From Senses to Decisions: The Information Flow of Auditory and Visual Perception in Multimodal LLMs

Wish Suharitdamrong[1*] **Muhammad Awais**[1,2] **Xiatian Zhu**[1,2] **Sara Atito**[1,2]

[1]Surrey Institute for People-Centred AI (PAI), University of Surrey, UK
[2]Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
{w.suharitdamrong,xiatian.zhu,muhammad.awais,sara.atito}@surrey.ac.uk

Code : https://github.com/peterwisu/senses2decision

Abstract

Multimodal Large Language Models (MLLMs) can listen and see, but how do audio and visual signals actually travel through the network to shape an answer? Despite their growing role in research and real-world applications, the internal pathways through which audio and visual tokens influence the final prediction remain poorly understood. In this study, we examine audio-visual information flow inside Audio-Visual Large Language Models (AVLLMs), tracing how AVLLMs route, utilize, and integrate audio and visual information across two input configurations, audio-visual video and multiple interleaved audio-visual items. We find that for audio-visual video, AVLLMs follow the sequential information flow pathway established for VLMs and VideoLLMs, with audio and visual contribution flowing along this pathway in proportion to the task's reliance on each modality. In settings with multiple interleaved audio-visual items, this routing shifts to different parallel streams. Furthermore, we demonstrate that audio-visual and other token types can be discarded once their information is transferred to LLM, with minimal impact on the model's prediction or even slight improvement, generalizing across multiple tasks and datasets, enabling more efficient inference. These findings hold across multiple models and scales, Qwen2.5-Omni and Video-SALMONN2 Plus at 3B and 7B scales, leading to hypotheses on why these flow structures emerge. Together, these results deliver the first coherent picture of how AVLLMs orchestrate sound and sight inside the network and lay the groundwork for the next wave of interpretability, design, and efficiency advances in audio-visual and broader MLLMs.

1 Introduction

Multimodal Large Language Models (MLLMs) [50, 21] have progressed rapidly, jointly processing auditory and visual information in models that can both listen and see, bringing machine perception closer to human perception. Earlier research developed each modality independently, leading to specialized Vision-Language models (VLMs) [33, 29, 2, 4, 52, 61, 53] and Audio-Language models (ALMs) [18, 48, 14, 15, 7, 1], each operating effectively within its target modality. Recent Audio-Visual Large Language Models (AVLLMs) [56, 47, 57, 12, 6, 51] integrate visual and auditory inputs to enable unified audio-visual understanding. These models can answer questions about audio-visual scenes and transcribe visually grounded speech, tasks requiring cross-modal reasoning across the audio and visual modalities. These models span input formats from single images, videos, or audio clips to audio-visual videos and multiple interleaved audio-visual items, reaching diverse real-world

*Corresponding author

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

scenarios. Around these models, an active research landscape has emerged, including benchmarks probing audio-visual understanding across rich and complex scenarios [58, 30, 63, 32], parameter-efficient fine-tuning methods [54], adaptation-based token compression [16, 9], and training-free token compression at inference [49, 28].

In parallel, mechanistic interpretability has made significant progress in uncovering the internal mechanisms inside LLMs [37, 10, 42, 13]. Similar techniques have recently been extended to MLLMs [5, 40, 38, 62, 26, 24, 43]. In particular, attention knockout has been used to trace how cross-modal information flow emerges from image inputs in VLMs [62] and how spatiotemporal information flow emerges from video inputs in VideoLLMs [26]. Beyond these mechanistic studies, multi-image input handling has been actively studied in VLMs, surfacing failure modes and motivating mitigation strategies. Cross-image information leakage has been identified as a core failure mode, where visual content from different images entangles in the output [41]. Delimiter tokens have been examined and exploited as a mechanism to limit this entanglement [27]. Performance on multi-image tasks has also been shown to degrade as the number of input images grows [8].

While AVLLMs introduce a new dimension to machine perception through the integration of sound and sight, the internal mechanisms underlying audio-visual integration remain largely unstudied. Concurrent work on AVLLMs [43] examines audio-visual captioning and reports that cross-modal integration concentrates in deep layers. In contrast, information flow studies in VLMs [62] and VideoLLMs [26] locate cross-modal integration at earlier-to-middle layers, where visual information flows to the prediction only through the language tokens. Whether the information flow in AVLLMs aligns with these VLM and VideoLLM findings or departs from them, and how AVLLMs distribute their reliance on audio versus visual inputs along this flow, remains an open question. In particular, no prior work has examined the role of audio along the information flow in MLLMs, and it is unclear whether audio behaves similarly to visual information or follows different pathways. Additionally, in the multi-input interleaved configuration, prior work in VLMs has characterized model behavior on multi-image inputs [41, 27, 8], but the underlying information flow has not been examined, neither for multi-image inputs nor for the broader case where audio items are interleaved alongside visual items. In this study, we trace how audio and visual information jointly flow through AVLLMs to form the prediction, mapping these pathways in both input configurations and characterizing how each modality contributes along the way. Our key findings are as follows:

- **Audio-visual information does not reach the deep layers:** Video attention in later layers is dominated by attention artifacts that disproportionately attract attention, making attention allocation an unreliable indicator of information flow.
- **Task requirements steer the model's audio-visual flow:** The contribution of each modality to the prediction and the strength of interaction between audio and video vary with what the task requires, with more visual, auditory, or audio-visual alignment content depending on which is needed to answer the question.
- **Multiple independent audio-visual inputs route through parallel paths:** Independent audio and visual items interleaved with text route information to the prediction along multiple parallel paths, rather than through a single sequential path as in single audio-visual videos.
- **Tokens can be discarded after their information is transferred:** Once a token's content has been passed on, it can be discarded with minimal impact on accuracy or even slight improvement. We show this across tasks and datasets, and across input configurations, with each token type discarded at the distinct layer where its information transfer completes.

2 Preliminary on audio-visual large language models (AVLLMs)

Multimodal tokenization and sequence construction: AVLLMs process a video with its audio track and a text instruction through an autoregressive transformer over an interleaved token sequence. Let the video frames be $\mathcal{V} \in \mathbb{R}^{T\times H\times W\times 3}$, with T frames at spatial resolution $H \times W$. The frames are passed through a vision encoder and projector to produce N_V video tokens of dimension d, the audio track is processed by an audio encoder into N_A audio tokens of the same dimension, and the text instruction is tokenized into N_T text tokens. For a single audio-visual video input, AVLLMs preserve temporal alignment by interleaving audio and video tokens within fixed temporal windows. Let C denote the number of windows, and let $\mathbf{V}_c$ and $\mathbf{A}_c$ be the visual and audio tokens within

the c-th window. With system prompt, video, audio, and question segments, the full input sequence to the language model is

$$\mathcal{I} = \Big[\underbrace{s_1, \ldots, s_{N_S}}_{\text{system}} ; \underbrace{\mathbf{V}_1, \mathbf{A}_1 ; \ldots ; \mathbf{V}_C, \mathbf{A}_C}_{\text{single audio-visual video}} ; \underbrace{q_1, \ldots, q_{N_Q}}_{\text{question}} \Big], \tag{1}$$

where $s_1, \ldots, s_{N_S}$ are system-prompt tokens, $q_1, \ldots, q_{N_Q}$ are question tokens, and the total sequence length is $N = N_S + N_V + N_A + N_Q$. Beyond this single audio-visual video setting, AVLLMs also process *multi-input* sequences with multiple independent audio and visual items interleaved with text, which we describe in Section 5.

Causal self-attention: At each transformer layer ℓ, the hidden states $\mathbf{H}^\ell \in \mathbb{R}^{N \times d}$ are projected into query, key, and value matrices $\mathbf{Q}^\ell = \mathbf{H}^\ell \mathbf{W}_Q^\ell$, $\mathbf{K}^\ell = \mathbf{H}^\ell \mathbf{W}_K^\ell$, $\mathbf{V}^\ell = \mathbf{H}^\ell \mathbf{W}_V^\ell$, where $\mathbf{W}_Q^\ell, \mathbf{W}_K^\ell, \mathbf{W}_V^\ell \in \mathbb{R}^{d \times d_h}$ are learnable and d_h is the per-head dimension. The attention output is

$$\text{Attention}(\mathbf{Q}^\ell, \mathbf{K}^\ell, \mathbf{V}^\ell) = \text{softmax}\left(\frac{\mathbf{Q}^\ell (\mathbf{K}^\ell)^\top}{\sqrt{d_h}} + \mathbf{M} \right) \mathbf{V}^\ell, \tag{2}$$

where $\mathbf{M} \in \mathbb{R}^{N \times N}$ is a causal mask enforcing autoregressive decoding.

3 What attention patterns reveal about information flow?

To trace how audio-visual information reaches the prediction, a natural starting point is to examine where the model directs its attention. We do this on multiple-choice question-answering (MCQ) tasks, where the prediction is a single token (the answer letter), using Qwen2.5-Omni [56] at 3B scale as our subject model. We inspect the attention allocation of this last token which is the first generated token where the prediction is formed. Specifically, we track its allocation across layers and across token categories (system prompt, video, audio, user instruction). Figure 1 (left) shows that throughout most of the network, the last token attends predominantly to language tokens (system prompt and user instruction), and attention to multimodal tokens fades through the layers. However, attention to video sharply spikes at layer 31 and remains elevated through the final layer.

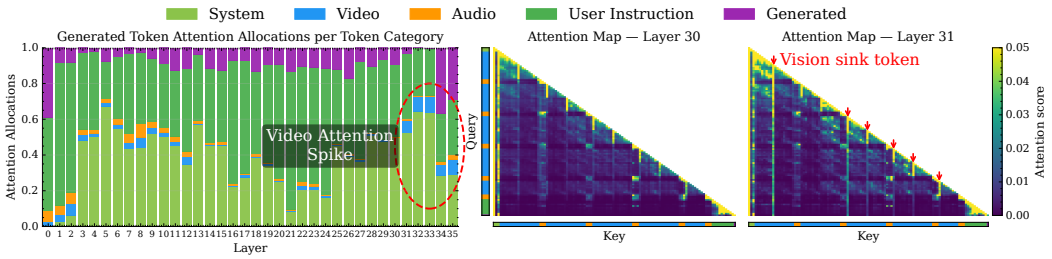

Figure 1: **Attention to video sharply spikes at layer 31.** (Left) Attention allocation of the last token per layer and token category. (Middle, Right) Attention maps at layers 30 and 31 of Qwen2.5-Omni 3B, with the vision sinks at layer 31 marked by red arrows.

To understand why this spike emerges, we examine the attention maps at layers 30 and 31 (Figure 1 middle and right). The spike is driven by a sparse set of visual tokens generally at the first visual position of frames, receiving concentrated attention at layer 31 but absent at layer 30. This behavior of visual tokens matches the visual attention sinks identified in [25, 35], and Figure 2 confirms the sink-token characteristic, with these tokens generally exhibiting much larger L_2 norms than the rest of the sequence and activating the same hidden dimensions as the language sink tokens [55, 46, 19] in the

Table 1: Masking attention to video and audio tokens at later layers (31–35) leaves AV-SpeakerBench accuracy unchanged or slightly improved.

Mask	Accuracy
Original Causal Mask	42.24
Mask video for last token	42.24
Mask video for all text	42.31
Mask video and audio for all text	42.52

system prompt. Therefore, the attention to these visual tokens is a mechanical artifact of their massive activation, not a sign of meaningful visual information. This motivates the central question: if video attention in the later layers is dominated by sinks, does any audio-visual information actually flow to the prediction through these layers? To answer this, we apply three masking conditions at layers 31 through the final layer and measure their impact on AV-SpeakerBench [39], an audio-visual video MCQ benchmark (Table 1). Across all three conditions, accuracy is unchanged or marginally improved. Despite the strong attention weights on the vision sinks, neither audio nor visual information flows through these later layers. This aligns with earlier evidence that attention weights do not necessarily reflect the importance of input tokens [22, 3], and our results identify vision sinks as a concrete source of this mismatch in AVLLMs.

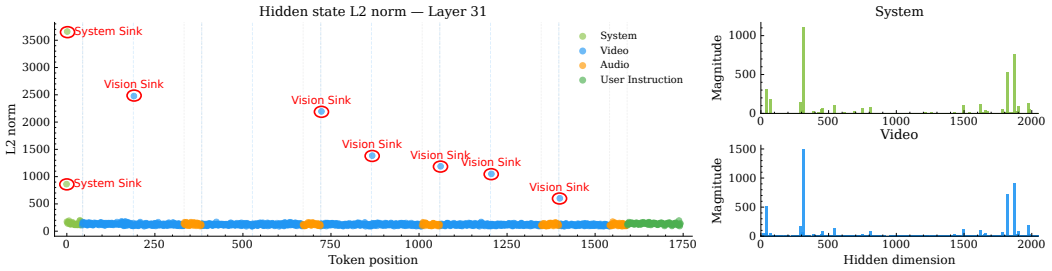

Figure 2: **Vision sinks share the same hidden-state activation as the language sinks.** (Left) Hidden state L_2 norm at layer 31, with vision sink tokens marked by red circles. (Right) Magnitude per hidden dimension for a system sink and a vision sink, with massive activation peaks at dimensions 318, 1874, and 1819 for both, on Qwen2.5-Omni 3B.

Finding 1: Attention allocation is not a reliable indicator of information flow in AVLLMs. Video attention in later layers is dominated by attention sinks, and audio-visual information does not flow through these deep layers.

4 How do audio and visual information flow in audio-visual videos?

In the previous section (Section 3), we showed that attention allocation does not reliably reveal information flow. To trace the information flow, we use causal interventions to investigate this, starting with the single audio-visual video input. Section 4.2 examines within- and cross-modal interactions, and Section 4.3 traces the route taken to the prediction. We first describe the experimental setup in Section 4.1. We then extend the analysis to the multi-input interleaved configuration in Section 5.

4.1 Experimental setup

We use *Attention Knockout* [13], a causal intervention that selectively blocks specific attention edges and measures the relative change in prediction probability. We apply it on AV-SpeakerBench [39], an audio-visual four-way MCQ benchmark. To ensure each knockout measures a degradation rather than a coincidental change, we run all interventions only on samples the model predicts correctly.

Attention knockout: Given a source set of token positions $\mathcal{S}$ (the key side) and a target set $\mathcal{T}$ (the query side), we modify the causal mask $\mathbf{M}$ from Section 2 at a chosen subset of layers $\mathcal{L}$ such that $\mathbf{M}^{\ell}_{i,j} = -\infty$ for all $i \in \mathcal{T}$, $j \in \mathcal{S}$, $\ell \in \mathcal{L}$, blocking query positions in $\mathcal{T}$ from attending to key positions in $\mathcal{S}$ while leaving all other attention edges intact. We measure the effect via the relative change in the model's probability of the predicted answer letter, $\Delta p = (p_{\text{knockout}} - p_{\text{base}})/p_{\text{base}}$, where p_{base} is the probability under the original causal mask and p_{knockout} is the probability after the mask modification. A large negative Δp indicates the blocked pathway carries information critical to the prediction, while $\Delta p \approx 0$ indicates it is not necessary. These effects therefore identify the pathways that are necessary for the model to form its prediction. Following [13, 62, 26], we localize where in the network a pathway operates by applying the knockout within a sliding window of k consecutive

Table 2: **Five representative task categories in AV-SpeakerBench**, with example questions and all four options per example (separated by ';'). Category color indicates the cross-modal direction, visual anchor → audio answer, audio anchor → visual answer, and mixed.

Task	Example question
Speech Recognition	What does the man in the blue shirt say just before he puts on a red jacket? *Options:* A. "No man"; B. "Good at games"; C. "Are you sure?"; D. "No way"
Speech Attributes	Among the people who speak, who speaks the most quietly overall? *Options:* A. The man with grey hair; B. The man with a mustache; C. The man with brown hair; D. The man with glasses
Visual Recognition	When does the man say "I have to tell you"? *Options:* A. Just before he sits down; B. Just after he sits down; C. Just before he takes off his glasses; D. Just after he takes off his glasses
Speaker Recognition	Who speaks right before the notebook is opened? *Options:* A. The man in black jacket says "it's a tiger"; B. The woman with blonde hair says "I can't"; C. The woman in blue jacket says "something to do with how I die"; D. The woman with blonde hair says "I tried to draw something different"
Speaker Detection	Does the man in the black shirt speak after the woman hands him the ring? *Options:* A. No, he only cries; B. No, he only laughs; C. Yes, he tries to defend himself; D. Yes, he tries to blame the woman

layers centered at each layer ℓ and sweeping ℓ across the full depth, yielding a curve of Δp versus center layer. We use $k = 7$ unless otherwise noted. We denote a knockout pathway as $\mathcal{S} \not\to \mathcal{T}$; for example, Video$\not\to$Question denotes blocking the question tokens from attending to the video tokens. We use $\mathcal{S} \not\leftrightarrow \mathcal{T}$ for bidirectional knockouts, where each set serves as both source and target of the other in the interleaved audio-video layout shown in Equation 1.

Dataset and tasks: AV-SpeakerBench is a speaker-centric audio-visual benchmark covering audio perception and visual understanding capabilities. Each task follows an *anchor–target* design [39] that requires cross-modal understanding, where the *anchor* is a cue in the question text pointing to an event in one modality (e.g., a visual action or a spoken phrase), and the *answer* must be read off the opposite modality at the moment the anchor identifies. We select a subset of the benchmark's tasks and group them into five representative categories (Table 2). Full details of the task selection and grouping are in Appendix D.1.

Models: We use Qwen2.5-Omni [56] at 3B scale as the main subject model throughout our analysis. Results for Qwen2.5-Omni 7B and Video-SALMONN2 Plus [47] at 3B and 7B scales are reported in Appendices I, J, and K. Full experimental details, including input settings and inference setup, are provided in Appendix C.

4.2 Do the modalities interact within themselves or with each other, and where?

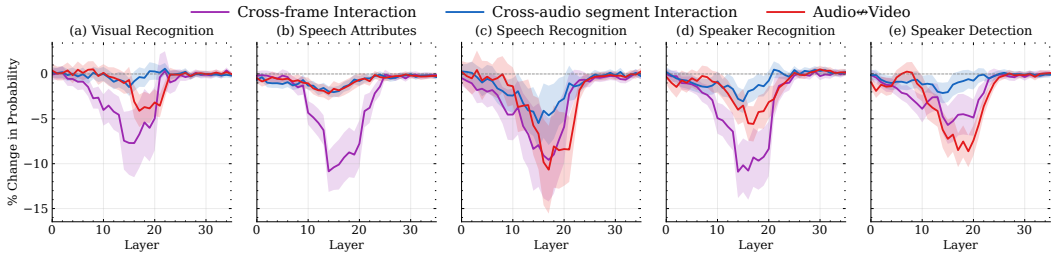

Figure 3: **Within- and cross-modal interactions concentrate at early-to-middle layers.** Change in prediction probability when disconnecting within-modality (Cross-frame, Cross-audio) and direct cross-modal (Audio$\not\leftrightarrow$Video) pathways, across layers and five AV-SpeakerBench tasks. Cross-frame attention contributes across all tasks, while cross-modal effects vary by task.

First, we investigate whether and where the two modalities interact within themselves and with each other. Following [26], we block cross-frame attention within the video stream, and we addition-

ally block cross-chunk attention within the audio stream and bidirection cross-modal interaction between both modalities (Audio$\not\leftrightarrow$Video). Figure 3 shows several patterns. First, all interactions are concentrated in the early-to-middle layers, with cross-modal exchange peaking shortly alongside within-modality interaction. Second, cross-frame attention contributes substantially to almost every task. This is expected for visually-grounded tasks like Visual Recognition and Speaker Recognition, but it also holds for tasks with audio-related answers. For example, in Speech Attributes, the question asks "who speaks the most quietly?" with the answer being a person's identity (e.g., "the man with glasses"). Although the answer concerns audio (relative loudness), the model must first identify *which speaker* corresponds to each visual descriptor, which is a visual task, before it can answer. Third, bidirectional cross-modal interaction varies by task. It carries substantial information for tasks that require fine-grained audio–visual alignment (Speech Recognition, Speaker Detection), where it operates shortly alongside cross-frame attention, but contributes little for tasks that can be solved through visual information alone. Fourth, cross-audio interaction has minimal impact across all tasks, likely because audio tokens within each chunk have already temporally interacted in the audio encoder before reaching the LLM, while video tokens lack such pre-interaction and rely on cross-frame attention within the LLM for temporal context.

