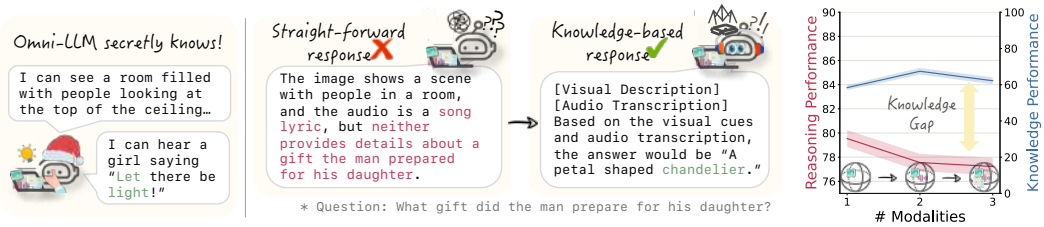
OMNI-MODAL LLMS KNOW MORE THAN THEY SAY

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni-modal large language models (Omni-LLMs) extend LLMs to process heterogeneous modalities (*e.g.*, vision, audio, and IMU signals) alongside language, enabling a broader range of multimodal tasks. Yet we find that these models often fail to express information from non-textual modalities even when they perceive it correctly. We term this disconnect *knows-but-cannot-say*. To investigate this phenomenon, we introduce Token-level Modality Alignment (TMA), which measures the alignment between individual modality representations and response representations. Our analysis shows that non-textual representations are less aligned with response tokens and receive less attention than textual representations, suggesting a potential explanation for uneven modality preferences. Motivated by these findings, we propose OMNI-CATALYST, a simple *perceive* → *infer* strategy that complements the original modality inputs with self-generated text to make non-textual knowledge more accessible. Across seven Omni-LLMs, OMNI-CATALYST consistently improves performance (+7.29%) and yields more balanced modality preferences, with further gains from diverse training strategies. Further analysis demonstrates that the knowledge extraction phases increase TMA, suggesting that improved alignment makes non-textual information more accessible during response generation. In summary, our findings uncover **how Omni-LLMs can better express modality knowledge through language**.



(a) Modality knowledge extraction (b) Multimodality

Figure 1: **The "knows but cannot say" problem in Omni-LLMs.** (a) Omni-LLM often embeds sufficient modality knowledge, yet it produces an incorrect answer when directly responding (left). By complementing the relatively weak modality, audio recovers the correct answer (right). (b) This phenomenon becomes more pronounced as the number of modalities increases.

1 INTRODUCTION

Omni-modal large-language models (Omni-LLMs) (Han et al., 2024; Cheng et al., 2024; Xie & Wu, 2024; Hurst et al., 2024; ChameleonTeam, 2024; Zhan et al., 2025; Ye et al., 2025; Li et al., 2025b; Inclusion et al., 2025; Xu et al., 2025c; Liu et al., 2025b; Chen et al., 2025a; Xu et al., 2025a; QwenTeam, 2026; LongcatTeam et al., 2026; GemmaTeam et al., 2026; NAVERTeam, 2026; Luo et al., 2026a; Nguyen et al., 2026; Dai et al., 2026) integrate diverse sensory modalities within a single language-centric reasoning system[1]. Their design broadly resembles the organization of human perception, in which specialized pathways process distinct sensory streams before integrating them for higher-level cognition (McClure, 1916; Stokes, 2021; Beck, 2025). Omni-LLMs encode inputs such as vision and audio through modality-specific encoders, then project these representations into a shared space for the underlying language model to generate a response. While this unified

[1]We use *multimodal* and *omni-modal* to refer to models supporting two and > 2 modalities, respectively.

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.

architecture enables richer interaction across sensory channels, Omni-LLMs often still underperform modality-specialized counterparts on the very tasks those models are designed for (Zhu et al., 2025). This raises a fundamental question: *why do Omni-LLMs fail to fully benefit from the additional modalities available to them?* Limited training data and suboptimal training strategies may partly explain this gap. However, another possible explanation may be that the limitation lies not only in perception itself, but also in how effectively Omni-LLMs can access perceived information during the subsequent generation process.

In this work, we investigate the gap between what Omni-LLMs perceive and what they use to generate an answer. Figure 1 presents a motivating illustration of this discrepancy and how it evolves as the number of input modalities increases. As shown in Figure 1a, the Omni-LLM can correctly extract relevant modality information but fails to use it when directly answering the multimodal reasoning question. Remarkably, once the same modality-specific information is explicitly elicited in textual form, the model recovers the correct answer. This suggests a *knows-but-cannot-say* problem (Orgad et al., 2025; Lu et al., 2026): the model *knows* the relevant information at the perception level, but that knowledge fails to reach the generation path unless it is re-expressed in text. Crucially, this problem intensifies as additional modalities are introduced (see Figure 1b), precisely the regime Omni-LLMs are built for. Downstream accuracy collapses with each added modality while perception accuracy holds, isolating the bottleneck in read-out rather than encoding.

