CHAIN OF MODALITY: UNDERSTANDING AND CONTROLLING MODALITY ORGANIZATION BIAS IN OMNI-MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni-modal Large Language Models (Omni-MLLMs) are designed to reason over diverse sensory streams within a unified model. However, we find that multimodal reasoning is not solely determined by sensory evidence, but also by how modality information is organized. Through systematic analysis across diverse reasoning scenarios, we show that different modality organizations exhibit distinct advantages and limitations, and no single topology is universally optimal. Rather than proposing a general-purpose improvement to Omni-MLLM accuracy, we ask a narrower question: can these topology-induced failures be systematically identified and corrected? Motivated by this, we propose Chain of Modality (CoM), a framework that dynamically reorganizes multimodal topology during inference and learns adaptive organization strategies. Experiments across five benchmarks, diverse architectures, and model scales show that CoM reliably recovers topology-sensitive failures, through both training-free planning and lightweight Planner-SFT, while preserving performance on the full benchmark.

1 INTRODUCTION

The emergence of Omni-modal Large Language Models (Omni-MLLMs) has established a promising paradigm for unified perception and reasoning across heterogeneous modalities, including audio, vision, and text Xu et al. (2025a); Li et al. (2025); Xu et al. (2025b); AI et al. (2025); Team (2026). By jointly modeling multiple sensory streams within a unified framework, recent Omni-MLLMs have demonstrated increasingly strong multimodal perception and reasoning capabilities. A substantial body of recent work has therefore focused on improving the ability of Omni-MLLMs to understand and reason over heterogeneous modalities.

Existing efforts mainly enhance multimodal reasoning through large-scale multimodal training, modality alignment, and post-training optimization. Specialized models enhance modality interaction through task-specific adaptation Du et al. (2025); Kim et al. (2025); Ye et al. (2024), and generalist Omni-MLLMs improve multimodal reasoning through large-scale training and architectural scaling Guan et al. (2026); Tao et al. (2025b); Li et al. (2026). Alignment-based approaches using reinforcement learning further optimize model behavior with preference- or task-level supervision Chen et al. (2025); Ye et al. (2025b); Kulkarni & Fazli (2025); Lu et al. (2025); Chen et al. (2026), and emerging agent frameworks improve complex task solving by orchestrating external tools such as ASR, captioning, and retrieval Tao et al. (2025a); Li et al. (2026); Zhou et al. (2025). More recently, Selvakumar et al. (2026) questioned whether audio-visual LLMs truly exploit their available modalities, revealing modality bias and incomplete modality utilization. However, these studies primarily analyze *which modalities* models rely on, while the role of *how modalities are organized* during inference has received substantially less attention. In particular, existing Omni-MLLMs typically adopt a predefined multimodal input organization, leaving modality ordering and structural arrangement outside the inference decision process.

In this work, we revisit multimodal inference from the perspective of modality organization as shown in Figure 1. Prior studies have examined modality bias and input-order sensitivity in multimodal models Tan et al. (2024); Zheng et al. (2025); Schilcher et al. (2025), but this evidence comes largely *from image-text settings only with sequential input*. We instead investigate whether the organization of audio-visual information affects multimodal reasoning in Omni-MLLMs. To systematically study

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



Figure 1: **Modality organization affects Omni-MLLM reasoning.** Across diverse reasoning scenarios, different topology structures exhibit distinct advantages, motivating adaptive modality organization during inference.

this question, we consider several representative multimodal reasoning scenarios, including single-modality sufficiency, cross-modal conflict, cross-modal interaction, and temporal reasoning. We examine each requirement under different modality organizations in Section 3, including Single, Parallel, Sequential, and Interleaved organizations. Our experiments reveal that the benefit of multimodal information depends on how it is organized: different reasoning requirements exhibit different preferences over modality topologies, and no single organization consistently performs best across cases. These observations suggest that modality organization is not merely an input-format choice, but an important factor in Omni-modal inference.

To investigate whether and to what extent this phenomenon can be controlled, we develop a topology-controlled inference framework, termed **C**hain **o**f **M**odality (CoM). CoM consists of a Planner and a Reasoner, with an optional Decider. The Planner constructs modality chains by determining modality combinations and their interaction topologies, including Single, Parallel, Sequential, and Interleaved structures. The Reasoner performs inference following the specified organization. When multiple parallel chains are generated, the Decider aggregates their predictions through logit-level fusion to obtain the final prediction.

To evaluate the role of modality organization, we construct topology-sensitive subsets from fine-grained audio-visual perception benchmarks Li et al. (2022), open-domain multimodal scenarios Zhou et al. (2025); Li et al. (2024); Hong et al. (2025), and cross-modal hallucination evaluation Sung-Bin et al. (2024). These subsets capture two types of topology-sensitive cases: (1) modality integration failures, where unimodal organization produces the correct answer, while all multimodal organizations fail; and (2) topology-induced prediction variance, where identical multimodal evidence produces different outcomes under different organizations. On these challenging samples, we investigate two adaptation settings: training-free topology reorganization and planner-only supervision. Training-free CoM shows that explicitly modifying modality organization can already improve multimodal inference, revealing the importance of topology selection itself. Further Planner-SFT supervision achieves additional improvements and generalizes across unseen benchmarks, suggesting that the planner learns transferable organization preferences rather than dataset-specific heuristics. **Our goal is not to improve general Omni-MLLM accuracy, but to establish that topology-induced failures are systematically identifiable and learnable.**

Our contributions are summarized as follows: **(1)** We reveal modality organization as an overlooked factor affecting Omni-MLLM reasoning, and show that different multimodal reasoning scenarios exhibit distinct preferences over modality topologies. **(2)** We perform controlled intervention experiments to analyze the structural factors underlying topology sensitivity and introduce Chain of Modality (CoM) as a controllable framework for explicitly modeling and studying modality organization in multimodal inference. **(3)** We investigate the controllability of modality organization through training-free topology reorganization and Planner-SFT, demonstrating that CoM can effectively learn and leverage adaptive organization strategies.

2 Related Work

Omni-modal Large Language Models. Omni-modal Large Language Models (Omni-MLLMs) integrate audio, vision, and text for general-purpose reasoning, spanning early domain-specific benchmarks Yang et al. (2022); Li et al. (2022) to open-domain, hallucination-aware evaluations Gong et al. (2024); Hong et al. (2025); Zhou et al. (2025); Sung-Bin et al. (2024), and progressing from proprietary systems Hurst et al. (2024); Comanici et al. (2025) to open-source models such as Qwen-Omni Xu et al. (2025a;b), Baichuan-Omni Li et al. (2025), and Ola Liu et al. (2025). In practice, these models organize audio and visual streams either sequentially AI et al. (2025); Liu et al. (2025) or as interleaved, timestamp-synchronized chunks Xu et al. (2025a;b), with the topology fixed in advance rather than adapted to the task. Separately, prior work has shown that multimodal models are sensitive to input order Tan et al. (2024); Zheng et al. (2025); Schilcher et al. (2025) and that multimodal fusion can produce cross-modal hallucination Sung-Bin et al. (2024); Leng et al. (2024), with concurrent efforts auditing order sensitivity more broadly Paruchuri et al. (2026); Xu et al. (2026). However, these studies mainly analyze the symptoms of modality bias or hallucination from image-text or multi-video settings, while the role of modality organization and interleaving topology remains largely unexplored. To our knowledge, no prior work studies whether these topology-dependent behaviors emerge in audio-visual Omni-MLLMs, where sequential or interleaved organizations are already adopted as default input paradigms.

Omni-modal Reasoning and Adaptation. Beyond developing Omni-MLLM architectures, recent studies have explored various strategies to improve the reasoning capability of omni-modal systems. Specialized models, such as Crab Du et al. (2025), QA-TIGER Kim et al. (2025), and CAT Ye et al. (2024; 2025a), improve modality-specific understanding through task-oriented adaptation. Large-scale reasoning-oriented models, such as ThinkOmni Guan et al. (2026), improve omni-modal reasoning through advanced inference strategies. Alignment-based approaches using DPO or reinforcement learning Chen et al. (2025); Ye et al. (2025b); Kulkarni & Fazli (2025); Lu et al. (2025); Chen et al. (2026) further optimize model behaviors through preference or task-level supervision. Meanwhile, omni-modal agent frameworks, such as OmniAgent Tao et al. (2025a), and OmniGaia Li et al. (2026), improve complex task solving by orchestrating external tools or specialized modules. However, these analyses mainly focus on whether modalities are effectively perceived and utilized, treating modality organization as a fixed preprocessing choice rather than a controllable reasoning factor. In contrast, our work investigates modality organization as an independent factor affecting Omni-MLLM reasoning and introduces the Chain of Modality (CoM) framework to explicitly model and study modality interactions.

