Who Says What: Symbolic Trimodal Binding Mechanisms in Audio-Visual LLMs

Jihoo Jung[1] **Youngjoon Jang**[1,2] **Joon Son Chung**[1]
[1]KAIST [2]VGG, University of Oxford

Abstract

Current Audio-Visual LLMs (AVLLMs) struggle with reasoning over videos featuring multi-speaker dialogues. In such videos, resolving "who says what" is crucial, which necessitates trimodal (text-audio-visual) binding. Motivated by these challenges, we systematically investigate how this trimodal binding is achieved in AVLLMs. Specifically, we identify emergent symbolic trimodal binding mechanisms in AVLLMs that utilize modality-specific symbolic variables. By encoding auditory and visual components into symbolic variables-capturing temporal utterance sequences and spatial entity coordinates, respectively-the model establishes cross-modal linking within this abstract space. Crucially, we reveal that when trimodal binding fails, the breakdown predominantly stems from misaligned audio-visual connections. To overcome this bottleneck, we introduce an audio-visual prompting method utilizing an off-the-shelf Active Speaker Detection (ASD) model. By simply overlaying visual bounding boxes on active speakers, this training-free approach yields immediate performance gains across four conversation-centric benchmarks. Moreover, lightweight fine-tuning of fewer than 300 steps on these ASD-prompted-videos extends these gains to three general AV benchmarks, suggesting the generalizability of our method. Our code is available at `https://github.com/kaistmm/Who-Says-What`.

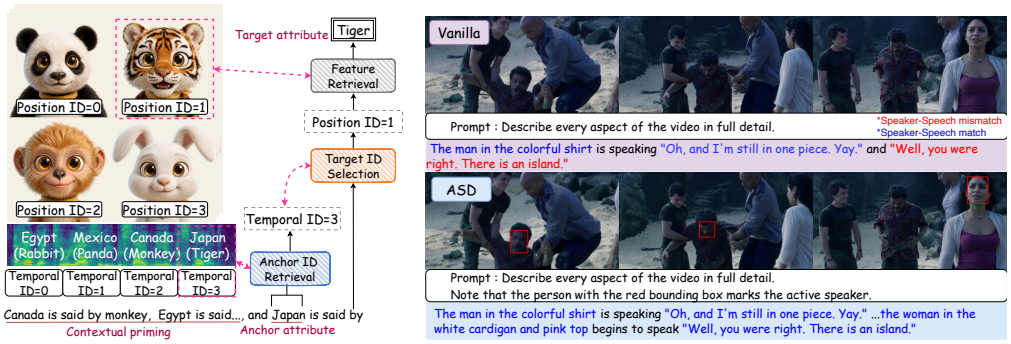
1 Introduction

Conversation-rich videos, including films, television shows, and daily vlogs, account for a significant fraction of contemporary video data. Comprehending such videos relies on the ability to resolve "who says what." This challenge can be framed as *binding problem*-the ability to associate multiple distinct features that belong to the same entity [16, 21]. For instance, resolving what "the man in the gray shirt" is saying requires the model to form robust in-context associations by jointly binding the textual description, the corresponding visual subject, and the target acoustic utterance.

The ability of neural networks to solve this binding problem has received substantial attention in the era of Large Language Models (LLMs) [14, 39, 40, 16, 17]: in LLMs it involves associating entities with their attributes in lengthy texts, while in Vision-Language Models (VLMs) it poses a cross-modal challenge of aligning textual descriptions with specific visual objects. A key finding is that models utilise content-independent symbolic IDs to track and parse objects [16, 17, 56, 1]: rather than processing associations directly in semantic space, models convert raw content into abstract symbolic representations. These IDs allow features belonging to the same entity to be linked together, thereby enabling the model to retrieve the necessary semantic details. For instance, [24] demonstrates that VLMs represent visual objects using symbolic IDs denoting spatial layout (e.g., left or right), independent of visual traits such as color. These symbolic IDs are invoked when processing textual descriptions to direct attention toward the corresponding spatial locations.

Despite recent progress in understanding unimodal and bimodal binding in LLMs and VLMs, it remains an open question how Audio-Visual LLMs (AVLLMs) resolve the more complex trimodal

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Symbolic trimodal binding mechanisms. (b) Audio-visual prompting with Active Speaker Detection.

Figure 1: (a) Trimodal binding consists of three stages: anchor ID retrieval, target ID selection, and feature retrieval. (b) The vanilla model fails to accurately associate the speaker to their corresponding utterance, whereas the ASD method successfully establishes this connection.

(text-audio-visual) binding. This challenge becomes markedly pronounced in conversation-rich, multi-speaker videos, where binding errors translate into speaker-attribution failures, such as assigning an utterance to the wrong visible person. Given that current AVLLMs struggle with reasoning over such videos [46, 7, 36], it is critical to uncover the mechanisms underlying trimodal binding. Understanding these mechanisms is essential for diagnosing the exact point of failure: whether the bottleneck stems from text-to-audio/visual grounding, or from direct audio-visual alignment.

In this work, we investigate the internal mechanisms of trimodal binding in conversation-rich multi-speaker videos using four recent AVLLMs: video-SALMONN2+ (7B) [45], Qwen2.5-Omni (3B, 7B) [53], and MiniCPM-o-4.5 (9B) [11]. We specifically focus on "single-shot" videos-where active and inactive speakers appear simultaneously within each frame-as opposed to videos showing only a single active speaker at a time. Such scenarios are of particular interest because the co-occurrence of multiple candidates complicates the binding task, causing AVLLMs to struggle significantly [7]. To systematically explore this, we build a toy dataset by simulating single-shot multi-speaker videos featuring four animal speakers, each delivering a separate utterance (Fig. 1a). Using this setup, we formulate a trimodal binding task: given a text prompt describing either an acoustic attribute (e.g., a spoken utterance) or a visual attribute (e.g., animal species) as the *anchor attribute*, the model is required to identify the corresponding *target attribute* in the opposite modality.

Our findings are three-fold. First, we demonstrate that AVLLMs, much like LLMs and VLMs, employ abstract symbolic IDs to resolve the binding task. Crucially, this abstraction varies by modality: acoustic attributes map to temporal IDs encoding when utterances occur, while visual attributes map to spatial IDs encoding where speakers are located. Second, we identify a three-stage trimodal binding mechanism: (i) anchor ID retrieval, where the text prompt describing an anchor attribute is mapped to its symbolic ID; (ii) target ID selection, where this ID is mapped to the symbolic ID of the target attribute; and (iii) feature retrieval, where the model extracts semantic features linked to the target symbolic ID. Finally, we show that binding failures primarily stem from the target ID selection stage, exposing a fundamental deficit in the audio-visual alignment of current AVLLMs.

Based on these findings, to compensate for the identified deficits, we propose a simple yet effective audio-visual prompting method using an off-the-shelf Active Speaker Detection (ASD) model. By leveraging the ASD model to overlay bounding boxes on active speakers before feeding the video into the AVLLMs (Fig. 1b), this training-free approach yields immediate performance gains across three conversation-centric benchmarks on Qwen2.5-Omni and MiniCPM-o-4.5. Furthermore, lightweight fine-tuning of fewer than 300 steps on these ASD-prompted-videos leads to further gains. Specifically, this tuning improves accuracy on conversation-centric datasets by 7.9% (video-SALMONN2+), 9.7% (Qwen2.5-Omni), and 3.5% (MiniCPM-o-4.5). Notably, our method generalizes beyond conversational settings, consistently improving performance across three additional general benchmarks.

2 Related Works

Audio-visual large language models. Building upon the foundational capabilities of LLMs, Audio-Visual Large Language Models (AVLLMs) have emerged to expand text-centric reasoning into the realms of audio and visual perception [60, 9, 10, 33, 58, 45, 20, 53, 54, 57]. Within this rapid wave

of diverse advancements, video-SALMONN2+ [45] targets detailed, holistic audio-visual captioning, while Qwen2.5-Omni [53] and MiniCPM-o-4.5 [11] elevate the field by introducing natively generated text and speech outputs. More recently, Qwen3-Omni [54] advances the landscape further through a Mixture-of-Experts structure. Alongside these modeling advances, a growing body of work has begun to analyze how AVLLMs internally process and integrate audio-visual inputs [28, 27, 59]. However, these models still frequently struggle in intricate, multi-speaker videos [46, 7, 36] that demand precise binding across modalities. Addressing these failures thus requires a closer examination of their internal cross-modal binding mechanisms.

Binding mechanisms in LLMs/VLMs. Recent studies have actively investigated how LLMs associate and bind textual entities with their corresponding attributes in long-context settings [16, 12, 21, 17, 40]. [16] first propose the existence of symbolic binding IDs that link entities to their attributes, while [40] further examine how language models bind each character's beliefs about the state of a given entity. This line of inquiry has also been extended to VLMs to explore multimodal binding mechanisms [1, 3, 22]. Recent works show that VLM failures on basic multi-object reasoning tasks, such as counting and localization, are closely linked to binding failures, and propose methods to mitigate them [5, 22, 24]. Building on this, our work extends the scope from unimodal and bimodal contexts to investigate *trimodal* binding-across text, audio, and vision-within AVLLMs.

Visual prompting in VLMs. Visual prompting [50] enhances VLMs by providing fine-grained instructions through markers such as circles [4, 55] and bounding boxes [15, 13, 34]. Recent works have explored the use of off-the-shelf vision modules to mitigate limitations of VLMs. For instance, to improve spatiotemporal reasoning, [47] leverage object tracking models [8, 43], providing the resulting object bounding-box coordinates as textual prompts. Similarly, to reduce hallucinations, [62] use object detection models [6, 61] as additional guidance. While often applied as training-free prompts [51, 30, 24], these techniques can also be paired with fine-tuning to enhance the model's understanding of the augmented prompts [51, 47, 32]. Inspired by these approaches, we introduce an audio-visual prompting method that leverages an off-the-shelf active speaker detection model to address the identified limitations of AVLLMs under both training-free and fine-tuning settings.

3 Symbolic Trimodal Binding Mechanisms in AVLLMs

In this section, we define the trimodal binding tasks (Sec. 3.1) and propose a three-stage symbolic mechanism (Sec. 3.2), validated through representational (Sec. 3.3) and causal mediation analyses (Sec. 3.4). We study four AVLLMs: video-SALMONN2+ (7B) [45], Qwen2.5-Omni (7B, 3B) [53], and MiniCPM-o-4.5 [11]. Unless otherwise specified, results are reported for video-SALMONN2+ in this section; results for other models are provided in Sec. A.2 of the supp. mat.

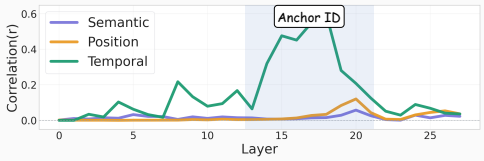
3.1 Trimodal Binding Task

Trimodal binding task evaluates an ability to associate textual, acoustic and visual information within single-shot, multi-speaker videos. We define a video as a sequence of n distinct speech events, where each event i is characterized by a 4-tuple (v_i, p_i, a_i, t_i). Specifically, the visual attribute $v_i \in V$ denotes the visual characteristics of the speaker, the spatial position $p_i \in P$ indicates their location on the screen (e.g., top-left, bottom-right etc.), the acoustic attribute $a_i \in A$ represents acoustic features such as the spoken content, and the temporal order $t_i \in T$ denotes the chronological speaking order of the utterance. The complete context of the video, c, is thus formulated as a set of bound tuples:

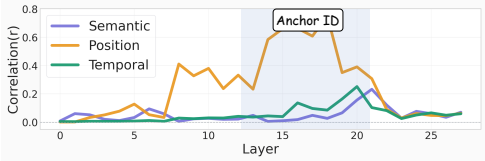
$$c = \{(v_1, p_1, a_1, t_1), (v_2, p_2, a_2, t_2), \ldots, (v_n, p_n, a_n, t_n)\}.$$

In our analysis setup, as illustrated in Fig. 1a, each video features four animal characters randomly placed in the four quadrants ($p \in \{0, 1, 2, 3\}$ each representing top-left, top-right, bottom-left and bottom-right position), each uttering a different country name in a random order ($t \in \{0, 1, 2, 3\}$). For instance, the video context in Fig. 1a is represented as, $c = \{(\text{panda}, 0, \text{Mexico}, 1), (\text{tiger}, 1, \text{Japan}, 3), (\text{monkey}, 2, \text{Canada}, 2), (\text{rabbit}, 3, \text{Egypt}, 0)\}$.

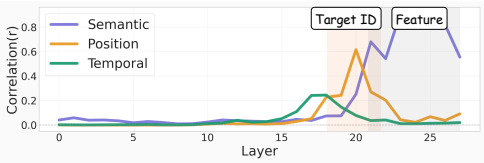
Given a context of video c, and a target event k, we provide the model with a text prompt describing one semantic attribute of the event, either the audio attribute a_k or the visual attribute v_k. This attribute serves as the *anchor*, and the model is asked to retrieve the corresponding *target* attribute from the complementary modality. This yields two sub-tasks: (1) Acoustically-Anchored Visual Retrieval (AAVR), relying on an acoustic anchor a_k to predict v_k, and (2) Visually-Anchored Audio



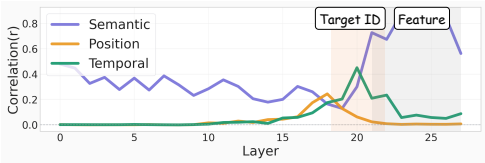
(a) RSA at anchor attribute token in AAVR task



(b) RSA at anchor attribute token in VAAR task



(c) RSA at last token in AAVR task



(d) RSA at last token in VAAR task

Figure 2: **RSA results at anchor attribute and last prompt tokens for AAVR and VAAR tasks.** At the anchor token, anchor attribute symbolic IDs strongly correlate in mid-to-late layers (anchor ID retrieval). At the last token, target attribute symbolic IDs correlate strongly in late layers (target ID selection), while semantic information peaks in the deepest layers (feature retrieval).

Retrieval (VAAR), relying on a visual anchor v_k to predict a_k. As illustrated in Fig. 1 for the AAVR task, the model is given the acoustic anchor $a_k =$ "Japan" in the text prompt "*Japan is said by the*" and is expected to complete the sentence by predicting the target visual attribute $v_k =$ "tiger".

