, which show that cross-modality coreference performance lags far behind single-modality coreference and model-internal thinking patterns yield only marginal gains, we regard the lack of coreference-aware thinking patterns as the main bottleneck. Therefore, we explore two approaches to enhance the coreference-aware thinking patterns for modality coreference alignment: a training-free approach based on In-Context Learning (ICL), and training-based approaches using Supervised Fine-Tuning (SFT) and Group Relative Policy Optimization (GRPO).

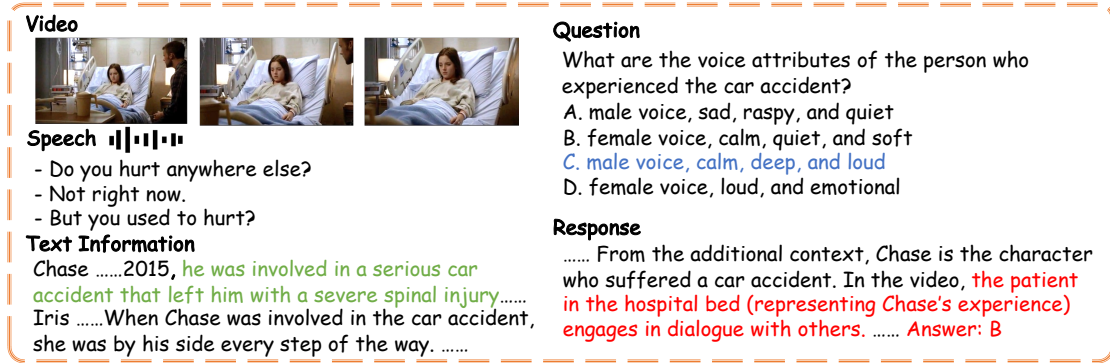


Figure 4: CoT case study on text→audio task. The green denotes important information, the blue denotes the right choice, and the red denotes the error rationales. The model fails to establish the correct bridge between the textual information (the person who experienced the car accident is a man) and the visual information (the man is sitting on the chair), and therefore cannot answer the question correctly. The full case is provide in Figure 6 in Appendix C.

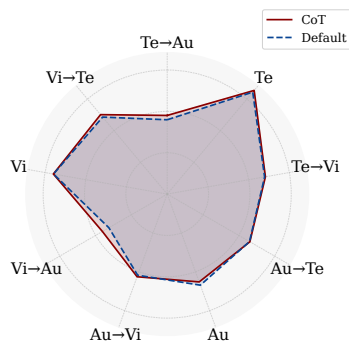


Figure 5: Accuracy comparison of 13 models under default and Chain-of-Thought prompting.

4.2.1 Training-free: In-Context Learning

Given that existing models can correctly extract information within a single modality but struggle to align it across modalities, we design a dedicated prompt to guide step-by-step cross-modal coreference alignment via In-Context Learning (ICL), as shown in Table 17 in Appendix D. As reported in Table 6, applying ICL yields substantial performance gains, particularly on cross-modal coreference tasks, with a 21% relative improvement. The performance gap between single-modality and cross-modality coreference shrinks from 24.37% to 11.01% under ICL. These results are consistent with our diagnosis that the primary limitation in cross-modal alignment lies in coreference-aware thinking patterns, indicating that future work should focus on explicitly learning and refining these thinking patterns.

4.2.2 Training-based: SFT and GRPO

Training Details. We use both SFT-based and GRPO-based training methods to enhance cross-modal coreference capability. For SFT, we perform LoRA fine-tuning and then apply GRPO on top of

Method	Single-modality	Cross-modality	Overall
Default	55.69	42.12	46.01
CoT	55.52 (−0.30%)	43.56 (+3.42%)	46.83 (+1.78%)
ICL	57.29 (+2.87%)	50.98 (+21.04%)	52.63 (+14.39%)

Table 6: Average performance of 13 models with CoT and ICL on CROSSOMNI.

the SFT checkpoint using 982 training instances balanced across modalities, with Qwen3-Next-80B-A3B-Instruct (Yang et al., 2025a) as the reward model. More details are provided in Appendix B.2.

Experimental Results. To further enhance coreference-aware thinking patterns for cross-modal alignment, we conduct experiments on Qwen2.5-Omni-3B and Qwen2.5-Omni-7B using SFT and GRPO training methods. To establish the coreference-aware thinking patterns, we apply CoT supervision during SFT and introduce a thinking-pattern reward during GRPO. Specifically, the GRPO objective combines rewards for output format, reasoning accuracy, contextual consistency, and logical coherence. The details are provided in the Table 18 in Appendix D. As shown in Table 7, both SFT and GRPO yield substantial performance gains, with GRPO achieving the largest improvements, particularly on cross-modal coreference tasks. We further present two representative cases in Figure 7 and Figure 8 in Appendix C, which illustrate the effectiveness of these patterns. Overall, these results demonstrate that explicitly training coreference-aware thinking patterns is crucial for improving cross-modal performance.

Necessity of Explicit Coreference-aware Thinking Patterns. To further isolate the contribution of the coreference-aware thinking patterns in the SFT and GRPO stages, we conduct ablations that

Model	Method	Audio-centric			Text-centric			Visual-centric			Overall
		Au	Au→Te	Au→Vi	Te	Te→Au	Te→Vi	Vi	Vi→Au	Vi→Te	
Qwen2.5-Omni-3B	Default	45.05	44.54	39.62	57.21	39.06	50.65	63.18	33.04	44.02	45.62
	CoT	44.09	42.53	42.81	50.23	37.81	48.92	66.11	33.91	39.38	44.57 (-2.30%)
	SFT	39.62	50.28	52.72	76.74	38.75	54.54	58.99	40.00	51.35	50.45 (+10.59%)
	GRPO	46.33	60.34	56.55	87.44	46.25	62.34	69.04	43.48	65.25	58.59 (+28.43%)
Qwen2.5-Omni-7B	Default	48.88	45.11	40.26	83.26	37.19	52.81	64.44	33.91	59.85	50.36
	CoT	43.77	46.26	37.70	83.26	36.56	55.84	63.18	37.39	58.69	49.84 (-1.03%)
	SFT	50.48	56.90	54.63	80.46	47.19	60.61	59.83	45.65	59.83	56.65 (+12.49%)
	GRPO	57.19	62.93	61.02	88.84	51.56	63.64	67.36	56.48	64.86	62.56 (+24.23%)

Table 7: Performance of Qwen2.5-Omni models across modalities and training methods. Au, Te, and Vi denote Audio, Text, and Visual, respectively. X→Y denotes source-to-target modality (e.g., Au→Te: Audio-to-Text).