3 Understanding Topology Preferences across Multimodal Reasoning Scenarios

To systematically study how modality organization interacts with different reasoning scenarios, we construct four controlled subsets: *single-modality sufficiency*, *cross-modal conflict*, *cross-modal interaction*, and *temporal reasoning*. Here, *single-modality sufficiency* refers to cases where the answer can be correctly determined from either audio or visual information alone. *Cross-modal conflict* refers to cases where the audio and visual streams describe different or inconsistent content. Moreover, *cross-modal interaction* refers to cases where the two modalities provide complementary evidence, and correctly answering the question requires jointly integrating information from both. Finally, *temporal reasoning* involves complementary multimodal evidence that additionally requires identifying or associating the relevant information with specific moments in time. Rather than following benchmark-level categories, we categorize samples according to the type of evidence required for correct reasoning. Specifically, we draw candidate samples from representative benchmarks that contain potential cases of these reasoning patterns: Music-AVQA Yang et al. (2022), AVHBench

Table 1: **Different multimodal reasoning scenarios exhibit distinct preferences for modality organization.** We fix the temperature and random seed across all experiments to reduce the impact of stochastic variation. "-" indicates unavailable results: Ola-7B does not support interleaved modality organization, and the cross-modal interaction subset only involves audio-image interaction.

Dataset	Subset	Single		Parallel	Sequential		Interleaved
		A	V		A→V	V→A	
Qwen-Omni-7B							
single-modality sufficiency	Audio	**60.3**	27.6	27.6	39.7	49.3	55.2
	Video	35.7	**88.6**	38.6	82.9	81.4	80.0
cross-modal conflict	-	57.0	67.0	**68.0**	59.5	61.0	52.5
cross-modal interaction	-	31.3	31.5	34.8	41.4	**47.5**	-
temporal reasoning	-	35.6	42.4	47.2	45.4	47.7	**48.5**
Ola-7B							
single-modality sufficiency	Audio	**57.4**	55.2	56.6	53.9	56.6	-
	Video	65.6	**85.3**	80.3	83.6	77.1	-
cross-modal conflict	-	67.2	68.3	**68.9**	56.7	60.0	-
cross-modal interaction	-	27.0	30.0	27.5	35.0	**37.0**	-
temporal reasoning	-	**44.3**	35.8	39.8	43.2	41.5	-

Table 2: **Attention redistribution under single-modality sufficiency.** Attention values denote question-to-modality attention.

Subset	Single	A+V	
Qwen-Omni-7B	Decisive Attn	Attn(A)	Attn(V)
A-sufficient	Attn(A)=0.0487	0.0386	0.0103
V-sufficient	Attn(V)=0.0367	0.0489	0.0181

Table 3: **Correct-answer average log-probability under cross-modal conflict.** We compare unimodal and fused log-probability under different organizations.

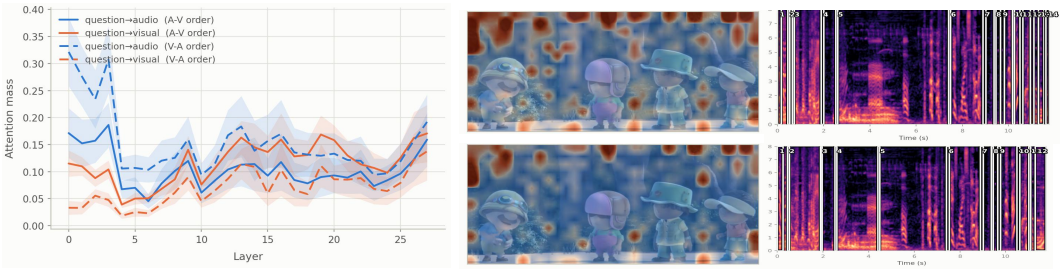
Subset	Single-A	Single-V	Parallel	Joint
Conflict	-1.05	-0.74	**-0.67**	-0.86

Sung-Bin et al. (2024), OmniBench Li et al. (2024), and DailyOmni Zhou et al. (2025). We first use Qwen-Omni-30B Xu et al. (2025b) to pre-annotate candidate samples according to the evidence required for answering each question, based on the question, ground-truth answer, and multimodal content, without exposing predictions under different modality organizations. The annotations are then manually verified using predefined criteria, and ambiguous cases are removed. For each reasoning requirement, we retain about 200 verified samples to ensure balanced evaluation across different modality organizations.

We evaluate each controlled subset under four modality organizations: *Single*, where only one modality is provided to inference; *Parallel*, where audio and visual modalities are reasoned over independently and their predictions are subsequently aggregated; *Sequential*, where modalities are concatenated sequentially following a specific order, such as Audio→Visual or Visual→Audio; and *Interleaved*, where audio and visual segments are alternated to preserve fine-grained cross-modal correspondence. Table 1 summarizes the performance across the four reasoning scenarios. We observe that no single organization consistently dominates across all cases. Instead, different reasoning requirements exhibit distinct preferences over modality organization, suggesting that multimodal reasoning is determined not only by the available modalities but also by how they are structurally organized during inference.

Single-modality Sufficiency As shown in Table 1, we compare single-modality and multimodal organizations on both video-sufficient and audio-sufficient samples. For each subset, the modality containing decisive evidence achieves the best single-modality performance, while introducing the non-essential modality does not provide additional benefits and may even degrade performance. We further analyze modality-question interactions before and after introducing the non-essential modality in Table 2. Across both audio- and video-sufficient samples, introducing the non-essential modality shifts attention toward irrelevant evidence and reduces attention allocated to the decisive modality. These results suggest that unnecessary modalities may interfere with effective evidence utilization by altering the model's attention allocation across modalities.

Cross-modal Conflict When audio and visual modalities provide conflicting evidence, Parallel organization achieves more reliable performance, as shown in Table 1. The conflict subset contains samples with different modality biases, and neither unimodal branch consistently dominates across all samples. Unlike single-modality inference, conflict cases require integrating competing evidence from multiple modalities rather than relying on a single sensory stream. We hypothesize that Parallel improves robustness by preserving modality-specific reasoning before evidence aggregation, allowing



(a) (b)

Figure 2: **(a) Layer-wise question-to-audio/video attention under different modality orders. (b) Attention visualization before and after modality reordering.** For audio inputs, white regions denote attended temporal regions. Up: $A \to V$, Down: $V \to A$

each modality to independently contribute its evidence. To analyze this effect, we measure the average log-probability assigned to the ground-truth answer tokens under each organization (Table 3). The two unimodal branches provide different confidence signals, and Parallel aggregation further improves the confidence of correct predictions. In contrast, joint organizations exhibit larger deviations after early cross-modal interaction, suggesting that premature fusion may cause conflicting evidence to interfere before modality-specific evidence is properly integrated.

Cross-modal Interaction When audio and visual modalities provide complementary evidence, we observe that Sequential organization exhibits strong order sensitivity, where reversing modality order leads to different reasoning outcomes despite identical multimodal content, as shown in Table 1. We analyze the effect of modality ordering from the perspective of RoPE-based attention. Following the formulation of RoPE Su et al. (2024), consider an unrotated query representation q_m at position m and an unrotated key representation k_n at position n. By the relative-position property of RoPE, their attention logit can be written as

$$\ell(q_m, k_n) = q_m^T R_{n-m} k_n = \underbrace{q_m^T k_n}_{\text{content interaction}} + \underbrace{q_m^T (R_{n-m} - I) k_n}_{\text{positional modulation}},$$

where $R_{n-m} = R_m^T R_n$ is the relative block-diagonal rotation matrix determined by the positional offset $n - m$. Thus, even when the underlying semantic content remains unchanged, the query-key interaction can be modulated by their relative positions. Following the Abel transformation used in the original RoPE analysis Su et al. (2024), the attention interaction is bounded by

$$\left| q_m^T R_{n-m} k_n \right| \leq \sum_{i=0}^{d/2-1} |S_{i+1}(n-m)| \, |h_{i+1} - h_i| \,,$$

where $h_i = q_m[2i : 2i+1] k_n[2i : 2i+1]$ and $S_{i+1}(n-m) = \sum_{j=0}^{i} e^{\mathrm{i}(n-m)\theta_j}$, with the boundary convention $h_{d/2} := 0$. We denote the right-hand side as $B(n-m)$. The RoPE analysis shows that $B(\cdot)$ exhibits a **long-term decaying** envelope as the relative distance increases. Hence,$|\ell(q_m, k_n)| \leq B(|n-m|)$, indicating that relative position imposes a distance-dependent upper envelope on the strength of query-key interaction. For our question-to-modality analysis, let $\Delta m_A = p_{Q,t} - p_{A,t}$ and $\Delta m_V = p_{Q,t} - p_{V,t}$ denote the relative temporal positions between the question and the audio or visual modality, respectively. Under a modality-ordering intervention, the relative positions vary with the modality order. Therefore, $A \to V \to Q : |\Delta m_V| < |\Delta m_A|$, and $V \to A \to Q : |\Delta m_A| < |\Delta m_V|$. A smaller relative distance to the question corresponds to a less restrictive attenuation envelope, leaving the modality's query-key interaction less constrained and potentially allowing greater attention allocation.