Following prior mechanistic interpretability work that conditions analyses on successful behaviors [35, 19, 18], we restrict our analysis to correctly predicted samples. Since models perform poorly in the standard setting (Fig. 6b), we use contextual priming (e.g., "Canada is said by monkey, Egypt is said by rabbit, Mexico is said by panda, and Japan is said by" for AAVR). This induces the correct binding and yields accurate predictions for controlled analysis (see Sec. A.1 of the supp. mat. for further discussion).

3.2 Symbolic Trimodal Binding Mechanisms

Consistent with recent findings in LLMs and VLMs [56, 1], we find that AVLLMs utilize symbolic IDs to perform trimodal binding. Rather than encoding semantic attributes (a_i, v_i) based on their content, models abstract them into content-independent symbolic IDs and bind modalities by aligning these IDs. Specifically, AVLLMs derive symbolic IDs from modality-specific structural cues: sequential speaking order (t_i) for parsing and tracking acoustic attributes, which we term *Temporal IDs*, and spatial locations (p_i) for tracking visual attributes, which we term *Position IDs*. We then identify a three-stage trimodal binding mechanism based on these modality-specific symbolic IDs.

Stage 1: Anchor ID retrieval. The text prompt describing the anchor attribute is converted into its corresponding symbolic ID (Temporal ID t_k for AAVR, or Position ID p_k for VAAR). We refer to the symbolic ID resulting from this stage as *anchor ID*. For instance, in the AAVR shown in Fig. 1a, the model represents the text prompt "Japan" by mapping it to its temporal order within the utterance sequence (Temporal ID $t_k = 3$).

Stage 2: Target ID selection. The anchor ID is then transformed into the symbolic ID of the target attribute (Position ID p_k for AAVR, or Temporal ID t_k for VAAR). We refer to this as *target ID*. In Fig. 1a, the Temporal ID $t_k = 3$ is converted into the corresponding Position ID $p_k = 1$.

Stage 3: Feature retrieval. Finally, the model uses the target ID to locate the target attribute and retrieve its semantic content. In Fig. 1a, the model attends to the spatial location indexed by the Position ID $p_k = 1$, recovers the corresponding visual attribute ("tiger"), and verbalizes it.

3.3 Representational Analysis

To examine whether the embeddings of the model encode the proposed symbolic IDs, we perform Representational Similarity Analysis (RSA) [29]. RSA compares two embedding spaces by computing pairwise similarity patterns within each space and measuring their Pearson correlation. We compare the hidden states of the model against three hypothesized embedding spaces, each independently encoding the Temporal ID t_k, Position ID p_k, or semantic content of the target attribute (a_k or v_k). Specifically, we extract hidden states from per-layer attention outputs at two key text prompt

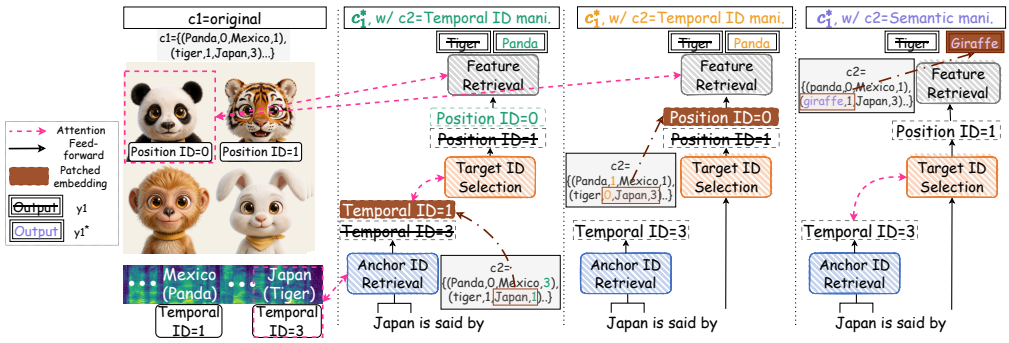


Figure 3: **Illustration of CMA for the AAVR task using three manipulated contexts.** We construct modified contexts c_2 via temporal ID, position ID, and semantic content manipulations, and illustrate the resulting patched run c_1^* obtained by patching c_2 outputs into c_1 after the anchor ID retrieval, target ID selection, and feature retrieval stages, respectively.

positions: the token specifying the anchor attribute and the last prompt token. Results are aggregated over 800 samples.

Results. Figs. 2a and 2b present the RSA results at the anchor token for the AAVR and VAAR tasks, respectively. In the middle-to-late layers, temporal information dominates in AAVR, while positional information dominates in VAAR. These results suggest the emergence of modality-specific symbolic IDs and anchor ID retrieval stage. At the last prompt token (Figs. 2c and 2d), late layers show prominent positional information in AAVR and temporal information in VAAR, reflecting target ID selection. In the deepest layers, semantic information becomes dominant, reflecting the feature retrieval stage. Similar patterns for other models are reported in Sec. A.2 of the supp. mat.

3.4 Causal Mediation Analysis

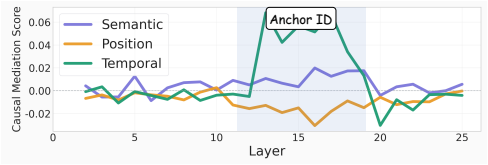
To investigate whether symbolic IDs causally affect the final output, we conduct Causal Mediation Analysis (CMA) [38, 48, 35, 56]. CMA estimates the causal effect of certain hidden states by copying selected states from one forward pass into the same locations of another forward pass and measuring how the original prediction changes. Specifically, it involves three forward passes: an original run under context c_1, producing the original output y_1; a modified run under context c_2, producing output y_2; and a patched run c_1^*, where selected embeddings from the c_2 run are copied (*patch*) into the corresponding locations of the original c_1 run, with expected output y_1^*, where $y_1^* \neq y_1$. We quantify the patching effect with the Causal Mediation Score (CM Score) [48]:

$$s = (M(c_1^*)[y_1^*] - M(c_1^*)[y_1]) - (M(c_1)[y_1^*] - M(c_1)[y_1]) \tag{1}$$

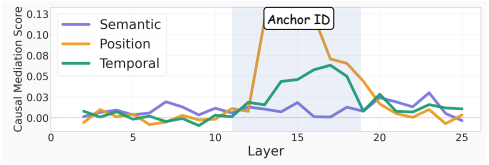
where $M(c)[y]$ denotes model M's logit for token y under context c. A higher score indicates that the patched embedding carries information that can steer the prediction toward y_1^*.

We apply three distinct manipulations to construct the modified context c_2 and define its corresponding expected answer y_1^*. For *temporal ID manipulation*, we swap the temporal order of the utterance for the target event k and another event j, resulting in $c_2 = \{\dots, (v_k, p_k, a_k, t_j), \dots, (v_j, p_j, a_j, t_k), \dots\}$. For *position ID manipulation*, we swap the spatial positions of the animal for events k and j, yielding $c_2 = \{\dots, (v_k, p_j, a_k, t_k), \dots, (v_j, p_k, a_j, t_j), \dots\}$. In both cases, the text prompt remains identical across c_1 and c_2, explicitly querying the target event k, yielding $y_1 = y_2 = v_k$ for AAVR, and $y_1 = y_2 = a_k$ for VAAR. In contrast, we set the expected patched answer y_1^* to the target attribute of event j (v_j for AAVR, or a_j for VAAR). Therefore, if this causal patching shifts the output to y_1^*, it indicates that these symbolic IDs are being actively utilized by the model. For *semantic content manipulation*, we alter the target attribute by replacing it with a novel semantic attribute not present in the original video. For instance, we substitute the target visual attribute v_k with an unseen $v^* \notin V$ in AAVR (yielding $c_2 = \{\dots, (v^*, p_k, a_k, t_k), \dots\}$), or the target acoustic attribute a_k with an unseen $a^* \notin A$ in VAAR. In both cases, the expected answer is set to this novel attribute ($y_1^* = v^*$ or a^*).

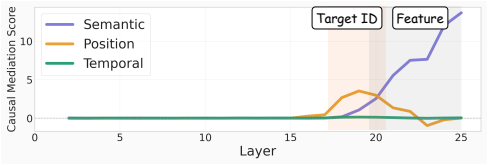
For both AAVR and VAAR, we perform CMA using three variants of the modified context c_2, patching them at two token positions: the anchor attribute token and the last prompt token. We illustrate, as an example, the expected effect of each manipulation when applied at each stage in the



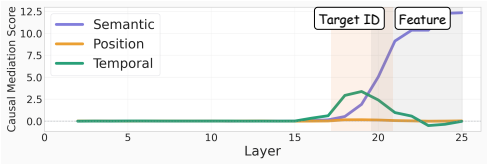
(a) CMA at anchor attribute token in AAVR task



(b) CMA at anchor attribute token in VAAR task

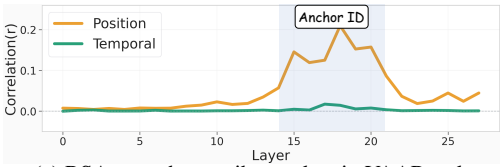


(c) CMA at last token in AAVR task

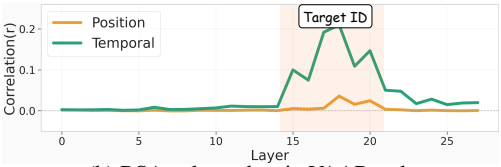


(d) CMA at last token in VAAR task

Figure 4: **CMA results at anchor attribute and last prompt tokens for AAVR and VAAR tasks.** At the anchor token, anchor ID manipulation exhibit high CM scores in mid-to-late layers (anchor ID retrieval). At the last token, target ID manipulation exhibit high CM scores in late layers (target ID selection), while semantic content manipulation peaks in the deepest layers (feature retrieval).



(a) RSA at anchor attribute token in VAAR task



(b) RSA at last token in VAAR task

Figure 5: **RSA results in the VAAR task for a real-world dataset.** The symbolic trimodal binding mechanism generalizes to real-world settings. Position IDs exhibit strong correlations at the anchor token (anchor ID retrieval), whereas temporal IDs dominate at the last token (target ID selection).

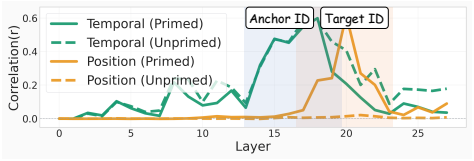
AAVR task in Sec. C.2 of the supp. mat. At each position, we identify which manipulation yields the highest CM Score. We use attention outputs as the patched embeddings and apply patching over layer windows spanning one-quarter of the total depth. We hypothesize the following behavior:

Anchor ID retrieval. When patching at the anchor attribute token, we expect manipulation of the anchor ID (temporal ID for AAVR, position ID for VAAR) to yield the highest CM scores in mid-to-late layers. As illustrated in Fig. 3, when c_2 is constructed via temporal ID manipulation in AAVR, the anchor ID retrieval stage extracts the manipulated temporal ID t_j. Patching this state into c_1 then directs the model's target ID selection stage to choose the matching position ID p_j. Ultimately, this leads to the extraction of v_j during feature retrieval stage, which is our expected answer y_1^*.

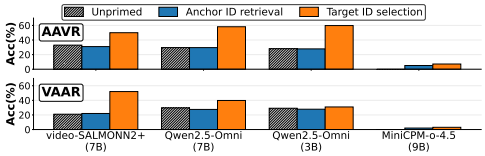
Target ID selection. Conversely, when patching at the last prompt token, we expect the manipulation of the target ID (position ID for AAVR, temporal ID for VAAR) to yield the highest CM scores in the late layers. As illustrated in Fig. 3, when c_2 is constructed via position ID manipulation in AAVR, the target ID selection stage for c_2 extracts the manipulated position ID p_j. Patching this into c_1 triggers the extraction of v_j during the feature retrieval stage, which is our expected answer y_1^*.

Feature retrieval. Finally, when patching at the deepest layers of the last prompt token, we expect semantic content manipulation to achieve the highest CM scores. As illustrated in Fig. 3, in the AAVR task, the feature retrieval stage for c_2 directly extracts the novel attribute v^*. Once this state is patched into c_1, the model outputs v^*, which is our expected answer y_1^*.

Results. Fig. 4 presents the CMA results averaged over 960 samples for the AAVR and VAAR tasks. At the anchor attribute token, Figs. 4a and 4b reveal that anchor ID manipulation yields significantly higher scores in the mid-to-late layers compared to other manipulations, consistent with the anchor ID retrieval stage. Moving to the last prompt token, Figs. 4c and 4d show that target ID manipulation produces the strongest effect in the late layers, consistent with the target ID selection stage. In the deepest layers, semantic content manipulation reaches its peak, marking the feature retrieval stage.



(a) RSA results for primed and unprimed setting



(b) Causal intervention analysis

Figure 6: **Target ID selection stage identified as the locus of failure.** (a) The primed-unprimed representational gap peaks at the target ID selection stage in AAVR task. (b) Patching correct symbolic IDs at target ID selection stage significantly improves accuracy.

3.5 Real-World Generalization

To verify the generalization of symbolic mechanisms to real-world settings, we conduct RSA on 2000 clips from the SocialOmni [52]. We focus on two-speaker videos to construct AAVR and VAAR tasks. As defining a semantic hypothesis space is infeasible in these unconstrained settings, we compare the embeddings against Temporal and Position IDs. Figs. 5a and 5b illustrate the VAAR results at the anchor and last token, respectively. Although the inherent variability of real-world data attenuates the overall correlations, the results support our symbolic mechanism: positional information is pronounced in mid-to-late layers at the anchor token (anchor ID retrieval), while temporal information peaks in late layers at the last token (target ID selection). Similar trends are observed across other models and tasks (Sec. A.5 of the supp. mat.)

4 Diagnosing and Mitigating Trimodal Binding Failures

In this section, we first locate the source of trimodal binding failures in the unprimed setting (Sec. 4.1). Building on this diagnosis, we propose a simple audio-visual prompting method using an off-the-shelf Active Speaker Detection (ASD) module to compensate for the identified bottleneck (Sec. 4.2).