4.3 How and where does audio-visual information travel to the prediction?

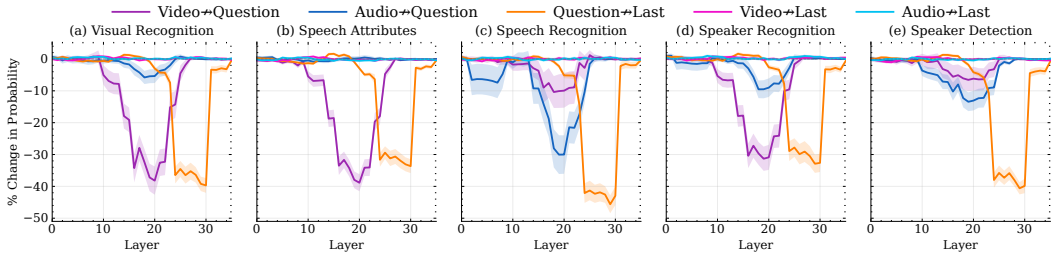

Figure 4: **Overall audio-visual information flow in AVLLMs.** Change in prediction probability across knockouts targeting the question and the last token. Source$\not\rightarrow$Target indicates blocking attention edges from source tokens to target tokens. The flow follows a single sequential pathway with no direct flow from the modalities to the last token.

Next, we trace how audio and visual information reach the prediction. Figure 4 reveals a clean indirect route, Modalities $\rightarrow$ Question $\rightarrow$ Last, aligning with prior findings on VLMs [62] and VideoLLMs [26]. At mid layers, video and audio transfer their information into the question tokens, picking up where the modality interactions of Section 4.2 left off. Since the question is positioned after the modalities in the sequence, it acts as the aggregation point for audio-visual content. The contribution of each modality along this route is shaped by the task requirements. Tasks that can be solved through visual contexts (Visual Recognition, Speech Attributes, Speaker Recognition) flow primarily through video, while tasks requiring fine-grained audio information (Speech Recognition, Speaker Detection) draw from both modalities simultaneously. At late layers, the question then carries this combined content to the last token, where the model prediction is formed. A deeper analysis of which question components (correct option, incorrect options, non-option question) carry the audio-visual content is provided in Appendix H.1.

Finding 2: Audio-visual information follows a single sequential pathway. Within- and cross-modal interactions at early-to-mid layers transfer audio-visual content into the question as the aggregation point, with their relative contribution shaped by the task requirements. Subsequently, the question carries this information to the model's prediction.

5 How does information flow across multiple interleaved inputs?

In Section 4, we traced information flow in the single audio-visual video input. However, real-world prompts often arrive as multiple independent images and audio clips interleaved with text instructions. We investigate this multi-input interleaved setting, where these independent audio-visual items are

Table 3: **Example multi-input interleaved prompts in AV-Odyssey.** The model matches a single reference in one modality against four candidates in the opposite modality, with the question text describing the task.

Direction	Example prompt
A Ref → I Cand	Which instrument illustrated in images in [img1][img2][img3][img4] do you think best matches audio [audio1] ? *Options:* second image; fourth image; third image; first image
I Ref → A Cand	Which audio among [audio1][audio2][audio3][audio4] best matches the scene shown in [img1] ? *Options:* first audio; second audio; fourth audio; third audio

interleaved with the question text. Section 5.2 traces how information from these independent sources flows to the prediction, and Section 5.3 reveals the information flow to the model's decision. We first describe the experimental setup in Section 5.1.

5.1 Experimental setup for multiple interleaved audio-visual inputs

Dataset and tasks: AV-Odyssey [17] is a multi-input audio-visual benchmark where the inputs contain multiple independent images and audio clips interleaved with the question text. We focus on the matching subset where the model matches a single *reference* (Ref) item in one modality against four *candidate* (Cand) items in the opposite modality, across two task directions, audio reference to image candidates (A Ref → I Cand) and image reference to audio candidates (I Ref → A Cand) as shown in Table 3. Within this subset, samples vary in the order of candidates, reference, and question text. We use the most common ordering in the dataset, *candidates, question, reference*, and report results averaged across the selected tasks. The knockout results for individual tasks, task selection details, and additional dataset information are in Appendices H.4 and D.3.1.

Input structure: Multi-input interleaved samples consist of several independent items (images and audio clips) interleaved with the question text without temporal alignment. We extend the notation from Section 2 as follows: $\mathbf{C}_c$ denotes the tokens of the c-th candidate, $\mathbf{Q}$ the tokens of the question text, $\mathbf{R}$ the tokens of the reference, and $\mathbf{O}_o$ the tokens of the o-th option letter. Unlike Equation 1 where the question text and option letters form a single block, the interleaved structure here separates them, with the reference appearing between the question text and the option letters. The full input sequence with candidates, question, and reference segments, is

$$\mathcal{I} = \Big[\underbrace{s_1, \ldots, s_{N_S}}_{\text{system}} ; \underbrace{\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3, \mathbf{C}_4}_{\text{candidates}} ; \underbrace{\mathbf{Q}}_{\text{question}} ; \underbrace{\mathbf{R}}_{\text{reference}} ; \underbrace{\mathbf{O}_1, \mathbf{O}_2, \mathbf{O}_3, \mathbf{O}_4}_{\text{options}} \Big]. \tag{3}$$

5.2 How does the model route information across multiple interleaved inputs?

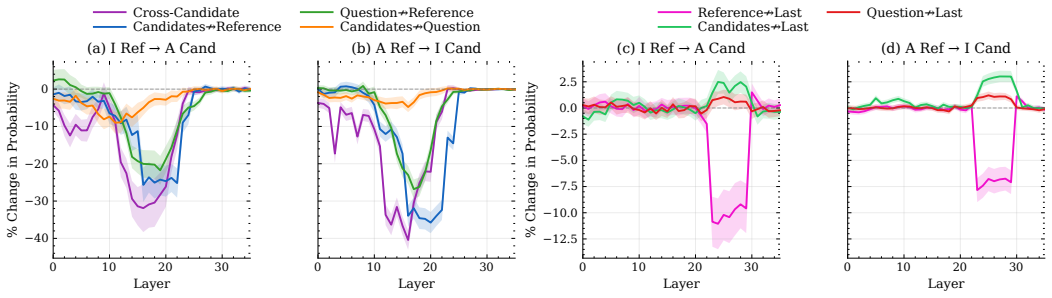

Figure 5: **Multi-input interleaved information aggregates at the late-positioned token.** At mid layers, candidates interact among themselves (Cross-Candidate), and both candidates and question transfer their content to the reference. At late layers, only the reference reaches the last token.

To trace this flow, we apply attention knockout (Section 4.1) across candidates, question, and reference. Figure 5 reveals the route Candidates + Question → Reference → Last. At mid layers, candidates exchange information among themselves (Cross-Candidate), then both the candidates and the question flow into the reference. The candidates and question reach the reference independently, with no flow between them. At late layers, only the reference flows to the model's prediction; blocking the candidates or question from the last token has negligible effect. The reference therefore plays the role that the question played in audio-visual videos (Section 4.3), with the late-positioned token serving as the aggregation point in both settings. Since the prediction is one of four answer-choice letters, how does the reference's content translate into a specific option? We answer this in Section 5.3.

5.3 How is the final answer selected?

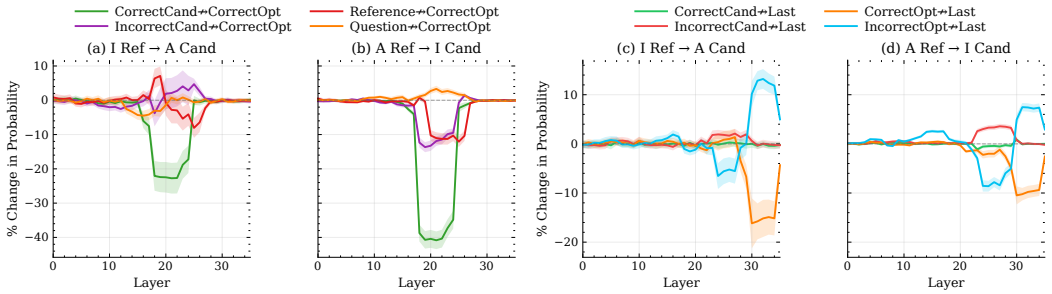

Figure 6: **The prediction flows through the option letters.** (a-b) At mid layers, the correct option letter (CorrectOpt) aggregates from the correct and incorrect candidates (CorrectCand, IncorrectCand) and the reference. (c-d) At late layers, the last token reads from both correct and incorrect option letters, with the competition between them driving the prediction.

Next, we trace how the model selects the correct option. We knock out pathways into the correct option (CorrectOpt) and from the option letters into the last token. At mid layers (Figure 6a-b), CorrectOpt draws primarily from the correct candidate, with smaller contributions from the incorrect candidates and reference, and no flow from the question. At late layers (Figure 6c-d), both correct and incorrect (IncorrectOpt) option letters flow to the last token. Blocking CorrectOpt $\not\to$ Last suppresses the correct prediction, while blocking IncorrectOpt $\not\to$ Last *increases* it, indicating the incorrect options actively compete with the correct one. The decision therefore flows through the option letters, not directly from the candidates. Together, Sections 5.2 and 5.3 reveal two parallel paths to the prediction: (1) Candidates + Question → Reference → Last and (2) Candidates → Option letters → Last. At mid layers, both paths aggregate candidate content, but only the reference receives question content, so the question reaches the prediction exclusively via the reference. At late layers, the last token integrates from both paths independently.

Finding 3: Unlike the single sequential path in audio-visual videos, multi-input interleaved information flows through two parallel paths to the prediction, each with its own late-positioned aggregation point, integrating independently at the last token.

6 Do we still need multimodal and text tokens after information transfer?

Sections 4 and 5 establish that audio, visual, and non-option question tokens transfer their content to the late-positioned aggregation token (with the non-option question analysis for video reported in Appendix H.1.2). Combined with the attention sink result of Section 3, this implies that these tokens become dispensable from the sequence once their content has been transferred and can be discarded. Previously, [62] demonstrated this token removal for image tokens in VLMs and [26] validated effective pathways in VideoLLMs through attention masking. We extend these findings to AVLLMs by discarding audio, visual, and non-option question tokens, with each token type discarded at the distinct layer where its information transfer completes, applying this to both the single audio-visual video and multi-input interleaved configurations. We evaluate discarding video, audio, and non-option

question tokens (single video) or candidates, reference, and non-option question tokens (multi-input), individually or all together. Each setting covers the task analyzed in Sections 4 and 5. We further evaluate generalization on cross-task and cross-dataset settings using AV-SpeakerBench [39], AV-Odyssey [17], and WorldSense [20] (details in Appendix D). Table 4 shows that discarding has minimal impact on model prediction and generalizes across tasks and datasets, improving model efficiency. The complete results on the cross-dataset WorldSense benchmark are provided in Appendix G.

Finding 4: Multimodal and text tokens can be discarded after their information is transferred, with each token type discarded once its information transfer is complete, with minimal impact on accuracy or slight improvement.

Table 4: **Effect of discarding multi-modal tokens on task accuracy and inference efficiency across analyzed task, cross-task, and cross-dataset settings.** L denotes the layer after which the tokens are discarded. Multi-input results are reported in both matching directions reference $\rightarrow$ candidates (I$\rightarrow$A, A$\rightarrow$I). Numbers in parentheses show change from baseline (green = improvement, red = drop); underlined indicates no change. Best per column in **bold**. **Sp.**: Speech; **Vis.**: Visual; **Rec.**: Recognition; **Count.**: Counting; **Vid.**: Video; **Aud.**: Audio; **Transp.**: Transportation; **Ques**: non-option Question tokens; **All**: all token types together.

Video *(AV-SpeakerBench knockout, Discard Video & Audio at $L = 26$, Ques at $L = 29$; cross-dataset from WorldSense)*

Config / Task	*Tasks in Knockout*		*Cross-task*		*Cross-dataset*		Avg. Prefill Latency (ms)
	Sp. Rec.	Vis. Rec.	Vis. Count.	Sp. Count.	Vid. Emotion	Aud. Change	
Baseline	50.25	46.58	43.9	26.39	66.67	42.22	2288.65
Discard Ques	50.25	46.83 (+0.25)	**44.39** (+0.49)	26.39	66.67	40.00 (-2.22)	2279.97
Discard Audio	50.25	**47.55** (+0.97)	**44.39** (+0.49)	**26.74** (+0.35)	66.67	40.00 (-2.22)	2232.45
Discard Video	**50.75** (+0.50)	46.10 (-0.48)	43.9	**26.74** (+0.35)	66.67	40.00 (-2.22)	2098.75
Discard All	49.75 (-0.50)	46.59 (+0.01)	42.93 (-0.97)	26.04 (-0.35)	66.67	**42.22**	2089.47

Multi-input *(AV-Odyssey knockout, Discard Cand at $L = 25$, Discard Ref at $L = 31$, Discard Ques at $L = 29$)*

Config / Task	*Tasks in Knockout*		*Cross-task*				Avg. Prefill Latency (ms)
	Animal Rec.		Bird Rec.		Transp. Rec.		
	A$\rightarrow$I	I$\rightarrow$A	A$\rightarrow$I	I$\rightarrow$A	A$\rightarrow$I	I$\rightarrow$A	
Baseline	61.00	38.00	29.41	33.67	46.67	23.16	558.75
Discard Ques	62.00 (+1.00)	38.00	30.39 (+0.98)	**34.69** (+1.02)	44.76 (-1.91)	24.21 (+1.05)	550.41
Discard Ref	**63.00** (+2.00)	**40.00** (+2.00)	29.41	33.67	46.67	**25.26** (+2.10)	552.11
Discard Cand	**63.00** (+2.00)	39.00 (+1.00)	**32.35** (+2.94)	31.63 (-2.04)	**47.62** (+0.95)	**25.26** (+2.10)	533.07
Discard All	**63.00** (+2.00)	38.00	**32.35** (+2.94)	33.67	46.67	24.21 (+1.05)	530.62

7 Discussion, future works and limitations

Discussion: We present the first comprehensive analysis of information flow in AVLLMs across single audio-visual video and multi-input interleaved configurations. In both configurations, modality content reaches the prediction by being aggregated into a token positioned later in the sequence. A plausible explanation for this aggregator emergence is causal attention, which makes tokens positioned later in the sequence structurally available to absorb upstream content, paired with the finding from [26] that modality tokens progressively align with linguistic embeddings at mid layers, which could plausibly provide the semantic basis for cross-modal information transfer. Concurrent work [43] reports that cross-modal integration concentrates in deep layers in captioning. Our analysis offers an alternative view in the question-answering setting, where attention to video tokens at the late layers is dominated by sinks (Section 3) and discarding different token segments at the distinct layers where their transfer completes has minimal impact on performance (Section 6). The actual integration takes place much earlier, at mid layers where modality tokens transfer to the aggregator. What appears as deep-layer integration in captioning may therefore reflect attention to sinks rather

than meaningful integration, though captioning and question-answering may engage distinct routing at late layers.

Additional analyses: Furthermore, we show in Appendix B that vision sinks emerge later in the network than the language sink. In addition, we show in Appendix F that our findings remain consistent across different frame rates and spatial resolutions, and in Appendix H.2 that they hold on VideoMME tasks that rely less on audio information. Finally, we show in Appendices H.3, I.2, J.2, and K.2 that the same information flow holds on WorldSense across all models.

Future work: Our findings open several research directions. First, our finding that tokens can be discarded after their information is transferred (Section 6) opens a new direction for AVLLM efficiency through token compression at the LLM's internal layers, complementing existing input-level methods that compress audio-visual tokens before they enter the LLM [16, 9, 49, 28]. Second, the task-dependent modality contribution we observe (Section 4) raises the question of whether steering AVLLMs to rebalance modality reliance could improve performance for tasks where one modality is underused. Third, visual bias in AVLLMs has been reported from [43] through counterfactual analysis where audio and video are intentionally mismatched. Extending our information flow analysis to these conditions is a natural next step for understanding where visual bias emerges along the routing pathway. Fourth, complementing attention knockout with other mechanistic techniques such as causal patching [36] could provide further evidence on how these pathways operate under the model's original inference conditions.

Limitations: Our analysis operates in the MCQ setting, where the prediction is a single answer letter. Open-ended generation tasks such as captioning or free-form dialogue may engage distinct pathways that our setup does not capture.

Acknowledgments

We acknowledge EuroHPC Joint Undertaking for awarding us access to MareNostrum5 as BSC, Spain. Use as many instances of the pattern MareNostrum5 at BSC, Spain as the number of systems awarded via EuroHPC. We also acknowledge the support of the Surrey Institute for People-Centred Artificial Intelligence (PAI) at the University of Surrey through a doctoral studentship.

References

[1] T. Alex, W. Suharitdamrong, S. Atito, A. Mustafa, P. J. Jackson, I. Razzak, and M. Awais. Pal: Probing audio encoders via llms–a study of information transfer from audio encoders to llms. *arXiv e-prints*, pages arXiv–2506, 2025.

[2] X. An, Y. Xie, K. Yang, W. Zhang, X. Zhao, Z. Cheng, Y. Wang, S. Xu, C. Chen, D. Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

[3] B. Bai, J. Liang, G. Zhang, H. Li, K. Bai, and F. Wang. Why attentions may not be interpretable? In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 25–34, 2021.

[4] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[5] S. Basu, M. Grayson, C. Morrison, B. Nushi, S. Feizi, and D. Massiceti. Understanding information storage and transfer in multi-modal large language models. *Advances in Neural Information Processing Systems*, 37:7400–7426, 2024.

[6] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

[7] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

[8] A. Das, A. Bulat, A. Baldrati, I. M. Metaxas, B. Schiele, G. Tzimiropoulos, and B. Martinez. More images, more problems? a controlled analysis of vlm failure modes. *arXiv preprint arXiv:2601.07812*, 2026.

[9] Y. Ding, Y. Ji, J. Li, X. Liu, X. Chen, J. Wu, B. Li, B. Zeng, Y. Shi, Y. Guan, et al. Omnisift: Modality-asymmetric token compression for efficient omni-modal large language models. *arXiv preprint arXiv:2602.04804*, 2026.

[10] N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

[11] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24118. IEEE, 2025.

[12] C. Fu, H. Lin, X. Wang, Y.-F. Zhang, Y. Shen, X. Liu, H. Cao, Z. Long, H. Gao, K. Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025.

[13] M. Geva, J. Bastings, K. Filippova, and A. Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, 2023.

[14] S. Ghosh, S. Kumar, A. Seth, C. K. R. Evuru, U. Tyagi, S. Sakshi, O. Nieto, R. Duraiswami, and D. Manocha. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6288–6313, 2024.

[15] A. Goel, S. Ghosh, J. Kim, S. Kumar, Z. Kong, S.-g. Lee, C.-H. H. Yang, R. Duraiswami, D. Manocha, R. Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*, 2025.

[16] C. Gong, D. Wang, Z. Wei, Y. Guo, H. Zhu, and J. Chen. Echoingpixels: Cross-modal adaptive token reduction for efficient audio-visual llms. *arXiv preprint arXiv:2512.10324*, 2025.

[17] K. Gong, K. Feng, B. Li, Y. Wang, M. Cheng, S. Yang, J. Han, B. Wang, Y. Bai, Z. Yang, et al. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? *arXiv preprint arXiv:2412.02611*, 2024.

[18] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*, 2023.

[19] X. Gu, T. Pang, C. Du, Q. Liu, F. Zhang, C. Du, Y. Wang, and M. Lin. When attention sink emerges in language models: An empirical view. *arXiv preprint arXiv:2410.10781*, 2024.

[20] J. Hong, S. Yan, J. Cai, X. Jiang, Y. Hu, and W. Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025.

[21] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[22] S. Jain and B. C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, 2019.

[23] J. Jung, C. Jung, J.-H. Kim, and J. S. Chung. Probing cross-modal information hubs in audio-visual llms. *arXiv preprint arXiv:2605.10815*, 2026.

[24] O. Kaduri, S. Bagon, and T. Dekel. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14549–14558, 2025.

[25] S. Kang, J. Kim, J. Kim, and S. J. Hwang. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*, 2025.

[26] M. Kim, T. Kim, and B. Han. Map the flow: Revealing hidden pathways of information in videollms. *arXiv preprint arXiv:2510.13251*, 2025.

[27] M. Lee, Y. Park, D. Hwang, Y. Kim, S. J. Oh, and J. Choe. Enhancing multi-image understanding through delimiter token scaling. *arXiv preprint arXiv:2602.01984*, 2026.

[28] B. Li and T. Huang. Dash: Dynamic audio-driven semantic chunking for efficient omnimodal token compression. *arXiv preprint arXiv:2603.15685*, 2026.

[29] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[30] C. Li, Y. Chen, Y. Ji, J. Xu, Z. Cui, S. Li, Y. Zhang, W. Wang, Z. Song, D. Zhang, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. *arXiv preprint arXiv:2510.10689*, 2025.

