BINDING VOICES TO CHARACTERS: DISENTANGLED CROSS-MODAL ALIGNMENT FOR MULTI-SPEAKER PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans effortlessly perceive crowded multi-speaker scenes, resolving *who* is speaking, *when*, *where*, and *what* they say in real time. Replicating this capability in machines, however, remains a fundamental open challenge rooted in cross-modal speaker ambiguity that existing end-to-end models fail to resolve. A natural remedy is to scale training through data engineering. However, scaling training data alone may not sufficiently constrain the correspondence between speech segments and visible speakers, motivating explicit structural supervision for speaker binding. To address these limitations, we propose **CAST** (disentangled **C**ross-modal **A**lignment for multi-**S**peaker percep**T**ion), a framework that endows model with genuine multi-speaker perception capability through two complementary structural interventions. First, Speaker-aware Audio Encoding (SAE) restructures the audio encoder to explicitly segment the audio stream by speaker identity. Second, an Audio-Conditioned Spatial Modulation (ACSM) module performs audio-visual contrastive alignment to generate a Speaker Heatmap Prior, whose spatial tokens are superimposed onto the visual tokens to produce speaker-conditioned visual representations. CAST is then trained end-to-end via supervised fine-tuning followed by reinforcement learning to align perception with structured reasoning. We further release **MSR-QA** (Multi-Speaker Reasoning QA), a large-scale multilingual dataset comprising over **8000+ hours** of effective speech video and **11 million** QA pairs across movies, TV dramas, variety shows, and interview programs. Extensive experiments demonstrate that CAST significantly outperforms existing approaches by up to **34.30** percentage points on cross-speaker temporal and spatial reasoning, while providing transparent and verifiable reasoning processes.

Web Demo: https://cast-iclr.github.io/CAST

1 INTRODUCTION

Matching each voice to the right face at every moment is central to audio-visual understanding, and it is hardest in multi-speaker Video QA, where models must jointly reason about *who* is speaking, *when*, *where*, and *what* they say. With multiple or overlapping speakers, such cross-modal correspondence becomes highly ambiguous, which global feature fusion cannot resolve.

Existing approaches fall short of addressing these ambiguities. Audio-visual QA models (e.g., MUSIC-AVQA (Li et al., 2022)) rely on global pooling, which entangles speaker-specific cues. MLLMs such as Qwen3-Omni (Xu et al., 2025) fuse modalities implicitly through adapters, without explicit or verifiable speaker localization. Localization methods (e.g., EZ-VSL (Mo & Morgado, 2022)) predict sound-source heatmaps but are decoupled from downstream reasoning, answering only *where*, not *who* or *what*. CoT prompting (Wei et al., 2022) yields free-text rationales without spatio-temporal or speaker grounding, and thus remains prone to hallucination.

These limitations highlight three fundamental challenges: **(1) Resolving cross-modal speaker ambiguity.** In complex scenes with multiple or overlapping speakers, machines struggle to bind voices to specific faces and locations. Global or coarse audio-visual alignment conflates sources, depriving models of the precise spatial awareness needed for speaker-specific reasoning. **(2) Ensuring interpretable and verifiable reasoning.** Even when models yield correct answers, their internal processes remain opaque. Ungrounded CoT rationales cannot be verified against temporal, spatial,

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

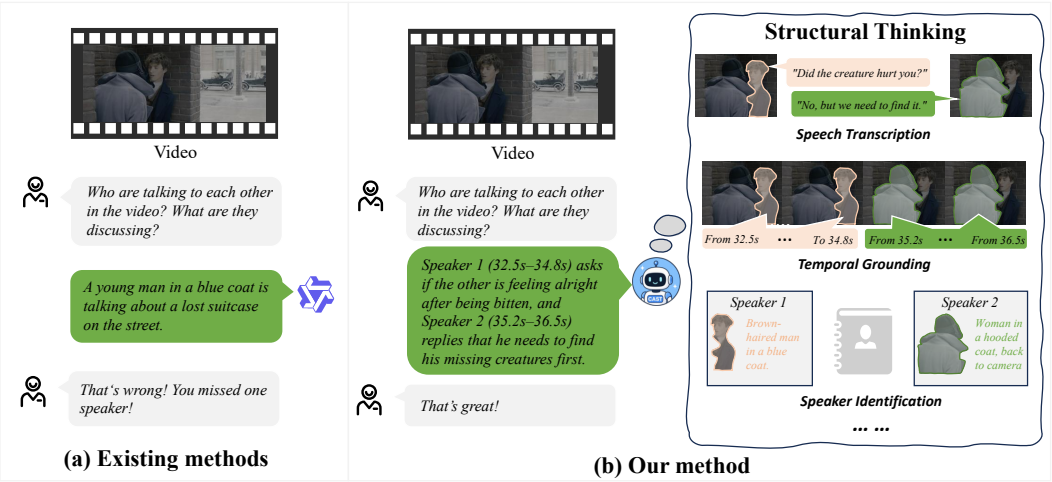

Figure 1: **Comparison between existing methods and CAST.** (a) Existing methods reason in a black-box manner, producing answers that mix up speakers and are hard to verify or debug. (b) CAST breaks multi-speaker understanding into interpretable parts: *temporal grounding* (`<timestamp>`), *speaker identification* (`<speaker>`), and *speech transcription* (`<asr>`), making the reasoning transparent and easy to check.

or acoustic facts, and standard next-token optimization fails to enforce structured multimodal reasoning. This opacity prevents reliable diagnosis and undermines trust in safety-critical applications. **(3) The absence of large-scale structured training data.** Existing datasets lack structured annotations spanning temporal, spatial, and conversational dimensions, depriving models of the fine-grained supervision signal needed to develop verifiable speaker-binding perception.