4.1 Locating the Failure Bottleneck in Trimodal Binding

Unlike preceding sections that use contextual priming, we now shift to an unprimed setting, where such contextual cues are absent and failures often occur. We identify the locus of these failures through representational and causal intervention analyses.

Representational Analysis To localize the failure point, we repeat the RSA from Sec. 3.3 in the unprimed setting; the stage with the largest primed-unprimed deviation indicates the likely source of failure. As shown in Fig. 6a for AAVR task, the anchor ID retrieval stage (green solid/dashed lines) shows negligible divergence, whereas the target ID selection stage (yellow solid/dashed lines) exhibits a substantial gap. This suggests that performance degradation stems from a failure in establishing audio-visual correspondence, rather than an inability to parse or textually ground the visual scenes or audio streams. This pattern holds across models and tasks (Sec. A.4 of our supp. mat.).

Causal Intervention Analysis To further localize the failure, we perform causal interventions by injecting correct symbolic IDs at each stage. We obtain correct symbolic representations from the primed setting and patch them into the unprimed setting. A substantial accuracy improvement after intervention would implicate the corresponding stage as the primary source of error. We first identify the specific head-layer pairs for intervention. To minimize confounding factors, we refine our earlier layer-level analysis to individual attention heads, following [56, 1]. By reapplying the CMA (Sec. 3.4) at this finer granularity, we select the top 10 pairs yielding the highest scores across a 100-sample subset for each binding stage. Then, to capture the correct symbolic ID, we compute a representative activation vector for each selected pair by averaging its outputs over 800 primed instances. Finally, we patch these cached representations into the unprimed setting to evaluate whether the intervention successfully restores model accuracy. As shown in Fig. 6b, intervening at the anchor ID retrieval stage has little effect, whereas intervention at the target ID selection stage improves accuracy. This suggests that target ID selection is the key bottleneck responsible for trimodal binding failures.

Table 1: **Results of the proposed methods.** We evaluate three recent AVLLMs across four conversation-centric and three general benchmarks. For Qwen2.5-Omni and MiniCPM-o-4.5, training-free ASD improves on conversation-centric tasks (For general benchmarks, which often lack speech or visible speakers, we do not apply ASD to draw bounding boxes, making ASD identical to the vanilla setting.). ASD-FT improves bounding box grounding capabilities, yielding consistent performance gains across all three AVLLMs on all seven benchmarks.

Model	Converse-Centric					General		
	AVSpeaker[36]	DailyOmni[63]	SocialOmni[52]	DiaDemBench[7]		OmniBench[31]	DAVE[41]	WorldSense[23]
	Acc(%) ↑	Acc(%) ↑	Acc(%) ↑	REF ↑	ASR ↑	Acc(%) ↑	Acc(%) ↑	Acc(%) ↑
Qwen2.5-Omni[53] (7B)	44.86	64.33	38.95	19.1±0.4	28.1±0.1	50.79	30.75	50.06
+ **ASD**	45.59	64.66	41.90	**22.5**±0.2	**33.0**±0.2	50.79	30.75	50.06
+ **ASD-FT**	**48.79**	**71.09**	**44.25**	20.7±0.2	32.3±0.1	**52.80**	**33.91**	**53.73**
MiniCPM-o-4.5[11] (9B)	49.41	61.99	46.05	6.3±0.5	7.4±0.7	47.99	55.80	55.73
+ **ASD**	50.34	62.49	**50.00**	9.7±0.4	10.8±0.1	47.99	55.80	55.73
+ **ASD-FT**	**51.77**	**63.41**	49.70	**24.8**±0.2	**30.9**±0.0	**51.84**	**57.97**	**57.37**
video-SALMONN2+[45] (7B)	44.74	65.13	43.49	14.0±0.3	19.3±0.1	37.04	35.28	45.04
+ ASD	44.65	62.57	43.89	11.8±0.3	19.0±0.0	37.04	35.28	45.04
+ **ASD-FT**	**49.10**	**69.13**	**49.62**	**19.7**±0.2	**26.7**±0.2	**44.92**	**39.65**	**46.38**

4.2 Audio-Visual Prompting via Active Speaker Detection

Identifying the target ID selection stage as the critical bottleneck suggests that existing AVLLMs struggle primarily with audio-visual binding. To address this limitation, we propose a simple audio-visual prompting method using an off-the-shelf Active Speaker Detection (ASD) model. As illustrated in Fig. 1b, we use a recent ASD model [37] to overlay a red bounding box on the detected active speaker in each frame before feeding into AVLLM. To ensure the model correctly interprets this visual cue, the accompanying text prompt indicates that the red box denotes the active speaker. While this approach can work training-free, we further introduce a fine-tuning method to solidify the model's ability to ground these visual cues. To this end, we use videos with these audio-visual prompts to perform supervised fine-tuning (SFT) via Low-Rank Adaptation (LoRA).

5 Experiments

5.1 Experimental Setup

Training datasets. For the LoRA fine-tuning dataset, we construct a specialized corpus of 400 synthetic videos by adapting the toy dataset framework from our preceding analysis. Specifically, we replace the original animal speakers with human speakers to bridge the domain gap, ensuring the active speaker is highlighted with a red bounding box. Each video is paired with two types of tasks: a video captioning task describing the appearances and utterances of all speakers, and multiple-choice questions evaluating AAVR and VAAR tasks. See Sec. B.2.1 of the supp. mat. for more details.

Implementation details. We adopt Qwen2.5-Omni (7B), video-SALMONN2+ (7B), and MiniCPM-o-4.5. All fine-tuning is conducted on a single NVIDIA RTX A6000 GPU. We apply LoRA with a rank of 16, employing a batch size of 1 and 8 gradient accumulation steps, with all models trained for fewer than 300 optimization steps. See Sec. B.2.2 of the supp. mat. for more details.

Evaluation datasets and protocol. We evaluate our approach on four conversation-centric benchmarks encompassing diverse task formats. For multiple-choice question answering, we adopt AVSpeaker [36], DailyOmni [63], and perceptual task subset of SocialOmni [52]. For multi-speaker dialogue captioning, we use DiaDemBench [7]. Performance is evaluated on two metrics-ASR for transcription accuracy and REF for speaker attribution consistency-with scoring conducted using gemini-2.5-Pro and gemini-2.5-Flash. To assess generalization beyond conversation-centric scenarios, we further evaluate on three broader audio-visual benchmarks: DAVE [41] for egocentric videos, OmniBench [31] for image-audio pairs, and WorldSense [23] for diverse audio categories including music and environmental sounds.

5.2 Experimental Results

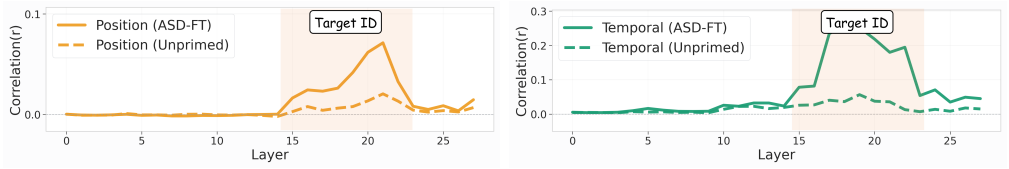
Quantitative results. Tab. 1 presents the quantitative results. The training-free ASD approach demonstrates clear effectiveness across three speech-centric datasets for Qwen2.5-Omni and MiniCPM-o-4.5. This suggests that the ASD intervention effectively compensates for the mod-

Table 2: **Comparison of training-free methods.** Our ASD surpasses both decoding- and prompting-based methods.

Type	Method	AVSpeaker
Vanilla	—	44.86
Decoding	AVCD [27]	42.82
	FMD [26]	45.27
	OutRo [59]	45.24
Prompting	NumPro [51]	45.08
	VISER [24]	44.55
	ASD(Ours)	**45.59**

Table 3: **Ablation on the prompting style of training-free ASD.** ▲and ▼indicate performance gain and drop compared to vanilla.

BBox	Text	Color	Thickness	AVSpeaker
	Vanilla			44.86
Inactive	✓	Red	Medium	44.77 ▼
Random	✓	Red	Medium	44.52 ▼
Active	✗	Red	Medium	44.80 ▼
Active	✓	Green	Medium	45.95 ▲
Active	✓	Red	Thick	45.70 ▲
Active	✓	Red	Thin	45.58 ▲



(a) AAVR task (b) VAAR task

Figure 7: **Target ID selection after ASD-FT in the unprimed setting.** Compared to the vanilla model (dashed), ASD-FT (solid) shows stronger correlation with the target ID.

els' inherent deficiencies in audio-visual binding. However, because this approach relies on a pre-existing ability to comprehend and ground bounding boxes, models for which this ability is less reliable in a zero-shot setting, such as video-SALMONN2+, benefit from explicit fine-tuning. After fine-tuning on ASD-prompted videos (ASD-FT), all three models improve across the three conversation-centric datasets. Interestingly, the ASD-FT models also show consistent gains on general datasets. These results suggest that ASD-FT strengthens audio-visual alignment beyond simple bounding-box comprehension, resolving the binding bottlenecks identified in our analysis. See Secs. B.3 and D of our supp. mat. for more discussions and qualitative results.

Comparison with other training-free methods. We compare our training-free ASD against existing training-free decoding methods for AVLLMs-AVCD [27], FMD [26], and OutRo [59]-and visual prompting methods-NumPro [51] and VISER [24] on Qwen2.5-Omni (7B) on the AVSpeaker [36]. Tab. 2 shows that ASD outperforms all baselines.

Ablation on prompting styles. To examine the effect of prompting styles in training-free ASD, we conduct an ablation study using Qwen2.5-Omni on the AVSpeaker (Tab. 3). First, applying the bounding box to the inactive speaker or a random region degrades performance, showing that ASD's gains stem from accurate active-speaker localization rather than generic visual saliency. Second, removing the explanatory text prompt that defines the bounding box degrades performance, confirming that explicit instructions are necessary for visual grounding. Third, the gains are robust to bounding box style variations, such as color or thickness.

Ablation on finetuning methods. We compare ASD-FT with a baseline fine-tuned on videos without bounding boxes (FT) using Qwen2.5-Omni (Tab. 4). While FT improves performance, ASD-FT yields larger gains 4 out of 5 datasets. This suggests that the ASD-prompted setup directs the model's learning capacity toward aligning audio with the correct visual region, rather than merely improving general task performance.

Table 4: **Ablation on finetuning methods.** ASD-FT yields greater gains than baseline FT.

Method	Converse-Centric		General		
	AV-Speaker	Daily-Omni	Omni-Bench	DAVE	World-Sense
Vanilla	44.86	64.33	50.79	30.75	50.06
+ FT	46.89	66.67	51.05	**35.15**	52.05
+ **ASD-FT**	**48.79**	**71.09**	**52.80**	33.91	**53.73**

Revisiting the target ID selection stage. We revisit the RSA in Sec. 4.1 to assess whether ASD-FT resolves the target ID selection bottleneck. Fig. 7 compares the vanilla model (dashed lines) and the ASD-FT model (solid lines) on unprimed AAVR and VAAR tasks. Notably, the ASD-FT model exhibits higher correlation with the target ID in the late layers (target ID selection stage). This suggests that ASD-FT mitigate the target ID selection errors observed in the vanilla model.

6 Conclusion

In this paper, motivated by the observation that AVLLMs struggle to resolve "who says what" in multi-speaker settings, we investigate the underlying trimodal binding mechanisms. Our investigation reveals that AVLLMs utilize modality-specific symbolic IDs and process information through a distinct three-stage mechanism. Additionally, we identify faulty audio-visual binding, rather than text-audio or text-visual grounding, as the primary bottleneck for overall binding failures. To address this, we propose a simple yet effective audio-visual prompting method utilizing an off-the-shelf Active Speaker Detection model (ASD). We envision that these findings will provide foundational insights for the architectural design and training strategies of next-generation AVLLMs.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2025- 02215122, Development and Demonstration of Lightweight AI Model for Smart Homes).

References

[1] Rim Assouel, Declan Campbell, Yoshua Bengio, and Taylor Webb. Visual symbolic mechanisms: Emergent symbol processing in vision language models. In *Proc. ICLR*, 2026.

[2] 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:2502.13923*, 2025.

[3] Nicolas Buzeta, Felipe del Rio, Cristian Hinostroza, Denis Parra, Hans Lobel, and Rodrigo Toro Icarte. Seeing to generalize: How visual data corrects binding shortcuts. *arXiv preprint arXiv:2602.15183*, 2026.

[4] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *Proc. CVPR*, 2024.

[5] Declan Iain Campbell, Sunayana Rane, Tyler Giallanza, C. Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor Whittington Webb. Understanding the limits of vision language models through the lens of the binding problem. In *Proc. NeurIPS*, 2024.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. ECCV*, 2020.

[7] Xinlong Chen, Weihong Lin, Jingyun Hua, Linli Yao, Yue Ding, Bozhou Li, Bohan Zeng, Yang Shi, Qiang Liu, Yuanxing Zhang, et al. Diadem: Advancing dialogue descriptions in audiovisual video captioning for multimodal large language models. *arXiv preprint arXiv:2601.19267*, 2026.

[8] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proc. CVPR*, 2024.

[9] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv:2406.07476*, 2024.

[10] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Meerkat: Audio-visual large language model for grounding in space and time. In *Proc. ECCV*, 2024.

[11] 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.

[12] Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Representational analysis of binding in language models. In *Proc. EMNLP*, 2024.

[13] Ronghao Dang, Jiangyan Feng, Haodong Zhang, Chongjian Ge, Lin Song, Lijun Gong, Chengju Liu, Qijun Chen, Feng Zhu, Rui Zhao, et al. Instructdet: Diversifying referring object detection with generalized instructions. In *Proc. ICLR*, 2024.

[14] Xander Davies, Max Nadeau, Nikhil Prakash, Tamar Rott Shaham, and David Bau. Discovering variable binding circuitry with desiderata. *arXiv preprint arXiv:2307.03637*, 2023.

[15] Zhizhao Duan, Hao Cheng, Duo Xu, Xi Wu, Xiangxie Zhang, Xi Ye, and Zhen Xie. Cityllava: Efficient fine-tuning for vlms in city scenario. In *Proc. CVPR*, 2024.