[31] Y. Li, X. Chen, S. Jiang, H. Shi, Z. Liu, X. Zhang, N. Deng, Z. Xu, Y. Ma, M. Zhang, et al. Uni-moe-2.0-omni: Scaling language-centric omnimodal large model with advanced moe, training and data. *arXiv preprint arXiv:2511.12609*, 2025.

[32] Y. Li, Y. Ma, G. Zhang, R. Yuan, K. Zhu, H. Guo, Y. Liang, J. Liu, Z. Wang, J. Yang, et al. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*, 2024.

[33] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[34] Z. Liu, Y. Dong, J. Wang, Z. Liu, W. Hu, J. Lu, and Y. Rao. Ola: Pushing the frontiers of omni-modal language model. *arXiv preprint arXiv:2502.04328*, 2025.

[35] J. Luo, W.-C. Fan, L. Wang, X. He, T. Rahman, P. Abolmaesumi, and L. Sigal. To sink or not to sink: Visual information pathways in large vision-language models. *arXiv preprint arXiv:2510.08510*, 2025.

[36] K. Meng, D. Bau, A. Andonian, and Y. Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

[37] N. Nanda, L. Chan, T. Lieberum, J. Smith, and J. Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

[38] C. Neo, L. Ong, P. Torr, M. Geva, D. Krueger, and F. Barez. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*, 2024.

[39] L. T. P. Nguyen, Z. Yu, S. L. Y. Hang, S. An, J. Lee, Y. Ban, S. Chung, T.-H. Nguyen, J. Maeng, S. Lee, et al. See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models. *arXiv preprint arXiv:2512.02231*, 2025.

[40] Y. Nikankin, D. Arad, Y. Gandelsman, and Y. Belinkov. Same task, different circuits: Disentangling modality-specific mechanisms in vlms. *arXiv preprint arXiv:2506.09047*, 2025.

[41] Y. Park, M. Lee, S. Chun, and J. Choe. Mitigating cross-image information leakage in lvlms for multi-image tasks. *arXiv preprint arXiv:2508.13744*, 2025.

[42] D. Rai, Y. Zhou, S. Feng, A. Saparov, and Z. Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024.

[43] R. Selvakumar, K. Jayakumar, S. Sakshi, S. Ghosh, R. Gao, and D. Manocha. Do audio-visual large language models really see and hear? *arXiv preprint arXiv:2604.02605*, 2026.

[44] L. Sharkey, B. Chughtai, J. Batson, J. Lindsey, J. Wu, L. Bushnaq, N. Goldowsky-Dill, S. Heimersheim, A. Ortega, J. Bloom, et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

[45] W. Suharitdamrong, T. Alex, X. Zhu, M. Awais, and S. Atito. The hidden evolution of disguised visual context inside the vlm. *arXiv preprint arXiv:2606.20077*, 2026.

[46] M. Sun, X. Chen, J. Z. Kolter, and Z. Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

[47] C. Tang, Y. Li, Y. Yang, J. Zhuang, G. Sun, W. Li, Z. Ma, and C. Zhang. video-salmonn 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

[48] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023.

[49] K. Tao, K. Shao, B. Yu, W. Wang, H. Wang, et al. Omnizip: Audio-guided dynamic token compression for fast omnimodal large language models. *arXiv preprint arXiv:2511.14582*, 2025.

[50] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[51] Q. Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

[52] S. Tong, E. Brown, P. Wu, S. Woo, M. Middepogu, S. C. Akula, J. Yang, S. Yang, A. Iyer, X. Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[53] W. Wang, Z. Gao, L. Gu, H. Pu, L. Cui, X. Wei, Z. Liu, L. Jing, S. Ye, J. Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[54] Y. Wei, Y. Miao, D. Zhou, and D. Hu. Moka: Multimodal low-rank adaptation for mllms. *arXiv preprint arXiv:2506.05191*, 2025.

[55] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

[56] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

[57] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

[58] Y. Yang, J. Zhuang, G. Sun, C. Tang, Y. Li, P. Li, Y. Jiang, W. Li, Z. Ma, and C. Zhang. Audio-centric video understanding benchmark without text shortcut. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6580–6598, 2025.

[59] H. Ye, C.-H. H. Yang, A. Goel, W. Huang, L. Zhu, Y. Su, S. Lin, A.-C. Cheng, Z. Wan, J. Tian, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*, 2025.

[60] Q. Ye, Z. Yu, R. Shao, X. Xie, P. Torr, and X. Cao. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In *European Conference on Computer Vision*, pages 146–164. Springer, 2024.

[61] B. Zhang, K. Li, Z. Cheng, Z. Hu, Y. Yuan, G. Chen, S. Leng, Y. Jiang, H. Zhang, X. Li, et al. Video-ollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

[62] Z. Zhang, S. Yadav, F. Han, and E. Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19781–19791, 2025.

[63] Z. Zhou, R. Wang, Z. Wu, and Y.-G. Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.

Appendix Contents

A Related Works . . . 15

A.1 Audio-Visual Large Language Models (AVLLMs) . . . 15

A.2 Mechanistic Interpretability of LLMs and MLLMs . . . 15

B Additional Visualizations of Attention Sink Tokens . . . 15

C Experimental Details . . . 16

D Dataset Details . . . 16

D.1 AV-SpeakerBench . . . 16

D.2 WorldSense . . . 17

D.3 AV-Odyssey . . . 18

D.4 VideoMME . . . 19

E Ablation on Window Size for Attention Knockout . . . 22

F Ablation on FPS and Spatial Resolution . . . 23

G Cross-Dataset Generalization of Token Discard on WorldSense . . . 25

H Additional Knockout Analyses on Qwen2.5-Omni 3B . . . 26

H.1 Extended Knockout Analyses on AV-SpeakerBench . . . 26

H.2 VideoMME . . . 28

H.3 WorldSense . . . 29

H.4 AV-Odyssey . . . 31

I Knockout Analyses: Generalization to Qwen2.5-Omni 7B . . . 33

I.1 AV-SpeakerBench . . . 33

I.2 WorldSense . . . 35

I.3 AV-Odyssey . . . 37

J Knockout Analyses: Generalization to Video-SALMONN2 3B Plus . . . 40

J.1 AV-SpeakerBench . . . 40

J.2 WorldSense . . . 42

K Knockout Analyses: Generalization to Video-SALMONN2 7B Plus . . . 44

K.1 AV-SpeakerBench . . . 44

K.2 WorldSense . . . 46

A Related works

A.1 Audio-visual large language models (AVLLMs)

Audio-visual large language models (AVLLMs) extend the multimodal LLM paradigm to jointly process audio and visual inputs. The first generation of AVLLMs [6, 47, 60, 34] couples dedicated audio and visual encoders with a language model to support text-based audio-visual question answering and dialogue. Building on this foundation, omni models [56, 57, 51, 59, 31, 12] introduce temporal alignment between audio and visual streams and unify perception with end-to-end speech generation, pushing AVLLMs toward real-time multimodal interaction. Alongside these architectural advances, a growing body of work targets the inference efficiency of AVLLMs by compressing audio and visual tokens prior to the language model [16, 9, 49, 28]. Distinct from these prior works, we perform a mechanistic interpretability study of AVLLMs in this study, tracing how audio and visual information flow through the model to form the prediction.

A.2 Mechanistic interpretability of LLMs and MLLMs

Mechanistic interpretability [44, 42, 37, 10, 13] is the study of how internal computations in neural networks give rise to their behavior. It has become an become a popular research direction that recently extends from LLMs to MLLMs. In LLMs, prior studies have uncovered the algorithms underlying grokking [37] and the attention head circuits behind transformer computation [10]. Among the methodological tools developed in this field, attention knockout [13] causally intervenes on attention pathways to identify the routes information takes through the model. Within MLLMs, recent work has examined information storage [5], modality-specific circuits [40],visual information processing [38, 24] and visual representations across different integration paradigms [45]. Most relevant to our study, [62] and [26] apply attention knockout to trace information flow in image VLMs and VideoLLMs, respectively. Concurrent work [43] investigates AVLLMs through counterfactual analysis on audio-visual captioning and [23] uses causal tracing to locate where cross-modal information is stored within token representations. Distinct from these prior works, we trace the internal mechanism of how audio and visual information flow inside AVLLMs in both audio-visual video and multi-input interleaved scenarios to form the prediction in this study.

B Additional visualizations of attention sink tokens

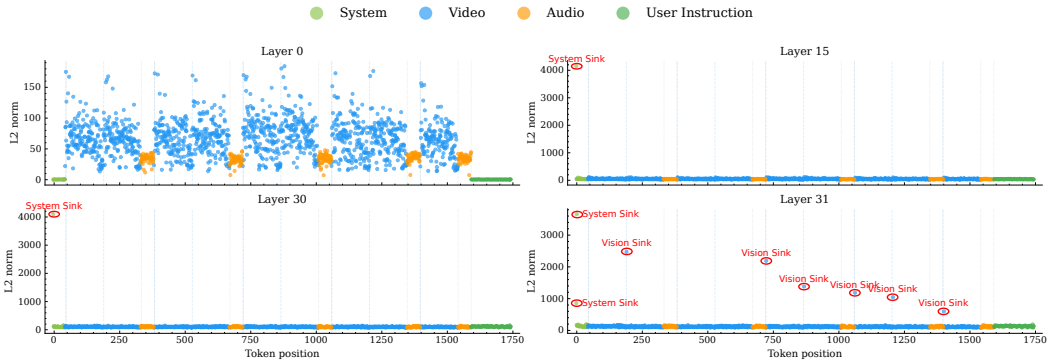

Figure 7: L2 norm distribution across token positions at four representative layers (0, 15, 30, 31). Tokens are colored by type (system, video, audio, user instruction). High-norm sink tokens are highlighted with red circles. The language sink emerges at layer 15 in the system prompt region, while vision sinks emerge later at layer 31 in the video sequence.

To trace when each type of sink emerges across layers, we analyze the L2 norm distribution at four representative layers (0, 15, 30, 31), shown in Figure 7. At layer 0, vision and audio tokens already have higher norms than language tokens, since they have already passed through the modality encoders before entering the LLM. By layer 15, a single high-norm token has emerged in the system prompt region, corresponding to the language sink described in [55]. This language sink persists through layer 30, while modality tokens remain at low norms with no vision sink yet present. At

layer 31, vision sinks emerge sharply at specific positions in the video sequence, marking the layer at which the visual anchors identified in Section 3 first appear. This progression confirms that vision sinks emerge later than the language sink, despite sharing the same mechanical signature.

C Experimental details

Models and inputs: We analyze Qwen2.5-Omni 3B [56] as the primary model, with additional results on its 7B variants and Video-SALMONN2 Plus 3B and 7B [47] in Appendix I, J and K. All models are loaded from their official Hugging Face checkpoints. Videos are sampled at 2 FPS with up to 128 frames per video, where each frame is represented by up to a 12×12 grid of visual tokens, and standalone images are represented by up to a 24×24 grid of visual tokens.

Inference setup: All experiments use greedy decoding (`do_sample=False`) for deterministic outputs and are conducted on a single NVIDIA H100 GPU. To ensure each knockout measures a real degradation rather than a coincidental change, we run all interventions only on samples that the model predicts correctly under the no-knockout baseline. A typical knockout run takes around 30 minutes for the audio-visual video setting and 15 minutes for the multi-input interleaved setting.

D Dataset details

This section provides additional details on the datasets used in our analysis. AV-SpeakerBench (Section D.1) is the audio-visual video benchmark used in the main paper, AV-Odyssey (Section D.3) is the multi-input audio-visual benchmark also used in the main paper, and WorldSense (Section D.2) is an additional audio-visual video benchmark used in this appendix to provide further evidence on the generality of our findings.

D.1 AV-SpeakerBench (audio-visual video)

AV-SpeakerBench [39] is a benchmark for evaluating audio–visual reasoning in models that jointly process video and speech. Each sample is a four-way multiple-choice question paired with an audio-visual video, and the benchmark contains 12 tasks across audio-centric, speaker-centric, and visual-centric domains.

Cross-modal anchor design: A key feature of AV-SpeakerBench is that each question is constructed around an explicit *anchor–target* structure that forces genuine audio–visual integration. The anchor is a cue described in the question text that points to a specific moment in one modality, and the answer must be read off the opposite modality at that moment. We summarize the three task types as follows:

- **Audio-centric tasks** (visual anchor $\rightarrow$ audio answer): the question describes a visual cue (e.g., *"after the man in the grey shirt wiggles his fingers"*), and the model must first locate that visual moment, then listen to the audio within that window to extract the answer (e.g., counting how often a phrase is spoken).
- **Visual-centric tasks** (audio anchor $\rightarrow$ visual answer): the question describes an audio cue, typically a spoken phrase (e.g., *"after the woman says This is very datable"*), and the model must locate that moment in the audio, then inspect the video at that timestamp to extract a visual answer (e.g., counting visible people).
- **Speaker-centric tasks** (mixed): the question may use either a visual or audio anchor, and the answer choices differ in the opposite modality (e.g., visually distinct speakers who utter different lines). Solving these requires jointly tracking identity, timing, and modality, which prevents unimodal shortcuts.

This anchor structure makes AV-SpeakerBench particularly well-suited for our analysis, since the cross-modal pathway is built directly into the question design rather than emerging incidentally from the content.

Task selection for attention knockout: For our analysis, we focus on tasks that exhibit the cross-modal anchor structure described above and exclude tasks where the question can be largely answered from a single modality without the anchor playing a meaningful role. The 8 selected tasks, totaling 2,281 samples, are organized into 5 categories used throughout Section 4. Table 5 (left) lists the selected tasks with their domain and category, and (right) reports the per-category sample counts. Categories with multiple tasks (Visual Recognition, Speech Attributes) merge tasks that share the same anchor direction and answer modality; the other three categories correspond to a single task each. The knockout results on AV-SpeakerBench are reported in Section 4 for Qwen2.5-Omni 3B, with extended analyses in Appendix H.1, and in Appendices I.1, J.1, and K.1 for Qwen2.5-Omni 7B and Video-SALMONN2 Plus 3B and 7B, respectively.

Table 5: **Left:** The 8 tasks selected from AV-SpeakerBench, with their domain, sample counts, and the category each task is grouped into for our analysis. **Right:** The 5 categories with their aggregated sample counts. Domain color indicates the cross-modal direction: visual anchor → audio answer, audio anchor → visual answer, and mixed.

Domain	Task	# Samples	Category
Audio-centric	Speech Intensity	206	Speech Attributes
	Speech Pitch	206	Speech Attributes
	Speech Rate	209	Speech Attributes
	Speech Recognition	201	Speech Recognition
Speaker-centric	Speaker Detection	427	Speaker Detection
	Speaker Recognition	422	Speaker Recognition
Visual-centric	Activity Recognition	206	Visual Recognition
	Attribute Recognition	204	Visual Recognition
Total		**2,281**	

Category	# Samples
Speech Attributes	621
Speech Recognition	201
Visual Recognition	410
Speaker Recognition	422
Speaker Detection	427
Total	**2,281**

Task selection for cross-task evaluation: In addition to the analysis tasks, for Section 6 we use Speech Counting and Visual Counting from AV-SpeakerBench. Table 6 reports the domain and sample count for each.

Table 6: Cross-task evaluation tasks from AV-SpeakerBench used in Section 6. Domain color indicates the cross-modal direction: visual anchor → audio answer, audio anchor → visual answer.

Domain	Task	# Samples
Audio-centric	Speech Counting	288
Visual-centric	Visual Counting	205
Total		**493**

D.2 WorldSense (audio-visual video)

WorldSense [20] is a benchmark of audio-visual videos paired with multiple-choice questions, designed to evaluate joint reasoning over visual, audio, and temporal information. The benchmark covers 26 task types organized into three task domains: Recognition (identifying entities or events), Reasoning (inferring causal or relational structure), and Understanding (comprehending temporal or contextual states).

Task selection for attention knockout: WorldSense contains videos ranging from a few seconds to over eight minutes. To keep the audio-visual stream short enough for tractable knockout sweeps and to ensure that all selected samples fit within a comparable input length, we filter to videos under one minute long. From this filtered set, we select the 10 task types marked under *Knockout* in Table 7, totaling 418 samples. The selected videos have a mean duration of 47.4 seconds (median 49 seconds, range 17–60 seconds). Compared to AV-SpeakerBench, WorldSense provides substantially fewer samples per task (typically 25–64 per task type, against 200–400 in AV-SpeakerBench). As a result,

the per-task knockout curves on WorldSense are visibly noisier than those on AV-SpeakerBench. The knockout results on WorldSense are reported in Appendices H.3, I.2, J.2, and K.2 for Qwen2.5-Omni 3B and 7B and Video-SALMONN2 Plus 3B and 7B, respectively.

Task selection for cross-dataset evaluation: For the cross-dataset evaluation of token discarding, we use all 26 task types of WorldSense under the same one-minute filter, totaling 1,047 samples. Section 6 reports Video Emotions and Audio Change, and the results on all tasks are reported in Appendix G.

Table 7: WorldSense tasks used in our analysis, restricted to videos under one minute. *Knockout* marks the tasks used for attention knockout. All tasks are used for the cross-dataset evaluation of token discarding in Section 6 and Appendix G.

Domain	Task	# Samples	Knockout
Recognition	Action Counting	46	
	Anomaly Recognition	27	
	Attribute Recognition	52	✓
	Audio Counting	38	✓
	Audio Recognition	88	
	Audio Source Localization	64	✓
	Event Recognition	30	✓
	Object Counting	47	
	Object Existence Recognition	31	
	Scene Recognition	27	✓
	Temporal Localization	56	
Reasoning	Attribute Reasoning	42	
	Audio Change	45	
	Causal Reasoning	37	
	Emotion Change	37	✓
	Hallucination	21	
	Object State Change	25	✓
	Relation Reasoning	20	
	Temporal Prediction	36	
Understanding	Event Sorting	46	✓
	Human Emotions	29	
	Human Interaction	36	
	Human-object Interaction	41	
	Spatial Relation	58	✓
	Text and Diagram Understanding	41	✓
	Video Emotions	27	
Total		**1,047**	**418**

D.3 AV-Odyssey (multi-input audio-visual interleaved)

AV-Odyssey [17] is a benchmark for audio-visual understanding that contains 26 tasks spanning multiple reasoning skills, where each sample interleaves multiple independent images and audio clips with text. Among these, we focus on the matching subset where the model selects an answer by comparing a single reference item in one modality against multiple candidate items in the opposite modality, which aligns with the multi-input setting analyzed in Section 5. This appendix provides additional details on how we process AV-Odyssey for our analysis: Section D.3.1 describes the procedure used to assign each sample an input structure label, and Section D.3.2 lists the specific tasks we select and reports the distribution of input structures across them.

D.3.1 Input structure assignment for AV-Odyssey

AV-Odyssey samples vary in how the candidate media, reference media, and question text are arranged within the prompt. To support per-ordering analysis, we automatically classify each sample's input structure through the following procedure:

1. **Parse media placeholders.** We scan the prompt to locate each media placeholder (e.g., `[img1]`, `[audio1]`) and identify the single reference media along with the candidate media in the opposite modality.
2. **Split into ordered segments.** Using the located media placeholders as boundaries, we split the prompt into an ordered sequence of segments alternating between media spans (candidates and reference) and text spans.
3. **Identify the question text.** Among the text segments, we assign the role of *question text* to the segment that describes the actual matching task. If the reference media is the final media in the prompt, we prefer the text segment immediately preceding it (which typically contains the matching prompt, e.g., *"which best matches"*); otherwise, we select the longest text segment by word count. All remaining text segments are treated as padding and excluded from the analysis.
4. **Construct the structure label.** The final segment ordering, for example candidates , question , reference , is recorded as the sample's structure label.

We use the resulting structure labels to bucket samples for both the main-paper analysis (which uses the most common ordering, candidates , question , reference) and the per-ordering breakdowns reported in this appendix.

D.3.2 Selected tasks from AV-Odyssey

Task selection for attention knockout: For our knockout analysis, we select 7 tasks from AV-Odyssey that use either a single reference media against multiple candidates of the opposite modality or vice versa. Table 8 lists the 7 selected tasks, totaling 1,304 samples. Within each task, the prompts can also vary in how the candidates , question text, and reference are ordered. Table 9 shows the aggregate distribution across all selected samples, and Table 10 provides the per-task breakdown. The dominant ordering is candidates , question , reference , which we use for the main paper analysis; the remaining orderings are analyzed individually in this appendix. The knockout results on AV-Odyssey are reported in Section 5 for Qwen2.5-Omni 3B averaged across tasks, with per-task results in Appendix H.4, and in Appendix I.3 for Qwen2.5-Omni 7B.

Task selection for cross-task evaluation: In addition to the tasks used for attention knockout, for Section 6 we use Bird Recognition and Transportation Recognition from AV-Odyssey. Table 11 reports the sample count and input layout distribution for each.