Method	Single-modality	Cross-modality	Overall
Qwen2.5-Omni-3B			
Default	55.15	41.82	45.62
SFT	58.45 (+5.98%)	47.94 (+14.63%)	50.45 (+10.59%)
SFT w/o CoT	50.15 (-9.06%)	42.55 (+1.75%)	45.22 (-0.88%)
Qwen2.5-Omni-7B			
Default	65.52	44.86	50.36
SFT	63.59 (-2.95%)	54.39 (+21.24%)	56.65 (+12.49%)
SFT w/o CoT	52.42 (-19.99%)	46.30 (+3.21%)	48.70 (-3.30%)

Table 8: Ablation study of SFT on Qwen2.5-Omni.

Method	Single-modality	Cross-modality	Overall
Qwen2.5-Omni-3B			
SFT	58.45	47.94	50.45
GRPO	67.60 (+15.65%)	55.70 (+16.19%)	58.59 (+16.13%)
GRPO w/o T-R	58.98 (+0.91%)	50.03 (+4.36%)	52.67 (+4.40%)
Qwen2.5-Omni-7B			
SFT	63.59	54.39	56.65
GRPO	71.13 (+11.86%)	59.58 (+9.54%)	62.56 (+10.43%)
GRPO w/o T-R	67.60 (+6.31%)	57.55 (+5.81%)	59.72 (+5.42%)

Table 9: Ablation study of GRPO on Qwen2.5-Omni. T-R denotes the thinking-pattern reward.

remove them from each training process. Specifically, we perform SFT without CoT supervision and GRPO without thinking patterns rewards. Tables 8 and 9 report consistent performance drops under both ablations, showing that CoT supervision and thinking-pattern rewards are effective for establishing coreference-aware reasoning. These results further indicate that explicitly modeling coreference-aware thinking patterns is crucial for effective cross-modal reasoning.

Generalization to OOD Benchmarks. To examine whether the cross-modal coreference alignment ability learned on CROSSOMNI generalizes, we evaluate the trained models on three out-of-distribution (OOD) benchmarks for collaborative reasoning and cross-modal alignment. As reported in Table 10, both training strategies achieve con-

Method	Daily-Omni	WorldSense	AVUT	Average
Qwen2.5-Omni-3B				
Default	40.52	43.50	48.10	44.04
SFT	47.45 (+17.10%)	39.24 (-9.79%)	48.79 (+1.43%)	45.16 (+2.54%)
GRPO	54.55 (+34.62%)	44.45 (+2.18%)	57.79 (+20.14%)	52.26 (+18.67%)
Qwen2.5-Omni-7B				
Default	47.45	45.40	48.90	47.25
SFT	55.97 (+17.96%)	41.36 (-8.90%)	49.77 (+1.78%)	49.03 (+3.77%)
GRPO	57.89 (+22.00%)	49.08 (+8.11%)	57.96 (+18.52%)	54.98 (+16.35%)

Table 10: Performance on Daily-Omni, WorldSense, and AVUT dataset. The details about these datasets are provided in Appendix B.3.

sistent gains on almost all benchmarks, with especially great improvements on collaborative reasoning tasks in the Daily-Omni and WorldSense benchmarks, highlighting the importance of coreference-aware reasoning for both cross-modal alignment and omni-modal collaborative reasoning. On WorldSense, SFT alone slightly degrades performance because of a mismatch in video scenarios between WorldSense and CROSSOMNI, whereas subsequent GRPO not only recovers this drop but also yields clear gains, indicating that GRPO is essential for coreference-aware training, as it learns robust cross-modal coreference alignment instead of overfitting to the CROSSOMNI distribution. Overall, these results demonstrate that cross-modal coreference alignment learned from our dataset generalizes to OOD settings and constitutes a critical capability for advancing Omni-LLMs.

5 Conclusions

To advance the reasoning capabilities of Omni-LLMs, we focus on cross-modal coreference alignment and introduce the CROSSOMNI dataset for evaluating and training this capability. Across 13 Omni-LLMs, we observe that unimodal competence does not directly translate into reliable cross-modal evidence transfer. Our analyses suggest that a key failure factor is the absence of

coreference-aware thinking patterns for aligning referents across modalities. Motivated by this, we propose a training-free In-Context Learning method and a training-based SFT+GRPO framework to induce such patterns, substantially improving performance on CROSSOMNI and generalizing to broader collaborative reasoning tasks. These results highlight cross-modal coreference as a crucial missing piece for improving the reasoning capabilities of Omni-LLMs and provide a concrete pathway toward more reliable omni-modal understanding.

Limitations

There are several limitations in this work. First, performance on audio-based tasks remains weaker than that on text- and visual-based tasks. Beyond multi-speaker interference, this gap may also be related to speech quality, understanding of clean single-speaker speech, and speaker discrimination across conversational turns. More fine-grained analysis is needed to identify whether the main bottleneck lies in auditory perception, speaker discrimination, or reasoning. Second, our annotation pipeline currently relies on human-provided subtitles, which limits automation, scalability, and extension to larger real-world corpora.

Ethical Considerations

All videos in our dataset are sampled from TVQA, which is publicly available for academic use. We follow its license terms, use the data solely for research, and do not redistribute raw video content. All models in our pipeline and evaluation are open-source, and our processing does not introduce additional personally identifiable information beyond what is present in TVQA, we only release derived annotations. College students provided human ratings without collecting personal characteristics or identifiers. We also used large language models as writing assistants for editing and polishing the paper. Overall, we believe our work poses minimal additional ethical or privacy risks beyond those already present in the underlying TVQA dataset.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62576209) and STCSM (No. 2025SHZDZX025G05).