To verify this hypothesis, we compare semantic relevance and attention allocation under different modality organizations. Specifically, $S(Q, A)$ and $S(Q, V)$ measure the semantic similarity, and $A(Q, A)$ and $A(Q, V)$ denote the attention. As shown in Table 4, reversing modality order does not change semantic relevance. However, attention allocation changes substantially. In the $A \to V$ organization, visual tokens receive higher attention despite lower semantic relevance, while in the $V \to A$ organization, attention shifts toward audio tokens after they are placed closer to the question representation. Layer-wise attention visualization in Figure 2(a) further shows that this

Table 4: **Semantic relevance vs. attention allocation between question and audio/visual on audio/video-bias subset on Qwen-Omni-7B.**

Subset	Order	$S(Q,A)$	$S(Q,V)$	$A(Q,A)$	$A(Q,V)$	Acc
Audio-Bias	$A \to V$	**0.234**	0.151	0.101	**0.117**	48.0
	$V \to A$	**0.234**	0.151	**0.152**	0.071	**48.8**
Video-Bias	$A \to V$	0.098	**0.114**	0.079	**0.129**	**50.0**
	$V \to A$	0.098	**0.114**	**0.126**	0.089	45.0

Table 5: **Temporal reasoning analysis.**

Organization	Local Attn.	Global Attn.	Acc.
Sequential	0.00010	0.00019	45.4
Interleaved	**0.00033**	0.00018	**48.5**

Table 6: **Temporal hallucination analysis.**

Organization	Local Attn.	Global Attn.	Acc.
Sequential	0.00021	0.00020	**75.0**
Interleaved	**0.00074**	0.00021	67.9

positional preference consistently emerges across Transformer layers. We additionally visualize the corresponding attended regions before and after reversing modality order, and observe that the model's focus shifts (Figure 2(b)). These results suggest that modality accessibility is not solely determined by semantic importance, but also influenced by positional organization. Aligning the modality containing more relevant evidence with closer relative positions may enable more effective evidence utilization, leading to improved reasoning performance.

Temporal Reasoning For temporal reasoning, we observe that Interleaved organization achieves the best performance, as shown in Table 1. Following the analysis above, the strength of cross-modal interaction between an audio chunk A_i and its temporally corresponding visual chunk V_i is subject to the distance-dependent upper bound $B\left(|n-m|\right)$. Therefore, the temporal organization directly determines the positional constraint imposed on their interaction. Consider T audio and visual chunks. Under the Sequential organization $A_1 \cdots A_T V_1 \cdots V_T$, the corresponding chunks satisfy $\Delta p_i^{\text{seq}} = p_{V_i} - p_{A_i} = T$, whereas under the Interleaved organization $A_1 V_1 \cdots A_T V_T$, $\Delta p_i^{\text{int}} = p_{V_i} - p_{A_i} = 1$. Thus,

$$|\Delta p_i^{\text{int}}| < |\Delta p_i^{\text{seq}}| \quad \Rightarrow \quad |B_{\text{int}}(1)| > |B_{\text{seq}}(T)|.$$

Hence, Interleaved organization may impose a less restrictive positional constraint on the temporally corresponding audio-visual chunks, allowing the model to allocate stronger attention to the local cross-modal evidence that is important for temporal reasoning.

To verify this hypothesis, we compare temporal attention patterns between Interleaved and Sequential organizations. Specifically, we measure the attention between temporally adjacent audio-video chunks. As shown in Table 5, Interleaved organization consistently exhibits higher local temporal attention than Sequential organization, which aligns with its superior performance on temporal reasoning samples. However, stronger local attention is not always beneficial. For temporally hallucination samples in Table 6, we observe that Interleaved still maintains higher local temporal attention but achieves lower performance than grouped organization, suggesting that excessive locality may introduce misleading temporal correlations. This suggests that the locality bias introduced by RoPE acts as a task-dependent inductive bias: it benefits fine-grained temporal alignment but may cause over-reliance on local evidence and impair global temporal reasoning.

Please note that the Finding and experimental subsets serve different purposes and are constructed independently: the former characterizes topology effects under specific reasoning scenarios, and the latter evaluates CoM's ability to identify effective modality organizations.

4 Chain-of-Modality: Topology-Controlled Dynamic Inference

Our empirical analysis reveals that modality organization serves as an important factor in Omni-MLLM inference: different reasoning requirements favor different modality topologies, and no single organization consistently achieves optimal performance. These findings motivate a natural question: can modality organization be dynamically controlled according to the underlying reasoning requirements of each input? To answer this question, we introduce **Chain-of-Modality (CoM)** (Figure 3), a controllable inference framework that adaptively reorganizes modality interactions without modifying the underlying Omni-MLLM parameters. CoM consists of three functional components: **Planner**, **Reasoner**, and **Decider**. Given a query with corresponding audio-visual inputs, the Planner selects an appropriate modality organization among *Single*, *Parallel*, *Sequential*, and *Interleaved*. The Reasoner performs topology-guided inference using the original Omni-MLLM

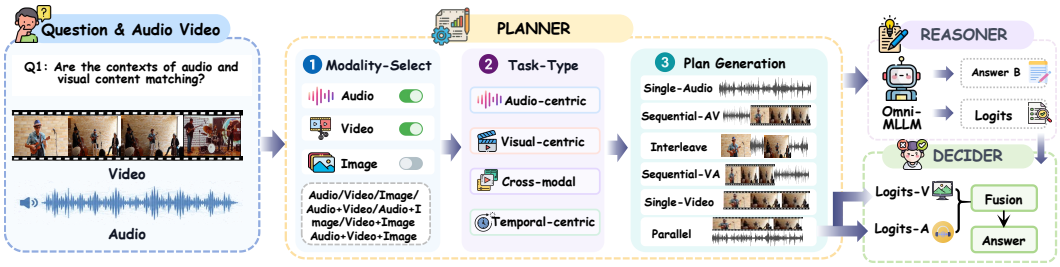


Figure 3: **Architecture of the CoM framework.** CoM is a topology-controlled inference framework, which consists of Planner, Reasoner, and Decider.

backbone. For Parallel organization, where modality-specific predictions are independently obtained, the Decider aggregates option-level logits to produce the final prediction. Formally, let f_θ denote the original Omni-MLLM backbone and f_ϕ denote the Planner. Different functional stages are instantiated through specialized system prompts, including p_{plan} for topology selection and p_{reason} for topology-guided reasoning.

4.1 Adaptive Modality Planning

The Planner explicitly models modality organization as an inference decision. Given a query $\boldsymbol{Q}$ and available modality set $\mathcal{M} = \{\boldsymbol{A}, \boldsymbol{V}, \boldsymbol{I}\}$, where $\boldsymbol{A}$, $\boldsymbol{V}$, and $\boldsymbol{I}$ denote audio, video, and image modalities, respectively, the Planner selects a task-dependent modality subset and organization strategy according to query requirements and available sensory evidence:

$$f_\phi(\boldsymbol{Q}, \mathcal{M}, p_{plan}) \rightarrow (t_{think}, \mathcal{M}_{min}, \mathcal{F}), \tag{1}$$

where t_{think} denotes the Planner reasoning output, $\mathcal{M}_{min}$ represents the selected modality subset, and $\mathcal{F} \in \{\text{Single}, \text{Parallel}, \text{Sequential}, \text{Interleaved}\}$ denotes the selected modality organization strategy. The selected organization determines how modality information is presented to the Reasoner. Specifically, interleaved organization constructs fine-grained cross-modal sequences, such as alternating modality representations $(\boldsymbol{A}_1, \boldsymbol{V}_1, \boldsymbol{A}_2, \boldsymbol{V}_2, \dots)$, and sequential organization arranges selected modalities according to explicit orders, such as $\boldsymbol{A} \rightarrow \boldsymbol{V}$ or $\boldsymbol{V} \rightarrow \boldsymbol{A}$. Parallel organization processes modalities independently and is suitable when modality-specific evidence can be aggregated without explicit cross-modal ordering. Single organization is selected when $\mathcal{M}_{min}$ contains only one modality, in which case the Reasoner performs unimodal inference directly following Equation 2. By separating modality organization from multimodal reasoning, CoM enables explicit investigation of how different organization strategies influence inference behavior and whether effective modality organizations can be discovered through learning.