Table 8: The 7 tasks selected from AV-Odyssey for attention knockout. A Ref → I Cand denotes audio reference to image candidates , and I Ref → A Cand denotes image reference to audio candidates .

Task	# Samples	A Ref → I Cand	I Ref → A Cand
Instrument Recognition	200	49%	51%
Animal Recognition	200	50%	50%
Material Recognition	200	100%	—
Hazard Recognition	108	100%	—
Action Recognition	196	100%	—
Music Genre Classification	200	52%	48%
Film and Music Matching	200	100%	—
Total	**1,304**	**77%**	**23%**

D.4 VideoMME (audio-visual video)

VideoMME [11] is a video understanding benchmark of four-way multiple-choice questions over videos of short, medium, and long duration. Most of its questions concern visual content, so the answer can be read from the video alone even when the audio track is provided. We include it to examine whether the task-dependent modality contribution observed in Section 4.3 holds on

Table 9: Aggregate distribution of prompt structures across all 1,304 selected samples.

Structure	# Samples	Proportion
candidates , question , reference	1,001	77%
question , reference , candidates	161	12%
reference , candidates , question	124	10%
reference , question , candidates	18	1%

Table 10: Per-task distribution of prompt structures across the 7 selected tasks. *Cand* = candidates , *Q* = question text, *Ref* = reference . Each row reports the sample counts for each structure observed in that task. Empty cells indicate the structure does not appear in that task.

	Structure			
Task	**Cand , Q , Ref**	**Q , Ref , Cand**	**Ref , Cand , Q**	**Ref , Q , Cand**
Instrument Recognition	46	154	—	—
Animal Recognition	200	—	—	—
Material Recognition	151	7	35	7
Hazard Recognition	97	—	—	11
Action Recognition	196	—	—	—
Music Genre Classification	164	—	36	—
Film and Music Matching	147	—	53	—
Total	**1,001**	**161**	**124**	**18**

video-reliant tasks, where audio is present in the input but not needed to answer. We use the videos with their audio tracks and without subtitles.

Task selection for attention knockout: To keep the input length comparable to the other benchmarks, we restrict the analysis to short videos, which are under two minutes long. From this set, we select the three vision-based task types listed in Table 12, totaling 421 samples. The knockout results on VideoMME are reported in Appendix H.2.

Table 11: Cross-task evaluation tasks from AV-Odyssey used in Section 6. A Ref → I Cand denotes audio reference to image candidates , and I Ref → A Cand denotes image reference to audio candidates.

Task	# Samples	A Ref → I Cand	I Ref → A Cand
Bird Recognition	200	51%	49%
Transportation Recognition	200	52%	48%

Table 12: The 3 tasks selected from VideoMME for attention knockout, restricted to short videos under two minutes.

Task	**# Samples**
Attribute Perception	122
Action Recognition	131
Object Recognition	168
Total	**421**

E Ablation on window size for attention knockout

Following [13], attention knockout is applied within a sliding window of k layers around each target layer. We ablate $k \in \{1, 3, 5, 7, 9, 11\}$ on the Speaker Recognition task to select the most informative window size. Figure 8 shows that small windows ($k = 1, 3$) produce shallow noisy drops as the narrow block is easily bypassed by remaining attention edges. Large windows ($k = 11$) instead blur the layer localization, with the Question$\not\rightarrow$Last drop spanning a much wider range and losing the recovery point where the pathway concludes. Both $k = 7$ and $k = 9$ give clean and well-localized drops, but we adopt $k = 7$ as it preserves the recovery point of the Question$\not\rightarrow$Last pathway, whereas $k = 9$ tends to merge the recovery into the broader drop similarly to $k = 11$. The choice of $k = 7$ over $k = 9$ also accounts for model scale. The 7B variants have only 28 layers compared to 36 in the 3B model, so a window of $k = 9$ would cover a relatively larger portion of the network in 7B and obscure the precise layer where each pathway operates. Adopting $k = 7$ across both scales preserves localization accuracy in 7B while remaining valid for the 3B model.

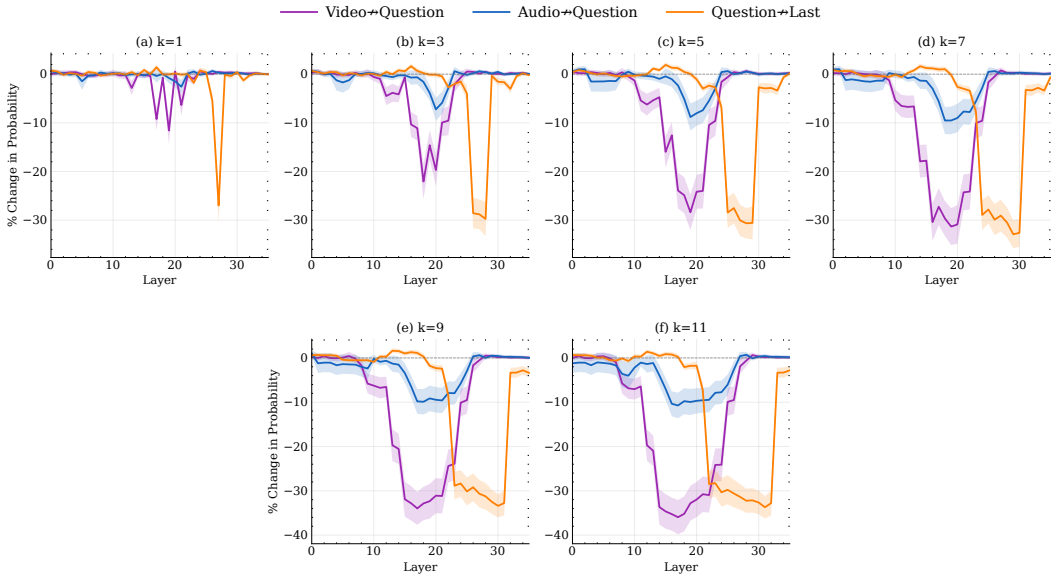

Figure 8: Window size ablation on the Speaker Recognition task. Each panel shows the relative change in prediction probability for three pathways (Video$\not\rightarrow$Question, Audio$\not\rightarrow$Question, Question$\not\rightarrow$Last) under a different window size k. Source$\not\rightarrow$Target indicates blocking attention edges from source tokens to target tokens.

F Ablation on FPS and spatial resolution

The main analysis samples videos at 2 FPS at the default spatial resolution (Appendix C). We vary the frame rate (Appendix F.1) and the spatial resolution (Appendix F.2) separately on AV-SpeakerBench with Qwen2.5-Omni 3B to verify that the information flow reported in Section 4.3 is robust to other input configurations. All other settings follow Section 4.1 and Appendix C.

F.1 FPS

The frame rate determines the temporal alignment granularity between audio and video tokens, as it sets how many frames each temporal window contains for the same audio segment. We evaluate frame rates of 1 FPS and 4 FPS at the default spatial resolution, halving and doubling the default rate. Figures 9 and 10 report the knockout results. At both frame rates, the route, the layers where each transfer localizes, and the task-dependent modality contribution all remain consistent with Section 4.3. Only the magnitudes vary, following how much visual content each frame rate provides. At 1 FPS, the video contribution to the question weakens across all three tasks, most clearly on Speech Recognition, where the Video$\not\to$Question drop falls from -10% at 2 FPS to -4.2%, consistent with sparser frames providing less visual content. At 4 FPS, the denser frames slightly strengthen the video contribution on the visually grounded tasks, while the audio contribution and the Question$\not\to$Last drop remain close to the default setting.

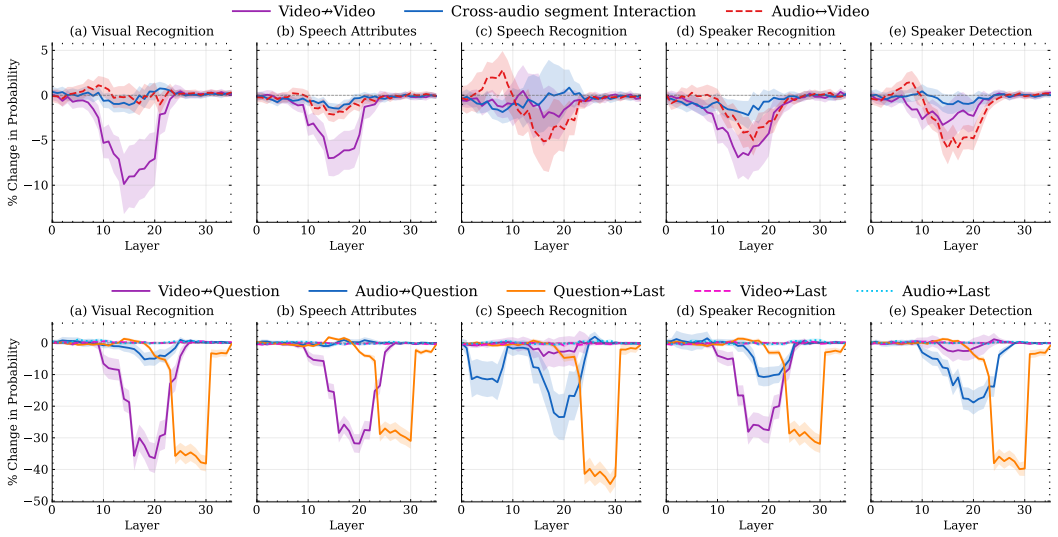

Figure 9: **Qwen2.5-Omni 3B on AV-SpeakerBench at 1 FPS.** *Top:* knockout of within- and cross-modal pathways (Cross-frame, Cross-audio, Audio$\not\leftrightarrow$Video). *Bottom:* knockout of modality and question pathways into the question and the last token (Video$\not\to$Question, Audio$\not\to$Question, Question$\not\to$Last, Video$\not\to$Last, Audio$\not\to$Last). Source$\not\to$Target indicates blocking attention edges from source tokens to target tokens.

F.2 Spatial resolution

The spatial resolution determines how much visual detail each frame carries. We reduce the resolution of each frame from 224×420 to 168×280 pixels, which reduces the grid of visual tokens per frame from 12×12 to 8×8 and halves the per-frame visual token budget, while keeping the frame rate at 2 FPS. Figure 11 reports the knockout results. The route, the layers where each transfer localizes, and the task-dependent modality contribution all remain consistent with Section 4.3. The magnitudes vary only slightly from the default setting, with differences of about 1% across tasks, suggesting that the coarser frames still carry the visual content these tasks rely on.

Together, these ablations show that the sequential Modalities → Question → Last route, the layers where each transfer localizes, and the task-dependent modality contribution reported in Section 4.3

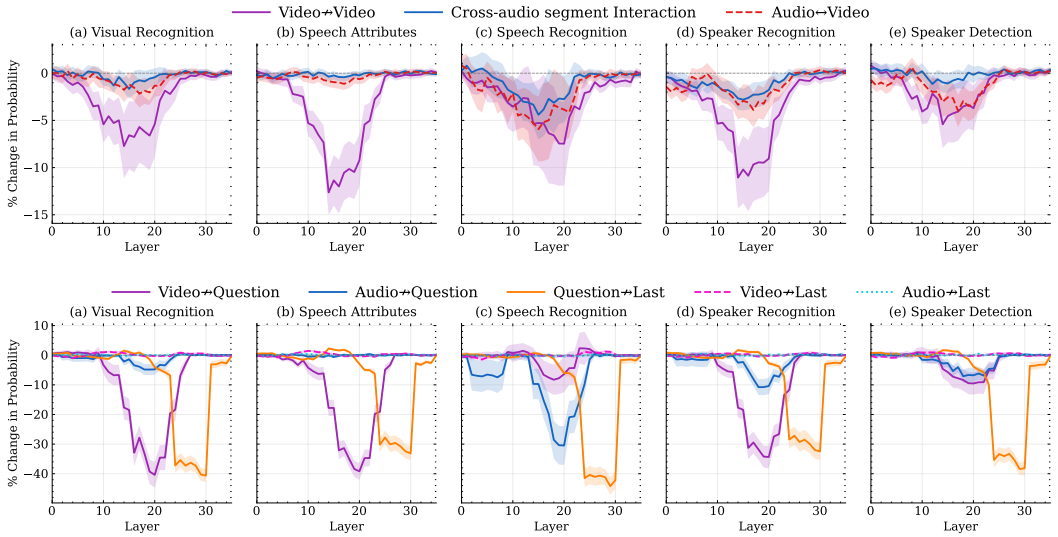

Figure 10: **Qwen2.5-Omni 3B on AV-SpeakerBench at 4 FPS.** *Top:* knockout of within- and cross-modal pathways (Cross-frame, Cross-audio, Audio↛Video). *Bottom:* knockout of modality and question pathways into the question and the last token (Video↛Question, Audio↛Question, Question↛Last, Video↛Last, Audio↛Last). Source↛Target indicates blocking attention edges from source tokens to target tokens.

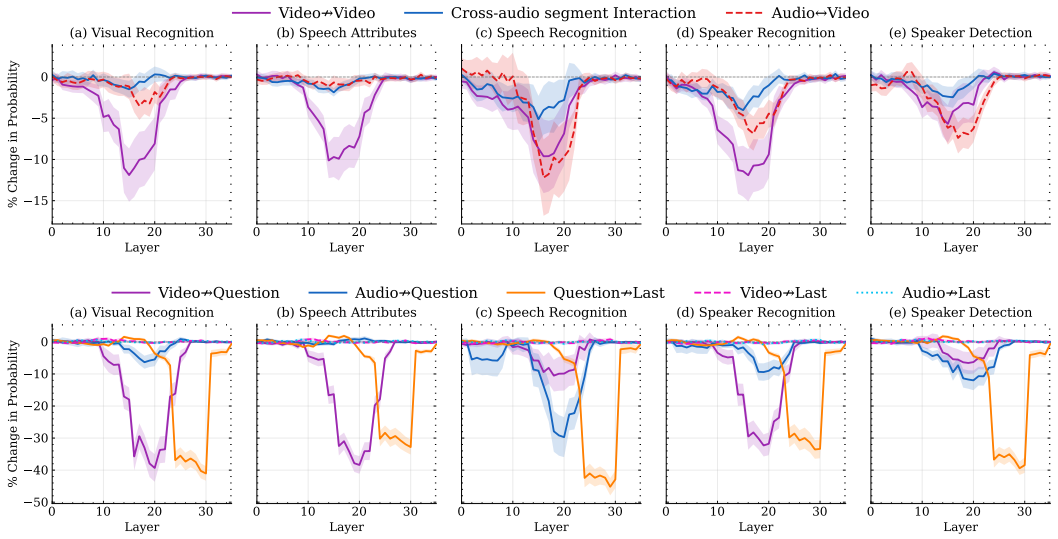

Figure 11: **Qwen2.5-Omni 3B on AV-SpeakerBench at 2 FPS with reduced spatial resolution** (168×280 pixels per frame). *Top:* knockout of within- and cross-modal pathways (Cross-frame, Cross-audio, Audio↛Video). *Bottom:* knockout of modality and question pathways into the question and the last token (Video↛Question, Audio↛Question, Question↛Last, Video↛Last, Audio↛Last). Source↛Target indicates blocking attention edges from source tokens to target tokens.

all hold across frame rates and spatial resolutions, with only the magnitudes varying according to the amount of visual content provided.

G More results on token discarding cross-dataset generalization on WorldSense

Table 4 reports two WorldSense categories, Video Emotions and Audio Change, as the cross-dataset evaluation of token discarding. Table 13 extends this evaluation to all 26 categories of WorldSense, using the same discarding configuration as in Section 6. The overall accuracy remains stable and slightly improves in all four configurations, from 45.75 at baseline to 46.23 when discarding all token types.

Table 13: **Per-category cross-dataset results on WorldSense.** Expanded breakdown of the cross-dataset evaluation in Table 4 (Video setting: Discard Video & Audio at $L = 26$, Ques at $L = 29$). Numbers in parentheses show change from baseline (green = improvement, red = drop); underlined indicates no change. Best per row in **bold**. **Ques**: non-option Question tokens; **All**: all token types together.

Category	Baseline	Discard Ques	Discard Audio	Discard Video	Discard All
Spatial Relation	24.14	24.14	24.14	24.14	24.14
Video Emotions	66.67	66.67	66.67	66.67	66.67
Event Sorting	56.52	**58.70** (+2.18)	56.52	54.35 (-2.17)	56.52
Event Recognition	46.67	46.67	50.00 (+3.33)	46.67	**53.33** (+6.66)
Attribute Recognition	51.92	51.92	51.92	51.92	51.92
Audio Recognition	63.64	62.50 (-1.14)	62.50 (-1.14)	**63.64**	**63.64**
Human-object Interaction	43.90	43.90	43.90	43.90	43.90
Audio Source Localization	53.12	53.12	53.12	51.56 (-1.56)	**54.69** (+1.57)
Audio Change	42.22	40.00 (-2.22)	40.00 (-2.22)	40.00 (-2.22)	**42.22**
Relation Reasoning	45.00	**50.00** (+5.00)	**50.00** (+5.00)	45.00	**50.00** (+5.00)
Action Counting	30.43	**32.61** (+2.18)	30.43	**32.61** (+2.18)	**32.61** (+2.18)
Text and Diagram Understanding	51.22	**51.22**	46.34 (-4.88)	48.78 (-2.44)	46.34 (-4.88)
Attribute Reasoning	47.62	**50.00** (+2.38)	47.62	47.62	47.62
Causal Reasoning	48.65	48.65	48.65	48.65	48.65
Human Emotions	37.93	37.93	**41.38** (+3.45)	37.93	37.93
Temporal Localization	33.93	35.71 (+1.78)	35.71 (+1.78)	**37.50** (+3.57)	35.71 (+1.78)
Anomaly Recognition	51.85	48.15 (-3.70)	**51.85**	**51.85**	48.15 (-3.70)
Audio Counting	39.47	42.11 (+2.64)	42.11 (+2.64)	**44.74** (+5.27)	42.11 (+2.64)
Human Interaction	30.56	30.56	30.56	30.56	30.56
Temporal Prediction	50.00	50.00	50.00	50.00	50.00
Hallucination	23.81	23.81	23.81	23.81	23.81
Object Counting	27.66	25.53 (-2.13)	25.53 (-2.13)	**27.66**	25.53 (-2.13)
Object State Change	56.00	56.00	56.00	56.00	56.00
Emotion Change	62.16	62.16	**64.86** (+2.70)	62.16	**64.86** (+2.70)
Scene Recognition	59.26	**62.96** (+3.70)	59.26	59.26	**62.96** (+3.70)
Object Existence Recognition	38.71	**41.94** (+3.23)	**41.94** (+3.23)	38.71	38.71
Overall	45.75	46.13 (+0.38)	45.94 (+0.19)	45.85 (+0.10)	**46.23** (+0.48)

H Additional results on Qwen2.5-Omni 3B

This appendix provides additional knockout analyses for Qwen2.5-Omni 3B, extending the AV-SpeakerBench [39] results from Section 4 and reporting attention knockout on WorldSense [20] and AV-Odyssey [17]. While AV-Odyssey is also analyzed in the main paper, the results there are averaged across tasks. Here we instead report per-task knockouts to show that the same pattern holds for each individual task.

H.1 Extended knockout analyses on AV-SpeakerBench (audio-visual video)

This subsection extends the AV-SpeakerBench analysis of Section 4 with additional knockouts on cross-modal directions, joint audio-visual routing, and the question's internal components.

H.1.1 Further analysis of cross-modal directions and joint audio-visual routing

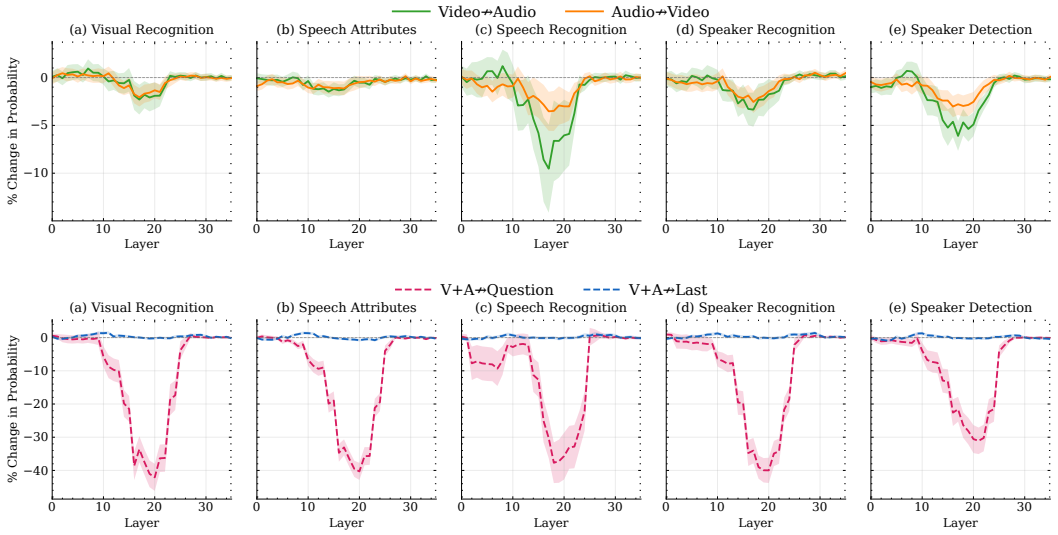

Figure 12: Additional knockout analysis on Qwen2.5-Omni 3B. *Top:* direction-specific cross-modal knockouts (Video↛Audio and Audio↛Video). *Bottom:* joint knockouts where audio and video serve as sources (V+A↛Question and V+A↛Last). Source↛Target indicates blocking attention edges from source tokens to target tokens.