To address these challenges, we propose **CAST** (disentangled **C**ross-modal **A**lignment for multi-**S**peaker percep**T**ion), a unified framework that endows model with genuine multi-speaker perception capability through two complementary structural interventions. At its core, **Speaker-aware Audio Encoding (SAE)** segments the acoustic stream by speaker identity using special tokens `<|spk|>`, establishing explicit cross-modal attribution between utterances and speaker profiles. Then, **Audio-Conditioned Spatial Modulation (ACSM)** produces pixel-level speaker localization maps through audio-visual token contrastive alignment. Acting as a Speaker Heatmap Prior, these maps dynamically reweight visual features toward active speakers, with the modulated spatial stream subsequently fused into a dual-path architecture alongside a complementary semantic branch. CAST enforces a structural perception CoT format, requiring explicit generation of `<timestamp>`, `<speaker>`, and `<asr>` tokens before the final `<answer>`. To align perception with reasoning, we introduce a Heterogeneous Perception Reward that jointly optimizes localization quality, cross-modal consistency, and QA accuracy. Finally, we release **Multi-Speaker Reasoning QA (MSR-QA)** dataset, a large-scale dataset with structured annotations to support multi-speaker perception research.

Our main contributions can be summarized as follows:

- We propose **CAST**, a multi-speaker perception framework with two structural interventions: SAE to segment the audio stream by speaker identity, and ACSM to produce speaker-conditioned visual representations. CAST outperforms existing approaches by up to 34.3 percentage points on cross-speaker temporal and spatial reasoning.
- We design a post-training pipeline comprising SFT with a Structural Perception CoT format that enforces explicit `<timestamp>`, `<speaker>`, and `<asr>` grounding, followed by RL with a Heterogeneous Perception Reward that jointly supervises temporal, spatial, speaker, and answer quality signals.
- We construct and release **MSR-QA**, a large-scale multilingual dataset comprising over 8000+ hours of effective speech video and 11 million QA pairs across movies, TV dramas, variety shows, and interview programs, providing the first dedicated resource for training and evaluating grounded multi-speaker perception at scale.

2 Related Work

Audio–visual QA and active-speaker detection. MUSIC-AVQA studies questions about audio–visual events (Li et al., 2022), Pano-AVQA introduces spatial grounding in panoramic video (Yun et al., 2021), and AVSD studies audio–visual scene-aware dialogue (Alamri et al., 2019). Speaker-specific questions additionally require associating an utterance with a visible person. AVA-ActiveSpeaker provides face tracks with speaking-activity labels (Roth et al., 2020), while TalkNet models temporal audio–visual evidence for active-speaker detection (Tao et al., 2021). EZ-VSL learns visual sound-source localization (Mo & Morgado, 2022). CAST incorporates related spatial cues into an open-ended QA framework, where attribution and conversational context jointly affect the answer.

Multi-speaker transcription and dialogue understanding. Qwen2.5-Omni and Qwen3-Omni support general audio–visual understanding (Jin Xu, 2025; Xu et al., 2025). Specialized systems model speaker-attributed speech more explicitly. TagSpeech combines multi-speaker ASR, diarization, and temporal grounding, including overlapping-speech settings (Huo et al., 2026). D-ORCA (Tang et al., 2026) introduces a bilingual multi-party dialogue corpus and optimizes speaker attribution, transcription, and temporal boundaries for dialogue-centric captioning (Tang et al., 2026). These methods establish important precedents for structured dialogue outputs. CAST focuses on coupling turn structure with visual feature modulation and using the resulting evidence for question answering.

Structured outputs and reinforcement learning. Chain-of-thought prompting and multimodal variants use intermediate outputs to support answer generation (Wei et al., 2022; Zhang et al., 2023; Shao et al., 2024a). CAST uses a structured evidence trace whose fields can be evaluated individually. GRPO and GSPO provide group-based optimization frameworks for scored model outputs (Shao et al., 2024b; Zheng et al., 2025). CAST applies GSPO to a combination of perception and answer-quality rewards. Temporal, transcription, and attribution objectives have direct precedents in D-ORCA (Tang et al., 2026); the focus here is their integration with turn-aware representations and audio-conditioned visual modulation.