References

- Iñigo Alonso, Gorka Azkune, Ander Salaberria, Jeremy Barnes, and Oier Lopez de Lacalle. 2025. Vision-language models struggle to align entities across modalities. *arXiv preprint arXiv:2503.03854*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Zhe Chen, Hongcheng Liu, and Yu Wang. 2024. **Dialogmcf: Multimodal context flow for audio visual scene-aware dialog**. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:753–764.
- Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, and 1 others. 2025. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*.
- li Guangyao, Wei Yake, Tian Yapeng, Xu Chenliang, Wen Ji-Rong, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Qingpei Guo, Kaiyou Song, Zipeng Feng, Ziping Ma, Qinglong Zhang, Sirui Gao, Xuzheng Yu, Yunxiao Sun, Tai-Wei Chang, Jingdong Chen, and 1 others. 2025. M2-omni: Advancing omni-mlm for comprehensive modality support with competitive performance. *arXiv preprint arXiv:2502.18778*.
- Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. 2025. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*.
- Inclusion, Bowen Ma, Cheng Zou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Chenyu Lian, Dandan Zheng, Fudong Wang, Furong Xu, and 1 others. 2025. Ming-flash-omni: A sparse, unified architecture for multimodal perception and generation. *arXiv preprint arXiv:2510.24821*.
- Xu Jin, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. **Qwen3-omni technical report**. *Preprint*, arXiv:2509.17765.
- Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, and 1 others. 2024. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*.
- Yan Li, Yifei Xing, Xiangyuan Lan, Xin Li, Haifeng Chen, and Dongmei Jiang. 2025. Alignmamba: Enhancing multimodal mamba with local and global

- cross-modal alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24774–24784.
- Hongcheng Liu, Zhe Chen, Hui Li, Pingjie Wang, Yanfeng Wang, and Yu Wang. 2024a. **Msg-bart: Multi-granularity scene graph-enhanced encoder-decoder language model for video-grounded dialogue generation**. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10516–10520.
- Hongcheng Liu, Yixuan Hou, Heyang Liu, Yuhao Wang, Yanfeng Wang, and Yu Wang. 2025a. **Vocalbench-df: A benchmark for evaluating speech llm robustness to disfluency**. *arXiv preprint arXiv:2510.15406*.
- Hongcheng Liu, Yusheng Liao, Siqv Ou, Yuhao Wang, Heyang Liu, Yanfeng Wang, and Yu Wang. 2024b. **Med-pmc: Medical personalized multi-modal consultation with a proactive ask-first-observe-next paradigm**. *arXiv preprint arXiv:2408.08693*.
- Hongcheng Liu, Pingjie Wang, Heyang Liu, Zhiyuan Zhu, Yusheng Liao, Yanfeng Wang, and Yu Wang. 2026. **Anchornet: Adaptive anchor token enhancement in video-grounded dialogue generation**. *IEEE Journal of Selected Topics in Signal Processing*, pages 1–11.
- Hongcheng Liu, Pingjie Wang, Yu Wang, and Yanfeng Wang. 2024c. **M2k-vdg: Model-adaptive multimodal knowledge anchor enhanced video-grounded dialogue generation**. *arXiv preprint arXiv:2402.11875*.
- Hongcheng Liu, Pingjie Wang, Zhiyuan Zhu, Yanfeng Wang, and Yu Wang. 2024d. **CE-VDG: Counterfactual entropy-based bias reduction for video-grounded dialogue generation**. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2958–2968, Torino, Italia. ELRA and ICCL.
- Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. 2025b. **Ola: Pushing the frontiers of omni-modal language model**. *arXiv preprint arXiv:2502.04328*.
- LongCat Meituan, Bairui Wang, Bin Xiao, Bo Zhang, Bolin Rong, Borun Chen, Chang Wan, Chao Zhang, Chen Huang, Chen Chen, and 1 others. 2025. **Longcat-flash-omni technical report**. *arXiv preprint arXiv:2511.00279*.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. **Gpqa: A graduate-level google-proof q&a benchmark**. In *First Conference on Language Modeling*.
- Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2025. **Mmsu: A massive multi-task spoken language understanding and reasoning benchmark**. *arXiv preprint arXiv:2506.04779*.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, and 1 others. 2024. **Mmlu-pro: A more robust and challenging multi-task language understanding benchmark**. *Advances in Neural Information Processing Systems*, 37:95266–95290.
- Zhenghao Xing, Xiaowei Hu, Chi-Wing Fu, Wenhai Wang, Jifeng Dai, and Pheng-Ann Heng. 2025. **Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning**. *arXiv preprint arXiv:2505.04623*.
- 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. 2025. **Qwen2.5-omni technical report**. *Preprint*, arXiv:2503.20215.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. **Qwen3 technical report**. *arXiv preprint arXiv:2505.09388*.
- Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. **Avqa: A dataset for audio-visual question answering on videos**. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3480–3491.
- Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaying Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. 2025b. **Humanomniv2: From understanding to omni-modal reasoning with context**. *arXiv preprint arXiv:2506.21277*.
- Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. 2025c. **Audio-centric video understanding benchmark without text shortcut**. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6580–6598.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. **Minicpm-v: A gpt-4v level mllm on your phone**. *arXiv preprint arXiv:2408.01800*.
- Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, and 1 others. 2025. **Omnivinci: Enhancing architecture and data for omni-modal understanding llm**. *arXiv preprint arXiv:2510.15870*.
- LI Yizhi, Ge Zhang, Yinghao Ma, Ruibin Yuan, King Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Noah Wang, Jian Yang, and 1 others. 2025. **Omnibench: Towards the future of universal omni-language models**.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. 2021. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2031–2041.

Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen, and Chunhua Shen. 2025. Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration. *arXiv preprint arXiv:2505.20256*.

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

A Related Works

Omni-LLMs Omni-LLMs are designed to jointly process textual information, auditory features, and visual content, and to generate accurate responses based on integrated omni-modal understanding and reasoning (Liu et al., 2024d; Chen et al., 2024; Ye et al., 2025; Meituan et al., 2025; Liu et al., 2026). However, compared with holistic multi-modal perception, effectively leveraging omni-modal information in complex scenarios remains a central challenge (Liu et al., 2024c; Inclusion et al., 2025; Liu et al., 2024a). For example, Baichuan-Omni (Li et al., 2024) adopts a specialized audio–visual alignment procedure via SFT, and Qwen3-Omni (Jin et al., 2025) applies GRPO to enhance collaboration among the three modalities. These models, however, mainly focus on aggregating multiple modalities and largely ignore fine-grained cross-modal alignment, which severely limits their overall reasoning performance. To bridge this gap, we formulate cross-modal coreference alignment as a core problem, requiring models to establish correct correspondences for the same referent across modalities and offering a new perspective for improving the overall performance of Omni-LLMs.

Omni-Modal Dataset The omni-modal dataset aims to enhance the capabilities of Omni-LLMs across diverse modalities. From a broader perspective, however, mainstream evaluation of Omni-LLMs still predominantly relies on single-modality benchmarks (Rein et al., 2024; Wang et al., 2025; Liu et al., 2025a), such as MMMU (Yue et al., 2024) for image understanding and MMLU (Wang et al., 2024) for text comprehension. In parallel, many works have been proposed to probe unimodal reasoning capabilities (Yizhi et al., 2025; Xing et al., 2025). For example, DailyOmni (Zhou et al., 2025) targets audio and visual reasoning in everyday scenarios, and WorldSense (Hong et al., 2025) focuses on assessing collaborative understanding and reasoning over omni-modal inputs. While these datasets are effective for measuring overall unimodal or holistic omni-modal performance, they typically do not isolate the key step of cross-modality coreference alignment, which limits their ability to diagnose why unimodal success does not translate into reliable omni-modal reasoning. To fill this gap, we introduce CROSSOMNI, a dataset that explicitly targets source-to-target coreference alignment, enabling systematic analysis of

cross-modal coreference failures, as well as providing training data to improve cross-modal coreference alignment.