[16] Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *Proc. ICLR*, 2024.

[17] Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Monitoring latent world states in language models with propositional probes. In *Proc. ICLR*, 2025.

[18] Constanza Fierro, Negar Foroutan, Desmond Elliott, and Anders Søgaard. How do multilingual language models remember facts? In *Findings of ACL*, 2025.

[19] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proc. EMNLP*, 2023.

[20] Yuxin Guo, Shuailei Ma, Shijie Ma, Xiaoyi Bao, Chen-Wei Xie, Kecheng Zheng, Tingyu Weng, Siyang Sun, Yun Zheng, and Wei Zou. Aligned better, listen better for audio-visual large language models. In *Proc. ICLR*, 2025.

[21] Yoav Gur-Arieh, Mor Geva, and Atticus Geiger. Mixing mechanisms: How language models retrieve bound entities in-context. In *Proc. ICLR*, 2026.

[22] Hosein Hasani, Amirmohammad Izadi, Fatemeh Askari, Mobin Bagherian, Sadegh Mohammadian, Mohammad Izadi, and Mahdieh Soleymani Baghshah. Uncovering grounding ids: How external cues shape multimodal binding. *arXiv preprint arXiv:2509.24072*, 2025.

[23] Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating real-world omnimodal understanding for multimodal LLMs. In *Proc. ICLR*, 2026.

[24] Amirmohammad Izadi, Mohammadali Banayeeanzade, Fatemeh Askari, Ali Rahimiakbar, Mohammad Mahdi Vahedi, Hosein Hasani, and Mahdieh Soleymani Baghshah. Visual structures help visual reasoning: Addressing the binding problem in LVLMs. In *Proc. NeurIPS*, 2025.

[25] Chaeyoung Jung, Suyeon Lee, Kihyun Nam, Kyeongha Rho, You Jin Kim, Youngjoon Jang, and Joon Son Chung. Talknce: Improving active speaker detection with talk-aware contrastive learning. In *Proc. ICASSP*, 2024.

[26] Chaeyoung Jung, Youngjoon Jang, Jongmin Choi, and Joon Son Chung. Fork-merge decoding: Enhancing multimodal understanding in audio-visual large language models. *arXiv preprint arXiv:2505.20873*, 2025.

[27] Chaeyoung Jung, Youngjoon Jang, and Joon Son Chung. AVCD: Mitigating hallucinations in audio-visual large language models through contrastive decoding. In *Proc. NeurIPS*, 2025.

[28] Jihoo Jung, Chaeyoung Jung, Ji-Hoon Kim, and Joon Son Chung. Probing cross-modal information hubs in audio-visual LLMs. In *Proc. ICML*, 2026.

[29] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

[30] Yeonkyung Lee, Dayun Ju, Youngmin Kim, Seil Kang, and Seong Jae Hwang. Vikey: Enhancing temporal understanding in videos via visual prompting. *arXiv preprint arXiv:2603.23186*, 2026.

[31] Yizhi LI, Ge Zhang, Yinghao Ma, Ruibin Yuan, King Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Moore Wang, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Yidan WEN, Yanghai Wang, Shihao Li, Zhaoxiang Zhang, Ruibo Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. Omnibench: Towards the future of universal omni-language models. In *Proc. NeurIPS*, 2025.

[32] Yuanze Lin, Yunsheng Li, Dongdong Chen, Weijian Xu, Ronald Clark, Philip Torr, and Lu Yuan. Rethinking visual prompting for multimodal large language models with external knowledge. *arXiv preprint arXiv:2407.04681*, 2024.

[33] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv:2306.09093*, 2023.

[34] Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *Proc. ECCV*, 2025.

[35] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Proc. NeurIPS*, 2022.

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

[37] Le Thien Phuc Nguyen, Zhuoran Yu, and Yong Jae Lee. Laser: Lip landmark assisted speaker detection for robustness. In *Proc. WACV*, 2026.

[38] Judea Pearl. Direct and indirect effects. *Probabilistic and Causal Inference: The Works of Judea Pearl*, page 373, 2001.

[39] Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *Proc. ICLR*, 2024.

[40] Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, and Atticus Geiger. Language models use lookbacks to track beliefs. In *Proc. ICLR*, 2026.

[41] Gorjan Radevski, Teodora Popordanoska, Matthew B. Blaschko, and Tinne Tuytelaars. DAVE: Diagnostic benchmark for audio visual evaluation. In *Proc. NeurIPS*, 2025.

[42] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*, 2023.

[43] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *Proc. ICLR*, 2025.

[44] David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

[45] Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN 2: Captioning-enhanced audio-visual large language models. *arXiv:2506.15220*, 2025.

[46] 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.

[47] Zitian Tang, Shijie Wang, Junho Cho, Jaewook Yoo, and Chen Sun. How can objects help video-language understanding? In *Proc. ICCV*, 2025.

[48] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

[49] Xizi Wang, Feng Cheng, and Gedas Bertasius. Loconet: Long-short context network for active speaker detection. In *Proc. CVPR*, 2024.

[50] Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*, 2024.

[51] Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. In *Proc. CVPR*, 2025.

[52] Tianyu Xie, Jinfa Huang, Yuexiao Ma, Rongfang Luo, Yan Yang, Qingchuan Ma, Wang Chen, Yuhui Zeng, Yixuan Zou, Zhiqiang Lu, et al. Socialomni: Benchmarking audio-visual social interactivity in omni models. *arXiv preprint arXiv:2603.16859*, 2026.

[53] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2.5-omni technical report. *arXiv:2503.20215*, 2025.

[54] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *arXiv:2509.17765*, 2025.

[55] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

[56] Yukang Yang, Declan Iain Campbell, Kaixuan Huang, Mengdi Wang, Jonathan D. Cohen, and Taylor Whittington Webb. Emergent symbolic mechanisms support abstract reasoning in large language models. In *Proc. ICML*, 2025.

[57] Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Zhen Wan, Jinchuan Tian, An-Chieh Cheng, Ligeng Zhu, Yuanhang Su, Yuming Lou, Yong-Xiang Lin, Dong Yang, Sreyan Ghosh, Zhijian Liu, Yukang Chen, Ehsan Jahangiri, Ambrish Dantrey, Daguang Xu, Ehsan Hosseini-Asl, Seyed Danial Mohseni Taheri, Vidya Nariyambut Murali, Sifei Liu, Yao Lu, Oluwatobi Olabiyi, Yu-Chiang Frank Wang, Rafael Valle, Bryan Catanzaro, Andrew Tao, Song Han, Jan Kautz, Hongxu Yin, and Pavlo Molchanov. Omnivinci: Enhancing architecture and data for omni-modal understanding LLM. In *Proc. ICLR*, 2026.

[58] Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In *Proc. ECCV*, 2024.

[59] Suho Yoo, Youngjoon Jang, and Joon Son Chung. On the nature of attention sink that shapes decoding strategy in mllms. *arXiv preprint arXiv:2603.14337*, 2026.

[60] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proc. EMNLP*, 2023.

[61] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *Proc. CVPR*, 2024.

[62] Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via image-grounded guidance. In *Proc. ICML*, 2025.

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

Appendix

Appendix Contents

A. Symbolic Binding Mechanisms

- A.1 Reasons to Introduce Priming
- A.2 Representational Analysis and Causal Mediation Analysis on Different Models
- A.3 Generalization of Binding Mechanism Beyond the Controlled Four-Speaker Setting
- A.4 Representational Analysis in Primed vs. Unprimed Settings
- A.5 Real-World Generalization Analysis on Different Models

B. Audio-Visual Prompting

- B.1 Details of Training-Free Prompting
- B.2 Details of Finetuning
- B.3 Discussion on the Experimental Results

C. Illustrations

- C.1 Illustration of the Symbolic Mechanism in VAAR
- C.2 Illustration of Causal Mediation Analysis for AAVR

D. Qualitative Results

E. Limitations

F. Computational Resource

G. Social Impact

A Symbolic Binding Mechanisms

A.1 Reasons to Introduce Priming

Our mechanistic analysis (Sec. 3.2) conditions on contextually primed inputs. This appendix first clarifies what information the prime provides, and then explains why such priming is methodologically necessary. We further address three potential concerns: whether the prime leaks the answer, whether the primed model merely emits residual audio-visual content, and whether priming induces a circuit unrelated to the unprimed setting.

Primed setting. In the primed setting, the model receives a prompt of the following form:

> *"There are four animals in the clip, each speaking one word. Continue the sentence with the single most likely next word. Output exactly one word. Do not add any explanation, extra words, or punctuation. Sentence: Germany is said by the panda, canada is said by the monkey, japan is said by the tiger, and egypt is said by the ___"*

The order of the three demonstration speakers in the prime is randomized across samples, so that no fixed positional cue is associated with the queried speaker. The prime serves two roles. First, it provides structural scaffolding by specifying the task format: what type of token to emit and in what grammatical slot. Second, it helps the model form the correct binding for the target speaker by demonstrating the form of the audio-visual binding computation using three speakers.

Priming is necessary for causal analysis. Mechanistic causal analyses are typically conducted on inputs for which the model produces the correct output [35, 19, 18]. This is necessary because patching effects are meaningful only when there is a reliable target behavior to explain. When model performance is near chance, causal effects are easily dominated by noise rather than by the computation of interest. In the unprimed setting, the AVLLMs we study do not provide such a success regime (Tab. 5): all four models achieve less than 30% accuracy. In contrast, priming raises accuracy to near-ceiling levels across all models, enabling us to measure causal effects on trajectories where the model reliably performs the intended audio-visual binding computation.

Table 5: Accuracy across the four AVLLMs in three conditions: unprimed, primed, and primed prompt only. All values are averaged over 800 samples per model.

Model	Unprimed	Primed	Primed prompt only
video-SALMONN2+ (7B)	27.0%	99.2%	1.00%
Qwen2.5-Omni (7B)	29.6%	100.0%	12.50%
Qwen2.5-Omni (3B)	28.7%	99.94%	12.50%
MiniCPM-o-4.5 (9B)	0.0%	99.31%	0.06%

Priming does not leak the answer. Although the prime demonstrates three speaker–word bindings, it does not specify the queried speaker. The correct answer therefore still depends on the audio-visual content of the clip. Moreover, the task is formulated as open-ended next-token generation rather than multiple-choice QA: the model generates from its full output vocabulary, and the candidate set is never enumerated in the prompt. Thus, text-only shortcuts based on selecting from an explicitly provided candidate list are not available.

We verify this empirically by removing the audio and video while keeping the primed prompt unchanged. As shown in Tab. 5, accuracy collapses to near-floor levels in the prompt-only condition: 1.0% for video-SALMONN2+, 12.5% for Qwen2.5-Omni(7B), and 0.06% for MiniCPM-o-4.5. The relatively higher prompt-only accuracy of Qwen2.5-Omni(7B) does not indicate successful binding from text alone. Rather, its outputs are largely driven by lexical priors over common country and animal names; some of these high-prior tokens, such as *japan* and *panda*, overlap with labels used in our dataset. Thus, prompt-only performance reflects coincidental lexical bias rather than recovery of the correct audio-visual binding.

Primed predictions are not residual content emission. Another concern is that, in the primed setting, the model may not form a binding with the anchor at all; instead, it may simply emit residual audio-visual content that remains salient after processing the prime. The CMA results in Sec. 3.4

argue against this interpretation. When we patch the symbolic ID at the anchor attribute token, the output logit shifts toward the target attribute associated with the manipulated ID. If the model were merely emitting residual content independently of the anchor, changing the symbolic representation patched at the anchor position should not systematically change the predicted target. The observed direction of the logit shift therefore indicates that primed predictions are produced through an active cross-modal binding computation.

Priming does not induce a separate circuit. Finally, one might worry that priming induces a circuit unrelated to the one used in the unprimed setting. Our causal intervention analysis in Sec. 4.1 suggests otherwise. We identify the top-10 attention heads associated with target-ID selection in the primed regime and patch their averaged activations into unprimed forward passes at the same locations. This intervention substantially improves unprimed accuracy: from 27.0% to 51.0% for video-SALMONN2+(7B), from 29.6% to 49.0% for Qwen2.5-Omni(7B), and from 28.7% to 45.4% for Qwen2.5-Omni(3B). If the unprimed model did not engage this circuit, injecting activations from these primed heads should have little effect.

Together, these results show that priming is an analytical device for obtaining reliable successful trajectories, not a source of a separate mechanism. The primed and unprimed regimes engage the same symbolic binding circuit: priming makes this circuit measurable, while unprimed failures arise when the same circuit produces incorrect symbolic IDs.

A.2 Representational Analysis and Causal Mediation Analysis on Different Models

A.2.1 Qwen2.5-Omni(7B)

Fig. 8 illustrates the representation and causal mediation analysis results for Qwen2.5-Omni-7B. These results corroborate our proposed mechanism. A minor distinction from video-SALMONN2+ is observed in the VAAR case, where target ID selection appears to initiate at the layer of prompt tokens. However, this is merely a model-specific variation and remains completely consistent with our proposed three-stage framework (anchor ID retrieval, target ID selection, and feature retrieval). Crucially, our analysis does not posit that Stage 1 occurs exclusively at the prompt token while Stages 2 and 3 occur exclusively at the last token. Rather, it demonstrates that this sequential, three-stage flow manifests layer-wise across the prompt and last tokens collectively.

A.2.2 Qwen2.5-Omni(3B)

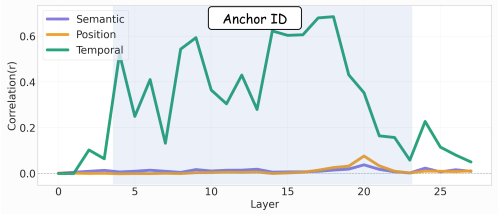
Fig. 9 presents the RSA and CMA results for Qwen2.5-Omni(3B), which exhibit trends consistent with our proposed three-stage mechanism. At the anchor attribute token, the temporal ID dominates in the AAVR task while the position ID dominates in the VAAR task across the mid-to-late layers, indicating anchor ID retrieval. At the last token, this pattern inverts in the late layers-position information becomes prominent in AAVR and temporal information in VAAR-consistent with target ID selection, before semantic information takes over in the deepest layers, marking feature retrieval. The CMA results corroborate these findings: anchor ID manipulation yields the largest causal effect at the anchor token in mid-to-late layers, target ID manipulation peaks at the last token in late layers, and semantic content manipulation dominates in the deepest layers.