Our study investigates the hidden *modality* knowledge problem through cross-modality representational alignment. Although Omni-LLMs are extensively trained to align heterogeneous modality representations (Liang et al., 2021; Wang et al., 2025; Zhao et al., 2025; Chowdhury et al., 2026; Li et al., 2026), irreducible misalignment persists (Liang et al., 2022; Fürst et al., 2022; Yaras et al., 2024; Fonseca et al., 2025; Li et al., 2025a; Grassucci et al., 2026; Schrodi et al., 2025; Eslami & de Melo, 2025; Yan et al., 2026a; Gröger et al., 2026; Chowers et al., 2026). To effectively characterize this hidden gap, we introduce **Token-level Modality Alignment (TMA)**, which complements existing global and local alignment measures by quantifying how input modality information interacts with language generation at a token-level. In §3, we find that TMA effectively explains why Omni-LLMs cannot fully access non-text modality representations. Notably, modality tokens with higher TMA show greater sensitivity in downstream performance, but not for knowledge extraction, and this effect is exacerbated with more modalities.

Such representational mismatch induces prevailing modality imbalance issue, in which Omni-LLMs rely on modalities that are easier to access rather than those most relevant to the query (Chen et al., 2025b; Wang et al., 2026; Yan et al., 2026b; Zheng et al., 2026; Yang et al., 2025c; Huang et al., 2022; Tjandrasuwita et al., 2025; Fang et al., 2026; Reilly et al., 2026). To address this issue, we explore in §4 a simple strategy that renders **modality-specific information into texts** before answer generation. We propose OMNI-CATALYST for both training-free and trainable settings and show that complementary, self-generated modality descriptions substantially improve omni-modal reasoning performance (from §5.1 to §6.2). Importantly, while OMNI-CATALYST is particularly beneficial for complementing weaker modalities in multimodal inputs (*e.g.*, audio given vision-audio), it also generalizes to single-modality settings, consistently improving performance across diverse modalities such as fMRI, IMU, and point clouds. The self-generated descriptions preserve modality-specific semantics while residing in the same representational space as the subsequent answering tokens, thereby bridging heterogeneous modality inputs and language generation. Overall, our results suggest that how much Omni-LLMs can make encoded information usable during generation matters. Our contributions can be summarized as follows:

- **Diagnosing omni-modal representations:** We trace modality imbalance in Omni-LLMs through representational misalignment between modality-specific inputs and generation representations, using our proposed Token-level Modality Alignment (TMA) metric.
- **Extracting modality knowledge into the generation space**: We propose OMNI-CATALYST, which self-generates descriptions for weaker modalities, steering Omni-LLMs toward the correct answer with their representations acting as an effective semantic and modality intermediate.
- **Attaining modality-balanced gains**: Across diverse Omni-LLMs and benchmarks, OMNI-CATALYST improves omni-modal reasoning by 7.29% and significantly increases TMA and attention in the knowledge-extraction phase.

2 BACKGROUND

Omni-LLMs. Omni-LLMs integrate heterogeneous modalities within a unified architecture to support multimodal understanding and reasoning (Zhu et al., 2025; Koutoupis et al., 2026; Cicchetti et al., 2026). Early efforts such as VAST (Chen et al., 2023) established large-scale vision-audio-subtitle-text pretraining, while One-LLM (Han et al., 2024) progressively incorporated additional modalities through mixture-of-experts projection. Subsequent studies show that modality-specific or joint omni-modal fine-tuning alone does not guarantee genuine omni-modal capability (Zhu et al., 2025), and that bimodal supervision provides limited transfer to omni-modal reasoning (Liu et al., 2026). NExT-OMNI (Luo et al., 2026b) extends the setting to any-to-any generation using discrete flow matching, More recently, inference-time methods have sought to improve modality use without additional training: OmniXR (Chen et al., 2025b) explicitly extracts modality information before reasoning and replaces the modalities, ThinkOmni (Guan et al., 2026) delegates reasoning to external text-only models (Yang et al., 2025a; Guo et al., 2025), and CoM (Luo et al., 2026c) decomposes inference into planner, reasoner, and decider roles. While these approaches improve omni-modal inference through architectural, training, or prompting strategies, our work instead investigates *why* non-textual information can remain underutilized despite being encoded by the model.

Knowing without saying. A growing body of work reveals a gap between information encoded internally and information expressed in model outputs. In language models, intermediate activations can decode answer-related information even when greedy decoding produces an incorrect answer (Orgad et al., 2025). Similar discrepancies arise in vision-language models: models may attend to correct visual evidence despite answering incorrectly (Liu et al., 2025a), encode spatial information that is linearly decodable but not reflected in predictions (Liao et al., 2026), or internally capture negation and spatial/multilingual information that does not reliably reach the output (Lu et al., 2026; An et al., 2026). Zhang et al. (2026) further shows that useful latent information can be silenced without influencing predictions. Complementarily, Gou et al. (2026) decouple perception from reasoning using an external captioner, demonstrating that richer visual descriptions can improve reasoning. These findings motivate our central question: whether Omni-LLMs similarly *know* information in non-textual representations but fail to effectively *use* it during generation.