B Experimental Details

B.1 Task Classification Details

We construct three single-modality coreference tasks and six cross-modality coreference tasks. For cross-modality coreference tasks that involve audio and text, we use the visual modality as an anchor to ground the referent and connect modalities through person-centric alignment. The details of each task are as follows:

Single-Modality Coreference Tasks.

- ▶ **Audio-only (Au).** Understanding speech content and audio characteristics.
- ▶ **Visual-only (Vi).** Visual understanding, including actions, attributes, spatial reasoning, and counting.
- ▶ **Text-only (Te).** Factual information from text, such as birthplace, career, education, and related attributes.

Cross-Modality Coreference Tasks.

- ▶ **Audio→Vision (Au→Vi).** Identify the person from the audio and answer about their visual information.
- ▶ **Audio→Text (Au→Te).** Identify the person from audio, ground via a visual anchor, then answer textual facts.
- ▶ **Vision→Audio (Vi→Au).** Identify the person from visual cues and answer about speech transcription or audio characteristics.
- ▶ **Vision→Text (Vi→Te).** Identify the person from visual cues and answer textual facts.
- ▶ **Text→Audio (Te→Au).** Identify the person from the text, ground via a visual anchor, then answer speech transcription or audio characteristics.
- ▶ **Text→Vision (Te→Vi).** Identify the person from the text and answer about the visual information.

B.2 Training Details

We use both SFT and GRPO training to enhance cross-modal coreference capability. For SFT, we perform LoRA fine-tuning with rank 16 and $\alpha = 32$, using a learning rate of 1×10^{-5} , a batch size of 32, and 1 training epoch. For GRPO, we initialize the policy from the SFT checkpoint and sample

982 instances from the training set, ensuring that each modality has a similar number of examples. We use a learning rate of 5×10^{-6} , a batch size of 16, a rollout number of 4, and 1 GRPO epoch, with Qwen3-Next-80B-A3B-Instruct (Yang et al., 2025a) as the reward model.

B.3 Datasets on Ablation Study

We conduct ablation studies on three omni datasets, and the detailed descriptions are as follows:

Daily-Omni: Daily-Omni (Zhou et al., 2025) is an audio-visual multiple-choice QA benchmark targeting temporally aligned cross-modal reasoning in daily-life scenarios. It contains 684 videos and 1,197 QA pairs, with clips segmented into 30 s or 60 s to probe temporal understanding under different context lengths. The questions are grouped into six task types, all designed to require integrating audio and visual evidence rather than relying on a single modality.

WorldSense: WorldSense (Hong et al., 2025) is a real-world omni-modal video understanding benchmark that evaluates models with jointly provided video and audio, emphasizing strong audio-visual coupling such that removing either modality degrades answerability. It contains 1,662 audio-visual synchronized videos with 3,172 multiple-choice QA pairs spanning 26 distinct tasks.

AVUT: AVUT (Yang et al., 2025c) is an audio-centric video understanding benchmark designed to evaluate auditory understanding and audio-visual interactions, while containing sufficient audio-visual cross-modality tasks. It includes 2,662 YouTube videos covering 18 audio-centric domains and 11,609 QA pairs over 8 tasks, with an average duration of 67.8 seconds. We choose the AV-Human as the test set.

B.4 Human Verification

We provide details of the human verification procedure used in data construction. Five college students are recruited to verify the quality of the annotations. For audio characteristics, each clip is annotated with a set of labels (e.g., voice quality, pitch, emotion), and we treat human labels as the reference. We treat a label as an error if it is predicted by the auto-labeling system but absent from the human annotations, and we measure agreement as the proportion of automatically predicted labels that are also present in the human-labeled set. For

video descriptions, we check whether the generated descriptions are inconsistent with the video content using a binary label (True/False). For textual biographies, we verify whether there is any inappropriate or illegal content, again with a binary label. For questions, we verify each question against the answer that is supported by the evidence, also using a binary label.

B.5 Further Discussion on the Main Results

From Table 5, we draw the following observations:

- 1. Model size helps, but is not sufficient.** Larger models generally perform better, but scale alone does not determine performance. For example, a smaller model can outperform a larger one (e.g., Qwen2.5-Omni-3B vs. Ola-7B), suggesting that training data and model design are at least as important as parameter count.
- 2. Architecture, especially the audio encoder, is critical.** Models equipped with stronger Whisper variants tend to achieve substantially better results. This is consistent with the fact that MiniCPM-o builds on Whisper-medium, whereas Qwen2.5-Omni adopts Whisper-large-v3.
- 3. The training paradigm also matters.** “Thinking”-enhanced omni variants (e.g., Qwen3-Omni-Thinking) exhibit stronger reasoning ability and better cross-modal alignment, indicating the value of explicit reasoning-oriented training.
- 4. Data scaling is essential.** Increasing alignment-focused training data consistently improves both SFT and GRPO in our experiments, underscoring the importance of sufficient and well-curated supervision signals.

Overall, strong cross-modal alignment appears to require not only sufficient model capacity, but also a high-quality audio front-end, explicit reasoning training, and adequate training data.

B.6 Detailed Results on CoT and ICL

The detailed results of Chain-of-Thought prompting and In-Context Learning are reported in Table 11 and Table 12. Most models benefit from In-Context Learning, with only two exceptions: Qwen2.5-Omni-3B shows degraded performance, potentially due to limited instruction-following

ability, and Qwen3-Omni-Thinking also underperforms, likely because it relies more heavily on its internal thinking patterns rather than externally provided demonstrations. The consistent advantage of in-context exemplars over a model’s native “thinking” behavior further suggests that the primary bottleneck of current omni-LLMs lies in the lack of effective coreference-aware thinking patterns for cross-modality coreference alignment.

C Case Study

To illustrate the effect of coreference-aware thinking patterns of training, we provide a CoT case in Figure 6 and two cases after training in Figure 7 and Figure 8.

D Prompts Details

For reproducibility, we show the prompts in Table 13-Table 18, which consist of data annotation, in-context learning, and GRPO.