3 Methodology

3.1 Overview

Given a video $\mathcal{V}$, its audio track $\mathcal{A}$, and a question $\mathcal{Q}$, CAST produces a structured evidence trace $\mathcal{Y}_{\text{think}}$ followed by an answer $\mathcal{Y}_{\text{ans}}$. A vision encoder extracts patch features $\mathbf{V} \in \mathbb{R}^{N\times D}$, and an audio encoder extracts acoustic features $\mathbf{A} \in \mathbb{R}^{T\times D}$ within a temporal context. SAE organizes acoustic features into speaker turns. ACSM consumes $\mathbf{V}$ and $\mathbf{A}$ to produce a speaker heatmap $\mathbf{H} \in \mathbb{R}^{h\times w}$ and a modulated visual representation $\hat{\mathbf{V}}$. The LLM backbone Φ_L then receives the fused multimodal context, defined as

$$\mathcal{C} = \text{Fusion}(\mathbf{V}, \hat{\mathbf{V}}) \oplus \tilde{\mathbf{A}} \oplus \mathcal{Q}, \tag{1}$$

where $\oplus$ denotes context assembly, $\tilde{\mathbf{A}}$ is the turn-aware audio sequence, and $\hat{\mathbf{V}}$ is the modulated visual stream. Retaining $\mathbf{V}$ allows the decoder to use objects, scene context, and non-speaking people when answering questions.

The output order gives the conditional factorization

$$p_\theta(\mathcal{Y}_{\text{think}}, \mathcal{Y}_{\text{ans}} \mid \mathcal{C}) = p_\theta(\mathcal{Y}_{\text{think}} \mid \mathcal{C})\, p_\theta(\mathcal{Y}_{\text{ans}} \mid \mathcal{Y}_{\text{think}}, \mathcal{C}). \tag{2}$$

This formulation connects perception and answering through a common generated sequence. The evidence block records predictions about the input, while the answer selects information relevant to the question. Both factors retain access to the multimodal context, so the evidence block is not an exclusive information bottleneck. In particular, the factorization specifies an output dependency, not a claim that the generated text reveals every computation used by the model.

3.2 Speaker-aware Audio Encoding

SAE provides an explicit interface between speech segmentation and multimodal decoding. Let S_k denote a predicted turn segment and $\mathbf{A}_{S_k}$ the corresponding acoustic features. A shared learnable

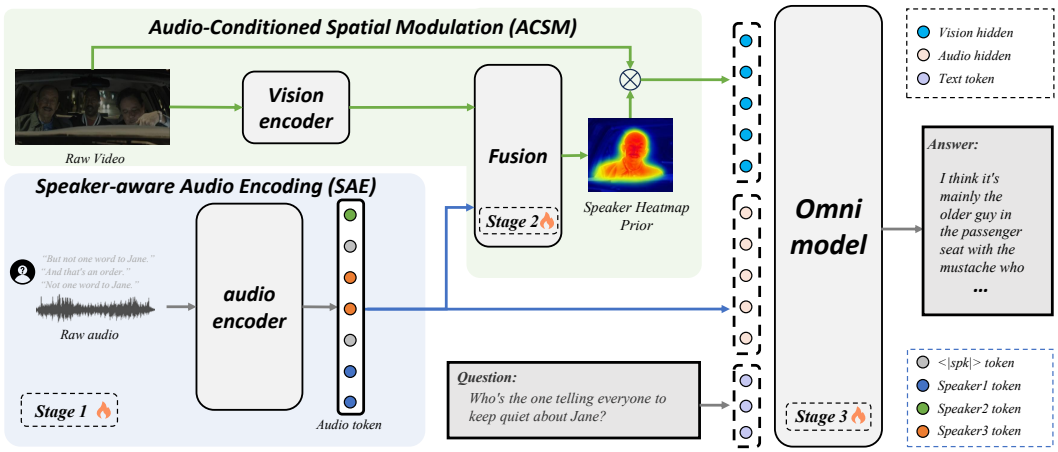

Figure 2: **Overview of the CAST framework and its three-stage training pipeline. Stage 1**: the SAE is continued pre-trained with supervised learning to segment audio by speaker identity and insert <|spk|> delimiter tokens for disentangled attribution. **Stage 2**: the ACSM module performs token-level audio-visual contrastive alignment to generate a Speaker Heatmap Prior, whose spatial tokens are superimposed onto ViT visual tokens to produce speaker-conditioned visual representations. **Stage 3**: the full model undergoes end-to-end post-training via SFT and RL with structured perception supervision.

delimiter $\mathbf{e}_{\text{spk}}$ marks the beginning of each segment:

$$\tilde{\mathbf{A}} = [\mathbf{e}_{\text{spk}}, \mathbf{A}_{S_1}, \mathbf{e}_{\text{spk}}, \mathbf{A}_{S_2}, \ldots]. \tag{3}$$

The resulting sequence preserves acoustic content while exposing boundaries that the language model can use to organize utterances. The supervision for these delimiter tokens is obtained from the speaker-attributed utterance annotations in our data construction pipeline, where a new $\langle|\text{spk}|\rangle$ token is inserted whenever a new speaker turn begins. Appendix C describes the representation interface and its relation to the output fields.