4.2 Modality-Organized Reasoning and Decision Fusion

After modality organization is determined by the Planner, the Reasoner performs inference according to the selected strategy. For sequential or interleaved organizations selected by $\mathcal{F}$, the Reasoner directly receives the organized modality input and generates the final prediction:

$$y = f_\theta(\mathcal{O}, \boldsymbol{Q}, p_{reason}), \tag{2}$$

where $\mathcal{O}$ denotes the organized modality input constructed according to the Planner output. For parallel organization, $\mathcal{M}_{min}$ is restricted to a pair of modalities, and CoM independently constructs two modality-specific branches $y_i = f_\theta(\mathcal{O}_i, \boldsymbol{Q}, p_{reason}), \quad i = 1, 2$, where each branch corresponds to an individual modality-specific input. When different branches produce consistent predictions, CoM directly adopts the shared prediction. Otherwise, the Decider resolves the disagreement by averaging the modality-specific prediction distributions:

$$p(y|\boldsymbol{Q}, \mathcal{M}) = \frac{1}{2} \sum_i p_i(y|\boldsymbol{Q}, \mathcal{O}_i, p_{reason}), \tag{3}$$

where $p_i(y|\boldsymbol{Q}, \mathcal{O}_i, p_{reason})$ denotes the prediction distribution from the i-th modality branch. This design retains the Omni-MLLM backbone for reasoning while enabling explicit control over modality organization and aggregating heterogeneous modality evidence.

Table 7: **Comprehensive performance comparison on our defined subset and full benchmark.**

Method	Music-AVQA			AVHBench		OmniBench	WorldSense	DailyOmni
	A	V	A-V	A-Hal	V-Hal			
Topology-sensitive Subset								
EchoInk	37.5	29.4	32.2	49.2	19.0	25.4	30.1	34.1
Omni-R1	32.5	19.6	32.2	30.9	21.9	30.6	34.7	37.6
AVATAR	29.0	18.2	28.9	30.9	21.8	31.0	34.4	36.5
Qwen-Omni-7B	31.5	21.0	28.5	16.2	25.9	23.4	27.8	29.0
+CoM	**35.8**	**25.8**	**30.3**	**40.4**	**32.7**	**29.0**	**37.3**	**29.7**
+CoM(SFT)	**39.5**	**40.0**	**39.7**	**51.7**	**49.2**	**35.9**	**42.6**	**38.3**
Full Benchmark								
Qwen-Omni-7B	72.6	74.3	57.8	73.2	67.1	49.1	**41.4**	55.6
+CoM(SFT)	**72.8**	**74.6**	**58.3**	**76.5**	**73.3**	**50.3**	41.3	**57.0**

5 EXPERIMENT

Evaluation Benchmarks. We evaluate our framework on five representative benchmarks: **Music-AVQA** Yang et al. (2022), **OmniBench** Li et al. (2024), **DailyOmni** Zhou et al. (2025), **WorldSense** Hong et al. (2025), and **AVHBench** Sung-Bin et al. (2024).

Implementation Details Our evaluation follows a two-stage paradigm: **(1)** We first evaluate a training-free setting across three representative Omni-MLLMs. **(2)** To investigate whether effective modality organization strategies can be learned rather than manually specified, we further train a lightweight Planner based on Qwen-Omni-7B using supervised fine-tuning. The trajectories are automatically generated by Qwen-Omni-30B Xu et al. (2025b) from OmniVideo-100K Cai et al. (2026), resulting in 5,791 modality planning samples. T We fine-tune only the Planner with LoRA for one epoch using a learning rate of 5×10^{-5}. This setting evaluates whether modality organization policies can be learned from external supervision and generalized to unseen tasks and benchmarks. For all experiments, we set the decoding temperature to $T = 0$ and fix the random seed to ensure deterministic evaluation and minimize stochastic variations.

5.1 MAIN RESULTS

To systematically analyze how modality organization affects different reasoning behaviors, we construct topology-sensitive subsets: (1) samples where unimodal inputs achieve correct predictions while multimodal integration leads to degraded performance, and (2) samples where identical multimodal content produces different predictions under different modality organizations. *These subsets are designed as a diagnostic evaluation to isolate topology-induced errors and quantify the extent to which our method can correct them. As our goal is not to improve general Omni-MLLM accuracy, but to establish that topology-induced failures are systematically learnable.* We first evaluate the training-free CoM strategy on these topology-sensitive subsets across five representative multimodal benchmarks. The vanilla baseline follows the A-V organization for consistency, as OmniBench Li et al. (2024) consists of audio-image tasks.

As shown in Table 7, training-free CoM consistently improves over the vanilla Qwen-Omni-7B baseline across different benchmarks and reasoning scenarios. These results indicate that multimodal reasoning failures are not solely caused by insufficient perception capability, but can also arise from suboptimal organization of available evidence. To investigate whether effective modality organization strategies can be learned, we further introduce a lightweight Planner-SFT stage. As shown in Table 7, CoM(SFT) further improves over the training-free strategy on most benchmarks, demonstrating that modality organization policies can be learned through external supervision and transferred to downstream multimodal reasoning tasks. We further compare CoM with existing omni-modal models built upon Qwen-Omni-7B Kulkarni & Fazli (2025); Xing et al. (2025); Zhong et al. (2025). We re-evaluate all models on the same topology-sensitive subsets using the same prompts and identical evaluation settings and CoM achieves competitive performance. Finally, we report results on the full benchmarks. CoM achieves comparable or better performance than the vanilla Qwen-Omni-7B, with gains comparable in magnitude to those reported by Omni-MLLMs trained with GRPO Kulkarni & Fazli (2025). CoM achieves these gains through inference-time modality organization without updating the underlying Omni-MLLM. The lack of gain on WorldSense may stem from its relatively lower sensitivity to modality organization at the full-benchmark level. The gap between unimodal and multimodal organizations is only around 3 points, compared with nearly 20 points on DailyOmni and

Table 8: **Generalization of CoM across different Omni-MLLM architectures and scales on topology-sensitive subset.**

Method	WorldSense	OmniBench	DailyOmni
Topology-sensitive Subset			
Omni-3B	28.3	27.4	30.2
+CoM	31.2	31.4	30.3
+7B Planner	**32.9**	**40.8**	**38.9**
Ola-7B	13.6	15.2	11.2
+CoM	13.3	17.8	16.8
+7B Planner	**28.5**	**32.2**	**22.2**

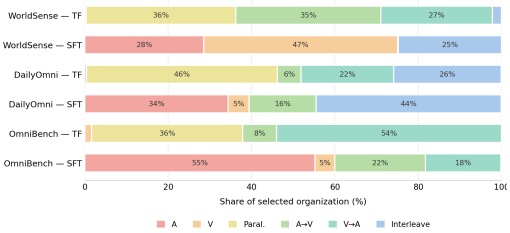


Figure 4: **Distribution of selected modality organizations across benchmarks under training-free (TF) and Planner-SFT (SFT) CoM.**

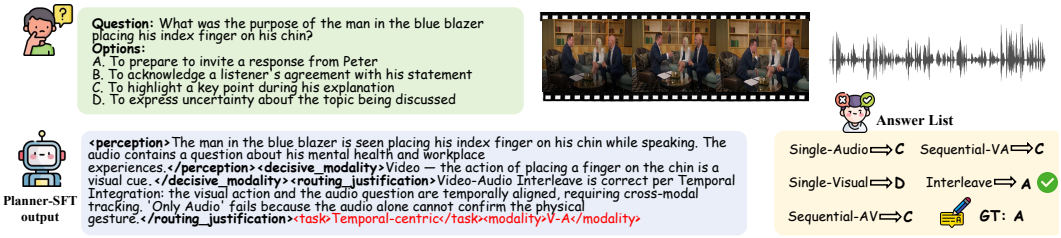


Figure 5: **Qualitative comparison of Planner-SFT and fixed modality organizations.** The results show that Planner-SFT can select task-adaptive modality sequences and improve multimodal reasoning.

OmniBench, leaving less room for topology optimization. Detailed qualitative examples of modality planning decisions and reasoning behaviors are provided in Figure 5.

5.2 Cross-Model Generalization Analysis

To investigate whether the effectiveness of CoM depends on a specific model architecture or scale, we evaluate our framework on diverse Omni-MLLMs, including Qwen-Omni-3B Xu et al. (2025a) and Ola-7B Liu et al. (2025), which follow different Omni-MLLM architectures. As shown in Table 8, training-free CoM improves the performance of both Qwen-Omni-3B and Ola-7B on the majority of evaluated benchmarks. These results demonstrate that the effectiveness of explicit modality organization is not restricted to a specific backbone architecture or model scale. We further investigate whether the learned planning strategy can generalize across different Omni-MLLMs. Specifically, we transfer the Planner trained on Qwen-Omni-7B to other Omni-MLLMs while keeping their reasoning backbones unchanged. As shown in Table 8, the transferred Planner further improves performance across multiple models and benchmarks, suggesting that the learned modality organization policy generalizes beyond the source model rather than capturing model-specific behaviors (Appendix B.2).