Model	Audio-centric			Text-centric			Visual-centric			Overall
	Au	Au→Te	Au→Vi	Te	Te→Au	Te→Vi	Vi	Vi→Au	Vi→Te	
Baichuan-Omni	37.70	29.60	26.20	56.28	26.25	27.71	32.22	21.74	27.80	31.24
Humanomni	33.23	41.09	46.96	54.88	31.56	45.45	56.07	31.30	39.00	41.53
M2-Omni	44.73	37.36	32.91	73.02	28.12	46.32	53.14	31.30	54.05	43.19
Ming-Omni	33.87	41.09	48.24	64.65	36.88	53.68	60.25	36.52	48.26	45.95
MiniCPM	36.42	37.93	28.11	76.28	28.44	35.50	49.79	24.35	51.35	39.67
Ola	49.52	48.56	44.73	56.74	40.94	48.92	54.39	37.39	47.88	47.41
Omni-r1	40.26	44.25	37.38	86.98	36.25	55.41	63.60	39.57	64.86	50.20
OmniVinci	53.04	56.03	56.55	56.28	42.50	47.62	60.67	34.35	42.47	50.20
Qwen2.5-Omni-3B	44.09	42.53	42.81	50.23	37.81	48.92	66.11	33.91	39.38	44.57
Qwen2.5-Omni-7B	43.77	46.26	37.70	83.26	36.56	55.84	63.18	37.39	58.69	49.84
Qwen3-Omni-30B-Cap	51.95	53.05	44.60	52.56	46.25	48.14	50.13	44.78	49.34	49.01
Qwen3-Omni-30B-Ins	55.46	53.33	47.48	64.65	45.31	50.74	54.31	45.65	59.38	52.58
Qwen3-Omni-30B-Thk	63.77	69.14	59.62	74.88	57.50	62.86	63.10	52.17	69.42	63.52
Average										
–	45.22	46.17	42.56	65.44	38.03	48.24	55.92	36.19	50.15	46.83

Table 11: Performance of different models across modalities via Chain-of-Thought. The Au, Te, and Vi are the abbreviations of Audio, Text, and Visual, respectively; X→Y denotes source-to-target modality (e.g., Au→Te: Audio-to-Text).

Model	Audio-centric			Text-centric			Visual-centric			Overall
	Au	Au→Te	Au→Vi	Te	Te→Au	Te→Vi	Vi	Vi→Au	Vi→Te	
Baichuan-Omni	36.42	50.29	46.65	64.65	43.12	45.45	45.61	41.74	49.81	46.64
Humanomni	42.17	44.25	58.47	60.47	38.44	46.75	61.92	41.30	49.03	48.62
M2-Omni	42.49	48.28	53.04	72.56	45.94	48.92	52.72	47.39	53.67	50.93
Ming-Omni	45.69	54.60	58.47	81.86	41.88	59.74	67.36	47.39	54.44	55.71
MiniCPM	45.05	52.87	49.84	80.47	39.38	52.38	47.70	37.39	54.83	50.36
Ola	47.28	55.17	48.56	58.60	45.00	51.08	52.30	39.57	51.35	49.80
Omni-r1	47.92	65.23	60.06	94.88	43.75	61.47	62.34	44.78	69.50	60.09
OmniVinci	48.56	53.45	52.72	51.63	46.56	55.41	56.07	38.70	41.70	49.51
Qwen2.5-Omni-3B	35.46	39.08	38.98	67.44	28.44	36.80	37.66	31.30	40.54	38.78
Qwen2.5-Omni-7B	43.77	58.05	53.35	78.14	40.00	53.68	57.32	40.43	62.55	53.40
Qwen3-Omni-30B-Cap	59.94	64.83	55.14	63.26	50.94	54.63	51.80	52.17	64.40	57.60
Qwen3-Omni-30B-Ins	61.53	65.11	56.74	66.98	55.94	60.26	53.47	50.00	66.72	59.75
Qwen3-Omni-30B-Thk	59.30	68.56	64.73	68.37	55.62	61.99	63.10	55.65	69.81	62.99
Average										
–	47.35	55.37	53.60	69.95	44.23	52.97	54.57	43.68	56.03	52.63

Table 12: Performance of different models across modalities via In-Context Learning. The Au, Te, and Vi are the abbreviations of Audio, Text, and Visual, respectively; X→Y denotes source-to-target modality (e.g., Au→Te: Audio-to-Text).

Video



Speech

- Do you hurt anywhere else? - Not right now. - But you used to hurt?

Text Information

Chase was born in 1985 in a small, rural town in Vermont.2015, he was involved in a serious car accident that left him with a severe spinal injury..... Iris was born in 1987 in a bustling urban neighborhood in Chicago.When Chase was involved in the car accident, Iris was by his side every step of the way. She became his primary caregiver.....

Question

what are the voice attributes of the person who experienced the car accident?

- A. male voice, sad, raspy, and quiet
- B. female voice, calm, quiet, and soft
- C. male voice, calm, deep, and loud
- D. female voice, loud, and emotional

Answer

To solve this problem, we analyze the voice attributes of the person who experienced the car accident (Chase) using the video, text, and provided options.

Step 1: Identify the Subject

From the additional context, Chase is the character who suffered a car accident (leading to a spinal injury). In the video, the patient in the hospital bed (representing Chase's experience) engages in dialogue with others.

Step 2: Analyze Voice Characteristics in the Video

The most relevant audio segment involves a patient in the hospital bed speaking. Key observations:

Gender: The voice is distinctly female (higher pitch, softer tonal quality typical of a young girl struggling to communicate).

Tone: The delivery is calm and quiet, not agitated, loud, or emotionally charged. The tone is measured and gentle, reflecting vulnerability without distress.

Intensity: The speaker's volume is quiet and soft, with no harshness or raspiness, aligning with her physical and psychological state (post-accident recovery).

Step 3: Evaluate the Options

Option A: Male voice, sad, raspy, and quiet -> Incorrect (voice is female, not male).

Option B: Female voice, calm, quiet, and soft -> Matches exactly (female gender, calm tone, quiet volume, and soft delivery).

Option C: Male voice, calm, deep, and loud -> Incorrect (voice is female, not male; volume is quiet, not loud).

Option D: Female voice, loud, and emotional -> Incorrect (voice is quiet, not loud, and lacks overt emotion).

Conclusion

Based on the video's audio and contextual clues, only Option B accurately describes the voice attributes of the person who experienced the car accident.