The index k records conversational order, not persistent speaker identity: one person may occupy several non-contiguous segments. The shared delimiter marks boundaries, leaving identity to the acoustic and visual features. This distinction matters for questions about repeated contributions and response order.

SAE does not separate mixed waveforms: simultaneous speakers require more than boundaries in a single sequence. This formulation describes the downstream representation after segmentation and does not assume differentiability of hard boundary decisions.

3.3 Speaker Heatmap Prior via Audio-Conditioned Spatial Modulation

Lightweight projection adapters map audio and visual tokens into a shared perceptual space, where scaled dot-product affinities are computed between every audio segment and every visual patch.

$$\mathbf{a}_i = g_A(\mathbf{A}_i), \qquad \mathbf{v}_j = g_V(\mathbf{V}_j), \qquad s_{ij} = \frac{\mathbf{a}_i^\top \mathbf{v}_j}{\sqrt{d}}. \tag{4}$$

Here d is the projected feature dimension. A symmetric contrastive objective encourages paired audio–visual tokens to align:

$$\mathcal{L}_{\text{align}} = -\frac{1}{T}\sum_{i=1}^{T} \log \frac{\exp(s_{i,j^+}/\tau)}{\sum_j \exp(s_{ij}/\tau)} - \frac{1}{N}\sum_{j=1}^{N} \log \frac{\exp(s_{i^+,j}/\tau)}{\sum_i \exp(s_{ij}/\tau)}, \tag{5}$$

where τ is a temperature and j^+ and i^+ denote assigned positive correspondences. The objective assumes these assignments as supervision inputs: sharing a clip and temporal window constrains candidate positives but does not select a unique speaking-person patch. Several patches can cover

one person, while other patches describe background. The loss therefore specifies conditional token alignment, not an automatic rule for discovering identity labels. The spatial diagnostics assess the resulting prior separately from answer accuracy.

Within each temporal window, the maximum audio affinity supplies the patch score:

$$h_j = \max_{i \in \{1,\dots,T\}} s_{ij}, \qquad \mathbf{H} = \text{reshape}(\mathbf{h}), \qquad \beta = \sigma(\bar{\mathbf{a}}), \tag{6}$$

where $\bar{\mathbf{a}}$ is pooled audio context and σ is an MLP with an output in $(0, 1)$. The maximum allows a patch to respond to a local acoustic segment without averaging that response across every turn. It also compresses the temporal evidence into one map, so $\mathbf{H}$ does not assign a separate persistent identity to each speaker. Appendix B further describes the alignment and modulation interfaces.

A gate $\beta \in (0, 1)$, predicted from pooled audio context, controls residual modulation:

$$\hat{\mathbf{V}} = \mathbf{V} + \beta \cdot (\mathbf{H}_{\uparrow} \odot \mathbf{V}), \tag{7}$$

Here $\mathbf{H}_{\uparrow}$ broadcasts patch scores to feature dimensions. The multiplier $1 + \beta H_j$ exceeds one for positive scores, lies between zero and one when $-1/\beta < H_j < 0$, and can become negative for more negative scores. Thus, an unrestricted signed affinity is not necessarily an amplification-only gate. The prior is neither a calibrated speaking probability nor a segmentation mask. It supplies a spatial cue while the original visual representation remains available to the decoder.

3.4 Structural Perception Chain-of-Thought

CAST does not directly emit an answer. Instead, it is constrained to first produce a structured thinking block $\mathcal{Y}_{\text{think}}$ surrounded by special tokens <think> and </think>, followed by the answer block $\mathcal{Y}_{\text{ans}}$.Inside the thinking block, the model outputs a sequence of turns:

```
<think>
  <turn id="1">
    <timestamp>({start}, {end})</timestamp>
    <speaker>{speaker}</speaker>
    <asr>{asr}</asr>
  </turn>
  ...
</think>
[answer based on think]
```

Each turn must contain exactly three fields. The `<timestamp>` grounds the utterance in time; the `<speaker>` field provides a visual attribute description that is verifiable against the video frame; and the `<asr>` field transcribes the spoken content. The answer $\mathcal{Y}_{\text{ans}}$ is then generated conditioned on this explicit evidence, rather than on raw multimodal embeddings alone.

During training, we enforce this structure through two mechanisms: **(i)** supervised fine-tuning on structured chain-of-thought demonstrations that exhibit the target turn-based format; **(ii)** a rule-based format reward that penalizes malformed or incomplete reasoning chains (see §3.5).

3.5 Heterogeneous Perception Reward

To close the gap between proxy training objectives and the complex multi-component demands of multi-speaker QA, we introduce a **Heterogeneous Perception Reward** that jointly evaluates temporal, semantic, descriptive, and inferential qualities of the generated output. As shown in Figure 3, the reward comprises two tiers: *structural reasoning rewards* derived from automatic metrics, and *naturalistic inference rewards* derived from LLM-based judges.