Modality imbalance. Omni-LLMs are designed to integrate heterogeneous modalities, yet their representations and inference behavior need not be balanced. Representationally, modality-specific embeddings can remain misaligned despite being mapped into a shared latent space, a phenomenon known as the modality gap. Prior work addresses this gap through alignment and uniformity objectives (U+A) (Wang & Isola, 2020), post-hoc embedding shifts (MG) (Liang et al., 2022), and representation standardization (I0T) (An et al., 2025). However, reducing representational gaps alone does not necessarily improve downstream performance. Behaviorally, Yan et al. (2026b) show that Omni-LLMs may strongly favor visual information even when audio, image, and text are equally available, while HumanOmni-Speaker (Bai et al., 2026) finds reliance on static visual cues over audio-visual alignment in multi-speaker settings. Together, these results highlight that representational alignment and modality utilization are distinct; our work connects the two by measuring token-level modality alignment and examining how it relates to downstream use.

3 TOKEN-LEVEL MODALITY ALIGNMENT AND GENERATION ACCESSIBILITY

In this section, we investigate how Omni-LLMs access encoded modality information during response generation. Two considerations guide our analysis. First, useful modality information must be accessible during *generation*, so alignment among input representations alone is insufficient. Second, answering a query requires selectively using relevant *tokens*; high average alignment only across samples, at a dataset-level, may obscure whether those tokens are accessible. Existing encoder-level multimodal model measures primarily characterize alignment between modalities in a shared latent space (Liang et al., 2022; Huh et al., 2024). However, such encoder-level alignment may not directly apply to generative Omni-LLMs, whose ultimate responses come from the decoder's internal representations and support more than two modalities. We therefore introduce a token-level diagnostic metric that measures modality alignment between arbitrary input modality representations and the representations underlying response generation, and examine what this alignment implies about how modality knowledge is accessed.

3.1 Modality Representational Misalignment

Evaluation of Omni-LLM modality representations. We first highlight the limitations of existing metrics. Representative modality alignment (or gap) metrics, such as centroid distance (CD) (Liang et al., 2022) and CKNNA (Huh et al., 2024) (Figure 2), are developed for dual encoders, where 1) each modality sample yields a full embedding; 2) paired embeddings (*e.g.*, an image and its caption) are trained to be semantically aligned. That means, each modality spans numerous tokens (often thousands for video); hence, a per-sample embedding may blur which tokens are task-relevant. Second, input modalities are complementary rather than semantically similar: an audio track is not a paraphrase of the video, nor is the text query. While Qin et al. (2025) emphasize alignment to the decoder, they aggregate the final alignment score across tokens. Thus, for generative Omni-LLMs, the key question is not how closely the input modalities align dataset-wise, but how strongly each modality token aligns with the representations underlying response generation. Yet, token-level granularity alone does not guarantee an effective measurement of modality alignment. For instance, Yin et al. (2025); Xiang et al. (2025) examine token-level alignment only across input modalities and aggregate it as a top-k mean (SEA in Figure 2). This motivates a new, fine-grained alignment metric that scores each modality token by its maximum similarity to response tokens and tracks its *full distribution* across modality tokens (TMA in Figure 2).

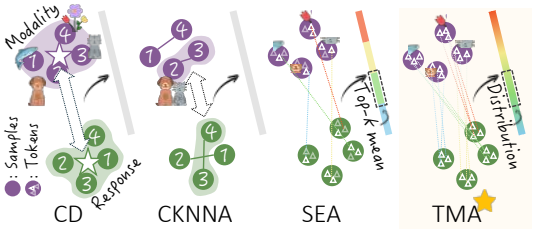


Figure 2: **Comparison of modality alignment metrics.** Unlike prior metrics, our TMA captures the token-level alignment distribution per sample.

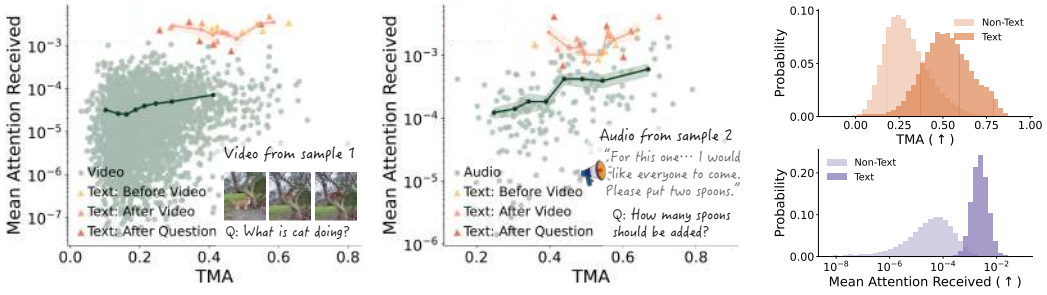
Token-level modality alignment (TMA) distributions. For a modality m, sample i, and modality token j, let $H_i^{(m)} = \mathbf{h}_{i,j}^{(m)}\, j = 1, ..., N_m$, and $H_i^{(r)} = \mathbf{h}_{i,k}^{(r)}, k = 1, ..., N_r$ denote the modality and response hidden states, respectively. We extract the Omni-LLM's last-decoder-layer hidden state at modality-specific placeholder positions (*e.g.*, `Qwen2.5-Omni`'s `<|video_pad|>`) and the generated response-token positions. For each modality token $\mathbf{h}_{i,j}^{(m)}$, we define its TMA score as

$$s_{i,j}^{m\to r} = \cos\left(\mathbf{h}_{i,j}^{(m)}, \mathbf{h}_{i,k_{i,j}^*(m\to r)}^{(r)}\right), \qquad \text{s.t.} \quad k_{i,j}^*(m \to r) = \arg\max_k \cos\left(\mathbf{h}_{i,j}^{(m)}, \mathbf{h}_{i,k}^{(r)}\right). \tag{1}$$