Answer: B

Correct Alignment Process

<Reasoning Details>:

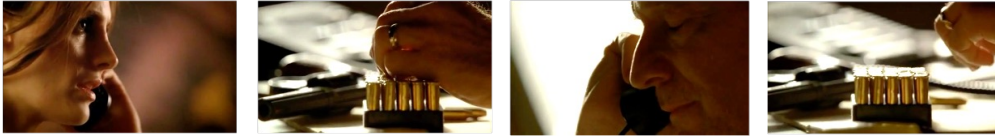
1. The question refers to the person who experienced a car accident, which the biography explicitly identifies as Chase. His accident in 2015 left him with a spinal injury, altering his physical and psychological state.
2. Two individuals are present in the scene: a woman lying on the bed and a man seated in a chair. Chase corresponds to the man seated in the chair.
3. Chase asks, "Do you hurt anywhere else?" and "But you used to hurt?", a calm, measured inquiries.

<Final Answer>:

male voice, calm, deep, and loud.

Figure 6: Case study on text→audio task via Qwen3-Omni-30B-A3B-Thinking via CoT. The green denotes important information, the blue denotes the right choice, and the red denotes the error rationales. The model fails to establish the correct bridge between the textual information (the person who experienced the car accident is a man) and the visual information (the man is sitting on the chair), and therefore cannot answer the question correctly. The correct process should first identify Chase in the text and extract that this person is a man, then locate his position in the video, and finally link his speech in the audio.

Video



Speech

I remember you, Detective Raglan.
Listen, I.

We need to talk about your mother's case.
There's something you don't know.

Text Information

Beckett was born in 1985 in the coastal city of Wilmington, North Carolina, to a family of artists and educators. Raglan was born in 1972 in the bustling city of Mumbai, India, to a family of modest means.

Question

What are the voice attributes of the person with blonde hair?

- A. female voice, calm, thick, and loud volume
- B. male voice, calm, thick, and soft volume
- C. female voice, calm, hoarse, and soft volume.
- D. male voice, calm, hoarse, and soft volume

Answer

<think>

The question asks about the voice attributes of the person with blonde hair. From the text, the dialogue reveals a conversation between two individuals: one with blonde hair and another with gray hair. The blonde-haired individual is the one speaking the first line, "I remember you, Detective Raglan." This line is delivered in a calm, soft voice, as indicated by the question description of the voice as "calm, soft volume." The blonde-haired person is a woman, as the text biography for Beckett states she was born in 1985 in Wilmington, North Carolina, and was the eldest of three siblings. Her mother, a former schoolteacher, and her father, a sculptor, shaped her into a thoughtful and introspective individual. Beckett's personality is described as someone who listens more than she speaks, with a serious expression during conversations, reflecting her thoughtful nature.

So, the voice attributes of the person with blonde hair are:

- Female voice - Calm - Soft volume

Therefore, the final answer is: C. female voice, calm, soft volume

</think>

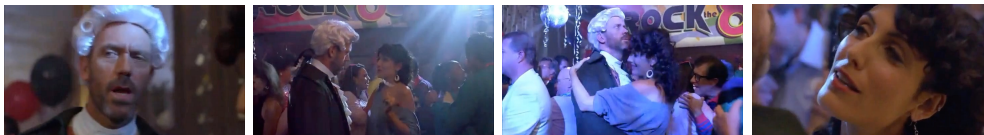
<answer>

The final answer is: C

</answer>

Figure 7: Case study on visual→audio task via Qwen2.5-Omni-7b after training. The blue option denotes the right choice.

Video



Speech

Sorry, we were just about to hit the floor.
Thank you.
Your leg?
I'm okay, as long as you don't expect rhythm.
You remember the last time we danced
Med school. The week we met.

Text Information

Dr. Gregory House was born in 1960 in New York City His education at Princeton University and later at the University of Chicago Medical School solidified his reputation as a brilliant but socially awkward physician.....Jordan: Born in 1985 in Tokyo, Japan, Jordan grew up in a modest family where her parents worked in the fashion industry. From an early age.....

Question

what medical school did the person attend who said 'med school. the week we met'?

- A. university of tokyo medical school
- B. university of mumbai medical school
- C. princeton university medical school
- D. university of chicago medical school

Answer

<think>

Based on the video and text information, the person who said \"med school the week we met\" is Dr. Gregory House, as indicated by the text. House is described as a character from the video who is wearing a white wig and a black coat, and he is interacting with a woman in a blue dress. The text also mentions that House attended the University of Chicago Medical School, which aligns with the information provided. Therefore, the correct answer is D. university of chicago medical school.

</think>

<answer>

D

</answer>

Figure 8: Case study on audio→text task via Qwen2.5-Omni-3b after training. The blue option denotes the right choice.

Prompt
Audio Annotation
<p>Is the voice smooth? Is the voice rough? Is the voice nasal? Is the pitch high-pitched? Is the pitch low-pitched? Does the speech sound happy? Does the speech sound angry? Does the speech sound calm?</p>
Single Video Frame Annotation
<p>Describe the sequence of images in a detailed and continuous narrative, focusing on the following key aspects:</p> <ol style="list-style-type: none"> 1. People: a. Describe the individuals in each image (clothing, appearance, expressions, body language). b. Track any changes in their appearance or actions across images. 2. Background and Environment: a. Include detailed descriptions of the setting (e.g., buildings, nature, objects). b. Provide spatial context (left, right, foreground, background). c. Mention any visible objects and their interaction with people. 3. Sequence and Time: a. Follow a chronological order, indicating how the scene evolves. b. Ensure the narrative flows naturally, showing changes over time. c. Do not describe this image by image. <p>Make sure the description is rich in detail, including both human and environmental elements, while maintaining clarity and coherence</p>
Multiple Video Frames Combination
<p>You will receive an overall description of a video along with descriptions of individual segments of the video. Your task is to integrate these into a comprehensive and detailed description. Follow the rules below:</p> <ol style="list-style-type: none"> 1. Overall Video Description: This description provides the general background, theme, and main content of the video. 2. Segment Descriptions: These provide detailed descriptions of each segment of the video, which may contain specific events or scenes. <p>Please generate the final, complete description as follows:</p> <ul style="list-style-type: none"> - Integrate the overall description with the individual segment descriptions, ensuring the final output covers all aspects of the video. - If there are any inconsistencies between the overall description and the segment descriptions, ignore the inconsistent parts and only retain the consistent and accurate information. - The final description should be coherent, and events should be presented in chronological order, ensuring a logical flow throughout.
Person Description
<p>There is a video with one person is speaking. Please describe the {external knowledge} in detail who is speaking, including characteristics and clothing.</p> <ol style="list-style-type: none"> 1. Do not describe the background. 2. Only output the {external knowledge}. 3. Use '{external knowledge}' as the beginning. 4. If there is no person speaking, directly output <failure>.
Text Biography Generation
<p>Please generate detailed biographies for the following characters based on their provided background information (name + apparel). For each character, create a unique and detailed life story, highlighting their experiences, relationships, challenges, and achievements, including but not limited to:</p> <ol style="list-style-type: none"> 1. Place of birth and upbringing: Where they were born and raised, family environment, education background, etc. 2. Significant life events: Major life events or turning points that shaped their life. 3. Career and achievements: Their career path, important achievements, and contributions. 4. Relationships: Key relationships with others, such as family, friends, enemies, or partners. 5. Personality and psychological development: Their personality traits and any psychological or emotional growth. <p>Ensure the biography covers all provided background details, and the name used in the biography must match the one provided.</p> <p>The format is: Name + Biography. <background information>:</p>

Table 13: Prompt for modality annotation.