5.3 Adaptive Modality Organization Analysis

Beyond overall performance improvements, we further analyze how CoM selects modality organizations under different settings. Figure 4 summarizes the organization distributions selected by training-free CoM and Planner-SFT CoM. For the training-free setting, the selected organizations vary substantially across benchmarks, suggesting that modality organization preferences vary with sample characteristics. After Planner-SFT, the organization distributions exhibit clearer sample-dependent preferences. Rather than maintaining a uniform routing strategy, the learned Planner shifts toward different organization patterns across samples. These observations demonstrate that CoM performs adaptive modality orchestration rather than fixed topology selection.

6 Conclusion

In this work, we reveal modality organization as an overlooked factor influencing Omni-MLLM inference. Our analysis shows that different reasoning scenarios favor different modality topologies, indicating that no single organization is universally optimal. We introduce Chain of Modality (CoM), which enables adaptive modality organization through inference-time planning. Experiments across benchmarks, architectures, and model scales establish that topology-induced failures are not merely stochastic artifacts, but exhibit systematic patterns that can be partially controlled.

AI Use Statement In this work, we used generative AI tools for aid or polish of writing, retrieval and discovery of related work, research ideation, code generation. We have reviewed all AI-assisted work and verified its accuracy and relevance. We take responsibility for the final content of this work, including all text, claims, experimental designs, results, and conclusions.

Ethics statement This work does not involve human subjects, personally identifiable information, or the collection of new data involving human subjects. We use publicly available datasets and pre-trained models in accordance with their respective licenses and intended usage. Our method focuses on analyzing and mitigating modality-organization failures in omni-modal large language models and does not introduce new capabilities for harmful or high-risk applications. The results reported in this work should not be interpreted as guarantees of factual correctness or safety in real-world deployment. We have no financial or other conflicts of interest related to this work.

Reproducibility statement We provide an anonymous implementation of our method at `https://anonymous.4open.science/r/anonymization_iclr-4F69buffcc/`. The main paper describes the proposed Chain of Modality (CoM) framework and inference procedures, and the appendix provides additional implementation details, experimental settings, and analyses. The anonymous repository includes the training and evaluation configurations, prompts, and relevant experimental scripts to facilitate reproduction of our results. All experiments are conducted using publicly available datasets and models, with the corresponding data processing and evaluation protocols described in the paper and supplementary materials.

REFERENCES

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025.

Xinyue Cai, Chaoyou Fu, Yi-Fan Zhang, Ran He, and Caifeng Shan. Omnivideo-100k: A dataset for audio-visual reasoning through structured scripts and evidence chains. *arXiv preprint arXiv:2606.14702*, 2026.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Chao Sun, Rongzhou Zhang, Guanyu Zhou, Lijie Wen, and Xuming Hu. Omnidpo: A preference optimization framework to address omni-modal hallucination. *arXiv preprint arXiv:2509.00723*, 2025.

Zhangquan Chen, Jiale Tao, Ruihuang Li, Yihao Hu, Ruitao Chen, Zhantao Yang, Xinlei Yu, Haodong Jing, Manyuan Zhang, Shuai Shao, et al. Omnivideo-r1: Reinforcing audio-visual reasoning with query intention and modality attention. *arXiv preprint arXiv:2602.05847*, 2026.

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

Henghui Du, Guangyao Li, Chang Zhou, Chunjie Zhang, Alan Zhao, and Di Hu. Crab: A unified audio-visual scene understanding model with explicit cooperation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18804–18814, 2025.

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

Yiran Guan, Sifan Tu, Dingkang Liang, Linghao Zhu, Jianzhong Ju, Zhenbo Luo, Jian Luan, Yuliang Liu, and Xiang Bai. Thinkomni: Lifting textual reasoning to omni-modal scenarios via guidance decoding. *arXiv preprint arXiv:2602.23306*, 2026.

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

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

Hongyeob Kim, Inyoung Jung, Dayoon Suh, Youjia Zhang, Sangmin Lee, and Sungeun Hong. Question-aware gaussian experts for audio-visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13681–13690, 2025.

Yogesh Kulkarni and Pooyan Fazli. Avatar: Reinforcement learning to see, hear, and reason over video. *arXiv preprint arXiv:2508.03100*, 2025.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *arXiv preprint arXiv:2410.12787*, 2024.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19108–19118, 2022.

Xiaoxi Li, Wenxiang Jiao, Jiarui Jin, Shijian Wang, Guanting Dong, Jiajie Jin, Hao Wang, Yinuo Wang, Ji-Rong Wen, Yuan Lu, et al. Omnigaia: Towards native omni-modal ai agents. *arXiv preprint arXiv:2602.22897*, 2026.

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

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

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model. *arXiv preprint arXiv:2502.04328*, 2025.

Lidong Lu, Guo Chen, Zhiqi Li, Yicheng Liu, and Tong Lu. Av-reasoner: Improving and benchmarking clue-grounded audio-visual counting for mllms. *arXiv preprint arXiv:2506.05328*, 2025.

Akshay Paruchuri, Sanmi Koyejo, and Ehsan Adeli. Same evidence, different answer: Auditing order sensitivity in multimodal large language models. *arXiv preprint arXiv:2606.26079*, 2026.

Patrick Schilcher, Dominik Karasin, Michael Schöpf, Haisam Saleh, Antonela Tommasel, and Markus Schedl. Characterizing positional bias in large language models: A multi-model evaluation of prompt order effects. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20643–20664, 2025.

Ramaneswaran Selvakumar, Kaousheik Jayakumar, S Sakshi, Sreyan Ghosh, Ruohan Gao, and Dinesh Manocha. Do audio-visual large language models really see and hear? *arXiv preprint arXiv:2604.02605*, 2026.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

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

Zhijie Tan, Xu Chu, Weiping Li, and Tong Mo. Order matters: Exploring order sensitivity in multimodal large language models. *arXiv preprint arXiv:2410.16983*, 2024.

Keda Tao, Wenjie Du, Bohan Yu, Weiqiang Wang, Jian Liu, and Huan Wang. Omniagent: Audio-guided active perception agent for omnimodal audio-video understanding. *arXiv e-prints*, pp. arXiv–2512, 2025a.

Keda Tao, Kele Shao, Bohan Yu, Weiqiang Wang, Huan Wang, et al. Omnizip: Audio-guided dynamic token compression for fast omnimodal large language models. *arXiv preprint arXiv:2511.14582*, 2025b.

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

Zhenghao Xing, Xiaowei Hu, Chi-Wing Fu, Wenhai Wang, Jifeng Dai, and Pheng-Ann Heng. Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning. *arXiv preprint arXiv:2505.04623*, 2025.

Huangchen Xu, Yuan Wu, and Yi Chang. A systematic evaluation of positional bias in multi-video summarization with mllms. *arXiv preprint arXiv:2606.04596*, 2026.

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

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

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 3480–3491, 2022.

Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In *European Conference on Computer Vision*, pp. 146–164. Springer, 2024.

Qilang Ye, Zitong Yu, Rui Shao, Yawen Cui, Xiangui Kang, Xin Liu, Philip Torr, and Xiaochun Cao. Cat+: Investigating and enhancing audio-visual understanding in large language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Qilang Ye, Wei Zeng, Meng Liu, Jie Zhang, Yupeng Hu, Zitong Yu, and Yu Zhou. When eyes and ears disagree: Can mllms discern audio-visual confusion? *arXiv preprint arXiv:2511.10059*, 2025b.

Xu Zheng, Chenfei Liao, Yuqian Fu, Kaiyu Lei, Yuanhuiyi Lyu, Lutao Jiang, Bin Ren, Jialei Chen, Jiawen Wang, Chengxin Li, et al. Mllms are deeply affected by modality bias. *arXiv preprint arXiv:2505.18657*, 2025.

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

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

A Appendix

B Dataset Construction and Statistics

To analyze modality organization failures in a controlled manner, we construct topology-sensitive subsets according to model predictions under different modality organizations. Given a sample x_i with ground-truth answer g_i, we denote the prediction under organization o as $y_i^o = f(x_i, o)$. We consider five representative organizations: unimodal audio (A), unimodal visual (V), sequential audio-visual fusion ($A \to V$), sequential visual-audio fusion ($V \to A$), and interleaved fusion (I). The multimodal organization set is defined as:

$$\mathcal{O}_m = \{A \to V, V \to A, I\}.$$

We consider two representative failure patterns.