A.2.3 MiniCPM-o-4.5

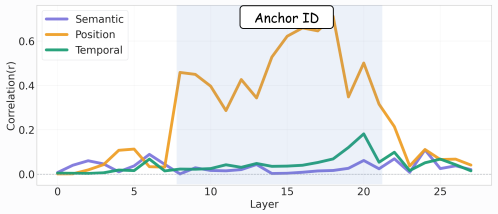
Fig. 10 reports the RSA and CMA results for MiniCPM-o-4.5 (9B), which again align with our proposed mechanism. At the prompt token, the modality-specific anchor ID (temporal for AAVR, position for VAAR) emerges as the dominant signal in the mid-to-late layers, reflecting anchor ID retrieval. At the last token, the target ID takes over in the late layers, followed by semantic content in the deepest layers, consistent with the target ID selection and feature retrieval stages. The CMA results largely echo these representational patterns, with only minor deviations at certain stages. Overall, MiniCPM-o-4.5 suggests that the symbolic trimodal binding mechanism generalizes across model families and parameter scales.

A.3 Generalization of Binding Mechanism Beyond the Controlled Four-Speaker Setting

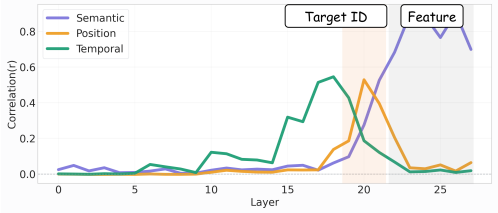
We further extend the synthetic analysis to five- and six-speaker settings.



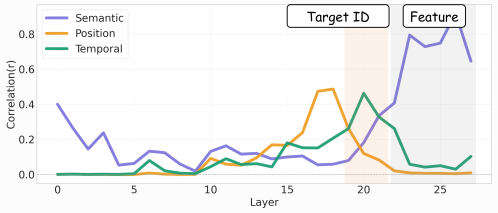
(a) RSA at anchor attribute token in AAVR task



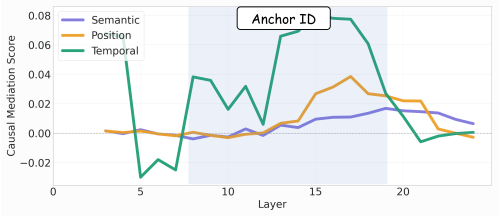
(b) RSA at anchor attribute token in VAAR task



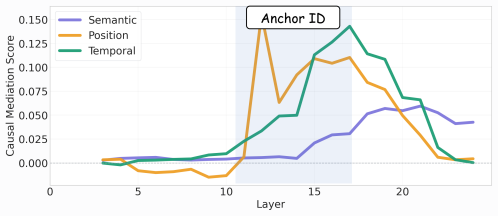
(c) RSA at last token in AAVR task



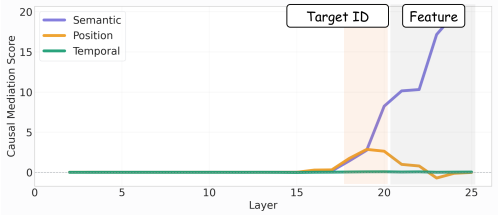
(d) RSA at last token in VAAR task



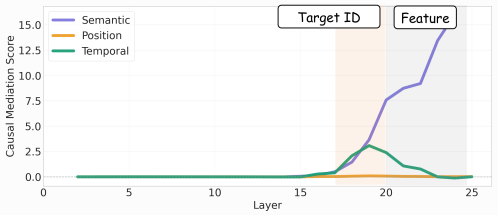
(e) CMA at anchor attribute token in AAVR task



(f) CMA at anchor attribute token in VAAR task



(g) CMA at last token in AAVR task



(h) CMA at last token in VAAR task

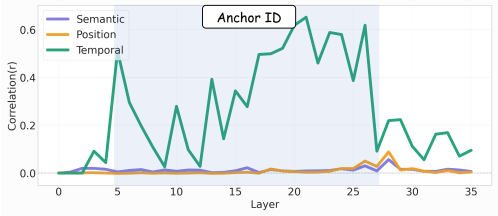
Figure 8: RSA and CMA results for Qwen2.5-Omni(7B)

Beyond the 2×2 grid, with acoustic noise. To demonstrate that our trimodal binding mechanism generalizes beyond the 2×2 grid, we applied Sec. 3.3 RSA to five- and six-speaker settings, including five-speaker variants with MUSAN noise [44] at 20 and 10 dB SNR to approximate realistic noisy conditions. In the five-speaker setting, three speakers are arranged in the top row and two in the bottom row; in the six-speaker setting, three speakers are arranged in each row. For each hypothesized representational space—semantic content, Position IDs, and Temporal IDs—we report the maximum correlation and its corresponding layer. Table 6 reports video-SALMONN2+ (7B) results and aligns with the proposed mechanism. At the anchor token in the middle layers, the modality-specific anchor ID (Temporal ID for AAVR; Position ID for VAAR) dominates (anchor ID retrieval). At the last prompt token, the complementary target ID (Position ID for AAVR; Temporal ID for VAAR) becomes dominant in later layers (target ID selection)—followed by semantic information in the deepest layers (feature retrieval).

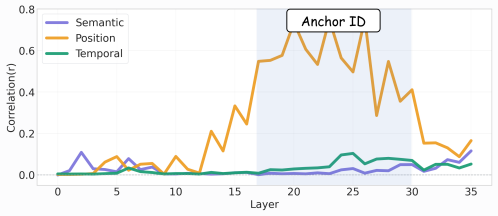
These results show that the trimodal binding mechanism is not merely an artifact of the specific 2×2 grid structure, but generalizes across broader synthetic and real-world settings.

A.4 Representation Analysis on Different Models on Primed vs. Unprimed Settings

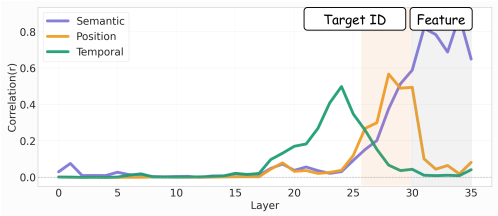
Figs. 11 to 14 present the representation analysis results comparing the unprimed setting and primed conditions for both the AAVR and VAAR tasks across four models: video-SALMONN2+, Qwen2.5-



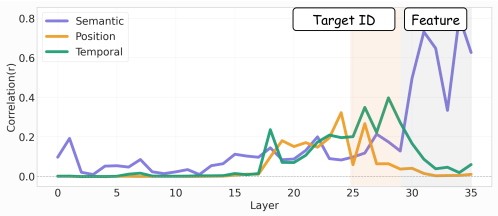
(a) RSA at anchor attribute token in AAVR task



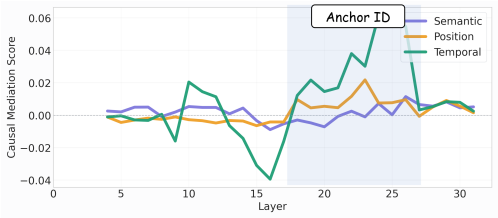
(b) RSA at anchor attribute token in VAAR task



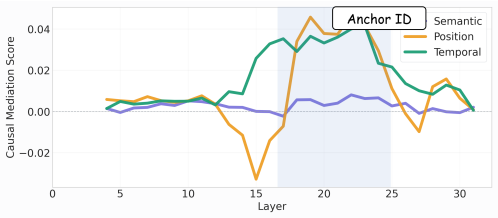
(c) RSA at last token in AAVR task



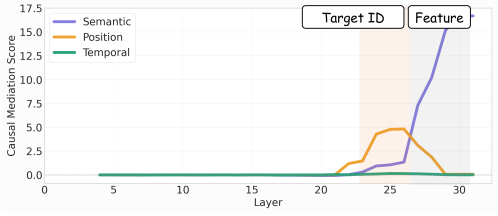
(d) RSA at last token in VAAR task



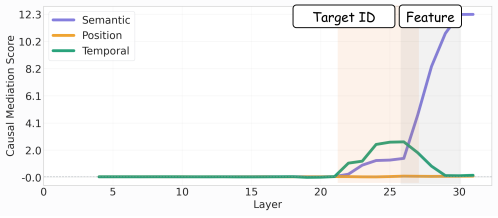
(e) CMA at anchor attribute token in AAVR task



(f) CMA at anchor attribute token in VAAR task



(g) CMA at last token in AAVR task



(h) CMA at last token in VAAR task

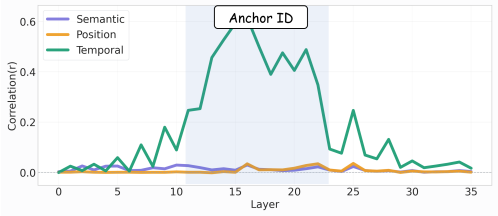
Figure 9: RSA and CMA results for Qwen2.5-Omni(3B)

Omni(7B), Qwen2.5-Omni(3B), and MiniCPM-o-4.5. Across all cases, a common trend emerges: during the anchor ID retrieval stage, there is no observable difference between the unprimed and primed settings. This indicates that this initial stage functions correctly even in the absence of explicit priming. However, a significant divergence occurs during the subsequent target ID selection stage, where the unprimed condition clearly struggles to perform effectively compared to the primed condition.

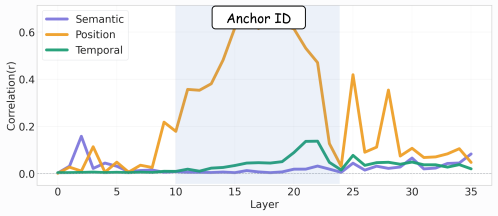
A.5 Real World Generalization Analysis on Different Models

We conduct RSA on approximately 2,000 SocialOmni [52] two-speaker clips. We extract segments in which two speakers are positioned left and right and each produces one utterance, then annotate speaker position and utterance order to construct RSA-compatible datasets. Although this forms a 2×1 layout, speaker locations vary freely across clips rather than occupying fixed grid positions. As in Section 3.5, we report Position ID and Temporal ID correlations.

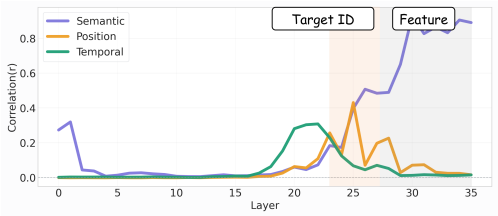
For our study, we utilize a subset restricted to videos containing exactly two individuals. We enrich this data by extracting explicit visual attributes. Specifically, we sample a single frame from each video and prompt GPT-4o-mini to generate detailed descriptions of both speakers' appearances and attire (e.g., "a man wearing a white shirt and a hat").



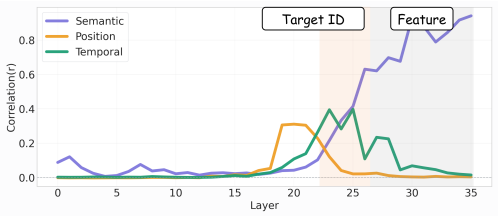
(a) Prompt tokens in AAVR case



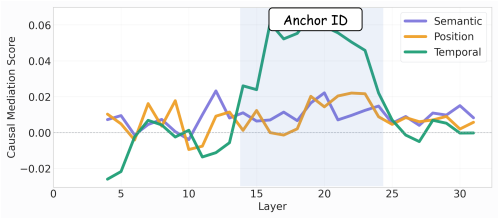
(b) Prompt tokens in VAAR case



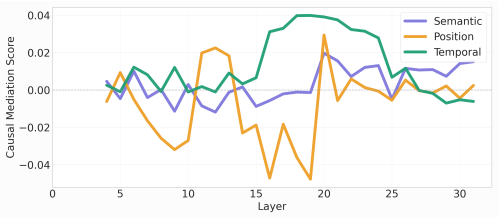
(c) Last tokens in AAVR case



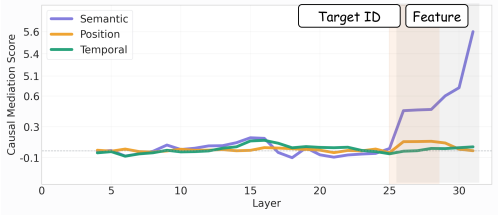
(d) Last tokens in VAAR case



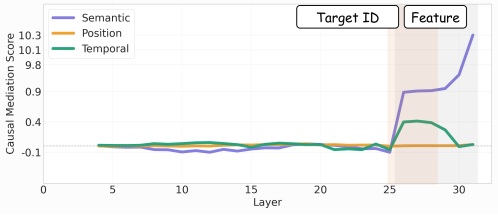
(e) Prompt tokens in AAVR case



(f) Prompt tokens in VAAR case

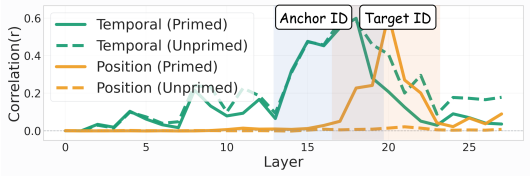


(g) Last tokens in AAVR case

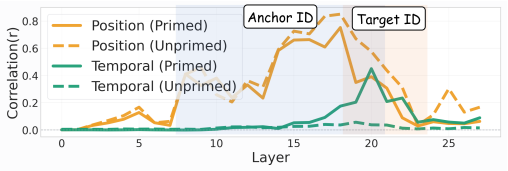


(h) Last tokens in VAAR case

Figure 10: RSA and CMA results for MiniCPM-o-4.5(9B)



(a) RSA results for primed and unprimed setting in AAVR task



(b) RSA results for primed and unprimed setting in VAAR task

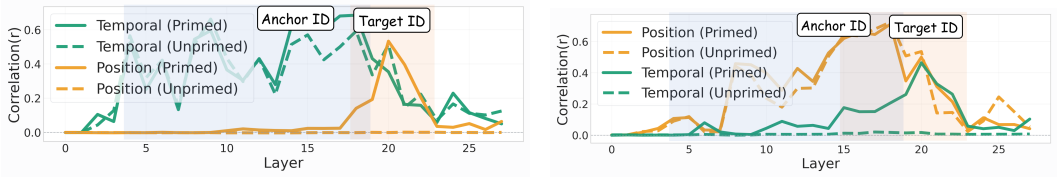
Figure 11: RSA results primed vs unprimed settings for video-SALMONN2+(7B)

Furthermore, we generate timestamped audio transcriptions using Whisper-large-v3 [42]. By aligning these transcriptions with the original active speaker bounding box annotations, we accurately attribute each spoken utterance to either the "left" or "right" speaker. Finally, we formulate our AAVR and VAAR tasks using the first two utterances from each speaker.