Higher TMA indicates stronger alignment between a modality token and the representations underlying the model's generation. We can examine the TMA distribution at multiple granularities beyond the token level. at the *sample level*, $\mathcal{S}_i^m = \{s_{i,j}^{m\to r}\}_{j=1}^{N_m}$ characterizes how alignment is distributed within an input; and at the *dataset level*, $\{\mathcal{S}_i^m\}_{i=1}^N$ captures variation across examples and modality conditions. Unlike aggregate alignment measures, TMA therefore preserves both within-sample and across-sample variation, providing a finer-grained view of how modality representations connect to generation. The TMA distribution also offers a lens into whether the model retains relevant knowledge and expresses it effectively in its output.

3.2 TMA Reflecting Limited Accessibility to Response Generation

TMA well captures how modality representations reach generation. We use *generation accessibility* to describe how an input representation participates in subsequent response generation, rather than whether it contains relevant information. Since TMA measures the compatibility of an input representation with the response representation space, we examine its relationship with the generation-time attention weights from the same last decoder layer and greedy generation pass used for TMA. For sample i, let $\mathcal{T}_i$ denote the set of decoding steps and M_i^m denote the input-token positions of modality m. The mean attention received by position $j \in M_i^m$ is $\bar{a}_{i,j}^m = \frac{1}{|\mathcal{T}_i| H} \sum_{t\in\mathcal{T}_i} \sum_{h=1}^H \mathrm{softmax}_j\left(\frac{(\mathbf{q}_{i,t}^h)^\top \mathbf{k}_{i,j}^h}{\sqrt{d_k}}\right)$, Across samples, higher TMA modality tokens tend to receive higher attention for both video and audio; however, the association is modest, and their absolute attention remains low (Figure 3a). This distinction is crucial, where TMA and attention capture related but different aspects of generation accessibility. TMA measures how closely an input representation aligns with representations underlying the generated response, whereas attention

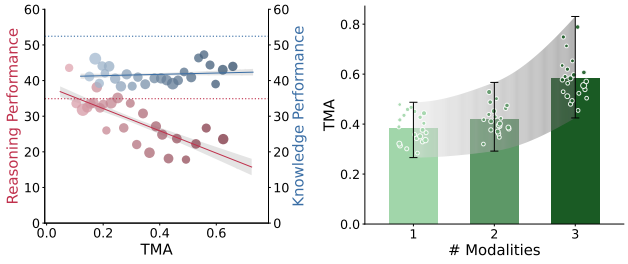


(a) Relation between TMA and attention received at the sample-level (b) Overall distribution

Figure 3: **TMA distributions and accessibility to the generation pathway.** (a) Within samples, input modality (*e.g.*, video and audio) tokens with higher TMA tend to receive higher attention weights from generated-token queries relatively, but exhibit low absolute TMA values. In contrast, text tokens yield higher attention weights and TMA values than non-text tokens, regardless of insertion position. (b) Across samples, text representations systematically exhibit higher TMA and attention than non-textual representations, revealing their greater accessibility to language-centric generation. Note that TMA measures not the amount of information, but its alignment with generation.

measures how much the decoder attends to that representation during generation. Importantly, the two measures exhibit different distributions: while attention strongly separates text (inserted at different positions) from non-text tokens, their TMA distributions substantially overlap (Figure 3b). Thus, non-textual representations can exhibit response-level alignment comparable to text representations even when they receive substantially less generation-time attention, a distinction that attention alone would obscure. Conversely, text tokens, even when semantically unrelated to the queries and inserted at different positions (*i.e.*, before modality 1, between modalities, or after modality 2), consistently receive substantially higher attention weights than non-textual tokens and also exhibit higher TMA overall (Figures 3a and 3b). Together, these patterns indicate that Omni-LLMs prefer textual representations during language-centric generation with higher generation accessibility.

TMA distinguishes what the model *knows* from what it *says*. We then investigate whether low generation accessibility of non-textual modality tokens implies that the model lacks the underlying modality information. To differentiate retained knowledge from its expression in downstream responses (Li et al., 2026), we evaluate both knowledge extraction and direct answering by masking tokens per sample. Masking tokens across TMA levels has little effect on knowledge performance, whereas overall downstream performance degrades substantially (Figure 4a), resulting in a noticeable gap between the two. This contrast suggests that modality information can remain recoverable even when its corresponding tokens are weakly aligned with the representations underlying generation. The TMA distribution therefore characterizes not whether the model holds information, but how strongly that information is aligned with its generation, suggesting a *know-but-cannot-say* gap.



(a) TMA explanation (b) Modality # effects

Figure 4: **TMA reveals a gap between what Omni-LLMs know and what they can express.** (a) Knowledge performance remains high regardless of the TMA of masked tokens, whereas answering performance is substantially lower and further degrades when higher-TMA tokens are masked. (b) Increasing the number of input modalities increases both the mean and variance of TMA, indicating more heterogeneous accessibility to response generation.