We extend the analysis of Section 4 with two follow-up analyses. The top panel of Figure 12 decomposes the bidirectional cross-modal knockout of Section 4.2 into its two unidirectional components, Video↛Audio and Audio↛Video. The Video→Audio direction is the dominant one on tasks that require fine-grained audio-visual alignment (Speech Recognition, Speaker Detection), accounting for most of the bidirectional cross-modal effect observed in Section 4.2, while Audio→Video plays a smaller role. The remaining tasks remain near zero in both directions, consistent with Section 4.2. We hypothesize that this asymmetry stems from the input ordering of the time-aligned sequence, where each video frame is followed by its corresponding audio segment. Audio tokens can therefore attend back to the time-aligned video frame, while video tokens can only attend to audio segments from earlier time steps and not to the time-aligned audio segment, weakening the Audio→Video direction. The bottom panel groups audio and video into a single source (V+A↛Question and V+A↛Last). V+A↛Question produces large mid-layer drops across all tasks, while V+A↛Last produces near-zero changes, confirming that audio-visual information reaches the prediction through the question rather than directly, in line with the Modalities → Question → Last route established in Section 4.3.

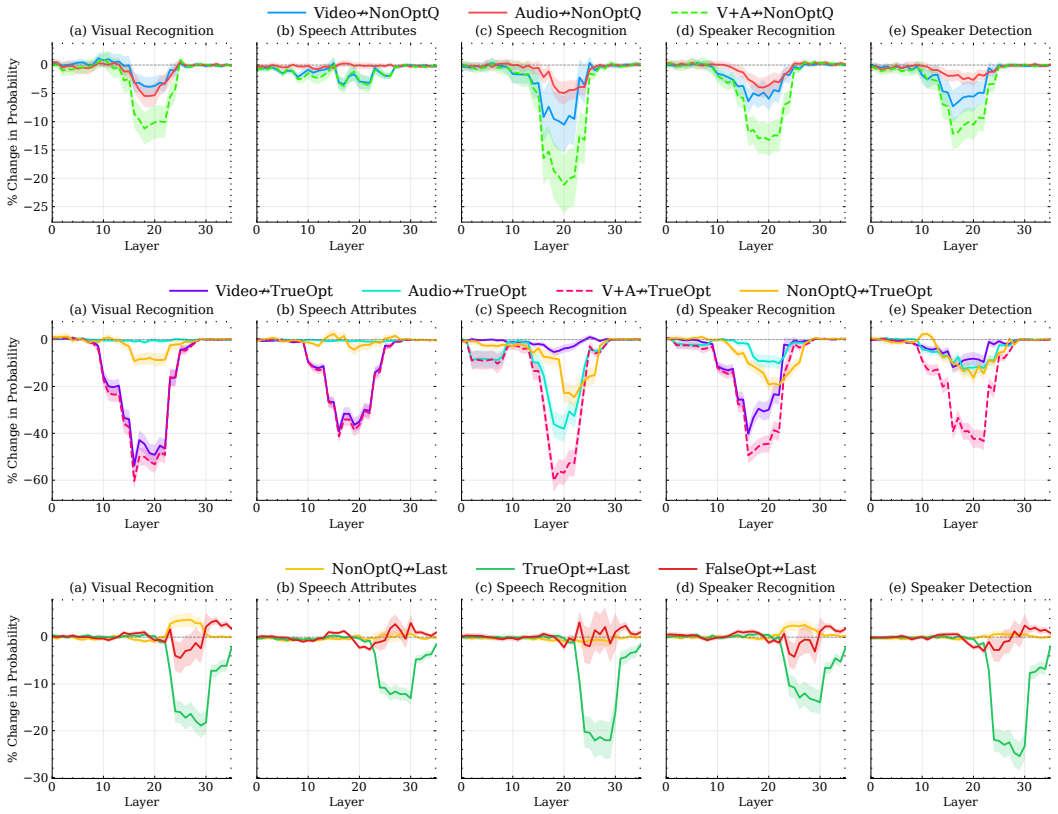

Figure 13: Component-level question-internal knockouts on Qwen2.5-Omni 3B. *Top:* Source↛NonOptQ. *Middle:* Source↛TrueOpt. *Bottom:* Source↛Last. Source↛Target indicates blocking attention edges from source tokens to target tokens.

H.1.2 Further analysis of how audio-visual information reaches the prediction via question components

Section 4.3 treats the question as a single aggregator. Here we decompose it into three components, the correct-option letter (TrueOpt), the incorrect-option letters (FalseOpt), and the non-option question text (NonOptQ), and trace how audio-visual content reaches the prediction within the question. Figure 13 reveals two routes converging at the correct-option letter. The direct route, Modalities $\rightarrow$ TrueOpt, sends modality content straight to the option letter. The indirect route, Modalities $\rightarrow$ NonOptQ $\rightarrow$ TrueOpt, first passes through the non-option question text before reaching the option letter. Both routes are active on Visual Recognition, Speech Recognition, Speaker Recognition, and Speaker Detection, while Speech Attributes relies on the direct route only, with negligible flow through the non-option question text. The correct-option letter therefore acts as the local aggregator, absorbing audio-visual content from these routes at mid layers before being read by the last token at late layers. The incorrect-option letters and the non-option question text never flow to the prediction directly, only through the correct-option letter. This component-level view is consistent with previous work in VideoLLMs [26], which also reports the option token as the decisive integration point with both direct and indirect routes from video to the option. Our analysis extends this observation to the audio-visual setting, where the same dual-route structure governs how both modalities reach the option letter.

H.2 VideoMME

We further examine whether our findings hold across benchmarks by repeating the analysis of Section 4 on VideoMME [11] at 2 FPS, using three vision-based categories, Attribute Perception, Action Recognition, and Object Recognition (details in Appendix D.4). Figures 14 and 15 report the knockout results. The route and the layers where each transfer localizes remain consistent with Section 4.3. All three categories can be answered from the video alone, and the flow follows accordingly, with the contribution coming almost entirely from video and little from audio, consistent with the task-dependent modality contribution observed on AV-SpeakerBench.

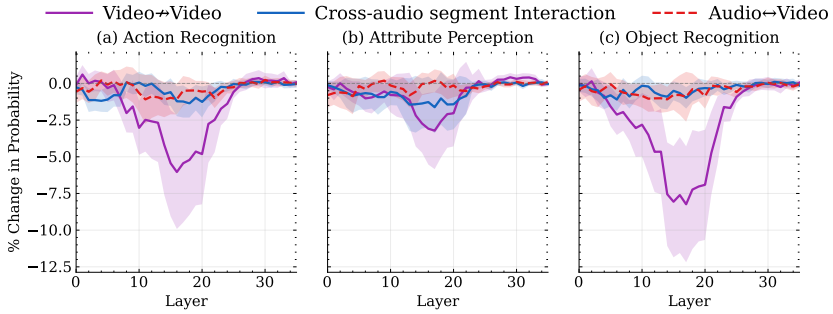

Figure 14: **Qwen2.5-Omni 3B on VideoMME.** Knockout of within- and cross-modal pathways (Cross-frame, Cross-audio, Audio↛Video). Source↛Target indicates blocking attention edges from source tokens to target tokens.

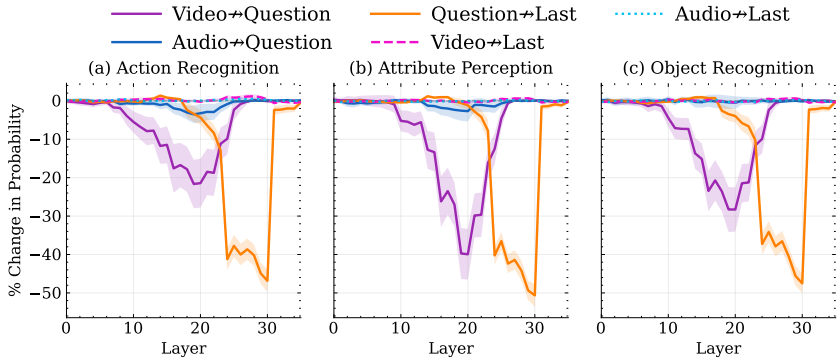

Figure 15: **Qwen2.5-Omni 3B on VideoMME.** Knockout of modality and question pathways into the question and the last token (Video↛Question, Audio↛Question, Question↛Last, Video↛Last, Audio↛Last). Source↛Target indicates blocking attention edges from source tokens to target tokens.

H.3 WorldSense (audio-visual video)

Figures 16 and 17 report the WorldSense knockout results. Figure 16 covers within- and cross-modal pathways together with modality and question pathways into the last token, while Figure 17 covers the question-internal pathways and the pathways into the correct option letter and the non-option question text.

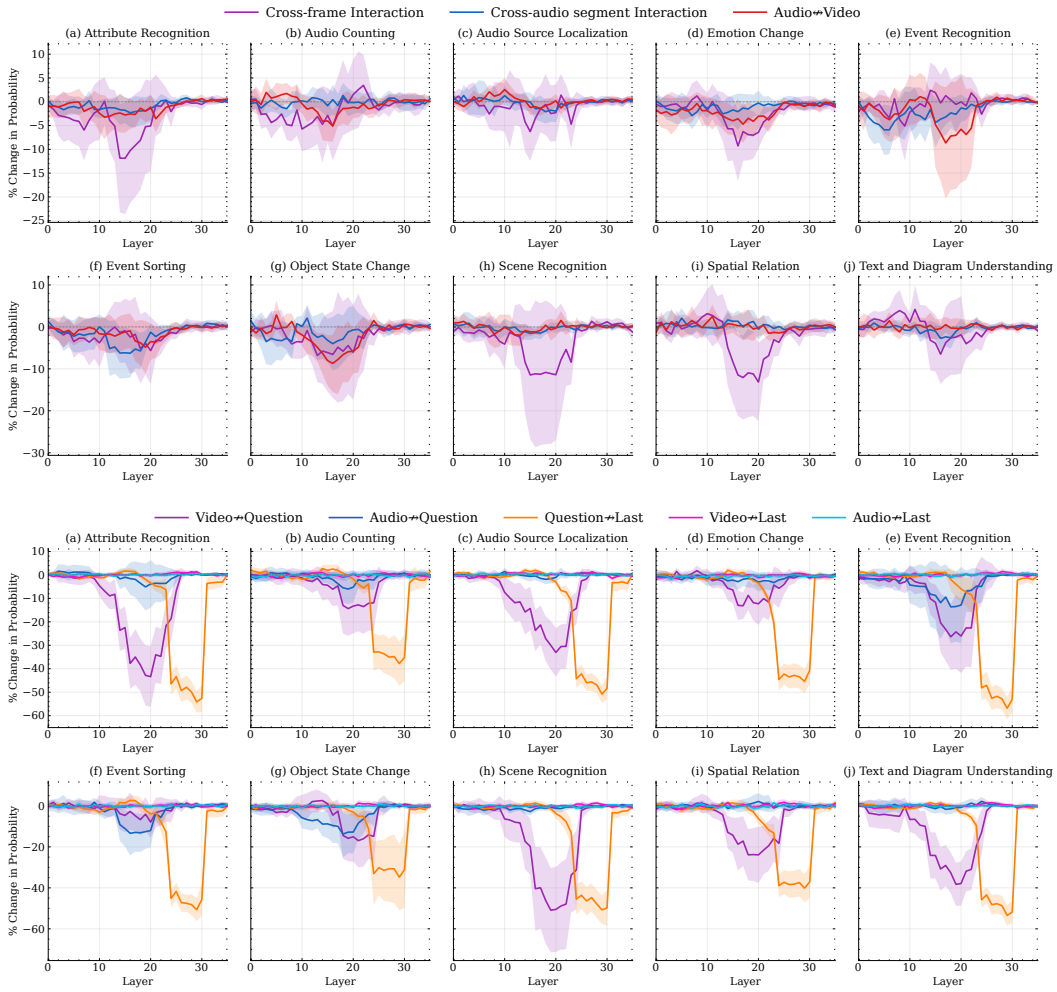

Figure 16: **Qwen2.5-Omni 3B on WorldSense.** Knockout of within- and cross-modal pathways (Cross-frame, Cross-audio segment, Audio↔Video) and of modality and question pathways into the last token (Video↛Question, Audio↛Question, Question↛Last, Video↛Last, Audio↛Last). Source↛Target indicates blocking attention edges from source tokens to target tokens.

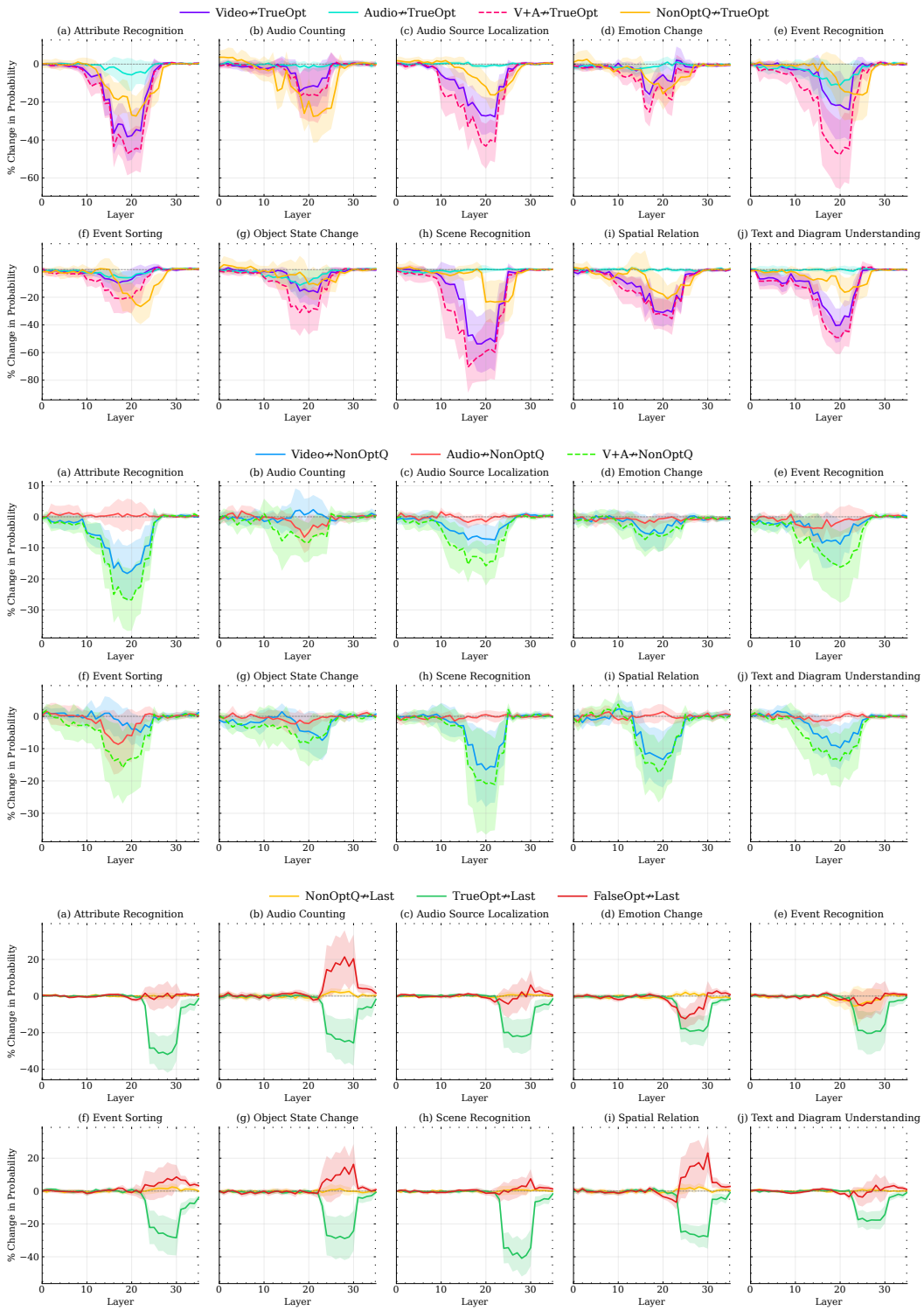

Figure 17: **Qwen2.5-Omni 3B on WorldSense.** Modality and question pathways into the correct option letter (Video↛TrueOpt, Audio↛TrueOpt, NonOptQ↛TrueOpt, V+A↛TrueOpt); modality pathways into the non-option question text (Video↛NonOptQ, Audio↛NonOptQ, V+A↛NonOptQ); and question-internal pathways into the last token (TrueOpt↛Last, FalseOpt↛Last, NonOptQ↛Last). Source↛Target indicates blocking attention edges from source tokens to target tokens.

H.4 AV-Odyssey (multi-input audio-visual interleaved)

This subsection reports per-task multi-input interleaved knockouts on AV-Odyssey. The main paper reports results averaged across tasks, while Figures 18 and 19 break results down by individual task and input ordering (I Ref → A Cand and A Ref → I Cand) to show consistency across the benchmark. Figure 18 covers pathways into the Reference, the Question, and the last token, while Figure 19 covers pathways into the correct option letter and pathways from the correct and incorrect candidates and option letters into the last token.

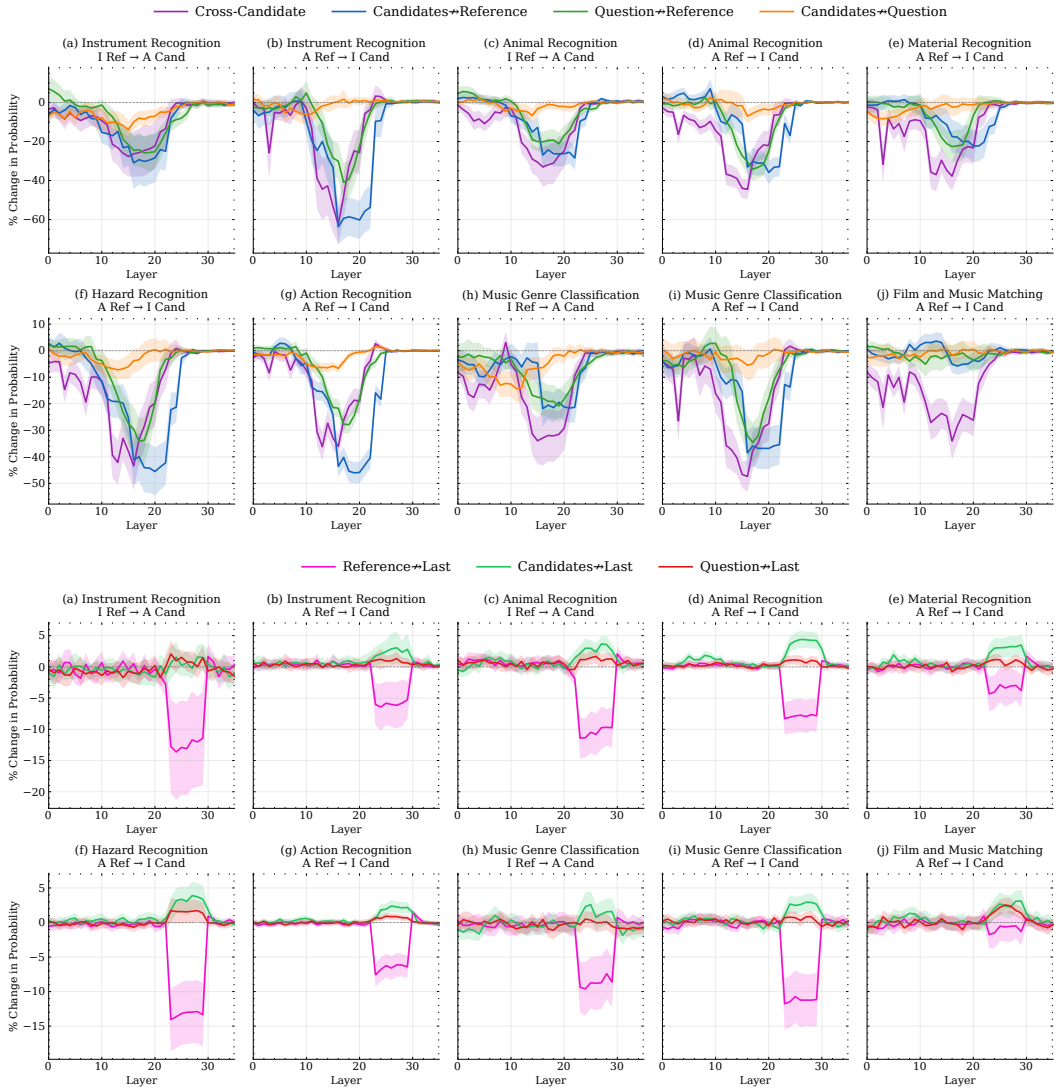

Figure 18: **Qwen2.5-Omni 3B per-task multi-input knockout (AV-Odyssey).** Pathways into the Reference and Question (Cross-Candidate, Candidates↛Reference, Question↛Reference, Candidates↛Question) and pathways into the last token (Reference↛Last, Candidates↛Last, Question↛Last). Each panel shows one task under one input ordering (I Ref → A Cand or A Ref → I Cand). Source↛Target indicates blocking attention edges from source tokens to target tokens.