Temporal-aware reward (R_{IoU}). For each predicted turn, we compute the temporal intersection-over-union between the predicted timestamp interval and the ground-truth interval. For multiple turns, a bipartite matching algorithm aligns predictions to ground truths before averaging (Appendix D).

Speech-aware reward (R_{WER}). The ASR transcript in each turn is compared against the reference via word error rate (Appendix D).

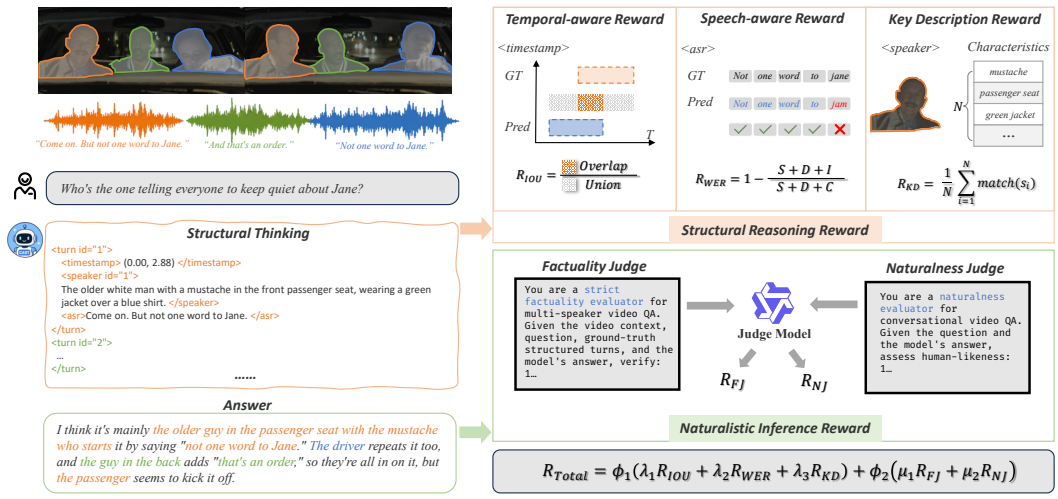

Figure 3: **Structured reasoning and reward design.** (Left) CAST outputs explicit `<timestamp>`, `<speaker>`, and `<asr>` turns grounded in video evidence before synthesizing the answer. (Right) The **Heterogeneous Perception Reward** combines automatic structural metrics (temporal IoU, WER, key-description) with judge-based naturalistic scores (factuality and naturalness).

Key-description reward (R_{KD}). Speaker descriptions are verified against a pre-defined attribute set (e.g., clothing color, facial hair, spatial position, relative age). A lightweight attribute parser extracts predicted attributes and tests semantic equivalence via an embedding similarity threshold. Collectively, these three terms form the **structural reasoning reward**:

$$R_{\text{struct}} = \lambda_1 R_{\text{IoU}} + \lambda_2 R_{\text{WER}} + \lambda_3 R_{\text{KD}}. \tag{8}$$

Factuality judge (R_{FJ}). Beyond automatic metrics, we employ a frozen judge LLM prompted as a strict factuality evaluator. Given the video context, the question, the ground-truth structured turns, and the model's complete output (thinking + answer), the judge scores whether every factual claim in the answer is supported by the provided evidence. The score is normalized to $[0, 1]$.

Naturalness judge (R_{NJ}). A second judge LLM, prompted as a naturalness evaluator, assesses whether the final answer reads as a natural, human-like response to the question, independent of its factual correctness.

Total reward. The final heterogeneous reward is:

$$R_{\text{total}} = \phi_1 R_{\text{struct}} + \phi_2 \big(\mu_1 R_{\text{FJ}} + \mu_2 R_{\text{NJ}}\big). \tag{9}$$

During policy optimization, we sample a group of candidate outputs for each training instance, compute their relative advantages against the group mean, and update the entire perception-reasoning pipeline—including the heatmap generator, spatial modulation gate, and LLM backbone—toward maximizing R_{total}. This end-to-end optimization ensures that the speaker heatmap, audio token boundaries, and reasoning chain co-adapt to serve the ultimate QA objective.

4 Experiments

4.1 Experimental Setup

We evaluate CAST on four multi-speaker benchmarks: AV-SpeakerBench (Nguyen et al., 2026), OmniMMI-SI (Wang et al., 2025), Omni-MMSI, and MSR-QA. Baselines include general-purpose omni-modal models (Qwen2.5-Omni-7B (Jin Xu, 2025), Qwen3-Omni-30B-Instruct Xu et al. (2025), MiniCPM-o-4.5 (Cui et al., 2026), Gemini-3.5-Flash, Gemini-3-Flash-Preview (Comanici et al., 2025), Nemotron-3-Nano-Omni Deshmukh et al. (2026), OmniVinci Ye et al. (2026)) and specialized audio-visual models (R1-Omni-0.5B (Zhao et al., 2025a), HumanOmni-7B (Zhao et al., 2025b),

Table 1: System-level QA comparison across AVSpeaker, OmniMMI-SI, Omni-MMSI, and MSR-QA Bench. Higher is better; bold indicates the best value.