When does the accessibility problem become worse? Finally, we examine how TMA changes as the number of input modalities increases. Both its mean and variance rise with additional modalities (Figure 4b). The higher mean indicates stronger average alignment with response representations, while the higher variance reveals greater differences among modality tokens. Consequently, although some representations become more accessible as modalities accumulate, others remain relatively *less accessible* to generation. Thus, introducing new modalities does not uniformly translate

additional modality information into generation; instead, it can widen the variation in how readily that information is expressed, contributing to a gap between what it knows and what it can express.

4 METHOD: OMNI-CATALYST

From TMA diagnosis to methodology design. Our analysis in §3 showed that Omni-LLMs struggle to use retained modality information during generation, especially as modalities increases. TMA characterizes this accessibility at the token level, highlighting the importance of selectively accessing modality information relevant to the response. Since text tokens align more closely with response representations, OMNI-CATALYST complements sensory inputs with self-generated textual descriptions. These descriptions express modality knowledge in the model's native generation space while preserving the original sensory representations, providing an additional pathway through which relevant information can inform the response.

Preliminary. Conventional generative Omni-LLMs (Yang et al., 2024; GemmaTeam et al., 2026) learn a function $f(\mathbf{X}, q) \to y$, where $\mathbf{X} = X_m, m \in \mathcal{M}$ denotes modality inputs, q the query, and $\mathbf{y} = (y_1, \ldots, y_T)$ the response. Each modality-specific encoder maps its input to a latent representation, which is then processed by a shared language model: $h_m = \mathcal{E}^m(X_m), \quad y = f_{\text{LLM}}(\{h_m\}_{m\in\mathcal{M}}, q)$, where $\mathcal{E}_m$ and h_m denote the encoder and latent representation for modality m.

4.1 SELF-GENERATING MODALITY DESCRIPTIONS

We introduce a simple training-free formulation of OMNI-CATALYST that complements each non-textual representation with a self-generated modality description: $d_m = f_{\text{LLM}}(X_m; p_{\text{desc}}), \quad \mathbf{y} = f_{\text{LLM}}(\{h_m, d_m\}_{m\in\mathcal{M}}, q; p_{\text{ans}})$, where p_{desc} and p_{ans} denote the prompts for description and answer generation, respectively. Both stages use the same underlying LLM. Crucially, the generated description does not replace (h_m); rather, it provides an additional textual pathway derived from the same modality information. Thus, the description need not exhaustively reconstruct the original input, but should preserve task-relevant information in a form that is more directly compatible with the LLM's language-based processing. From the perspective of TMA, this intervention exposes modality information through a representation that is more compatible with subsequent language generation, while preserving the original sensory representation.

Generating descriptions in a separate pass incurs additional inference cost. We therefore also consider a single-pass procedure in which the modality descriptions and the final response are generated jointly: $(\mathbf{d}, y) = f(\mathbf{X}, q; p_{\text{desc}} \circ p_{\text{ans}})$, where $\mathbf{d} = \{d_m\}_{m\in\mathcal{M}}$; both $\mathbf{d}$ and $\mathbf{y}$ are generated conditioned on the same modality inputs and query. Thus, the two variants of OMNI-CATALYST only differ in whether modality descriptions are generated separately or jointly with the final response. Most existing Omni-modal models employ LLMs as their foundation models and therefore naturally support text generation, which our method relies on. Models based on alternative foundation architectures (Xie et al., 2025; Zhou et al., 2025; Luo et al., 2026b) could likewise adapt our method by generating modality "descriptions" in a form compatible with their respective generation processes.

4.2 REFINING MODALITY DESCRIPTIONS

Self-generated modality descriptions provide additional textual pathways for modality information, but their usefulness for downstream prediction is not guaranteed. We therefore develop a general framework to improve both their generation and use. Let $\mathbf{X} = X_m, m \in \mathcal{M}$ denote the non-textual modality inputs, q a multiple-choice question with candidate options, and y the correct option. We denote self-generated and ground-truth descriptions by $\hat{\mathbf{d}} = \{\hat{d}_m\}_{m\in\mathcal{M}}$ and $\mathbf{d} = \{d_m\}_{m\in\mathcal{M}}$.

Our framework considers two complementary objectives: *using* modality descriptions and *generating* them. The former trains the model to use self-generated descriptions for downstream prediction: $\mathcal{L}_{\text{U}} = -\log p_\theta\big(y \mid \mathbf{X}, \hat{\mathbf{d}}, q\big)$. Here, $\hat{\mathbf{d}}$ is generated from the same input before the prediction step, following the two-pass training-free formulation of OMNI-CATALYST (single-pass OMNI-CATALYST simulation in the Appendix). The latter trains the model to generate modality descriptions from the raw inputs: $\mathcal{L}_{\text{G}}(\mathbf{z}) = -\sum_{t=1}^{T_{\mathbf{z}}} \log p_\theta(z_t \mid z_{<t}, \mathbf{X}, q_{\text{d}})$, where q_{d} denotes a target modality description and q_{d} specifies the description-generated instruction.