Prompt
Audio Double Check
<p>Analyze the descriptions and identify any audio that stands out due to unique features such as:</p> <ol style="list-style-type: none"> 1. Gender: Unusually male or female voice. 2. Emotion: Distinct emotional tone (e.g., anger, sadness). 3. Audio Quality: Notable differences in speed, pitch, or volume. 4. Other Unusual Traits: Background noise, interruptions, etc. <p>If there are any audio is different, output this name(s) with the distinct feature and briefly explain what makes it different. The format is: <name + order> + <distinct feature> + <reason>.</p> <p>If all the descriptions are similarly, output <failure>.</p>
Video Double Check: Video Description
<p>Describe the video in a detailed and continuous narrative, focusing on the following key aspects:</p> <ol style="list-style-type: none"> 1. People: a. Describe the individuals in each image (clothing, appearance, expressions, body language). b. Track any changes in their appearance or actions across images. 2. Background and Environment: a. Include detailed descriptions of the setting (e.g., buildings, nature, objects). b. Provide spatial context (left, right, foreground, background). c. Mention any visible objects and their interaction with people. <p>Make sure the description is rich in detail, including both human and environmental elements, while maintaining clarity and coherence</p>
Video Double Check: Double Check
<p>You are a strict Video Description Consistency Checker. You will be given two texts, description1 and description2, which are opposed to describe the same video. Your job is to determine whether they are consistent at the core factual level. If there is any fundamental mismatch, output exactly failure. Otherwise output exactly pass.</p> <p>Definition of Fundamental Conflict (any one ⇒ failure):</p> <ol style="list-style-type: none"> 1) People conflicts: significantly different number of main people; clearly contradictory identity/type (e.g., adult man vs. child girl); mutually exclusive key actions/events (e.g., driving vs. cooking at a desk). 2) Background conflicts: clearly different main scene type (e.g., kitchen vs. highway); strongly contradictory environment (e.g., heavy rain outdoors vs. quiet indoor office); incompatible main objects/interactions. 3) Narrative conflicts: overall story/sequence cannot reasonably align as the same video; descriptions refer to different scenarios/storylines. <p>Allowed Differences (do NOT count as fundamental conflicts):</p> <ol style="list-style-type: none"> 1) Different level of detail (one richer, one concise). 2) Minor omissions or reordering without contradicting main events. 3) Small secondary-detail differences (e.g., unmentioned background object, ambiguous colors/positions) when core people/setting/actions match.
Person Double Check
<p>Analyzes the consistency of descriptions for each name and outputs whether they are correct or not.</p> <p>Rules:</p> <ol style="list-style-type: none"> 1. If the same name has multiple completely different descriptions, choose the description that appears most frequently as the final description. 2. If any description contains terms like 'failed', 'cannot describe', or anything that indicates the description is not valid, directly output <Failed>. 3. If different names have the same description (roughly similar), directly output <Failed>. 4. If there no failed casess, output the name and its description. The format is: name + description.

Table 14: Prompt for modality annotation double check.

Prompt

QA Generation

Design a question where the model infers **factual description** from **visual actions**, using the following information:

- **Overall Descriptions**
- **person biography**
- **object biography**

Consider these aspects:

1. **Visual Actions**:

- These include body language, gestures, and facial expressions.

2. **Factual description**:

- These refer to stated details such as names, dates, locations, occupations, and major life events (e.g., education, career milestones, relationships).

Design a question to infer **factual description** from **visual actions**.

If you cannot design an appropriate question, directly output “<failure>”.

QA Double Check: Unique

Task: Read the Question, the text biography. The Question ask somethings about a person or a object. Your job is to find the exact name(s) of the character(s) in the visual description who satisfy the description.

Input:

Question: {question}

Visual description: {video des}

Person description: {person des}

Output whether only one the person or object who match the description in the Question.

Output: yes or no

QA Double Check: Answerable

Task: Read the Question, the Answer, the text biography, the Person description, the Visual description. Your job is to determine whether this Answer is correct based only on the provided Context.

Input:

Question: {question}

Answer: {answer}

text biography: {person biography}

Visual description: {video des}

Person description: {person des}

Rules:

(1) You should first indentify the person or object according to the question, visual description.

(2) Secondly, you should extract the name of the person according to the person description or object according to the visual description.

(3) Thirdly, you should find this person in text biography or object in object biography.

(4) Then, you should answer this question based on the biography.

(5) Finally, you should compare the similarity between two answers and determine whether the answer can be correct to this quesiton.

Output:

correct or non-correct.

Option Generation

Task: Read the Question, the Answer, the text biography, the Person description, the Visual description. Your job is to determine whether this Answer is correct based only on the provided Context.

Input:

Question: {question}

Answer: {answer}

text biography: {person biography}

Visual description: {video des}

Person description: {person des}

Rules:

(1) You should first indentify the person or object according to the question, visual description.

(2) Secondly, you should extract the name of the person according to the person description or object according to the visual description.

(3) Thirdly, you should find this person in text biography or object in object biography.

(4) Then, you generate other three <incorrect answer>, the answer can be obtained from the biography or not from.

(5) Finally, you should ensure the <incorrect answer> are wrong for this question.

Output:

<incorrect answer>: xxx

<incorrect answer> xxx

<incorrect answer> xxx

Table 15: Prompt for QA generation. We use the visual→text questions in visual actions as a case study.