Methods	AVSpeaker ↑	OmniMMI-SI ↑	Omni-MMSI ↑	MSR-QA Bench ↑
Ming-Lite-Omni	41.28	55.38	25.68	14.80
Qwen2.5-Omni-7B	45.92	75.38	26.92	36.75
Qwen3-Omni-30B-Instruct	53.02	75.38	32.35	58.05
MiniCPM-o-4.5	48.72	78.46	25.68	25.24
Gemini-3.5-Flash	49.47	74.62	39.60	55.30
Gemini-3-Flash-Preview	53.80	76.92	42.03	58.00
R1-Omni-0.5B	0.16	0.15	0.10	0.00
HumanOmni-7B	40.88	27.69	11.49	0.22
Nemotron-3-Nano-Omni	47.14	67.69	25.96	1.82
OmniVinci	41.10	60.00	25.05	23.64
D-ORCA-8B	35.93	61.54	24.83	46.33
VibeVoice multi-talker + Qwen3-Omni	51.27	69.50	40.79	–
Qwen3-Omni + MSR-QA	56.14	78.50	43.74	62.37
CAST	**59.14**	**84.50**	**52.74**	**67.37**

Table 2: **Detailed results on MSR-QA.** ↑ indicates higher is better, ↓ indicates lower is better. The best results are highlighted in **bold**.

Methods	Speaker		Content		Temporal	Overall
	SDM ↑	SAA ↑	WER ↓	TAJ ↑	tIoU ↑	Score ↑
Ming-Lite-Omni	20.18	4.78	93.81	37.31	3.25	14.80
Qwen2.5-Omni-7B	21.20	57.60	95.95	81.62	17.90	36.75
Qwen3-Omni-30B-Instruct	45.52	48.75	46.74	95.97	21.98	58.05
MiniCPM-o-4.5	20.39	4.09	80.67	42.40	1.81	25.24
Gemini-3.5-Flash	51.81	51.75	51.45	93.83	25.59	55.30
Gemini-3-Flash-Preview	62.66	55.28	51.94	91.21	29.85	58.00
R1-Omni-0.5B	0.00	0.00	100.00	0.00	0.00	0.00
HumanOmni-7B	0.00	0.12	99.85	0.68	0.08	0.22
Nemotron-3-Nano-Omni	10.05	0.50	99.60	3.94	0.42	1.82
OmniVinci	10.62	35.61	89.54	60.02	3.71	23.64
D-ORCA-8B	54.67	19.78	47.37	93.26	22.23	46.33
CAST	**64.22**	**65.88**	**46.32**	**94.48**	**43.54**	**67.37**

D-ORCA-8B (Tang et al., 2026), VibeVoice multi-talker (Peng et al., 2025)+Qwen3-Omni). General-purpose models are evaluated under a unified one-shot protocol with a format-specifying demonstration; specialized models are run with their native inputs and prompting formats from the official implementations. To accommodate differences in native output formats across models, we use model-specific output parsing and answer extraction while applying the same underlying correctness criteria, thereby avoiding false negatives caused purely by formatting differences. Benchmark scores are in Table 1; Table 2 further reports SDM, SAA, WER, TAJ, and tIoU on MSR-QA, with the Overall Score judged by GPT-4o. Dataset and evaluation details are in Appendix E and D.

4.2 Main Results

Benchmark comparison. Table 1 shows that CAST reaches 67.37 on MSR-QA Bench, exceeding D-ORCA by 21.04 points and Qwen3-Omni by 9.32 points. On AVSpeaker, CAST improves over D-ORCA by 23.21 points; on OmniMMI-SI, it exceeds Qwen2.5-Omni by 9.12 points. The results indicate consistent gains across all four benchmarks.

Evidence quality. CAST's WER is 46.32, compared with 46.74 for Qwen3-Omni and 47.37 for D-ORCA, corresponding to reductions of 0.42 and 1.05 points (Table 2). Its tIoU is 43.54, exceeding Qwen3-Omni by 21.56 points, while its TAJ is 94.48, 1.49 points lower than Qwen3-Omni. These diagnostics show stronger transcription and temporal localization performance, while maintaining competitive answer consistency.

Table 3: **Component ablation study of CAST on MSR-QA.** We remove each key component from the full model to evaluate its contribution. ↑ indicates higher is better, ↓ indicates lower is better.

Variant	Speaker		Content		Temporal	Overall
	SDM↑	SAA↑	WER↓	TAJ↑	tIoU↑	Score↑
CAST (Full)	**64.22**	**65.88**	**46.32**	**94.48**	**43.54**	**67.37**
w/o SAE	52.15	42.36	58.47	85.23	35.18	58.92
w/o ACSM	56.38	48.72	54.19	87.65	26.43	60.15
w/o SFT	59.47	56.23	51.88	90.15	37.62	62.48
w/o RL	61.85	62.47	48.56	89.37	40.28	64.92

Table 4: **Reward ablation study of CAST on MSR-QA.** We remove each reward component from the full heterogeneous perception reward to evaluate its effect on structural reasoning and final-answer quality. ↑ indicates higher is better, ↓ indicates lower is better.