Method	AVHBench A-Hal	AVHBench V-Hal	OmniBench	WorldSense	*Avg.*
Specialist Models					
EchoInk*	72.4	68.3	46.8	39.3	56.7
Omni-R1*	75.8	65.8	47.6	39.7	57.2
HumanOmniV2*	81.7	73.6	43.0	40.5	59.7
Training-free Models					
`Qwen2.5-Omni-3B`	77.8	62.2	43.8	39.1	55.7
CoM (Luo et al., 2026c)	73.5$_{-4.3}$	**80.5**$_{+18.3}$	43.8$_{-0.0}$	40.7$_{+1.6}$	59.6$_{+3.9}$
OMNI-CATALYST	**77.9**$_{+0.1}$	**80.5**$_{+18.3}$	**50.6**$_{+6.8}$	**44.6**$_{+5.5}$	**63.2**$_{+7.5}$
`Qwen2.5-Omni-7B`	73.2	66.8	46.5	39.2	56.4
ThinkOmni (Guan et al., 2026)	34.1$_{-39.1}$	46.2$_{-20.6}$	43.6$_{-2.9}$	33.4$_{-5.8}$	39.3$_{-17.1}$
CoM (Luo et al., 2026c)	**78.2**$_{+5.0}$	79.7$_{+12.9}$	49.4$_{+2.9}$	41.9$_{+2.7}$	62.3$_{+5.9}$
OMNI-CATALYST	76.4$_{+3.2}$	**82.0**$_{+15.2}$	**51.7**$_{+5.2}$	**45.7**$_{+6.5}$	**64.0**$_{+7.6}$

Table 1: **Comparison across different training-free methods.** OMNI-CATALYST consistently outperforms corresponding baselines without an external reasoning model (ThinkOmni) and without an adaptive agentic framework (CoM).

Method	Music	Sound Event	Speech	*Avg.*
`gemma-4-e2b-it`	25.5	16.2	22.4	21.4
Self-CoT	29.3$_{+3.8}$	18.1$_{+1.9}$	26.7$_{+4.3}$	24.7$_{+3.3}$
OMNI-CATALYST	**43.4**$_{+17.9}$	**20.4**$_{+4.2}$	**34.0**$_{+11.6}$	**32.6**$_{+11.2}$
`gemma-4-e4b-it`	28.3	23.4	29.7	27.1
Self-CoT	34.0$_{+5.7}$	25.7$_{+2.3}$	34.5$_{+4.8}$	31.4$_{+4.3}$
OMNI-CATALYST	**48.1**$_{+19.8}$	**31.3**$_{+7.9}$	**42.2**$_{+12.5}$	**40.5**$_{+13.4}$
`Qwen2.5-Omni-3B`	48.1	44.9	38.4	43.8
Self-CoT	**51.9**$_{+3.8}$	41.9$_{-3.0}$	38.7$_{+0.3}$	44.2$_{+0.4}$
OMNI-CATALYST	49.6$_{+1.5}$	**53.8**$_{+8.9}$	**48.3**$_{+9.9}$	**49.6**$_{+5.8}$
`Qwen2.5-Omni-7B`	49.1	51.0	39.6	46.6
Self-CoT	**55.7**$_{+6.6}$	48.7$_{-2.3}$	40.0$_{+0.4}$	48.1$_{+1.5}$
OMNI-CATALYST	50.9$_{+1.8}$	**51.7**$_{+0.7}$	**52.5**$_{+12.9}$	**51.7**$_{+5.1}$

Table 2: **Comparison against Self-CoT.** OMNI-CATALYST generally improves performance over the baselines across different audio types.

When ground-truth descriptions are available, we can directly supervise their generation: $\mathcal{L}_{\mathrm{P}} = \mathcal{L}_{\mathrm{G}}(\mathbf{d})$. When they are unavailable, we instead use self-generated descriptions as targets, $\mathbf{z} = \hat{\mathbf{d}}$, and apply a self-training or refinement strategy to select or improve these targets. These objectives can be combined to first improve description generation and then refine their downstream use.

5 EXPERIMENTAL RESULTS

This section examines whether the benefit of explicitly rendering modality-specific information into text extends across model families, tasks, and modality types.

Setup. We primarily apply OMNI-CATALYST on `gemma-4-E2B/E4B-IT` (GemmaTeam et al., 2026), `Qwen2.5-Omni-3B/7B` (Xu et al., 2025a), and `OneLLM` (Han et al., 2024), across multimodal reasoning (OmniBench (Li et al., 2026); WorldSense (Hong et al., 2025)), modality-conditioned reasoning (AVHBench (Sung-Bin et al., 2025)), modality preference (OmniPreference (Yan et al., 2026b)), and single-modality-to-text tasks (Appendix). For training-based variants, we use 91k training samples from VALID (Nguyen et al., 2024). We instantiate three existing self-training strategies–STaR (Zelikman et al., 2022), R3V (Cheng et al., 2025), and SD-Zero (He et al., 2026)–using modality descriptions (rather than reasoning traces) for self-improvement. More details are provided in the Appendix.

5.1 GENERALIZATION ACROSS TASKS

Compared to prior competitive specialists (Xing et al., 2025; Yang et al., 2025b; Zhong et al., 2026) and training-free methods (Guan et al., 2026; Luo et al., 2026c), OMNI-CATALYST (§ 4.1) consistently attains superior performance across benchmarks (Table 1). On OmniBench and WorldSense, where successful inference requires jointly using visual and audio information, CoM yields only 0–3% gains, whereas OMNI-CATALYST produces +5–7 improvements ($p < 0.001$). This advantage also holds across audio categories (Table 2), especially consistently improving `gemma` performance by 8–18%, approximately 2–14% better than Self-CoT.