Prompt
QA Generation
<p>Task: You will produce detailed step-by-step reasoning (“Reasoning Details”) that explains and leads to the correct Answer. Very important constraints:</p> <ul style="list-style-type: none"> - You must pretend that ALL useful information comes ONLY from: <ul style="list-style-type: none"> (a) what is seen in the video (visual content), and (b) what is known from the biographies (person biography and object biography). - You may explicitly refer to the Question and to the person/object biographies. - You may internally use the Visual description and Person description to help you reconstruct what could be seen in the video and how it connects to the biographies, but you MUST NOT mention or imply that you used any descriptions, text, metadata, or annotations as special inputs. - You MUST NOT say that the Answer was given to you in advance. - You secretly know the correct Answer from the input, and your reasoning must naturally lead to exactly this Answer, but it should look like you inferred it only from watching the video and reading the biographies. <p>Inputs:</p> <ul style="list-style-type: none"> - Question: {question} - Answer: {answer} - text biography: {biography} - Visual description: {video des} - Person description: {person des} <p>Interpretation Rules:</p> <ol style="list-style-type: none"> 1) You MUST act as if all visual information (who/what appears, appearance, position, actions, environment, etc.) comes directly from watching the video. 2) You MUST act as if all background/identity information about people comes from the person biography. 3) You MUST act as if all background/identity information about objects comes from the object biography. 4) The Visual description and Person description are only internal tools to help you reconstruct what WOULD HAVE BEEN seen in the video and how it connects to the biographies. Do not mention or allude to them. 5) In your reasoning, always frame information as observations from: <ul style="list-style-type: none"> - watching the video, and - reading the relevant biography (person or object). <p>Reasoning Details Rules:</p> <ol style="list-style-type: none"> 1) First, determine whether the Question is about a person or an object, and identify which specific person or object it refers to, as if you are using only the video content (internally you may rely on the Visual description, but never mention it). 2) Second, if the Question is about a person, infer their name and identity as if you know it from the person’s biography (internally using the Person description + person biography); if it is about an object, infer its identity/type as if you recognize it from the video and the object biography. 3) Third, match this person to the person biography, or this object to the object biography, and explain how their background or properties are relevant to the Question. 4) Then, using the relevant biography information (person or object) together with what is seen in the video, logically derive and justify the correct Answer step by step. Make the reasoning explicit and multi-step, not just one short sentence. 5) Your final conclusion MUST exactly match the hidden Answer above, but you MUST NOT reveal that this Answer was given to you. 6) The Reasoning Details must be entirely in English. <p>Output format:</p> <p><Reasoning Details>: Your step-by-step reasoning here</p> <p><Final Answer>: State the final answer here, matching the hidden Answer</p>

Table 16: Prompt for CoT generation. We use the visual→text questions as a case study.

Prompt

In-Context Learning

Answer the question based on the video and the text information in options A/B/C/D.

You are REQUIRED to first think step by step and write out your reasoning, and THEN give the final answer.

You MUST strictly follow this output format:

<think>

Your detailed step-by-step reasoning in natural language. Explain how you identify the person or object from the question and video, describe its visual appearance, find the Additional Information using that description, and then use the Additional Information to choose the correct option.

</think>

Final answer only: A / B / C / D

Here are the rules you MUST follow:

(1) Carefully read the question and observe the video. First, use them to identify the person or object mentioned in the question.

(2) Secondly, describe the visual appearance of this person or object in the video.

(3) Thirdly, find the Additional Information about this person or object according to the visual description.

(4) Then, answer the question based on the Additional Information.

(5) All identification, description, retrieval of Additional Information, and reasoning steps MUST be written ONLY inside the <think> section.

(6) Outside the <think> tags, you MUST output ONLY the final answer (for example, just "A", "B", "C", or "D") with no extra words, no explanation, and no additional text.

(7) Do not output anything before <think> or after the final answer. Follow exactly this structure:

<think> ... </think>

A/B/C/D

If you do NOT follow this format, your answer will be considered incorrect.

Table 17: Prompt for in-context learning. We use the visual→text questions as a case study.

Prompt
Contextual Consistency
<p>You are assessing how well the 'hypothesis' text covers the key information from the 'reference' text. Differences in wording or extra details in the 'hypothesis' are fine if the 'reference's' main points are included.:</p> <p>Score based on this coverage:</p> <p>5 points : Hypothesis clearly and accurately reflects significant core themes or key aspects of the reference. It demonstrates a good understanding of a substantial part of the reference material.</p> <p>4 points : Hypothesis reflects some important themes or aspects of the reference. The connection is evident, though perhaps not as comprehensive or central as a 5.</p> <p>2 points : Hypothesis shows a recognizable connection to themes or aspects of the reference, but it might be more superficial, focus on less central points, or only partially grasp a key aspect.</p> <p>1 points : Hypothesis has a tenuous or very limited connection to the reference. It might touch on a peripheral detail or a heavily reinterpreted aspect, but largely misses the main substance.</p> <p>0 points : Hypothesis does not reflect any significant themes or key aspects of the reference, or is on a completely different topic.</p> <p>Example analysis process:</p> <p>Identify main themes and key aspects in 'reference'. Determine if 'hypothesis' connects to or discusses any of these themes/aspects from 'reference'.</p> <p>Judge the strength and relevance of this connection. Is a core part of the 'reference' reflected?</p> <p>Differences are expected; evaluate if the 'hypothesis' still meaningfully reflects some key part of the 'reference'.</p> <p>Assign score based on how well a significant aspect is reflected.</p> <p>reference: {reference}</p> <p>hypothesis: {hypothesis}</p> <p>only return the score number:</p>
Logical Coherence
<p>Please evaluate whether the reasoning path performs effective cross-modal reasoning from visual video content to text (Additional Information), and give a score from 0 to 3.</p> <p>Scoring criteria (1 point for each satisfied item):</p> <ol style="list-style-type: none"> 1. Question-guided visual localization (1 point): The reasoning interprets the question and uses it to identify and localize the relevant person/object in the video, including describing its visual appearance, rather than directly searching the text. 2. Text retrieval via visual description (1 point): Using this visual description, the reasoning explicitly finds the corresponding Additional Information about this person/object in the text (e.g., matching name or described appearance), clearly linking video appearance to a specific text entry. 3. Answer grounded in retrieved text (1 point): The reasoning then uses the retrieved textual Additional Information as the key evidence to select the final option A/B/C/D, with the answer clearly supported by that text and without major unsupported assumptions or contradictions. <p>Scoring rule:</p> <ul style="list-style-type: none"> - Award 1 point for each criterion that is clearly satisfied. - The final score is an integer from 0 to 3. <p>reasoning path: {hypothesis}</p> <p>Only return the score number (an integer from 0 to 3):</p>

Table 18: Prompt for GRPO. We use the visual→text questions as a case study.