First, *modality integration failure* captures cases where unimodal evidence is sufficient for solving the task, but multimodal organizations fail to effectively utilize such evidence. Specifically, a sample is selected when at least one unimodal organization produces the correct prediction, while all multimodal organizations produce incorrect predictions:

$$S_i^{int} = \mathbb{1}\left[(y_i^A = g_i \vee y_i^V = g_i) \wedge \bigwedge_{o \in \mathcal{O}_m} (y_i^o \neq g_i) \right].$$

Second, *topology-induced prediction variance* captures cases where the same multimodal evidence leads to different predictions under different organizations. Specifically, a sample is selected when multimodal organizations produce inconsistent predictions and at least one organization obtains the correct answer:

$$S_i^{topo} = \mathbb{1}\left[\exists o_1, o_2 \in \mathcal{O}_m, y_i^{o_1} \neq y_i^{o_2} \wedge \exists o \in \mathcal{O}_m, y_i^o = g_i \right].$$

Finally, the topology-sensitive subset is constructed by combining the two failure patterns:

$$\mathcal{D}_{topo} = \{x_i \mid S_i^{int} = 1 \vee S_i^{topo} = 1\}.$$

B.1 Topology-sensitive subset distribution

We further characterize the constructed subsets in terms of their prevalence and prediction consistency from Qwen-Omni-7B. Table 9 reports the proportion of topology-sensitive samples in the full datasets of DailyOmni, OmniBench, and WorldSense, together with the relative contribution of the two failure patterns. Overall, the selected samples constitute a non-negligible portion of each benchmark, indicating that topology-related failures are not isolated cases.

To further analyze the composition of the constructed subsets, we report the prediction accuracy under different modality organizations. Specifically, we evaluate unimodal audio (A), unimodal visual (V), sequential audio-visual fusion (A→V), sequential visual-audio fusion (V→A), and interleaved organizations. As shown in Table 10, the selected samples exhibit diverse prediction patterns across different organizations rather than concentrating on a specific modality order. The relatively balanced accuracy distributions indicate that our subset construction does not introduce a preference toward a particular organization, but instead captures diverse cases where modality utilization varies under different structural arrangements.

Table 9: **Statistics of topology-sensitive subsets across different benchmarks.**

Dataset	Full	S^{int}	S^{topo}	Total
DailyOmni	1179	97	182	279
WorldSense	3172	305	371	676
OmniBench	1142	114	134	248

Table 10: **Prediction accuracy of different modality organizations on topology-sensitive subsets.**

Dataset	A	V	A→V	V→A	Interleave
DailyOmni	44.4	41.9	38.7	27.6	29.0
WorldSense	51.0	43.3	28.4	25.0	27.8
OmniBench	43.2	0.0	23.4	0.0	-

B.2 MODEL CONSISTENCY ANALYSIS

To understand why a Planner trained on Qwen-Omni-7B Xu et al. (2025a) can generalize to other Omni-MLLMs, we investigate whether different models share effective modality organizations on the same reasoning cases. For a sample x, we define the set of effective organizations for model M as

$$\mathcal{O}_M(x) = \{o \in \mathcal{O} \mid f_M(x; o) = y^*\},$$

where $\mathcal{O} = \{A, V, A \to V, V \to A\}$ denotes the candidate modality organizations considered for these models, as Ola-7B does not support interleaved organization. We regard a sample as cross-model shared if

$$\mathcal{O}_{M_1}(x) \cap \mathcal{O}_{M_2}(x) \neq \emptyset,$$

i.e., the two models admit at least one common organization that produces the correct answer. We then measure the proportion of such shared samples over the entire dataset. This metric captures whether different models admit a common effective organization for the same reasoning case, rather than requiring them to have a unique or identical preferred organization. A high overlap as shown in Table 11 therefore suggests that the effective organization required to solve a reasoning case is shared across models, providing a possible explanation for the cross-model generalization of the Planner trained on Qwen-Omni-7B.

Table 11: **Overlap of Effective Modality Organizations across Different Omni-MLLMs.**

Model Pair	Best Topology Overlap
DailyOmni	
Qwen-Omni-7B vs Qwen-Omni-3B	97.19
Qwen-Omni-7B vs Ola-7B	90.05
Qwen-Omni-3B vs Ola-7B	89.45
OmniBench	
Qwen-Omni-7B vs Qwen-Omni-3B	94.67
Qwen-Omni-7B vs Ola-7B	93.72
Qwen-Omni-3B vs Ola-7B	93.90

C VALIDATION OF LEARNED MODALITY ORGANIZATION POLICY

To verify whether Planner-SFT learns meaningful modality organization policies rather than obtaining improvements from random topology exploration, we compare CoM(SFT) with random organization selection. Table 12 reports the organization selection accuracy under different strategies. Compared with random selection, CoM(SFT) consistently achieves higher accuracy across all evaluated benchmarks, improving from 31.3% to 42.6% on WorldSense, from 33.0% to 38.3% on DailyOmni, and from 26.2% to 35.9% on OmniBench. Interestingly, random organization selection can outperform the original Omni-MLLM with a fixed topology, since random exploration has a chance to select more suitable organizations for individual samples. However, CoM(SFT) further improves over random selection, demonstrating that the Planner learns sample-dependent organization preferences rather than relying on stochastic topology exploration.

Table 12: **Organization selection accuracy under different topology selection strategies.**

DataSet	WorldSense	DailyOmni	OmniBench
Qwen-Omni-7B	27.8	29.0	23.4
Random	31.3	33.0	26.2
CoM(SFT)	42.6	38.3	35.9

D MORE MODALITY ORGANIZATION EVALUATION ON FULL BENCHMARK

We further compare CoM(SFT) with random organization selection on full benchmark. Although random selection provides a simple exploration strategy and may occasionally choose more suitable

modality organizations for individual samples, it lacks the ability to adapt organization choices according to sample characteristics. In contrast, CoM(SFT) consistently achieves better or comparable performance, indicating that the improvements mainly arise from the learned sample-dependent modality organization policy rather than stochastic topology exploration. Compared with the strong prompt baseline, CoM(SFT) achieves additional gains without modifying the underlying Omni-MLLM parameters. These results suggest that explicitly modeling and controlling modality organization provides complementary benefits beyond conventional prompt-based reasoning enhancement. For heuristic routing, we construct a rule-based baseline that maps modality organization according to question keywords: audio-related queries are assigned to Audio-only organization, vision-related queries to Video-only organization, and queries involving both modalities to sequential Audio-Video organization. Compared with heuristic routing, the learned Planner considers multimodal requirements beyond explicit query semantics and enables more adaptive organization selection.

Table 13: **Full benchmark evaluation of different modality organization strategies.**

DataSet	WorldSense	DailyOmni	OmniBench
Qwen-Omni-7B	41.4	55.6	49.1
Strong Prompt	34.3	48.9	43.8
Random Select	38.5	49.7	39.2
Heuristic Routing	37.0	51.1	44.3
CoM(SFT)	41.3	**57.0**	**50.3**

E Planner Evaluation and Routing Quality

We further analyze the reliability of the learned Planner by reporting four metrics: answer accuracy (ACC), oracle planning performance (OP), invalid topology outputs (ITO), and fallback rates (FR).

Answer accuracy (ACC) measures the final task performance achieved by CoM after applying the organization predicted by the Planner. Since the Reasoner remains unchanged, the improvement of CoM mainly depends on whether the Planner can identify an effective modality organization for each sample. To characterize the potential upper bound of topology selection, we report Oracle Planning (OP), which assumes that the optimal organization among all candidate topologies can be selected for each sample. Specifically, OP represents the maximum achievable accuracy when the most effective modality organization is known in advance. We additionally evaluate the execution reliability of the Planner using invalid topology output rate (ITO). ITO measures the proportion of cases where the Planner generates organizations outside the predefined candidate topology space. In our implementation, invalid topology outputs directly trigger the fallback mechanism, where CoM resorts to the default inference strategy. Therefore, the invalid topology output rate is equivalent to the fallback rate.

These metrics provide complementary views of the learned organization policy: ACC reflects the practical performance, OP indicates the remaining optimization potential, and ITO/FR quantify the robustness of topology generation.