We further evaluate OMNI-CATALYST's generalizability across diverse modalities, including fMRI, IMU, and point clouds, in Figure 5. OMNI-CATALYST significantly enhances performance (†: $p < 0.1$; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$), especially for the *weaker* modalities showing relatively lower baseline performance, such as fMRI and IMU. Representing modality-specific information in text provides a general mechanism for making heterogeneous modality information more accessible to language generation, without requiring external reasoning models (Guan et al., 2026) or complex inference configurations (Luo et al., 2026c).

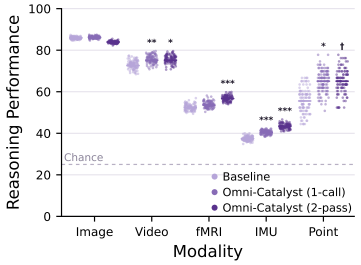


Figure 5: **Modality extension.** OMNI-CATALYST is applicable to various modalities.

5.2 APPLICABILITY TO DIFFERENT TRAINING METHODS

As shown in Figure 6, OMNI-CATALYST can be internalized in the model weights across training strategies (§ 4.2), improving over the untrained baseline by 2–3%. Notably, training the model to explicitly *perceive* modality information ($\mathcal{L}_{\mathrm{P}}$) yields a larger gain (+2.8%) than directly optimizing inference ($\mathcal{L}_{\mathrm{U}}$), even though the latter closely matches the test-time format. Perception-only training also outperforms self-training methods (STaR, R3V, and SD-Zero). Applying training-free OMNI-CATALYST (§ 4.1) on top of each approach yields an additional 5.1–6.1% gain on average, bringing all variants to a comparable performance range of 42.4–43.6%. Finally, applying the overall best self-training method (STaR; +6.1%) to the perceive-only-trained model does not further improve performance, although it produces higher-quality out-of-distribution audio captions (Drossos et al., 2020) (Appendix) Overall, these results suggest that OMNI-CATALYST complements rather than depends on any particular training strategy, while highlighting the value of improving modality perception before optimizing downstream inference.

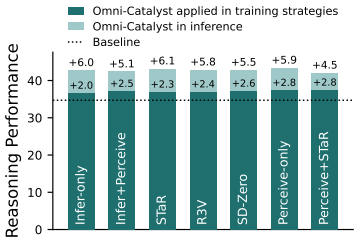
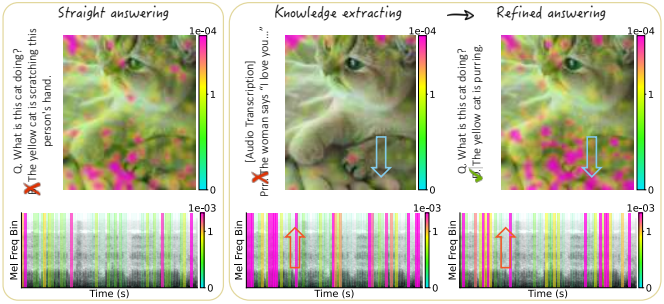


Figure 6: **OMNI-CATALYST adaptability.** Applying OMNI-CATALYST during both training and inference further improves Omni-LLM performance.

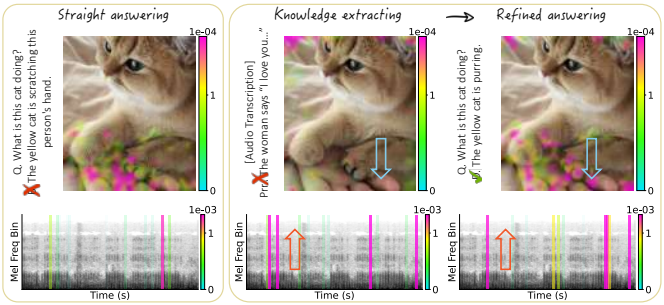
6 EMPIRICAL ANALYSES

6.1 HOW DOES OMNI-CATALYST BRIDGE THE DIAGNOSED MODALITY–GENERATION GAP?

We investigate how explicitly extracting modality information influences accessibility during generation. Figure 7 reveals that direct answering often under-attends to the weaker modality (audio), whereas the knowledge-extraction step distributes attention more broadly over audio and converts its information into text. The refined answering stage can then rely on this textualized audio information while focusing visual attention on the query-relevant region. Importantly, TMA distinguishes where attention is merely allocated from where modality information aligns with the generated response: filtering attention by TMA ($s_j \geq \mathrm{median}(s)$) removes less relevant modality tokens and makes these response-aligned regions substantially clearer. The aggregate distributions in Figure 8 further show that knowledge extraction selectively increases both TMA and attention for the weaker modality. Together, these results suggest that OMNI-CATALYST improves access to underused modality information by textualization, while TMA characterizes how this information aligns with subsequent generation.