Variant	Speaker		Content		Temporal	Overall
	SDM↑	SAA↑	WER↓	TAJ↑	tIoU↑	Score↑
Full Reward	**64.22**	**65.88**	**46.32**	**94.48**	**43.54**	**67.37**
w/o Temporal-aware Reward	63.15	64.36	47.18	92.23	30.47	64.82
w/o Speech-aware Reward	63.48	64.72	55.63	93.15	42.38	64.58
w/o Key Description Reward	55.37	52.46	47.88	92.48	41.56	64.23
w/o Naturalistic Inference Reward	63.85	64.58	46.89	85.37	42.68	64.92

4.3 Ablation Study

Component Ablation. Table 3 reports the contribution of each CAST component. Removing SAE causes the most severe degradation in speaker attribution (SAA drops from 65.88 to 42.36), confirming that explicit speaker delimiter tokens are essential for disentangling multi-speaker audio. Without ACSM, temporal grounding collapses (tIoU drops from 43.54 to 26.43) and speaker description accuracy declines (SDM -7.84), validating that the Speaker Heatmap Prior provides critical spatial evidence for active-speaker localization. Omitting SFT uniformly degrades all metrics, indicating that supervised pre-training on structured thinking chains provides necessary initialization before RL optimization. Removing RL also hurts performance, particularly TAJ (-5.11) and Overall Score (-2.45), demonstrating that the heterogeneous perception reward is required to align perception with reasoning beyond what SFT alone can achieve.

Reward ablation. Table 4 dissects the contribution of each tier in the heterogeneous perception reward. Removing the Key Description Reward causes the most severe degradation in speaker-related metrics (SDM -8.9, SAA -13.4), confirming that direct optimization of visual attribute accuracy is necessary to ground speaker identities. Without the Naturalistic Inference Reward, thinking-answer faithfulness collapses (TAJ drops 9.1 points), validating that judge-based supervision is essential to prevent hallucination beyond generated reasoning evidence. Omitting the Temporal-aware Reward primarily degrades temporal grounding (tIoU -13.1), while removing the Speech-aware Reward causes significant ASR degradation (WER $+9.3$). Each reward term addresses a distinct failure mode: Key Description grounds *who*, Temporal-aware grounds *when*, Speech-aware grounds *what*, and Naturalistic Inference enforces end-to-end faithfulness.

Table 5: **Evaluation of the Speaker Heatmap Prior.** We measure active-speaker selection accuracy, energy overlap with GT box, peak pointing accuracy, and heatmap-GT IoU.

Method	Speaker Sel. Acc. ↑	Heatmap Energy Overlap ↑	Pointing Acc. ↑	Heatmap IoU ↑
Visual-only Heatmap	19.82	19.27	23.15	9.31
Audio-shuffled Heatmap	16.94	19.04	21.33	9.17
ACSM	**38.43**	**38.54**	**35.65**	**56.27**

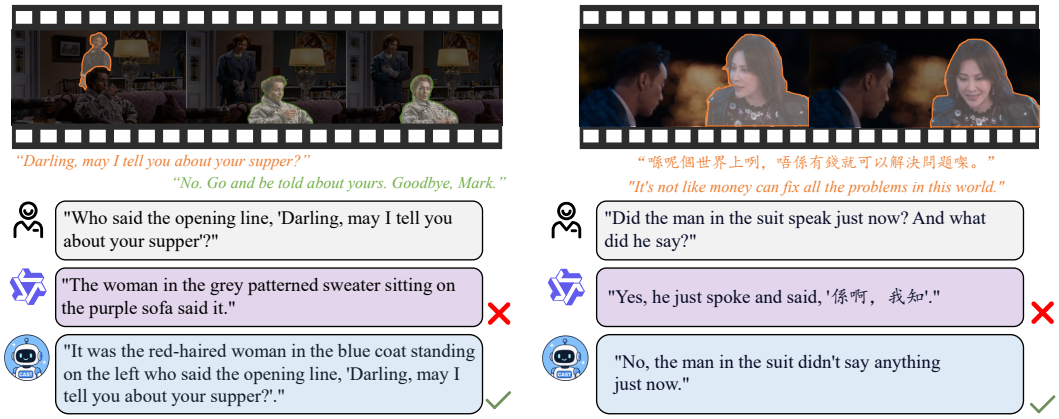

Figure 4: **Qualitative comparison.** (Left) Baseline misattributes speech to a silent but visually salient speaker; CAST correctly grounds via causal alignment. (Right) Baseline hallucinates speech for a non-speaking subject; CAST reports no utterance, enforcing audio-visual consistency.