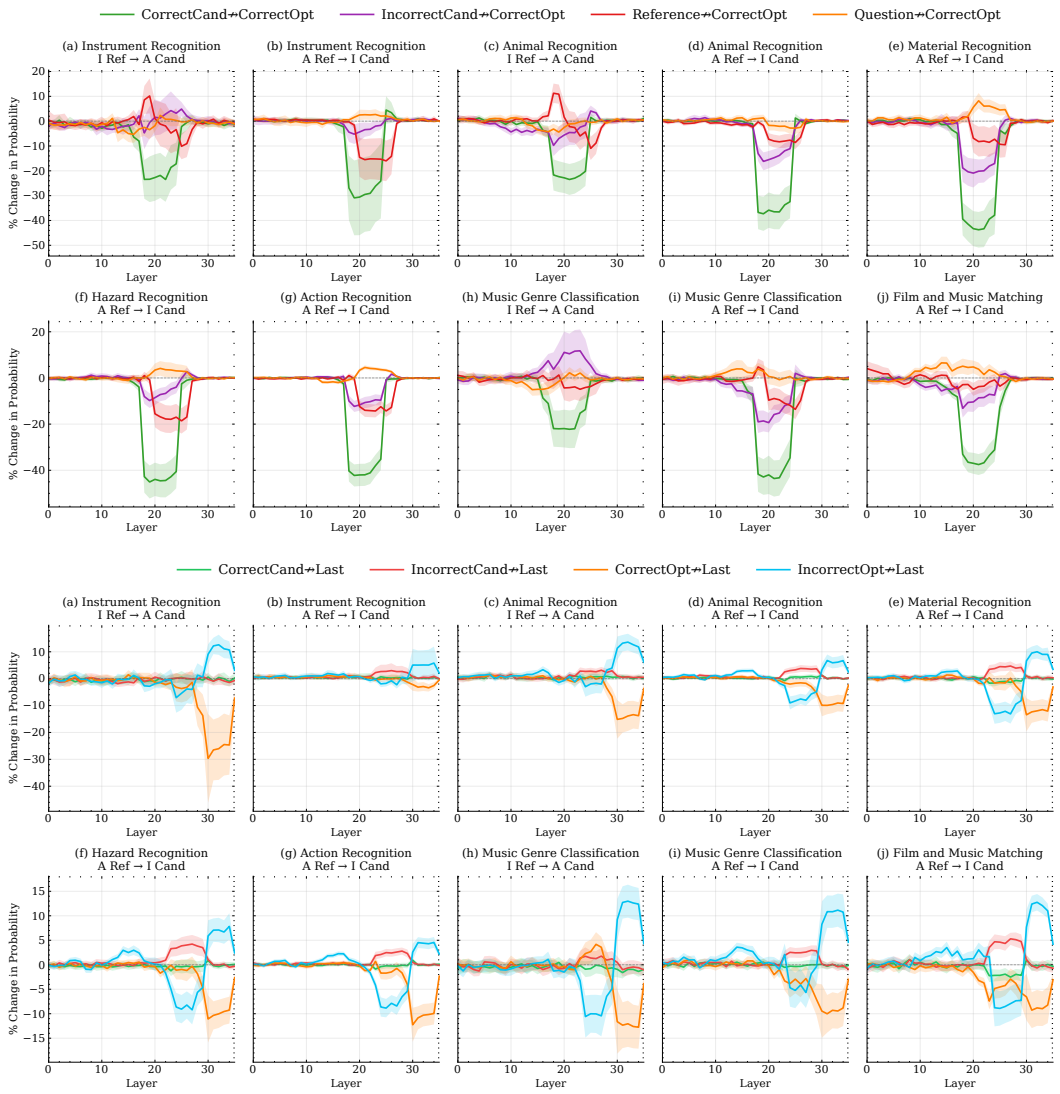

Figure 19: **Qwen2.5-Omni 3B per-task multi-input knockout (AV-Odyssey).** Pathways into the correct option letter (CorrectCand↛CorrectOpt, IncorrectCand↛CorrectOpt, Reference↛CorrectOpt, Question↛CorrectOpt) and finer-grained pathways into the last token (CorrectCand↛Last, IncorrectCand↛Last, CorrectOpt↛Last, IncorrectOpt↛Last). Each panel shows one task under one input ordering (I Ref → A Cand or A Ref → I Cand). Source↛Target indicates blocking attention edges from source tokens to target tokens.

I Generalization to Qwen2.5-Omni 7B

This appendix reports knockout analyses for Qwen2.5-Omni 7B across all three datasets used in the paper, AV-SpeakerBench [39], WorldSense [20], and AV-Odyssey [17]. The setup mirrors the analyses on Qwen2.5-Omni 3B in the main paper and Appendix H, allowing a direct comparison across model scales. For AV-Odyssey, we report both the averaged knockouts (matching the main paper's reporting style) and the per-task knockouts for completeness.

I.1 AV-SpeakerBench (audio-visual video)

Figures 20 and 21 report the AV-SpeakerBench knockout results. Figure 20 covers within- and cross-modal pathways together with modality and question pathways into the last token, while Figure 21 covers the question-internal pathways and the pathways into the correct option letter and the non-option question text.

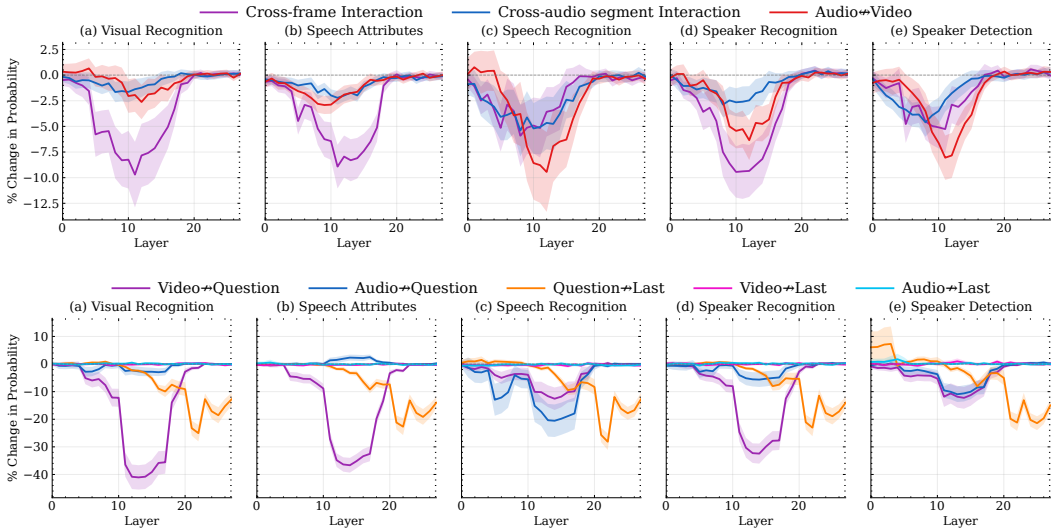

Figure 20: **Qwen2.5-Omni 7B on AV-SpeakerBench.** Knockout of within- and cross-modal pathways (Cross-frame, Cross-audio segment, Audio↔Video) and of modality and question pathways into the last token (Video↛Question, Audio↛Question, Question↛Last, Video↛Last, Audio↛Last). Source↛Target indicates blocking attention edges from source tokens to target tokens.

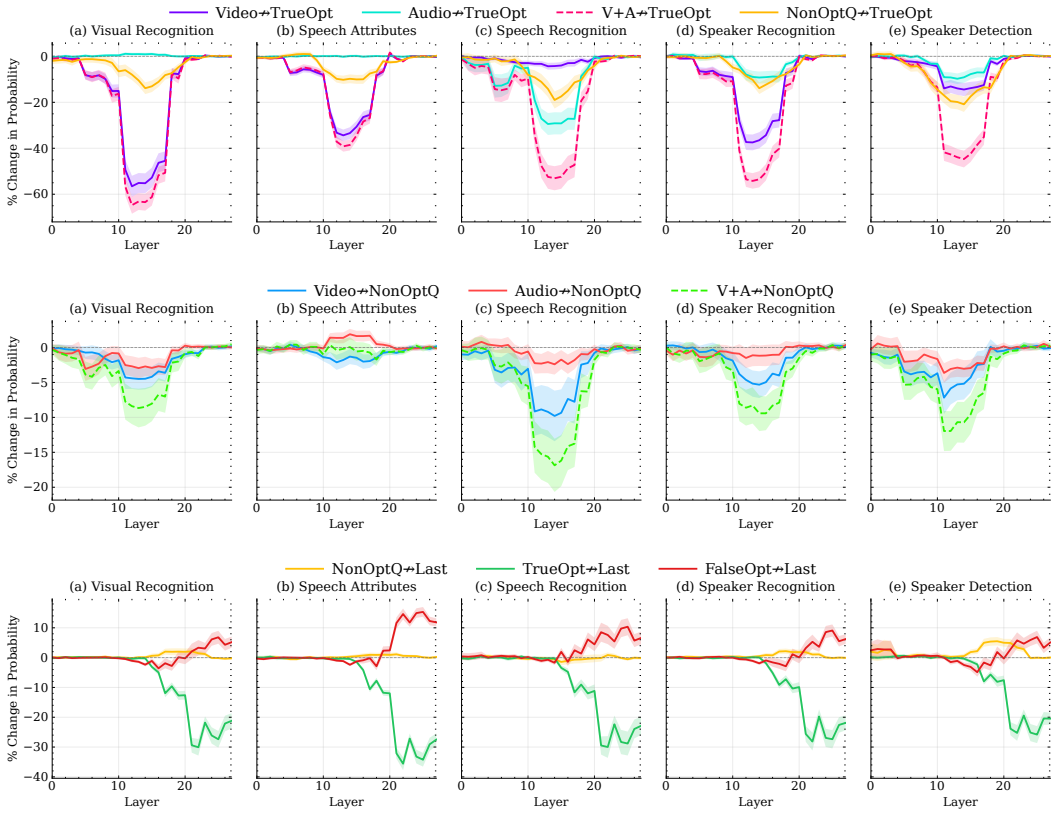

Figure 21: **Qwen2.5-Omni 7B on AV-SpeakerBench.** Modality and question pathways into the correct option letter (Video↛TrueOpt, Audio↛TrueOpt, NonOptQ↛TrueOpt, V+A↛TrueOpt); modality pathways into the non-option question text (Video↛NonOptQ, Audio↛NonOptQ, V+A↛NonOptQ); and question-internal pathways into the last token (TrueOpt↛Last, FalseOpt↛Last, NonOptQ↛Last). Source↛Target indicates blocking attention edges from source tokens to target tokens.

I.2 WorldSense (audio-visual video)

Figures 22 and 23 report the WorldSense knockout results. Figure 22 covers within- and cross-modal pathways together with modality and question pathways into the last token, while Figure 23 covers the question-internal pathways and the pathways into the correct option letter and the non-option question text.

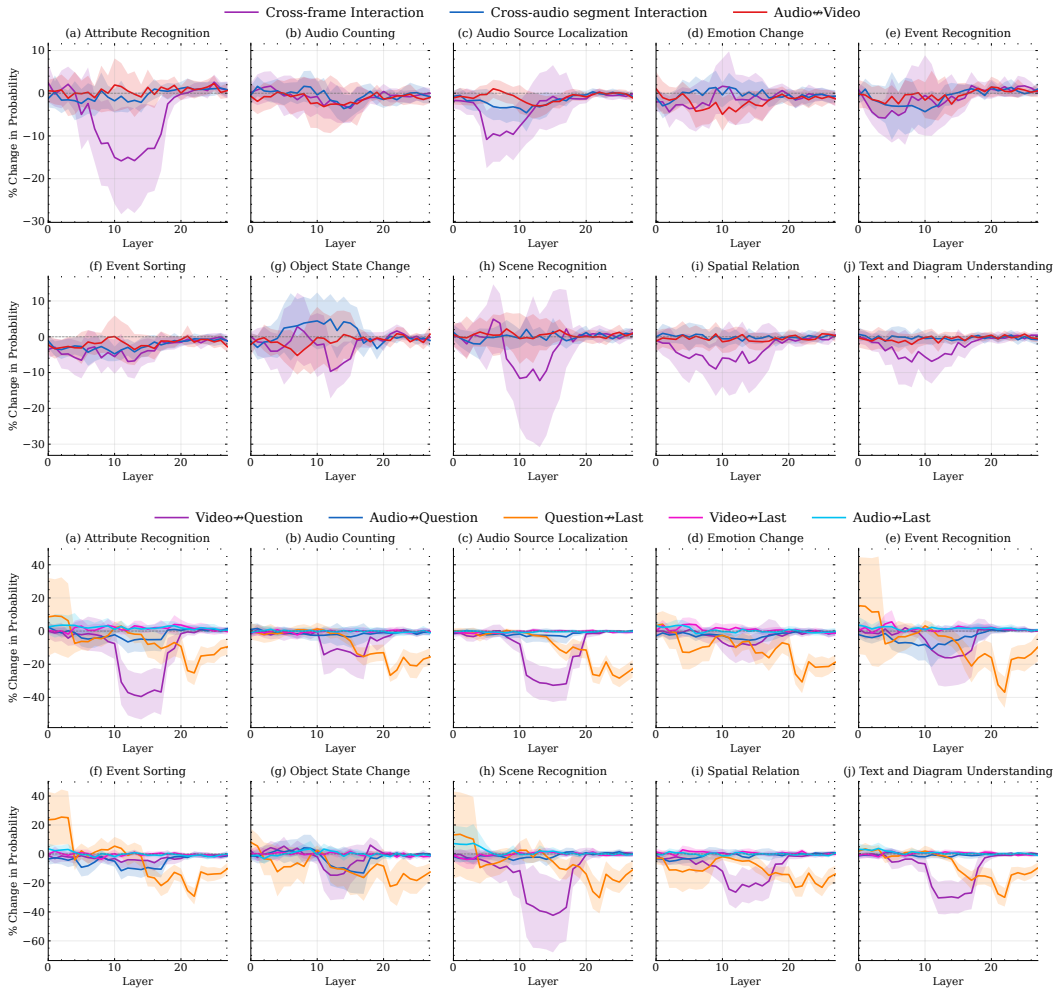

Figure 22: **Qwen2.5-Omni 7B on WorldSense.** Knockout of within- and cross-modal pathways (Cross-frame, Cross-audio segment, Audio↔Video) and of modality and question pathways into the last token (Video↛Question, Audio↛Question, Question↛Last, Video↛Last, Audio↛Last). Source↛Target indicates blocking attention edges from source tokens to target tokens.

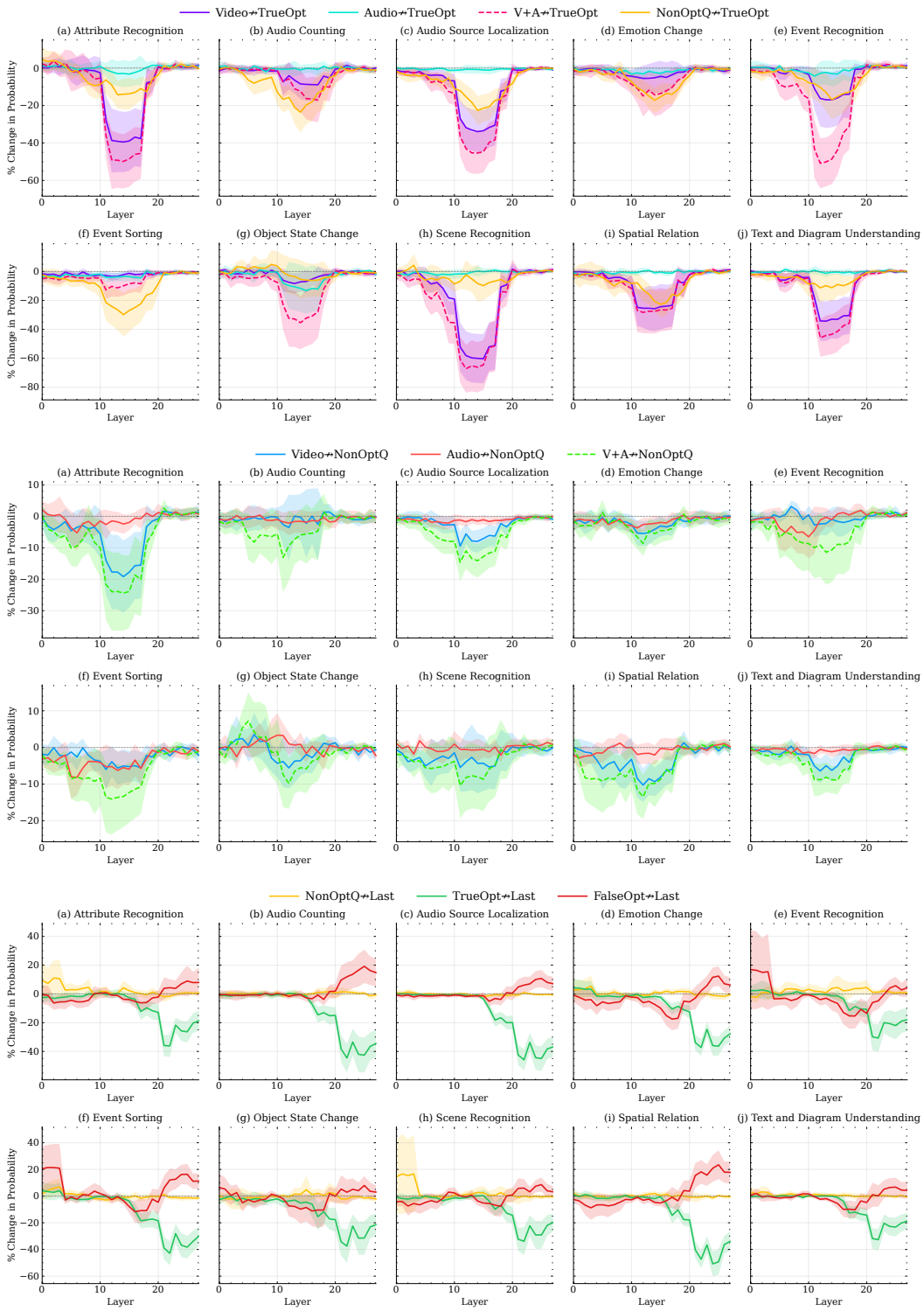

Figure 23: **Qwen2.5-Omni 7B on WorldSense.** Modality and question pathways into the correct option letter (Video↛TrueOpt, Audio↛TrueOpt, NonOptQ↛TrueOpt, V+A↛TrueOpt); modality pathways into the non-option question text (Video↛NonOptQ, Audio↛NonOptQ, V+A↛NonOptQ); and question-internal pathways into the last token (TrueOpt↛Last, FalseOpt↛Last, NonOptQ↛Last). Source↛Target indicates blocking attention edges from source tokens to target tokens.

I.3 AV-Odyssey (multi-input audio-visual interleaved)

This subsection reports multi-input knockouts on AV-Odyssey for Qwen2.5-Omni 7B in two views, averaged across tasks (Appendix I.3.1) and broken down per task (Appendix I.3.2).

I.3.1 Averaged Across Tasks

Figure 24 reports the multi-input knockout averaged across tasks, matching the reporting style of the main paper for Qwen2.5-Omni 3B. The same set of pathways into the Reference, Question, correct option letter, and last token are shown.