Table 14: **Evaluation of Planner performance under training-free and Planner-SFT settings.**

DataSet	WorldSense	DailyOmni	OmniBench
Training-free CoM			
ACC	37.3	29.7	29.0
OP	100	100	100
FR	0.0	0.0	0.0
Planner-SFT CoM			
ACC	42.6	38.3	30.5
FR	0.0	0.0	0.0

F PROMPT OF TRAINING-FREE COM

We provide the specific prompt template utilized by the Planner in training-free pattern, which is detailed below:

Listing 1: Prompt template for Planner stage.

```
"You are a master multimodal strategist. Create an optimal execution
    ↪ plan.\n\n"

    "**Available Modalities:** {modalities}\n\n"

    "**QUESTION TYPE CLASSIFICATION:**\n"
    "- **Audio-centric:** Pure audio comparison. Example: 'Is X louder
        ↪ than Y?'\n"
    "  Best route: Audio only.\n\n"
    "- **Visual-centric (static):** Static objects/properties captured
        ↪ in a single frame. Example: 'How many instruments?', 'What
        ↪ color?'\n"
    "  Best route: Image only.\n\n"
    "- **Temporal-centric (dynamic video):** Timing, sequence, duration,
        ↪ action tracking. Example: 'Is X always playing?'\n"
    "  Best route: Video (may also need Audio for cross-modal sync).\n\n"
    "- **Cross-modal:** Audio-visual sync/mapping. Example: 'Which
        ↪ instrument makes the sound?'\n"
    "  Best route: Both modalities together.\n\n"

    "**CRITICAL COGNITIVE INSIGHT:**\n"
    "Omni-modal models often suffer from static visual bias (treating
        ↪ video sequences merely as static frames). "
    "For dynamic video inputs, single-frame snapshots LOSE all critical
        ↪ motion, temporal transitions, and synchronization. "
    "When Video is present, prefer interleaved Audio+Video fusion over
        ↪ parallel independent analysis.\n\n"

    "**AVAILABLE PLAN TYPES:**\n"
    "Choose the EXECUTION_FORMAT for each plan:\n\n"

    "--- PARALLEL (Independent) ---\n"
    "  Each modality is processed INDEPENDENTLY (cannot see the other). "
    "Each modality alone produces its own answer/logprobs, then results
        ↪ are fused.\n"
    "  Format: Numbered steps, each with a SINGLE modality.\n"
    "  YOU CHOOSE THE ORDER, whichever modality is more important goes
        ↪ first.\n\n"
    "  Examples:\n"
    "    Audio-first: '1) Audio - Why: Audio alone can detect the
        ↪ sound\\n"
    "                 2) Image - Why: Image alone can confirm
        ↪ visually'\n"
    "    Image-first: '1) Image - Why: Visual alone provides key info\\n"
    "                 2) Audio - Why: Audio alone supplements'\n\n"

    "--- SEQUENTIAL (Interleaved / Fused, use '+') ---\n"
    "  Modalities are FUSED in one input, they CAN see and interact with
        ↪ each other. "
    "Best when the answer depends on cross-modal interaction.\n"
    "  Format: One step with '+' between fused modalities.\n\n"
    "  Examples:\n"
    "    '1) Audio+Image - Why: Must compare sound with visual
        ↪ simultaneously'\n"
    "    '1) Image+Audio - Why: Visual context guides audio
        ↪ interpretation'\n\n"

    "--- SINGLE (One modality only) ---\n"
```

```
    "  Only one modality is sufficient. Ignore the others.\n\n"
    "  Examples:\n"
    "    '1) Audio - Why: Question can be answered from audio alone'\n"
    "    '1) Image - Why: Question can be answered from visual content
        ↪ alone'\n\n"

  "**OUTPUT FORMAT:**\n"
   "<question_analysis>\n"
   "    <question_type>[Audio-centric | Visual-centric |
       ↪ Temporal-centric | Cross-modal]</question_type>\n"
   "    <core_task>[What exactly needs to be answered?]</core_task>\n"
   "    <strategy>[SEQUENTIAL|PARALLEL]</strategy>\n"
   "</question_analysis>\n\n"

   "<execution_plan>"
   "For PARALLEL strategy (independent modalities):"
   "1) [Modality1] - Why: [reason why this modality alone/first can
       ↪ help answer]"
   "2) [Modality2] - Why: [reason]"
   "3) [Modality3] - Why: [reason]"
   "[Only list NECESSARY modalities in order of importance/certainty]"

   "For SEQUENTIAL strategy (modalities must be fused):"
   "1) [Modality1]+[Modality2] - Why: [reason why they MUST be fused in
       ↪ this order]"
   "2) [Modality3] - Why: [reason]"
   "[Only list NECESSARY combinations. Order by: most critical modality
       ↪ first]"
   "</execution_plan>"

    "**YOUR TASK:**\n"
   "Question: {question}\n\n"
```

G Attention Analysis under Different Modality Organizations

To better understand how Chain of Modality (CoM) affects multimodal reasoning, we analyze the change of cross-modal attention patterns before and after modality reorganization. Specifically, we focus on samples where the original modality organization produces incorrect predictions, while CoM successfully recovers the correct answers. For each sample, we compare the attention distributions under the original topology and the topology selected by CoM, while keeping the underlying audio-visual evidence unchanged.

As illustrated in Figure 6, different modality organizations induce distinct cross-modal interaction patterns. Under the original topology, the model exhibits topology-dependent attention behaviors, where the information exchange between modalities varies according to their structural arrangement. These patterns are consistent with our previous findings that modality organization affects evidence accessibility and cross-modal integration.

After applying CoM, the reorganized modality topology leads to substantial changes in cross-modal attention distribution. In particular, CoM modifies both the direction and strength of modality interactions, enabling different information pathways compared with the original organization. For example, certain cross-modal interactions are suppressed while others are enhanced after topology reorganization, suggesting that modality organization serves as an implicit mechanism controlling how multimodal information is propagated inside the model.

Although attention weights alone do not provide a complete explanation of model decisions, the consistent attention redistribution observed in successful CoM cases demonstrates that topology reorganization can alter the internal information flow of Omni-MLLMs. These observations provide additional evidence that CoM improves multimodal reasoning not by modifying the underlying model parameters, but by controlling the structural pathway through which modality information is accessed and integrated during inference.

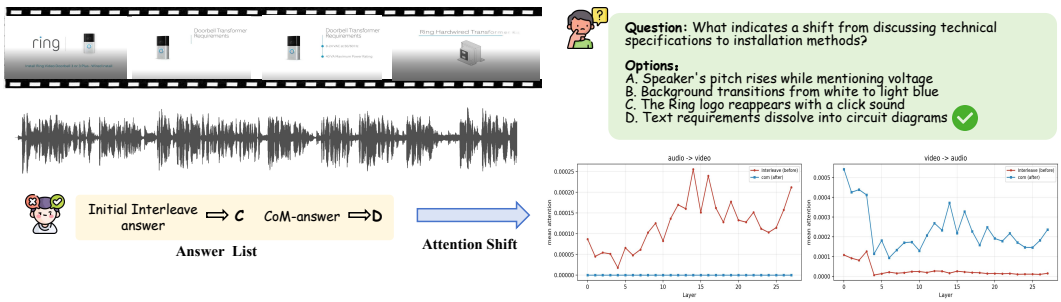


Figure 6: **Cross-modal attention redistribution induced by modality reorganization.**CoM modifies the interaction pathways between modalities without changing the underlying model parameters.

H Construction of Planner-SFT Training Trajectories

To train the Planner in CoM, we construct organization-level supervision trajectories from existing Omni-MLLM inference behaviors. Different from conventional instruction tuning, where the model learns to generate answers directly, Planner-SFT aims to learn which modality organization should be selected before reasoning. Therefore, the supervision focuses on the relationship between input modality topology and the resulting reasoning effectiveness.

Given a video-question pair x_i with ground-truth answer g_i from OmniVideo-100K Cai et al. (2026), we first employ Qwen-Omni-7B as the frozen Reasoner to perform inference under different candidate modality organizations: $\mathcal{O} = \{A, V, A \rightarrow V, V \rightarrow A, I\}$. For each organization $o \in \mathcal{O}$, the Reasoner produces a prediction $y_i^o = f(x_i, o)$. The effective organizations are identified according to whether they produce the correct prediction $\mathcal{O}_i^+ = \{o \in \mathcal{O} | y_i^o = g_i\}$.Samples without any effective organization are discarded, since they cannot provide reliable supervision for learning modality organization.