Figs. 15 to 18 present the results of replicating the real-world dataset experiments (Sec. 3.5) on video-SALMONN2+(7B), Qwen2.5-Omni (7B, 3B), and MiniCPM-o-4.5, respectively. Overall, while the patterns are less pronounced than those seen in the toy dataset, the observed trends remain consistent with our proposed mechanism. In the AAVR case, the temporal ID exhibits a strong

Table 6: RSA in five-/six-speaker settings (video-SALMONN2+, 7B). Each cell reports the peak correlation and its layer. **Bold**: highest among the three spaces; underline: second highest.

Data	Token	AAVR			VAAR		
		Semantic	Position	Temporal	Semantic	Position	Temporal
5spk	Anchor	0.03@L20	0.07@L20	**0.52@L18**	0.15@L20	**0.66@L18**	0.17@L20
5spk	Last	**0.83@L23**	0.47@L20	0.27@L18	**0.81@L23**	0.32@L18	0.34@L20
6spk	Anchor	0.03@L20	0.06@L20	**0.45@L18**	0.09@L20	**0.57@L18**	0.10@L20
6spk	Last	**0.65@L24**	0.34@L20	0.22@L18	**0.74@L23**	0.32@L18	0.26@L20
5spk, 20dB	Anchor	0.04@L20	0.08@L20	**0.52@L18**	0.15@L20	**0.66@L18**	0.18@L20
5spk, 20dB	Last	**0.82@L23**	0.47@L20	0.24@L18	**0.79@L23**	0.32@L18	0.34@L20
5spk, 10dB	Anchor	0.04@L20	0.08@L20	**0.52@L18**	0.15@L20	**0.66@L18**	0.18@L20
5spk, 10dB	Last	**0.82@L23**	0.47@L20	0.23@L18	**0.78@L23**	0.31@L18	0.33@L20



(a) RSA results for primed and unprimed setting in AAVR task

(b) RSA results for primed and unprimed setting in VAAR task

Figure 12: RSA results primed vs unprimed settings for Qwen2.5-Omni(7B)

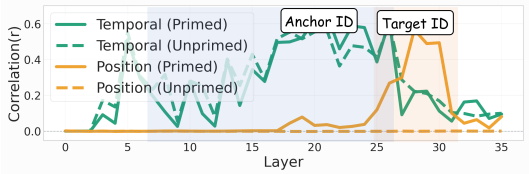
correlation at the anchor attribute token within the mid-to-late layers. Subsequently, at the last token—which is expected to encode the target entity from the opposite modality—the correlation with position IDs become stronger than at the anchor token, whereas the correlation with temporal IDs weakens. Conversely, in the VAAR case, while the temporal ID correlation at the anchor attribute token appears somewhat stronger in the early layers compared to the toy dataset, the position ID correlation remains distinctly dominant in the mid-to-late layers. Furthermore, at the last token, the temporal ID correlation becomes even stronger than at the anchor attribute token, whereas the position ID correlation noticeably weakens.

B Audio-Visual Prompting

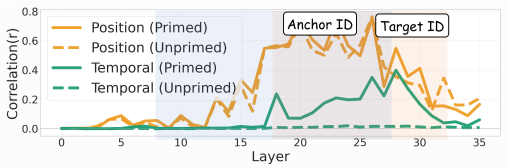
B.1 Details of Training Free Prompting

To leverage the zero-shot reasoning capabilities without additional parameter updates, we employ a training-free prompting strategy that explicitly guides the model's spatial and temporal attention toward the active speaker by incorporating visual cues directly into the text prompt. We modify the original prompts according to the specific task: for Question Answering (QA) tasks, where precise grounding is crucial, we prepend the instruction, *"At each moment, focus on the person with the red bounding box, as the red box marks the person that is currently speaking. [Original Prompt]"*. For Captioning tasks, we ensure narrative consistency by appending the note, *"[Original Prompt] Note that the person with the red bounding box at each moment marks the person that is currently speaking."* By defining the red bounding box as a active speaker, this prompting mechanism minimizes identity ambiguity in multi-speaker scenarios and effectively aligns the model's linguistic output with the target speaker's actions and speech.

Sensitivity to ASD Tab. 1 already uses imperfect annotations from a SOTA ASD model [37] Even with two weaker ASD models [49, 25]. ASD-FT reaches 48.4%/48.6% on AVSpeaker—close to the 49.1% with [37] and well above vanilla 44.7%—confirming robustness.

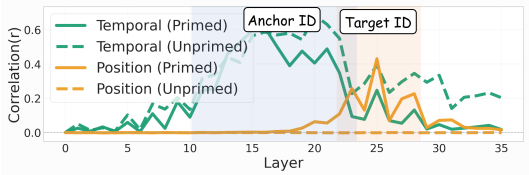


(a) RSA results for primed and unprimed setting in AAVR task

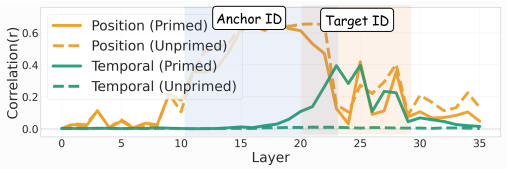


(b) RSA results for primed and unprimed setting in VAAR task

Figure 13: RSA results primed vs unprimed settings for Qwen2.5-Omni(3B)



(a) RSA results for primed and unprimed setting in AAVR task



(b) RSA results for primed and unprimed setting in VAAR task

Figure 14: RSA results primed vs unprimed settings for MiniCPM-o-4.5

B.2 Details of Finetuning

B.2.1 Training Datasets

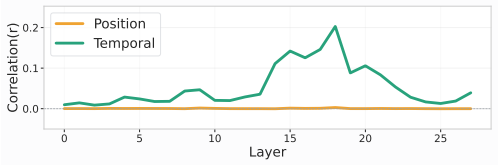
We construct the training corpus from a synthetic, human-character video dataset in which one or more speakers each utter a single country name (*Canada*, *Germany*, *Japan*, or *Egypt*). The active speaker at every moment is highlighted by a red bounding box, providing an unambiguous visual–audio–text alignment supervision signal. As illustrated in Fig. 19, the corpus contains three video configurations that vary along two axes: the number of people on screen and whether all speakers appear together in a single shot or are presented as a sequence of shots (multi-shot).

Asset pool. All clips are composed from a small, fixed asset pool, which makes per-axis ablation analyses tractable while still producing diverse compositions. The pool consists of (i) eight human *characters* – four female and four male – whose reference images are generated with Gemini and used consistently throughout the dataset; (ii) four *country names* that constitute the entire spoken vocabulary, namely *Canada*, *Germany*, *Japan*, and *Egypt*; and (iii) a set of TTS-generated *voices* used to synthesize the spoken country names. Each video is built by sampling a layout of characters, a per-character country word, and a per-character voice from these pools, so that visual identity, lexical content, and acoustic identity are independently controlled.

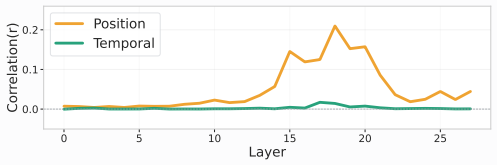
Video configurations. Tab. 7 summarizes the three configurations. The two-people single-shot setting (Fig. 19b) shows two speakers simultaneously in one continuous shot, with one word spoken per person. The four-people single-shot setting (Fig. 19a) extends this to four simultaneously visible speakers. The two-people multi-shot setting (Fig. 19c) again involves two speakers, but they appear in separate consecutive shots rather than together, requiring the model to track identity across shot boundaries.

Table 7: Video configurations in the training set. Each unique video yields multiple training samples by pairing it with different question types and target utterances.

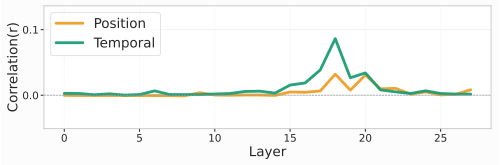
# People	Shot type	# Videos
2	Single-shot	200
4	Single-shot	100
2	Multi-shot	100
Total		400



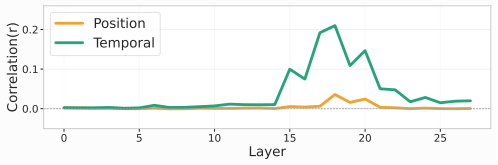
(a) RSA at anchor attribute token in AAVR task



(b) RSA at anchor attribute token in VAAR task

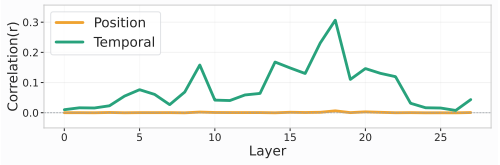


(c) RSA at last token in AAVR task

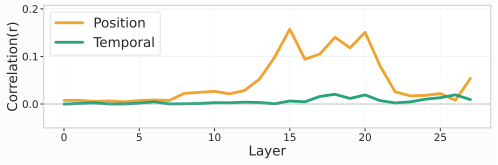


(d) RSA at last token in VAAR task

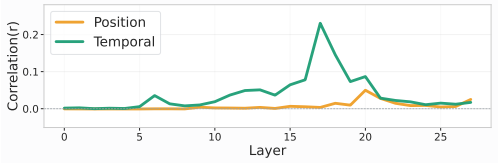
Figure 15: RSA results for video-SALMONN2+(7B) on real-world data.



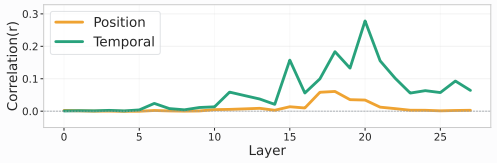
(a) RSA at anchor attribute token in AAVR task



(b) RSA at anchor attribute token in VAAR task



(c) RSA at last token in AAVR task



(d) RSA at last token in VAAR task

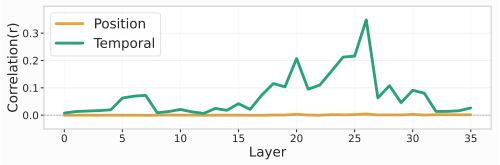
Figure 16: RSA results for Qwen2.5-Omni(7B) on real-world data.

Question types. Each video is paired with three kinds of questions, yielding 2,040 training samples in total. AAVR, VAAR, **Captioning**(It requires a free-form summary that names every speaker by appearance and reports the word each one utters). Tab. 8 reports per-setting counts.

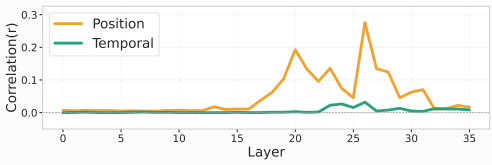
Table 8: Distribution of training samples by video setting and question type. AAVR and VAAR are 2-way multiple choice for the two-people settings and 4-way multiple choice for the four-people setting; Captioning is open-ended.

Setting	AAVR	VAAR	Captioning	Total
Two people, single-shot	400	400	13	813
Four people, single-shot	400	400	13	813
Two people, multi-shot	200	200	14	414
Total	1,000	1,000	40	2,040

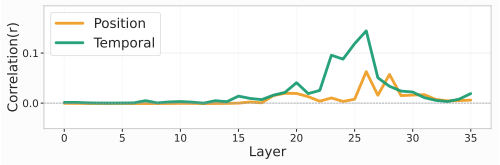
Question examples. For all multiple-choice questions, the prompt is preceded by a fixed instruction explaining that the red bounding box marks the current speaker, and the model is asked to respond with a single option letter. Representative prompts and gold answers from each question type are shown below.



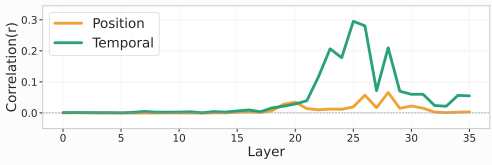
(a) RSA at anchor attribute token in AAVR task



(b) RSA at anchor attribute token in VAAR task

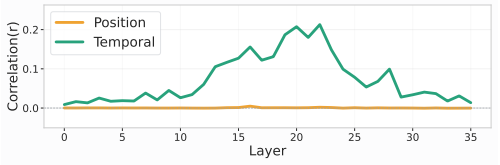


(c) RSA at last token in AAVR task



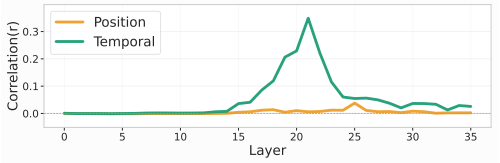
(d) RSA at last token in VAAR task

Figure 17: RSA results for Qwen2.5-Omni(3B) on real-world data.

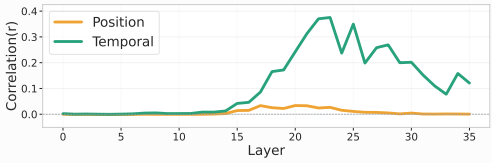


(a) RSA at anchor attribute token in AAVR task

(b) RSA at anchor attribute token in VAAR task



(c) RSA at last token in AAVR task



(d) RSA at last token in VAAR task

Figure 18: RSA results for MiniCPM-o-4.5(9B) on real-world data.