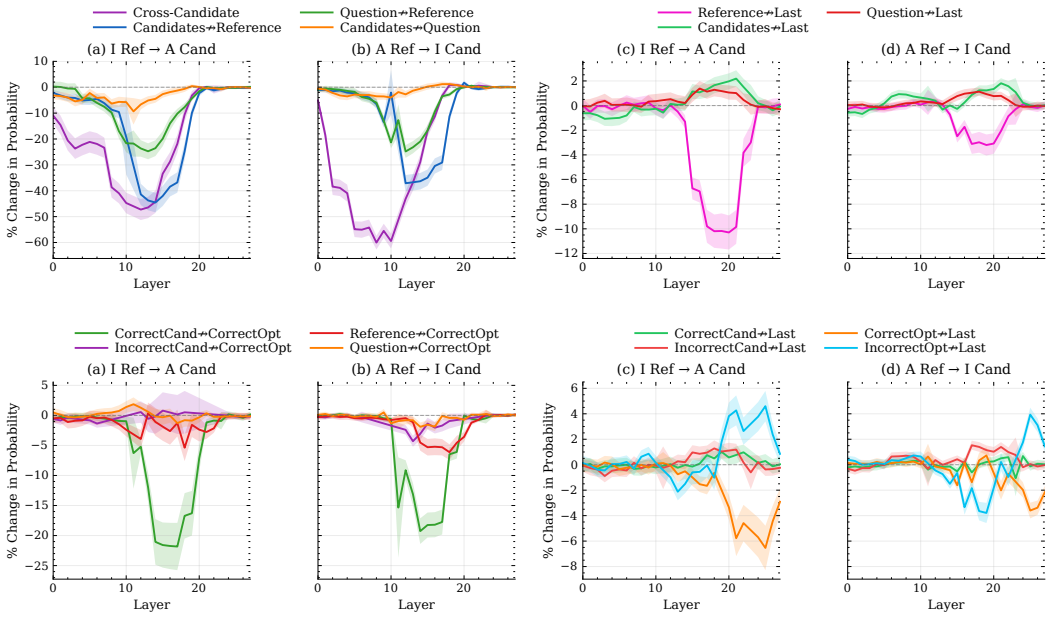

Figure 24: **Qwen2.5-Omni 7B multi-input knockout (AV-Odyssey), averaged across tasks.** Pathways into the Reference, Question, correct option letter, and last token, mirroring the analyses in the main paper for Qwen2.5-Omni 3B. Source↛Target indicates blocking attention edges from source tokens to target tokens.

I.3.2 Per-task breakdown

Figures 25 and 26 break the same knockouts down by individual task and input ordering (I Ref → A Cand and A Ref → I Cand) for completeness. Figure 25 covers pathways into the Reference and Question and pathways into the last token, while Figure 26 covers pathways into the correct option letter and the finer-grained pathways into the last token.

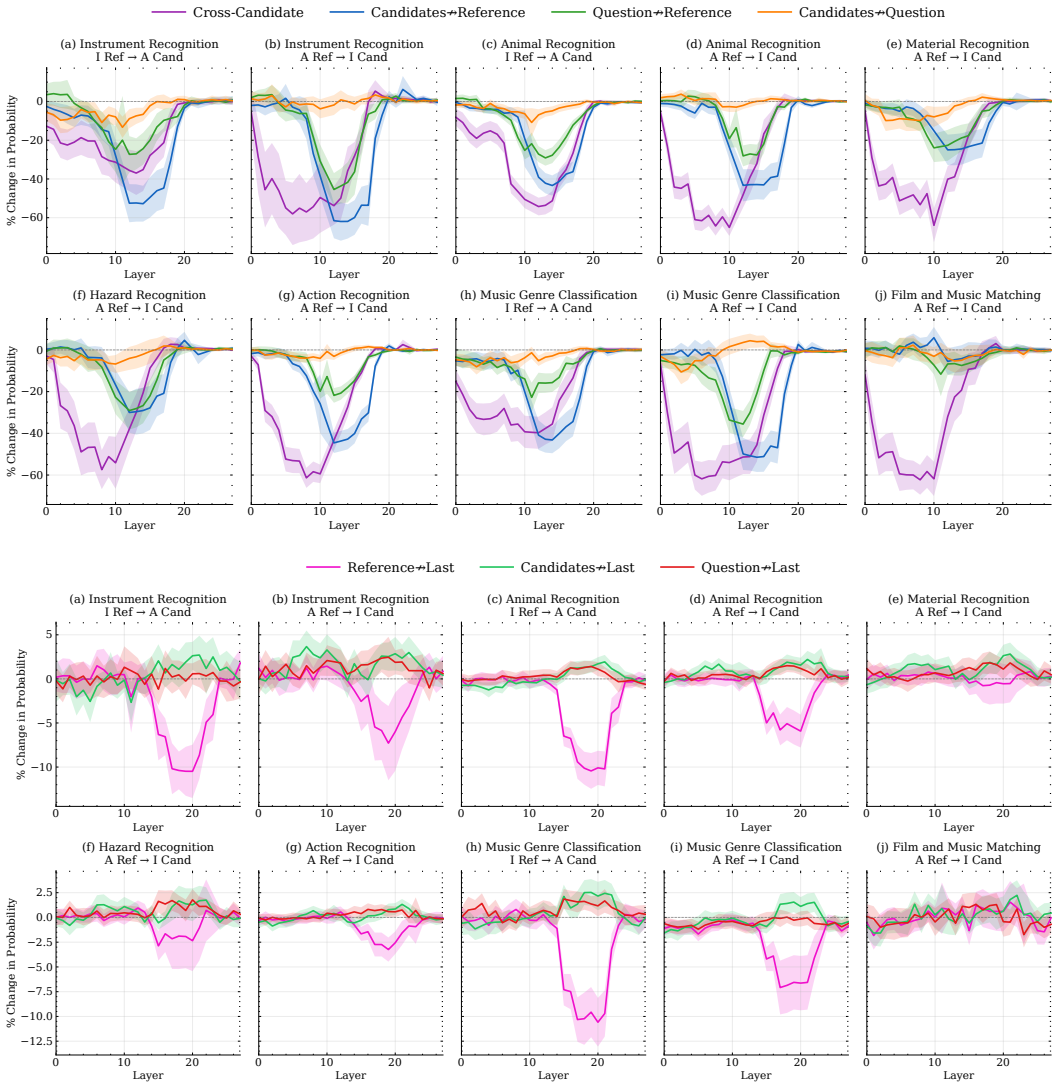

Figure 25: **Qwen2.5-Omni 7B per-task multi-input knockout (AV-Odyssey).** Pathways into the Reference and Question (Cross-Candidate, Candidates↛Reference, Question↛Reference, Candidates↛Question) and pathways into the last token (Reference↛Last, Candidates↛Last, Question↛Last). Each panel shows one task under one input ordering (I Ref → A Cand or A Ref → I Cand). Source↛Target indicates blocking attention edges from source tokens to target tokens.

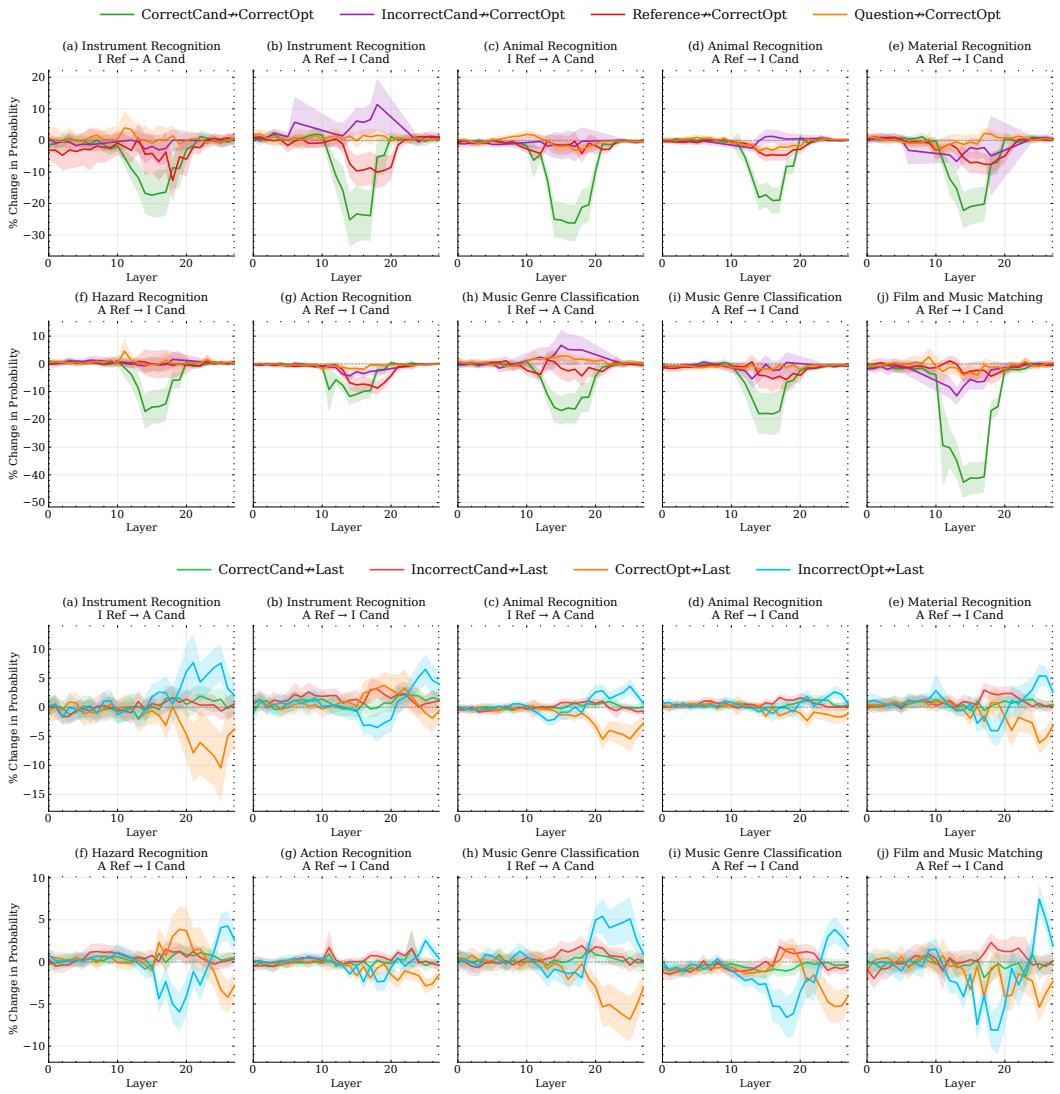

Figure 26: **Qwen2.5-Omni 7B per-task multi-input knockout (AV-Odyssey).** Pathways into the correct option letter (CorrectCand↛CorrectOpt, IncorrectCand↛CorrectOpt, Reference↛CorrectOpt, Question↛CorrectOpt) and finer-grained pathways into the last token (CorrectCand↛Last, IncorrectCand↛Last, CorrectOpt↛Last, IncorrectOpt↛Last). Each panel shows one task under one input ordering (I Ref → A Cand or A Ref → I Cand). Source↛Target indicates blocking attention edges from source tokens to target tokens.

J Generalization to Video-SALMONN2 3B Plus

This appendix reports knockout analyses for Video-SALMONN2 3B Plus on AV-SpeakerBench [39] and WorldSense [20], mirroring the analyses on Qwen2.5-Omni 3B in the main paper and Appendix H.

J.1 AV-SpeakerBench (audio-visual video)

Figures 27 and 28 report the AV-SpeakerBench knockout results. Figure 27 covers within- and cross-modal pathways together with modality and question pathways into the last token, while Figure 28 covers the question-internal pathways and the pathways into the correct option letter and the non-option question text.

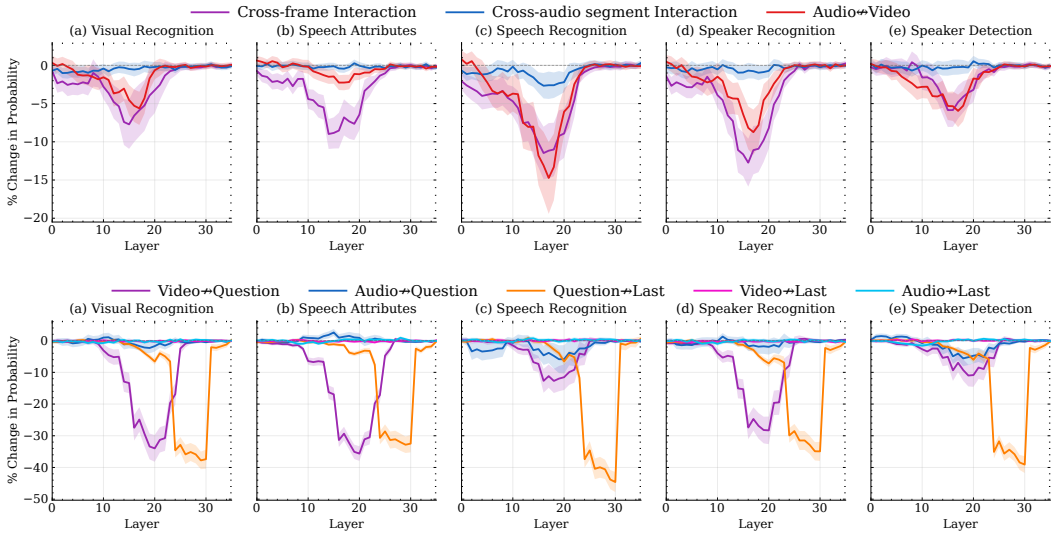

Figure 27: **Video-SALMONN2 3B Plus on AV-SpeakerBench.** Knockout of within- and cross-modal pathways (Cross-frame, Cross-audio segment, Audio↔Video) and of modality and question pathways into the last token (Video↛Question, Audio↛Question, Question↛Last, Video↛Last, Audio↛Last). Source↛Target indicates blocking attention edges from source tokens to target tokens.

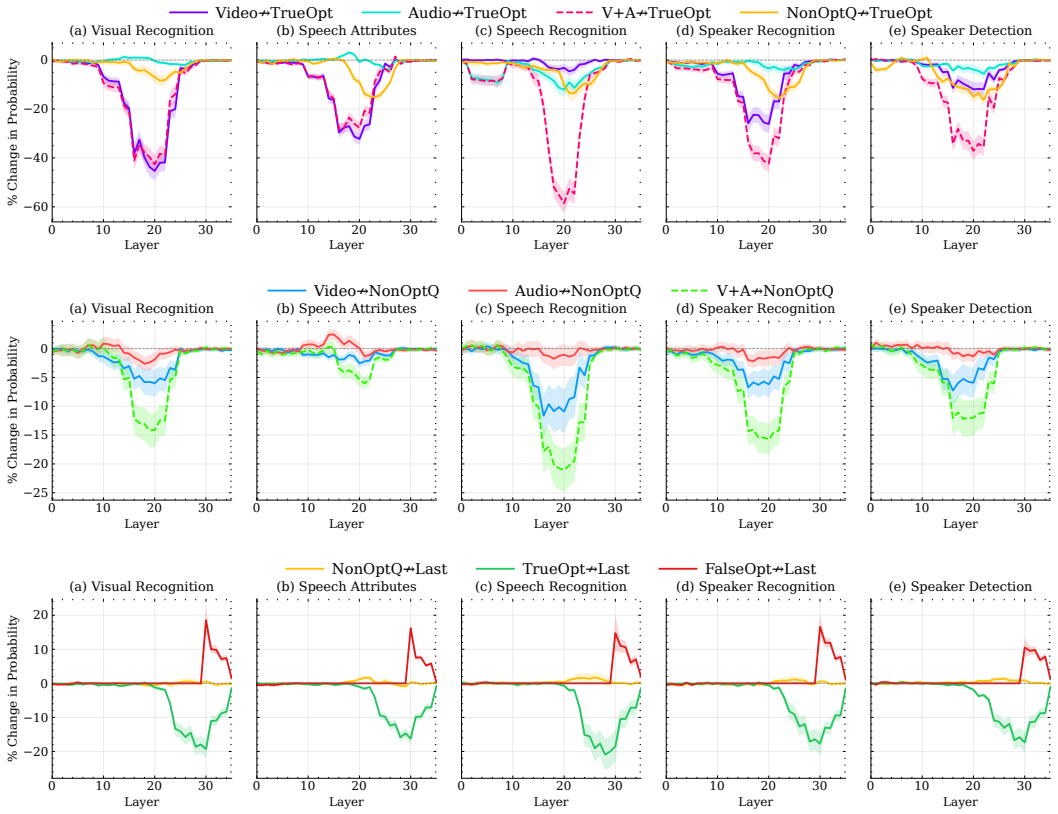

Figure 28: **Video-SALMONN2 3B Plus on AV-SpeakerBench.** Modality and question pathways into the correct option letter (Video↛TrueOpt, Audio↛TrueOpt, NonOptQ↛TrueOpt, V+A↛TrueOpt); modality pathways into the non-option question text (Video↛NonOptQ, Audio↛NonOptQ, V+A↛NonOptQ); and question-internal pathways into the last token (TrueOpt↛Last, FalseOpt↛Last, NonOptQ↛Last). Source↛Target indicates blocking attention edges from source tokens to target tokens.

J.2 WorldSense (audio-visual video)

Figures 29 and 30 report the WorldSense knockout results. Figure 29 covers within- and cross-modal pathways together with modality and question pathways into the last token, while Figure 30 covers the question-internal pathways and the pathways into the correct option letter and the non-option question text.

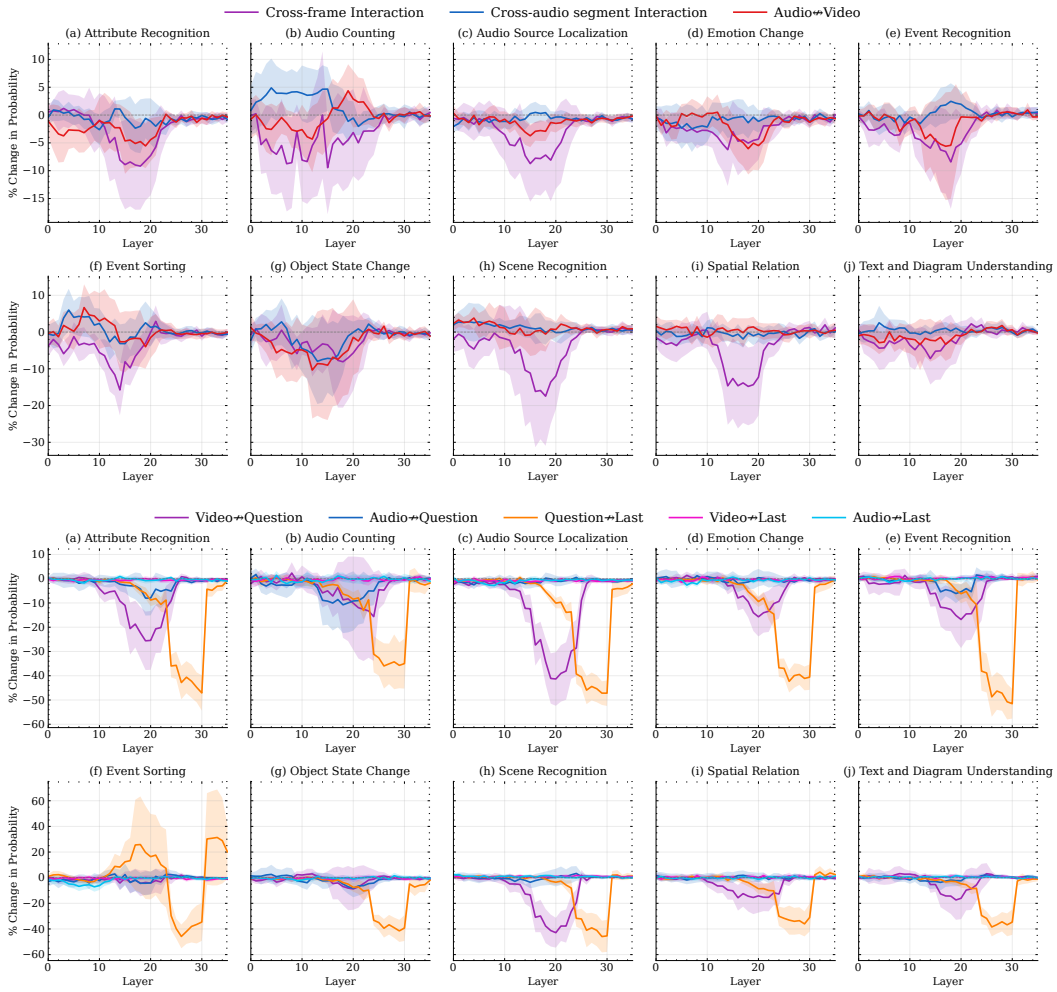

Figure 29: **Video-SALMONN2 3B Plus on WorldSense.** Knockout of within- and cross-modal pathways (Cross-frame, Cross-audio segment, Audio↔Video) and of modality and question pathways into the last token (Video↛Question, Audio↛Question, Question↛Last, Video↛Last, Audio↛Last). Source↛Target indicates blocking attention edges from source tokens to target tokens.