4.4 Analysis of Speaker Heatmap Prior

Table 5 evaluates the ACSM module against two control variants: Visual-only Heatmap, which removes audio conditioning, and Audio-shuffled Heatmap, which disrupts the correspondence between audio and visual inputs. ACSM achieves the highest performance across all four metrics, reaching 38.43 Speaker Selection Accuracy, 38.54 Energy Overlap, 35.65 Pointing Accuracy, and 56.27 Heatmap IoU. The Audio-shuffled variant performs worse than the Visual-only variant on all four metrics, indicating that misaligned audio does not provide useful localization cues and may even interfere with speaker selection. Meanwhile, the Visual-only variant remains substantially below ACSM, showing that audio conditioning provides complementary information for identifying and localizing active speakers.

4.5 Qualitative Results

Figure 4 shows two representative failure cases. In the left scene, the baseline incorrectly attributes the opening line to the woman in the grey patterned sweater sitting on the purple sofa," conflating visual salience with actual speech; CAST grounds it instead to the red-haired woman in the blue coat standing on the left," consistent with speaker identity, spatial position, and acoustic evidence. In the right scene, the baseline hallucinates speech from the man in the suit that is unsupported by the audio. CAST answers that he "didn't say anything just now," showing that the heterogeneous reward enforces cross-component consistency: when no audio segment is attributed to a detected speaker, the answer must reflect this negative evidence. Together, these cases show that CAST's causal alignment prevents both speaker misattribution and content hallucination in multi-speaker scenes.

5 Conclusion

We present CAST, a framework that unifies causal audio-visual alignment with structured chain-of-thought reasoning for multi-speaker video QA. At its core, a Speaker-Aware Audio Encoding disentangles multi-speaker audio through explicit speaker delimiters; an Audio-Conditioned Spatial Modulation module generates a causal heatmap prior that grounds acoustic signals to precise visual locations; and a structural perception chain-of-thought enforces explicit reasoning over *who*, *when*, *where*, and *what* before producing an answer. The entire pipeline is optimized end-to-end through a heterogeneous perception reward that jointly evaluates temporal grounding, semantic fidelity, speaker description, and answer faithfulness. To enable rigorous evaluation, we introduce MSR-QA, a large-scale benchmark with structured annotations for multi-speaker reasoning. Extensive experiments demonstrate that CAST establishes a new state of the art on multi-speaker video understanding benchmarks, substantially outperforming existing approaches across structured perception, speaker attribution, and question answering. All the code and dataset will be released in *Github* and *Huggingface*.

REFERENCES

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7550–7559. IEEE, 2019.

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. *arXiv preprint arXiv:2502.13923*, 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.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Arushi Goel, et al. Nemotron 3 nano omni: efficient and open multimodal intelligence. *arXiv preprint arXiv:2604.24954*, 2026.

Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, et al. Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario. *arXiv preprint arXiv:2104.03603*, 2021.

Mingyue Huo, Yiwen Shao, and Yuheng Zhang. Tagspeech: End-to-end multi-speaker asr and diarization with fine-grained temporal grounding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 41847–41862, 2026.

Jinzheng He Hangrui Hu Ting He Shuai Bai Keqin Chen Jialin Wang Yang Fan Kai Dang Bin Zhang Xiong Wang Yunfei Chu Junyang Lin Jin Xu, Zhifang Guo. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19086–19096. IEEE, 2022.

Jiaxuan Liu, Yang Xiang, Han Zhao, Xiangang Li, and Zhenhua Ling. Funcineforge: A unified dataset toolkit and model for zero-shot movie dubbing in diverse cinematic scenes. *arXiv preprint arXiv:2601.14777*, 2026.

Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *European Conference on Computer Vision*, pp. 218–234. Springer, 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2272–2283, 2026.

Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, Li Dong, Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, et al. Vibevoice technical report. *arXiv preprint arXiv:2508.19205*, 2025.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4492–4496. IEEE, 2020.

Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. Canary-1b-v2 & parakeet-tdt-0.6 b-v3: Efficient and high-performance models for multilingual asr and ast. *arXiv preprint arXiv:2509.14128*, 2025.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024a.

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

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Changli Tang, Tianyi Wang, Fengyun Rao, Jing Lyu, and Chao Zhang. D-orca: Dialogue-centric optimization for robust audio-visual captioning. *arXiv preprint arXiv:2602.07960*, 2026.

Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 3927–3935, 2021.

Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. Omni-immi: A comprehensive multi-modal interaction benchmark in streaming video contexts. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18925–18935. IEEE, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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*, 2025.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Zhen Wan, Jinchuan Tian, An-Chieh Cheng, Ligeng Zhu, Yuanhang Su, Yuming Lou, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. In *International Conference on Learning Representations*, volume 2026, pp. 56101–56138, 2026.

Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, et al. M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6167–6171. IEEE, 2022.

Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360 videos. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2011–2021. IEEE, 2021.

Youliang Zhang, Zhaoyang Li, Duomin Wang, Deyu Zhou, Zixin Yin, Xili Dai, Gang Yu, Xiu Li, et al. Speakervid-5m: A large-scale high-quality dataset for audio-visual dyadic interactive human generation. In *International Conference on Learning Representations*, volume 2026, pp. 117896–117926, 2026.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

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.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025. URL https://arxiv.org/abs/2507.18071.