AAVR (4-way)

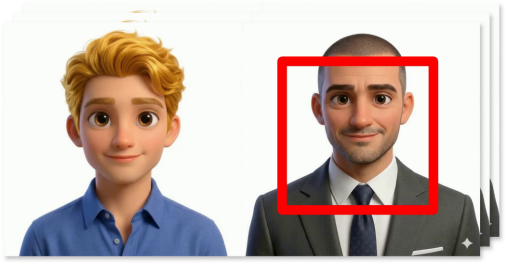
Prompt: Which person speaks the 2nd utterance in the clip (the word "Egypt")? Note that the person with the red bounding box at each moment marks the person that is currently speaking.
A. A man with black hair wearing a gray T-shirt
B. A long-haired blond woman dressed in blue
C. A man dressed formally in a suit with short hair
D. A brown-haired woman in a black shirt
Answer: B

VAAR (2-way)

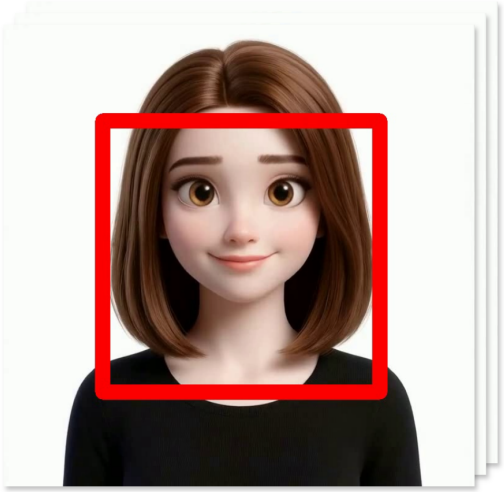
Prompt: What single word does a woman with long blond hair in a blue shirt speak? Note that the person with the red bounding box at each moment marks the person that is currently speaking.
A. Egypt
B. Japan
Answer: A



(a) Four people talking single-shot video



(b) Two people talking single-shot video



(c) Two people talking multi-shot video

Figure 19: Examples of the three video configurations in our training corpus. The active speaker at each moment is marked with a red bounding box.

Captioning (open-ended)

Prompt: Provide a caption for this video, including the people present and the words they speak. Note that the person with the red bounding box at each moment marks the person that is currently speaking.
Answer: A sequence of four speakers is shown, each contributing one word. It starts with a red-haired woman in a red shirt saying "Japan". A blond-haired man in a blue shirt then says "Egypt". Then, a woman with long yellow hair wearing a blue shirt continues with "Germany". Finally, a black-haired man in a gray shirt says "Canada".

B.2.2 Implementation Details

Due to its rapid convergence, video-SALMONN2+ (7B) was trained for 140 steps, whereas the remaining models were trained for 250 steps. Furthermore, because our training data is synthetic, we bypassed off-the-shelf active speaker detection models; instead, we directly generated the exact bounding boxes during the data synthesis process. Evaluation is performed with greedy decoding.

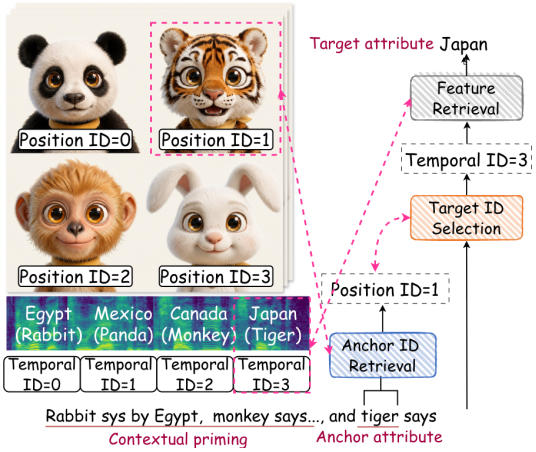


Figure 20: Illustration of trimodal symbolic binding mechanisms in VAAR task.

The $\pm$ values reported in Tab. 1 reflect variance across multiple Gemini-judge evaluations of the same model outputs.

B.3 Discussion on the Experimental Results

Training-Free ASD Results The effectiveness of training-free ASD relies heavily on an AVLLM's zero-shot visual grounding capability to interpret overlaid markers-such as red bounding boxes-as indicators of the active speaker. Qwen2.5-Omni successfully leverages these overlays, supported by its explicit reports on referring expression comprehension and grounding benchmarks [53]. Similarly, MiniCPM-o-4.5 effectively utilizes this approach, as its training incorporates a relevance-aware masking strategy specifically designed to encourage the model to focus on visually grounded content [11]. Conversely, video-SALMONN2+ does not benefit from training-free ASD. Although built upon Qwen2.5-VL [2], its downstream optimization is heavily tailored toward holistic audio-visual comprehension, leaving the zero-shot interpretation of fine-grained visual markers largely unaddressed. This indicates that without specific provisions to maintain or elicit these grounding skills during adaptation, training-free ASD becomes ineffective. When this prerequisite is unmet, our lightweight fine-tuning (ASD-FT) successfully bridges the gap, effectively teaching the model to bind the visual marker to the active speaker.

C Illustrations

C.1 Illustration of Symbolic Mechanism in VAAR task

Fig. 20 illustrates how the proposed symbolic mechanism operates within the VAAR task. Unlike the AAVR task, the anchor attribute here is a visual attribute (e.g., the "tiger"). Accordingly, during the anchor ID retrieval stage, the model retrieves the position ID, which serves as the binding ID for the visual modality. In the subsequent target ID selection stage, this position ID is mapped to its corresponding temporal ID (e.g., temporal ID = 1). Finally, during the feature retrieval stage, the model attends to the segment associated with this temporal ID to extract the target attribute's feature, which in this case is "Japan."

C.2 Illustration of Causal Mediation Analysis for AAVR task

The following figures illustrate the patching configurations used in our causal mediation analysis on the AAVR task. For each of the three-stage mechanisms-*anchor ID retrieval*, *target ID selection*, and *feature retrieval*-we patch the residual stream with a c_2 obtained from one of three manipulations: temporal ID manipulation, position ID manipulation, or semantic content manipulation. This yields a 3×3 grid of patching configurations (Fig. 21–Fig. 29).

Our three-stage hypothesis makes a precise prediction about which manipulation should drive the model's output toward the counterfactual target y_1^* at each stage. Specifically, we predict that:

- Patching at the **anchor ID retrieval** stage (anchor attribute token at mid-late layers) flips the prediction to y_1^* only when c_2 is constructed via *temporal ID manipulation* (Fig. 21), while the other two manipulations leave the output essentially unchanged (Figs. 22 and 23).
- Patching at the **target ID selection** stage (last prompt token at late layers) flips the prediction to y_1^* only when c_2 is constructed via *position ID manipulation* (Fig. 25), with the other two manipulations being ineffective (Figs. 24 and 26).
- Patching at the **feature retrieval** stage (last prompt token at deepest layers) flips the prediction to y_1^* only when c_2 is constructed via *semantic content manipulation* (Fig. 29), while temporal and position ID manipulations have no comparable effect (Figs. 27 and 28).

These predictions are consistent with the empirical results reported in the main text, supporting the view that each stage encodes a distinct, causally separable type of information.

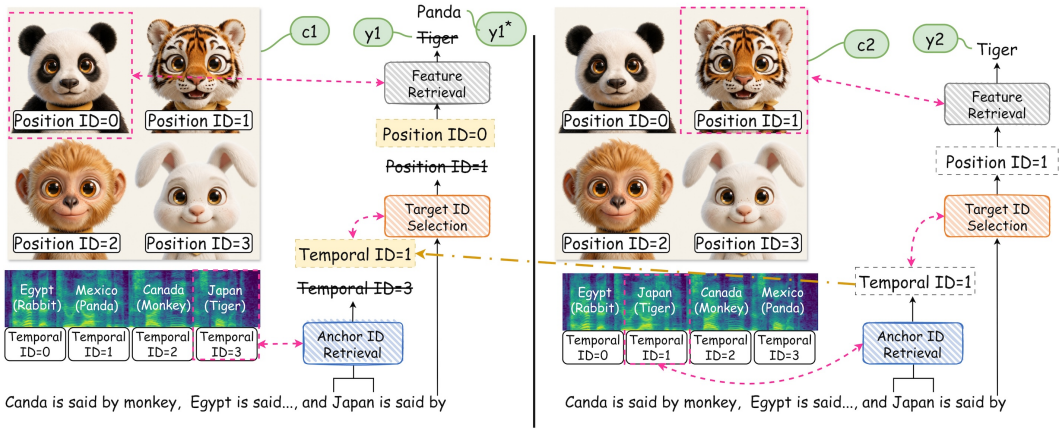


Figure 21: c_2=temporal ID manipulation, patching at the anchor ID retrieval stage

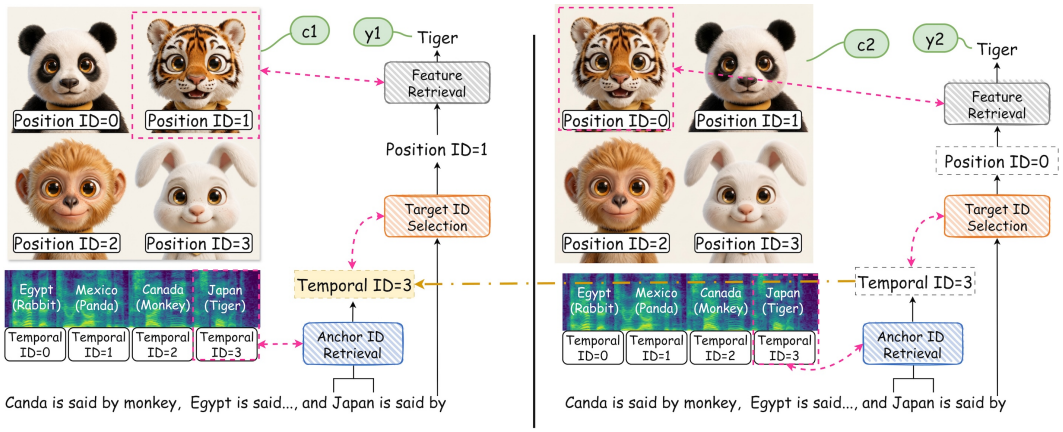


Figure 22: c_2=position ID manipulation, patching at the anchor ID retrieval stage

D Qualitative Results

We present qualitative comparisons between the vanilla Qwen2.5-Omni (7B) and our ASD/ASD-FT variants on conversation-rich videos. As illustrated in Fig. 30, the vanilla model frequently misattributes utterances to the wrong speaker on multi-speaker QA examples, reflecting the audio-visual binding bottleneck identified in our analysis. By overlaying a red bounding box on the active

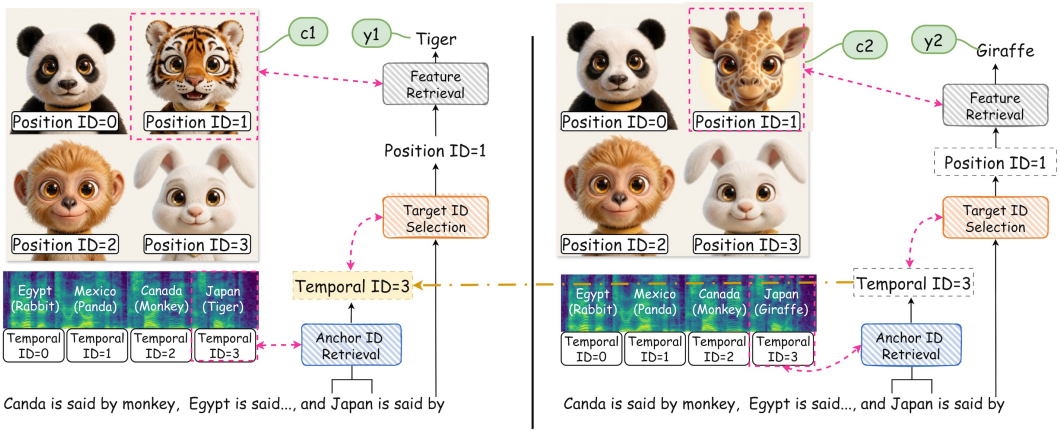


Figure 23: c_2=semantic content manipulation, Patching at the anchor ID retrieval stage

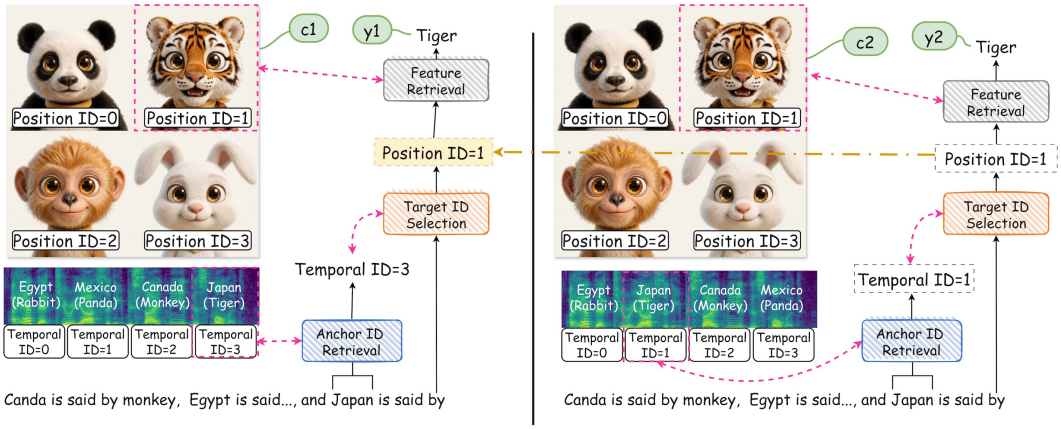


Figure 24: c_2=temporal ID manipulation, patching at the target ID selection stage