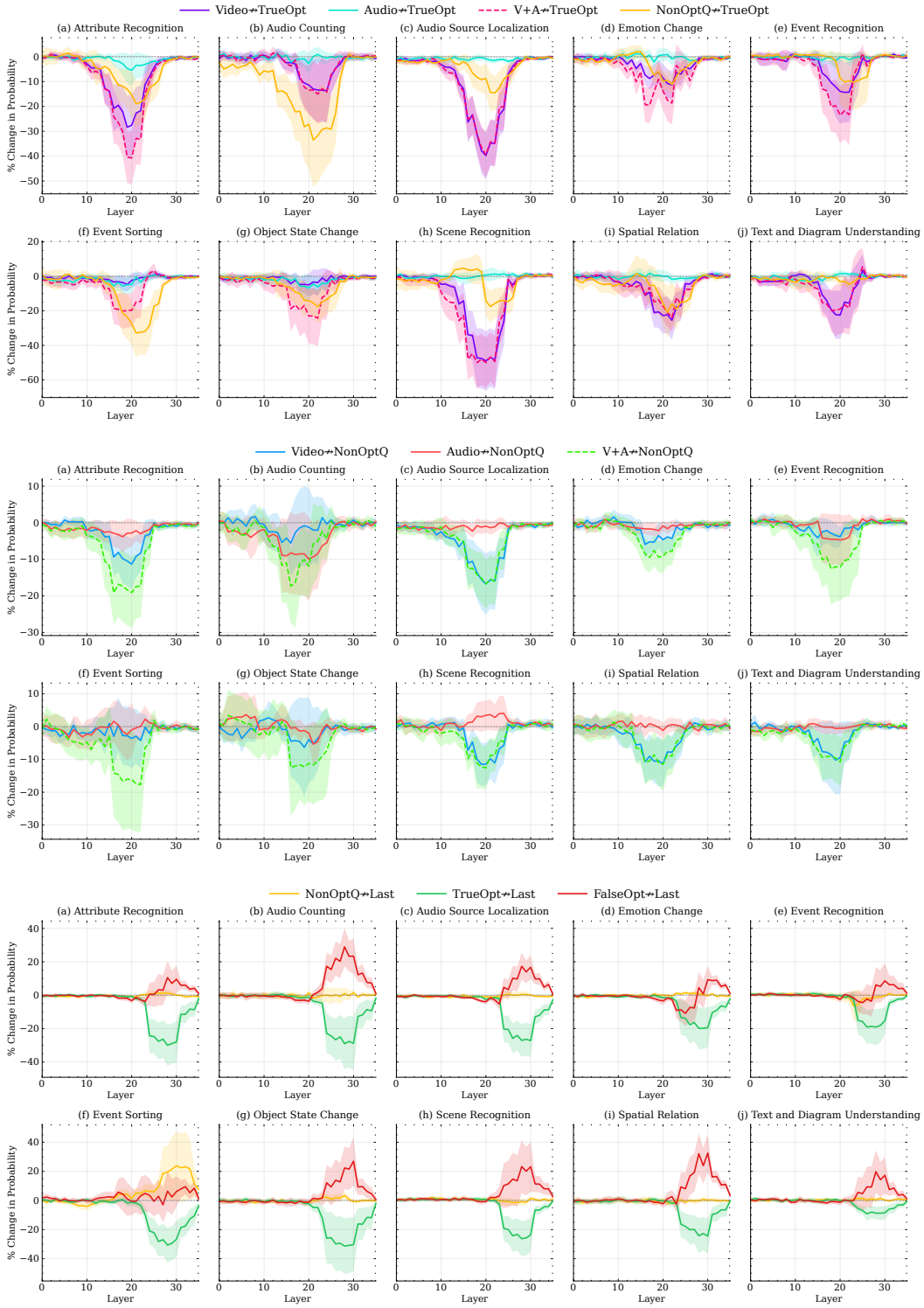

Figure 30: **Video-SALMONN2 3B Plus on WorldSense.** Modality and question pathways into the correct option letter (Video↛TrueOpt, Audio↛TrueOpt, NonOptQ↛TrueOpt, V+A↛TrueOpt); modality pathways into the non-option question text (Video↛NonOptQ, Audio↛NonOptQ, V+A↛NonOptQ); and question-internal pathways into the last token (TrueOpt↛Last, FalseOpt↛Last, NonOptQ↛Last). Source↛Target indicates blocking attention edges from source tokens to target tokens.

K Generalization to Video-SALMONN2 7B Plus

This appendix reports knockout analyses for Video-SALMONN2 7B Plus on AV-SpeakerBench [39] and WorldSense [20], mirroring the analyses on Qwen2.5-Omni 3B in the main paper and Appendix H.

K.1 AV-SpeakerBench (audio-visual video)

Figures 31 and 32 report the AV-SpeakerBench knockout results. Figure 31 covers within- and cross-modal pathways together with modality and question pathways into the last token, while Figure 32 covers the question-internal pathways and the pathways into the correct option letter and the non-option question text.

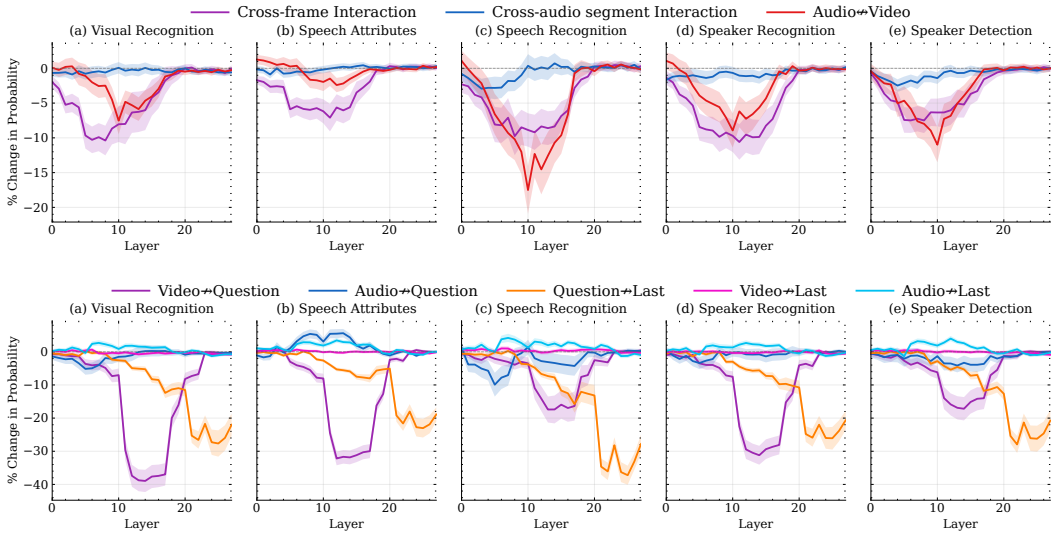

Figure 31: **Video-SALMONN2 7B Plus on AV-SpeakerBench.** Knockout of within- and cross-modal pathways (Cross-frame, Cross-audio segment, Audio↔Video) and of modality and question pathways into the last token (Video↛Question, Audio↛Question, Question↛Last, Video↛Last, Audio↛Last). Source↛Target indicates blocking attention edges from source tokens to target tokens.

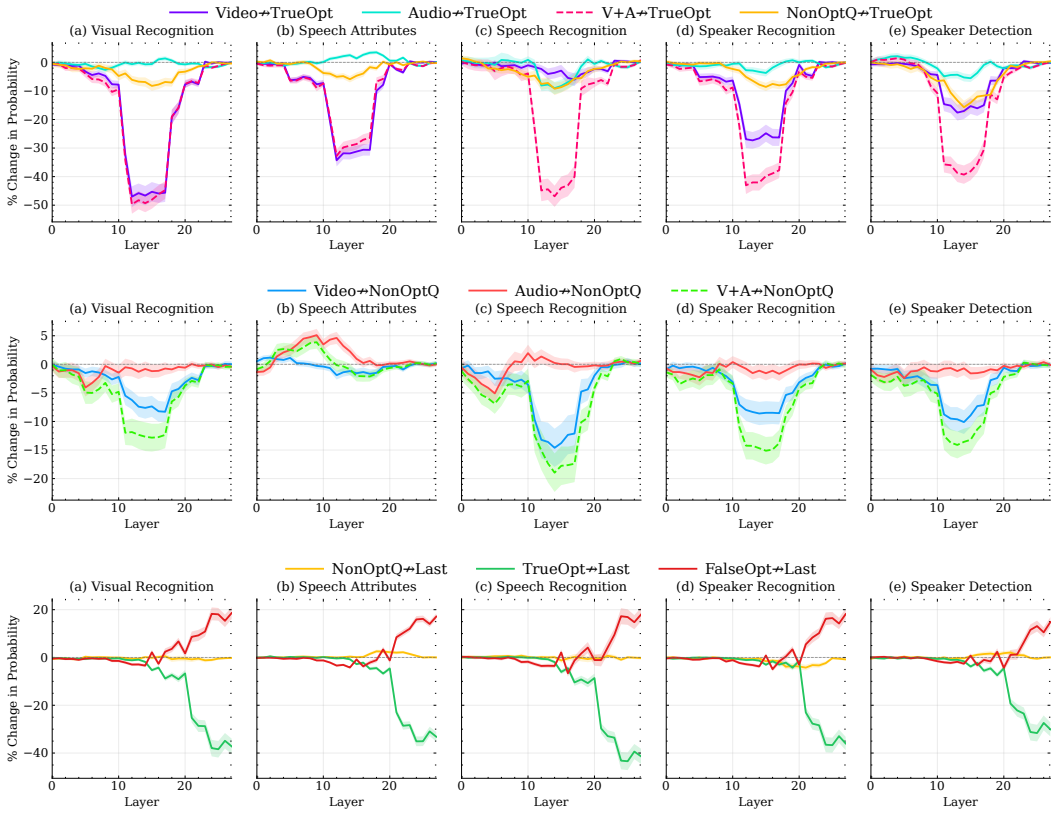

Figure 32: **Video-SALMONN2 7B Plus on AV-SpeakerBench.** Modality and question pathways into the correct option letter (Video↛TrueOpt, Audio↛TrueOpt, NonOptQ↛TrueOpt, V+A↛TrueOpt); modality pathways into the non-option question text (Video↛NonOptQ, Audio↛NonOptQ, V+A↛NonOptQ); and question-internal pathways into the last token (TrueOpt↛Last, FalseOpt↛Last, NonOptQ↛Last). Source↛Target indicates blocking attention edges from source tokens to target tokens.

K.2 WorldSense (audio-visual video)

Figures 33 and 34 report the WorldSense knockout results. Figure 33 covers within- and cross-modal pathways together with modality and question pathways into the last token, while Figure 34 covers the question-internal pathways and the pathways into the correct option letter and the non-option question text.

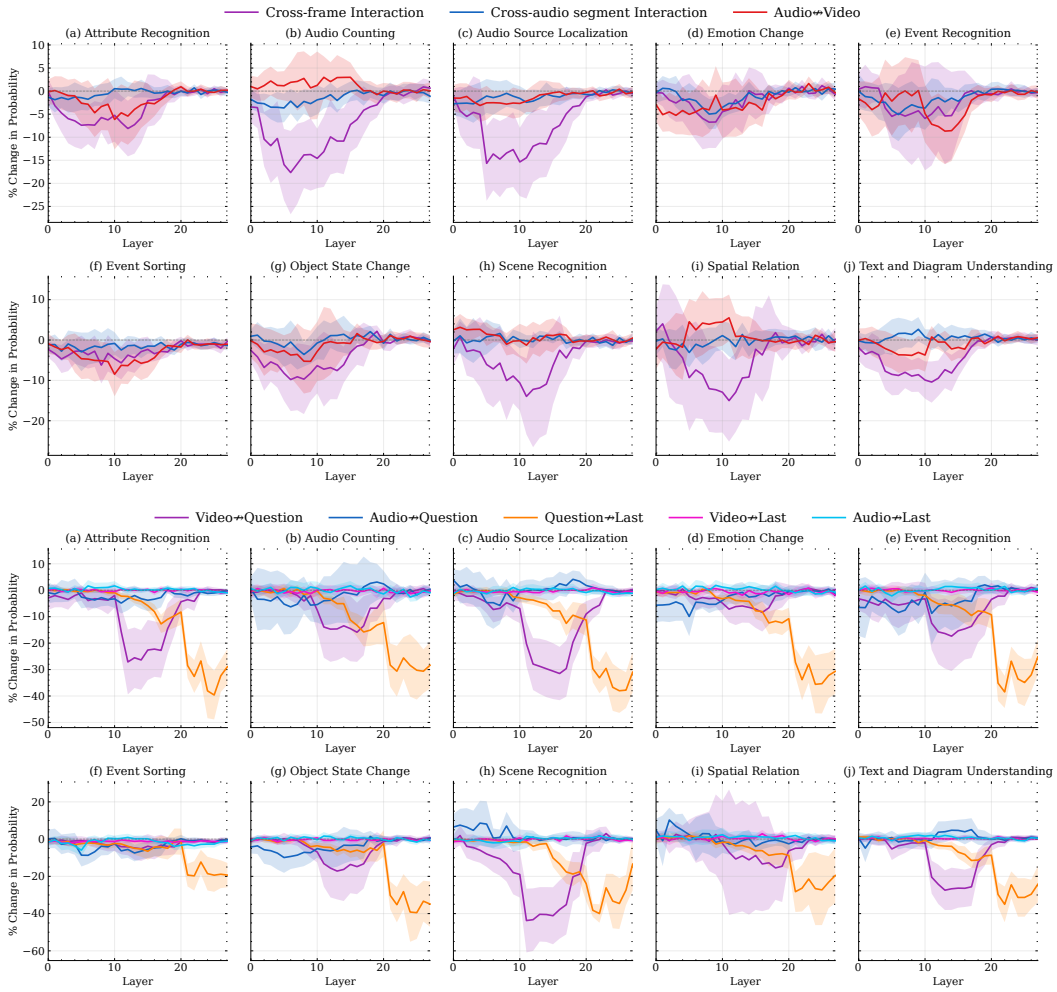

Figure 33: **Video-SALMONN2 7B Plus on WorldSense.** Knockout of within- and cross-modal pathways (Cross-frame, Cross-audio segment, Audio↔Video) and of modality and question pathways into the last token (Video↛Question, Audio↛Question, Question↛Last, Video↛Last, Audio↛Last). Source↛Target indicates blocking attention edges from source tokens to target tokens.

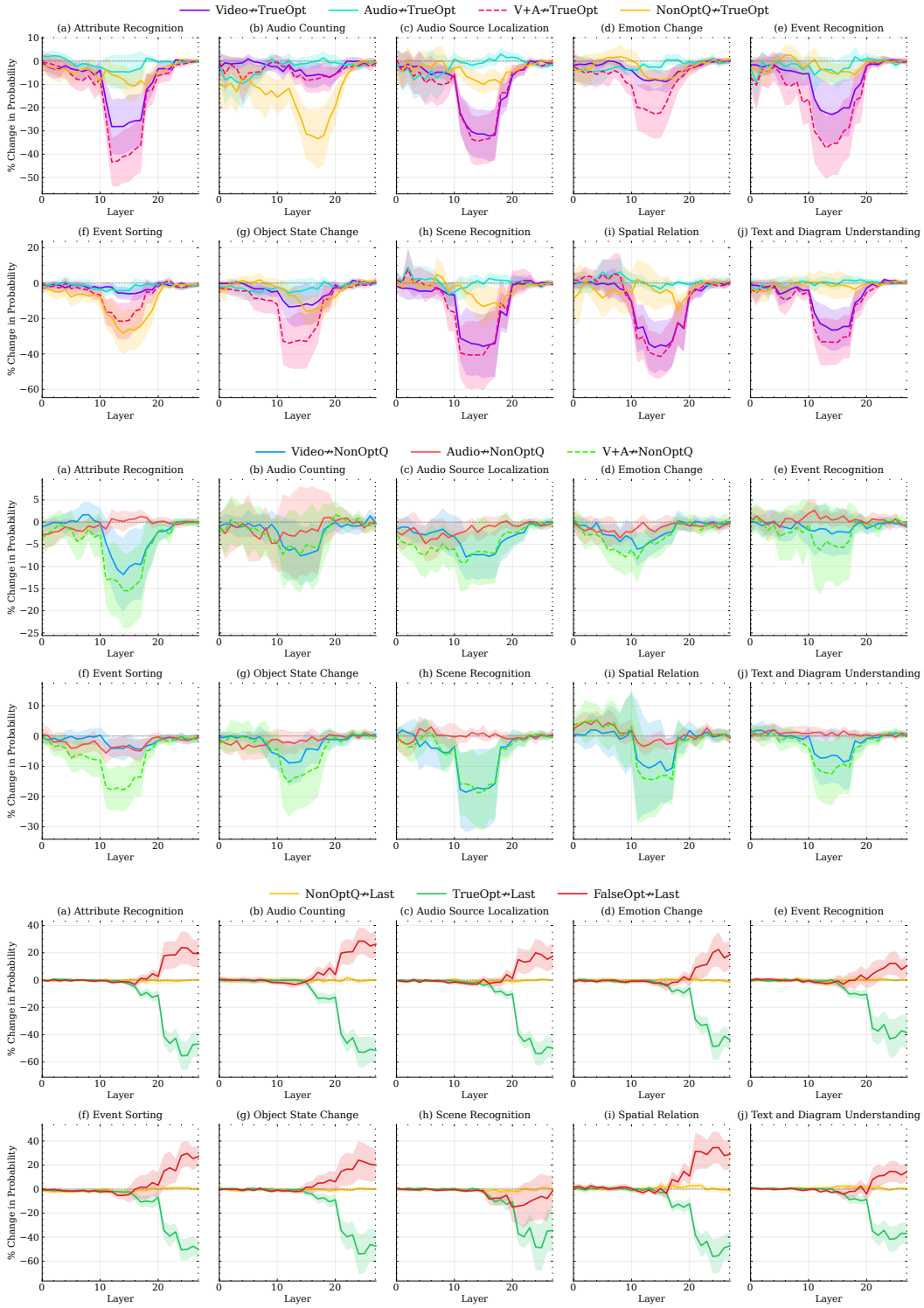

Figure 34: **Video-SALMONN2 7B Plus on WorldSense.** Modality and question pathways into the correct option letter (Video↛TrueOpt, Audio↛TrueOpt, NonOptQ↛TrueOpt, V+A↛TrueOpt); modality pathways into the non-option question text (Video↛NonOptQ, Audio↛NonOptQ, V+A↛NonOptQ); and question-internal pathways into the last token (TrueOpt↛Last, FalseOpt↛Last, NonOptQ↛Last). Source↛Target indicates blocking attention edges from source tokens to target tokens.

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: The abstract and introduction state the main claims, which are reflected in Sections 4, 5, and 6.

   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: In Section 7 discusses the limitations scope of the study.

   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: The main paper provides a comprehensive description of the experimental setup, and the Appendix includes the additional details necessary to reproduce the main experimental results.

   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: [No]

Justification: Code will be released publicly upon publication. All datasets and models used in this work are already publicly available and properly cited. The experimental methodology is described in Section 4, and full experimental details, including model checkpoints, inference setup, dataset processing, and task selection, are provided in the Appendix to support reproduction.

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: Experimental details are described across the paper, and specifically in the Appendix on Experimental Details.

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: [No]

Justification: The experiments use deterministic decoding on the full dataset in a single execution, so no statistical variability is reported.

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: Compute resources are described in the Experimental Details appendix.

   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: The paper does not deal with any research that might deviate from the NeurIPS 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: [N/A]

    Justification: The paper does not introduce new models or technologies with potential malicious uses.

    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: The paper does not release any models or datasets that pose risks of misuse.

    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: All models and datasets used are publicly available and properly cited. The assets are released under the Qwen Research License, Apache 2.0, and CC BY-NC 4.0. Each asset is used in accordance with its license and intended research use.

    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 are not releasing 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: The paper does not deal in research with 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: The paper does not deal in research with 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: [N/A]

    Justification: We used LLM only for writing, editing, or formatting purposes, and does not impact the core methodology.

    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.