To generate such rationales, we employ Qwen-Omni-30BTeam (2026) as an offline annotator. The annotator receives the video-question pair, the predictions obtained from different modality organizations, and the corresponding correctness information. It is prompted to analyze why the selected organization successfully utilizes the available evidence and why alternative organizations fail under the same multimodal input. We provide the specific prompt template utilized by Qwen-Omni-30B, which is detailed below:

Listing 2: Prompt template for Planner stage.

```
"You are an expert in Omni-Multimodal Large Language Models and a "
"Senior Causal Logic Analyst."
"Your task is to inspect the raw audio and video/image inputs, then "
"justify why my provided optimal modality routing is correct
    ↪ compared "
"to ONE representative incorrect alternative."
""
"# Routing Principles"
"1) Modality Pruning: If a single modality is sufficient, the other "
"modality introduces noise and must be ignored."
"2) Importance Ranking: If both modalities contain valid cues, "
"prioritize by causal reliability."
"3) Temporal Integration: If cross-timestamp tracking is required, "
"interleaved layout is mandatory."
""
"# Analysis Pipeline (Strict Execution Order, BE CONCISE)"
"Step 1 (Content Perception): Briefly describe ONLY the visual/audio
    ↪ "
"details directly relevant to answering the question. Skip unrelated
    ↪ "
"scene description. If speech is present and relevant, extract it. "
"(2-3 sentences max)"
"Step 2 (Question Evaluation): State which modality provides the "
"decisive evidence for the answer. (1-2 sentences)"
```

```
    "Step 3 (Routing Justification): Explain why the optimal routing "
    "{optimal} is correct, citing ONE principle above. Then explain why "
    "{representative_wrong} fails. (2-3 sentences)"
    ""
    "# Output Format (STRICT TOTAL LENGTH: under 120 words)"
    "<perception>...</perception>"
    "<decisive_modality>...</decisive_modality>"
    "<routing_justification>...</routing_justification>"
    ""
    "Question:{question}"
    "Options:{option}"
    "Optimal sequence:{optimal}"
    "Representative wrong sequence:{representative_wrong}"
    ""
    "Example:"
    "<perception>The video shows a vacant seat with a 'Suriname' placard
        ↪ "
    "during the initial pan. The audio mentions 'Philippines' and
        ↪ 'Nauru' "
    "near the end, unrelated to the vacant seat.</perception>"
    "<decisive_modality>Visual_the question asks about a static scene "
    "detail at a specific timestamp.</decisive_modality>"
    "<routing_justification>Video-Audio (V-A) sequence is optimal per "
    "Importance Ranking: it anchors the visual evidence first. Audio-"
    "centric sequences fail because the salient spoken country names "
    "create an alignment trap, causing the model to shortcut to an "
    "audio-mentioned country instead of the visually vacant
        ↪ seat.</routing_justification>"
    ""
    "Now generate the answer concisely."
```

The generated rationale provides additional supervision beyond the organization label, enabling the Planner to learn organization-level decision patterns rather than memorizing fixed topology choices.

H.1 Trajectory Design Ablation

Since multiple organizations may occasionally lead to correct predictions for the same sample, organization supervision can contain inherent ambiguity. Directly treating all successful organizations as equivalent candidates may prevent the Planner from learning a clear organization preference. Specifically, we compare two trajectory construction strategies: (1) Multi-optimal supervision, where all successful organizations are treated as equivalent targets; (2) Single-optimal supervision, where one organization is selected as the target together with its rationale. As shown in Table 15, Multi-optimal supervision obtains the worse performance. Although multiple organizations may produce correct answers for the same sample, treating them as equivalent targets introduces conflicting supervision signals. The Planner is required to map an identical input to multiple possible organizations without explicit preference, making it difficult to learn a deterministic organization policy. Single-optimal supervision achieves the better performance and is therefore adopted in our final Planner-SFT training. This result suggests that providing a unique organization target together with its corresponding rationale enables the Planner to learn more consistent modality organization preferences.

Table 15: **Effect of organization supervision design for Planner-SFT.**

Dataset	WorldSense	DailyOmni	OmniBench
Qwen-Omni-7B	41.4	55.6	49.1
Multi-optimal supervision	41.4	56.3	49.3
Single-optimal supervision	41.3	57.0	50.3

H.2 Failure Cases Analysis

Although CoM can adaptively reorganize modality information, some samples remain challenging when the evidence required for answering is inherently ambiguous. As illustrated in Figure 7,

neither modality provides a clear, decisive cue. In such under-constrained scenarios, the optimal inference topology itself becomes unidentifiable from the available evidence, thereby bottlenecking the benefits of dynamic topology adaptation. Specifically, although this question represents a standard audio-visual temporal reasoning task, the key audio action verb ("open", spanning only 0.3s) and the visual action graphic (downward arrow + envelope) exhibit temporal phase shifts and duration asymmetry across the timeline. When confronted with such asynchronous signals, cross-modal attention mechanisms fail to establish fine-grained temporal anchoring; instead, they misalign the long-duration, high-energy audio context ("mass emails") with the strong visual bias symbol (red downward arrow), ultimately steering the model toward Option A. In such cases, the optimal organization is itself difficult to determine from the available evidence, limiting the benefit of topology adaptation.

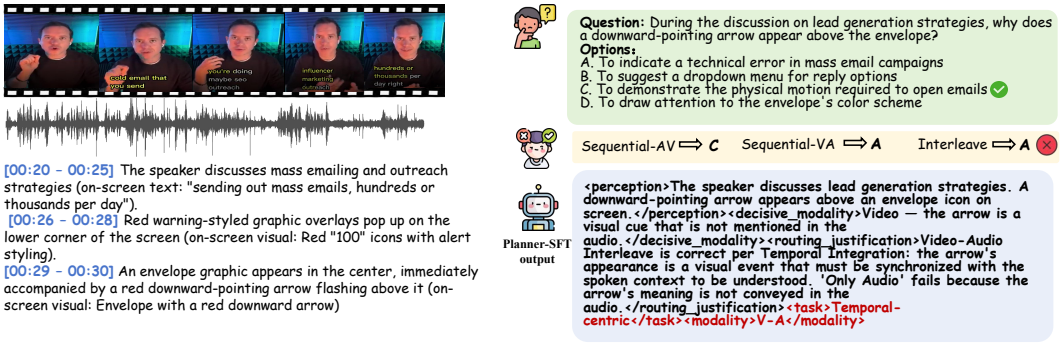


Figure 7: **A failure case caused by cross-modal temporal asynchrony.**

I Additional Qualitative Results

To further illustrate the effectiveness of CoM, we provide additional qualitative examples across diverse multimodal scenarios. These examples demonstrate how CoM dynamically selects modality organizations according to the evidence requirements of different questions, rather than relying on a fixed modality preference.

As shown in Figure 8 and Figure 9, each example compares the predictions under different modality organizations with the organization selected by CoM. These examples highlight that modality organization itself can affect multimodal reasoning outcomes. Although all organizations access the same underlying multimodal inputs, inappropriate modality arrangements may lead to evidence misalignment or over-reliance on irrelevant cues. By explicitly modeling modality organization before reasoning, CoM enables the Reasoner to access more task-relevant evidence and improves the final prediction quality.

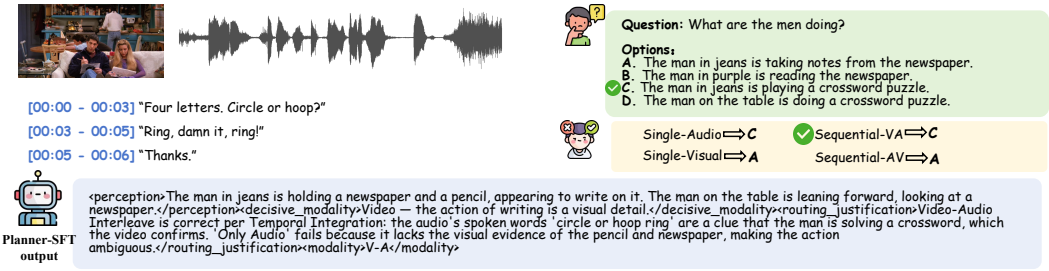


Figure 8: **Qualitative comparison of Planner-SFT and fixed modality organizations.**

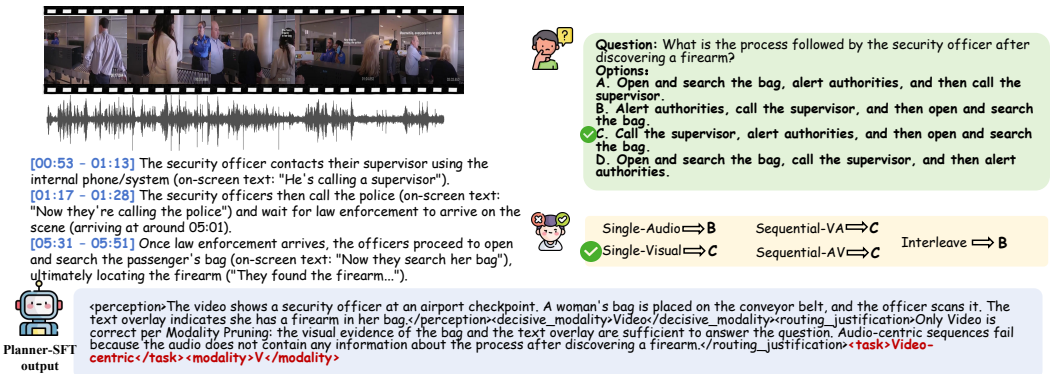


Figure 9: **Qualitative comparison of Planner-SFT and fixed modality organizations.**