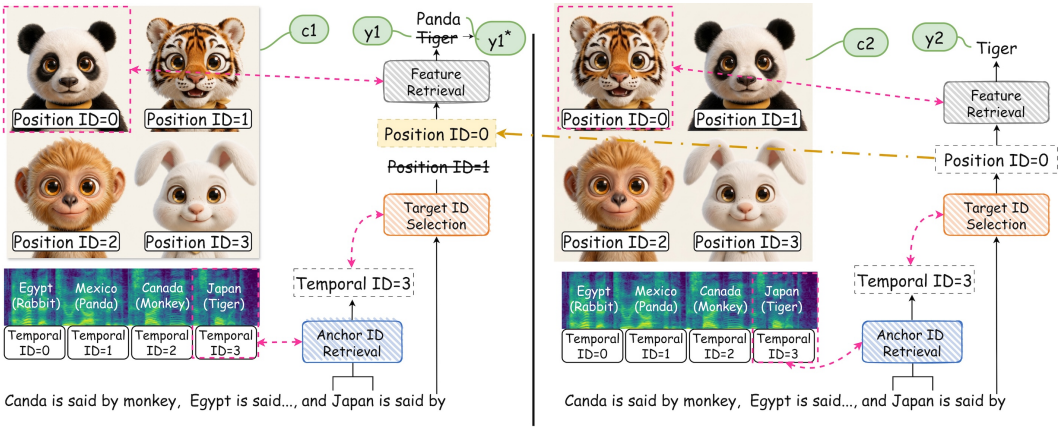


Figure 25: c_2=position ID manipulation, patching at the target ID selection stage

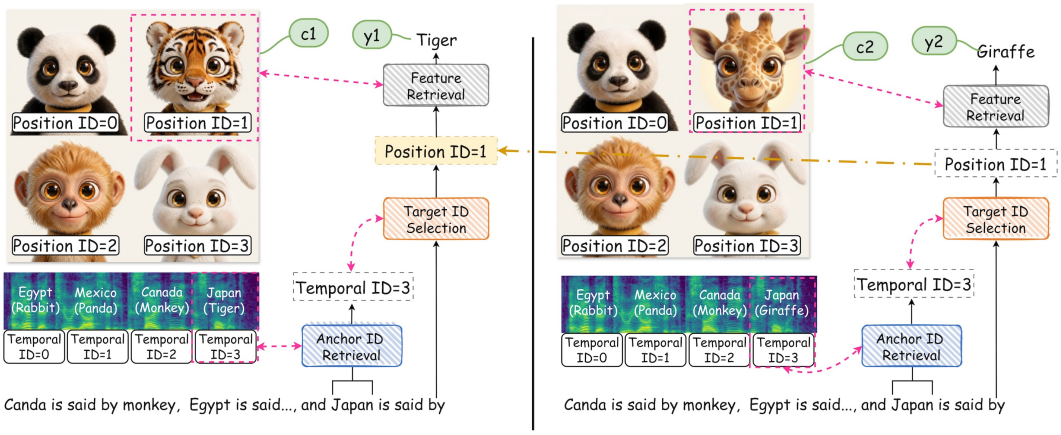


Figure 26: c_2=semantic content manipulation, patching at the target ID selection stage

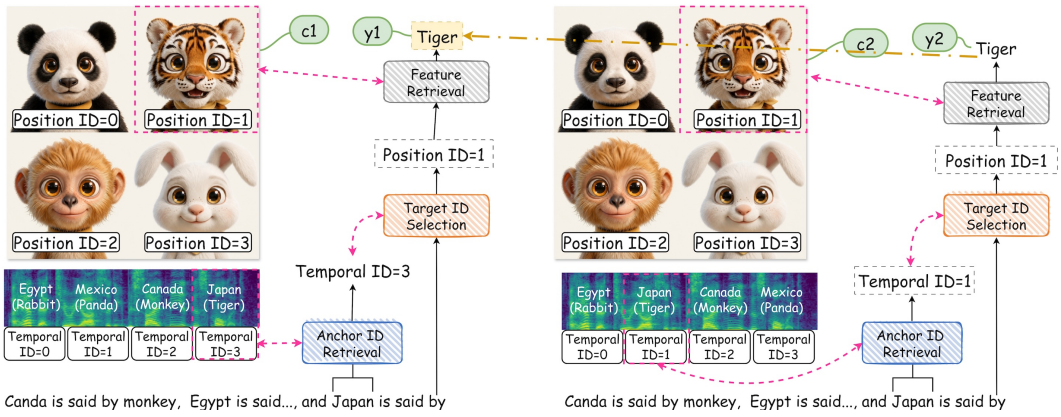


Figure 27: c_2=temporal ID manipulation, patching at the feature retrieval stage

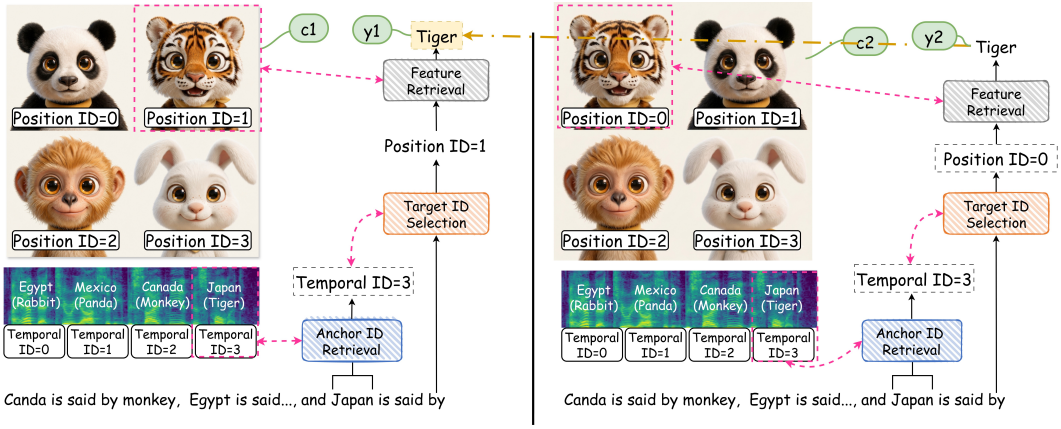


Figure 28: c_2=position ID manipulation, patching at the feature retrieval stage

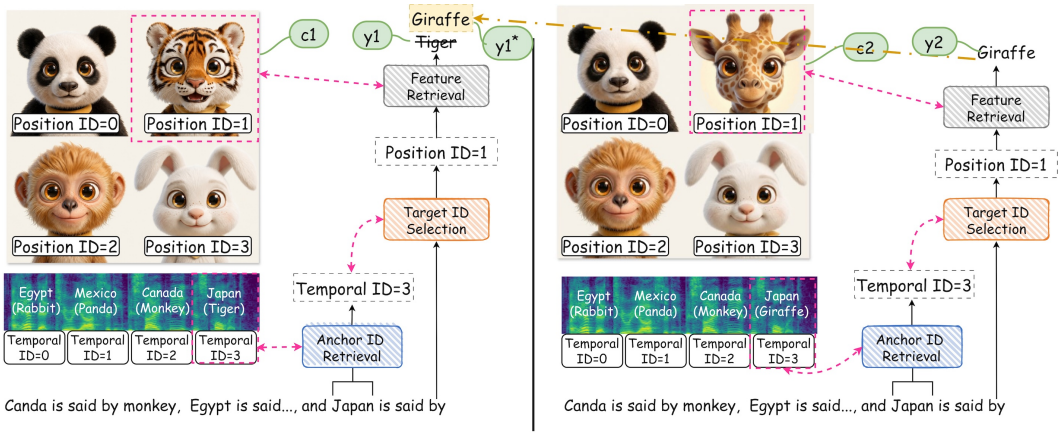


Figure 29: c_2=semantic content manipulation, patching at feature retrieval stage

speaker, our method allows the model to correctly align each utterance with its corresponding speaker. This effect also extends to free-form captioning on DiaDemBench (Fig. 31), where the ASD variant produces speaker–utterance pairings that are consistently faithful to the scene, demonstrating that the gains from audio-visual prompting generalize across both discriminative and generative settings.

E Limitations

While our study demonstrates generalizability to real-world datasets, our main analysis relies on synthetic toy datasets. This choice enables a controlled examination of the target phenomenon, and similar toy-dataset analyses have been widely adopted in prior work to isolate specific model behaviors [56, 1]. Nevertheless, such datasets do not fully capture the diversity and complexity of real-world audiovisual scenes, and thus our conclusions should be viewed as a focused analysis of the phenomenon rather than an exhaustive evaluation across all practical conditions. Following [7], we focus exclusively on the more challenging single-shot setting, where performance is typically worse than in multi-shot settings in which a single active speaker is often zoomed in per frame. As a result, we do not include an analysis of multi-shot scenarios. Additionally, while we propose a simple, efficient, and generalizable method that leverages off-the-shelf active speaker detection to mitigate the limitations of AVLLMs, our approach is not designed to modify or improve the intrinsic capabilities of the AVLLM itself. We hope that our findings serve as a stepping stone toward future work on training methodologies and model architectures that more directly address this issue.

F Computational Resource

We run most experiments on a machine equipped with an AMD EPYC 7513 32-core CPU and a single NVIDIA RTX A6000 GPU. For the WorldSense [23] evaluation, which requires larger memory, we use an NVIDIA B6000 GPU.

G Social Impact

Our work advances the ability of Audio-Visual LLMs to accurately resolve "who says what" in multi-speaker videos, which can benefit a range of socially valuable applications such as automatic captioning and transcription for the deaf and hard-of-hearing community, improved accessibility tools for educational and media content, and more reliable assistive technologies for analyzing conversational video. The mechanistic insights presented in this paper also contribute to the broader interpretability literature, helping researchers and practitioners better understand and diagnose failure modes in multimodal systems-an important step toward building more transparent and trustworthy AI. However, since our method depends on an off-the-shelf Active Speaker Detection model, any biases or demographic disparities in that module could propagate into downstream predictions, potentially affecting certain populations unequally. We encourage future work to pair these capabilities with ap-



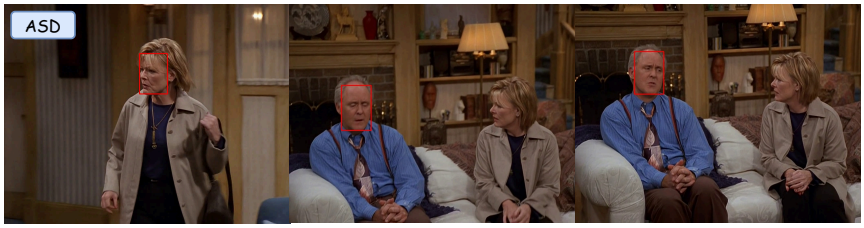
Figure 30: **Qualitative comparison on AVSpeaker[36] with Qwen2.5-Omni (7B).** The vanilla model selects the wrong answer, while the ASD/ASD-FT variant answers correctly.



Prompt : Describe every aspect of the video in full detail.

*Speaker-Speech mismatch
*Speaker-Speech match

Vanilla : The video begins with a woman with short blonde hair, wearing a beige trench coat over a dark blue top, adorned with a necklace. She is seen walking through a room with wooden walls and a staircase in the background. As she moves, she appears to be in a hurry, possibly indicating a sense of urgency or concern. The room is well-lit, with a lamp and some furniture visible, suggesting a domestic or office setting.\n\nAs the woman continues to walk, she eventually sits down on a white couch. The camera then shifts to show a man sitting next to her. The man has gray hair and is wearing a blue shirt with a tie. He appears to be engaged in a conversation with the woman, as indicated by his facial expressions and gestures. The man's body language suggests that he is explaining something to the woman, possibly sharing important information or news. The setting remains consistent with the previous scene, maintaining the same room and lighting conditions.\n\nThroughout the video, the dialogue between the two characters is crucial. The man speaks first, saying, \"What's wrong?\" This question indicates that he is concerned about the woman's well-being or is trying to understand what is troubling her. The woman responds with, \"There's something I haven't told you, Mary.\" This statement suggests that she has been withholding information from the man, and she is now revealing something significant to him. The mention of \"Mary\" implies that the woman is addressing the man by his name, which could indicate a close relationship between them, such as a family member or a close friend.\n\nOverall, the video captures a moment of tension and revelation between two characters. The woman's initial urgency and the man's concern create a sense of anticipation, while the dialogue reveals a deeper emotional connection and a need for trust and understanding between them. The setting, with its domestic or office-like atmosphere, adds to the realism and relatability of the scene.



Prompt : Describe every aspect of the video in full detail.
Note that the person with the red bounding box marks the active speaker.

ASD : The video begins with a scene set in a cozy, well-decorated living room. The room features a mix of traditional and rustic elements, including wooden furniture, a fireplace, and various decorative items on the shelves. The lighting is warm and inviting, creating a comfortable atmosphere.\n\nIn the first segment, a woman with short, light brown hair is seen walking through the room. She is wearing a light-colored trench coat over a dark top and has a necklace around her neck. Her expression appears concerned or worried as she moves through the space. The background includes a wooden door and a curtain, adding to the cozy ambiance of the room.\n\nAs the scene progresses, the woman sits down on a couch next to an older man. The man is wearing a blue shirt and a tie, and he has a serious expression on his face. The woman looks at him with a concerned expression, and the man begins to speak. The dialogue is as follows:\n\n**Woman:** \"What's wrong?\" \n**Man:** \"There's something I haven't told you, Mary.\"\n\nThe man's expression and body language suggest that he is about to reveal something important or sensitive to the woman. The woman's concerned look indicates that she is eager to hear what he has to say. The overall mood of the scene is tense and serious, with the characters' expressions and the setting contributing to the sense of anticipation and concern.

Figure 31: **Qualitative comparison on a multi-speaker captioning example from DiaDem-Bench [7] with Qwen2.5-Omni (7B).** The vanilla model produces speaker-speech mismatches, attributing utterances to the wrong characters in the scene. In contrast, the ASD variant generates a caption in which each utterance is consistently aligned with the correct speaker.

propriate consent frameworks, fairness audits across demographic groups, and dataset documentation practices to mitigate these risks.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: This is stated in the last three paragraphs of the Introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe the details of our experiments and we will release the code.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Yes. We include code in the supplementary materials and will make public upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We include the experimental setting and the details in the main paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report both the mean and standard deviation for the gemini-assisted evaluation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include the computational resources in the Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We include social impacts in the Appendix G.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper does not pose any foreseeable risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly use open source assets following their license.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We don't release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We don't use human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We don't include human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We used GPT-4o-mini to generate speaker appearance descriptions from sampled frames for constructing the real-world AAVR/VAAR analysis prompts. This usage was limited to auxiliary annotation for analysis and was not part of the proposed ASD prompting or ASD-FT method.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.