(a) Raw attention weights

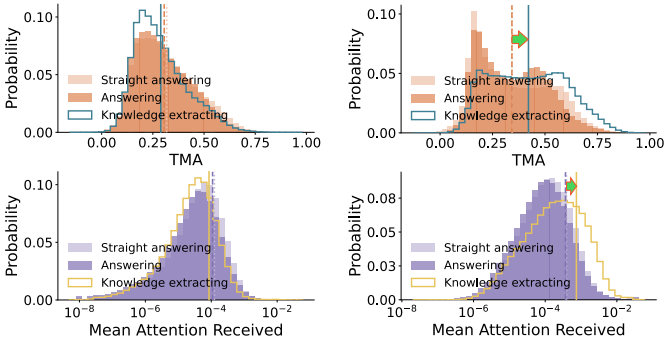


(b) Attention weights + TMA filter

Figure 7: **Probe of how Omni-LLM attention changes.** The knowledge-extraction step in OMNI-CATALYST enables the model to "hear" better, which in turn affects the refined answering stage, showing more balanced "seeing" and "hearing." Our TMA filter, applied on top of the raw attention weights on the original modalities, shows clearer visualizations of where the generated response aligns (*compare* (a) and (b)).

6.2 DOES BRIDGING THE GAP IMPROVE REASONING AND MODALITY BALANCE?

We first examine whether making modality-specific information accessible during generation improves omni-modal reasoning and modality balance. In Figure 9, the pre-trained Omni-LLM (Base-



(a) Video distribution (b) Audio distribution

Figure 8: **TMA and attention distribution for each modality.** The knowledge extraction step of OMNI-CATALYST selectively shifts TMA and attention for the weaker modality (here, audio).

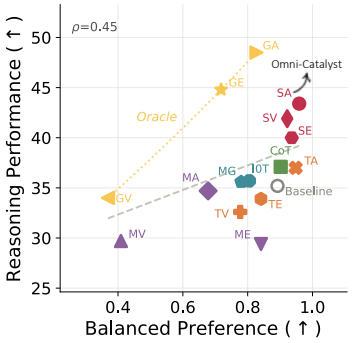


Figure 9: **Performance and preference relation.** Reasoning performance correlates positively with modality preference balance.

line) achieves low accuracy (35%) with an unbalanced modality preference, while existing modality-gap mitigation methods, MG (Liang et al., 2022) and I0T (An et al., 2025), leave both performance and modality preference largely unchanged. This suggests that reducing representational disparity alone does not necessarily make heterogeneous modality information equally accessible during generation. As an oracle, ground-truth audio descriptions (GA) move substantially toward higher performance and more balanced preference, yielding a 13% performance gain. OMNI-CATALYST follows the same direction: its single-pass (SA, SV, SE) and two-pass (TA, TV, TE) variants generally improve both reasoning performance and preference balance. The effect is particularly pronounced for audio descriptions (SA/TA), which shift modality preference toward a more balanced distribution (+10%) while improving accuracy by up to 8% (for SA).

These gains depend on preserving the original modality information rather than replacing it with text, since masking the original non-textual modalities with their descriptions (MA, MV, ME) (Chen et al., 2025b) degrades performance. Meanwhile, Self-CoT, which gives the model a second chance to self-correct the initial answer, yields only marginal improvements in performance (+3%) and preference balance (+1%), suggesting that OMNI-CATALYST's gains are not simply due to additional inference. Taken together, OMNI-CATALYST helps Omni-LLMs better express information from heterogeneous modalities through language.

7 CONCLUSION

Omni-LLMs can know more than they say. We find that non-textual modalities can encode useful information that models fail to use effectively when generating answers. TMA makes this gap visible at the token-level: non-textual representations are less aligned with generated response tokens and receive less attention. This suggests that modality imbalance is not necessarily a problem of missing knowledge, but of making that knowledge accessible to the language space used for answer generation. OMNI-CATALYST addresses this gap with a simple intervention by self-generating textual descriptions of non-textual inputs. These descriptions increase modality alignment and make information already present in the inputs more available during answer generation. Across models and benchmarks, audio descriptions yield particularly consistent gains in both performance and modality balance. We can automate the procedure in a learning process, where perceive-only training followed by StaR-based self-refinement produces the strongest descriptions and performance among the strategies we evaluate.

Limitations and future work. Our experiments cover only the modalities supported by current Omni-LLMs. Nevertheless, TMA and OMNI-CATALYST are nevertheless modality-agnostic and can extend to emerging modalities such as tactile (Tu et al., 2026), visual odometry (Hilscher et al., 2026), and olfactory (Kim et al., 2026). Our study also leaves the root causes of modality imbalance unresolved. Possible sources include the LLM-centric architecture, multimodal pretraining data, and the difficulty of aligning heterogeneous representations. More broadly, our findings suggest that progress toward omni-modal intelligence depends on how effectively models use information across modalities, beyond how many modalities they support.

REFERENCES

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Na Min An, Eunki Kim, James Thorne, and Hyunjung Shim. I0T: Embedding standardization method towards zero modality gap. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27182–27199. Association for Computational Linguistics, July 2025. ISBN 979-8-89176-251-0.

Na Min An, Inha Kang, Minhyun Lee, and Hyunjung Shim. Blind to position, biased in language: Probing mid-layer representational bias in vision-language encoders for zero-shot language-grounded spatial understanding, 2026.

Detao Bai, Zhiheng Ma, and Xihan Wei. Humanomni-speaker: Identifying who said what and when, 2026. URL https://arxiv.org/abs/2603.21664.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Jacob Beck. Between perception and thought. *Philosophy and Phenomenological Research*, 110 (1):294–301, 2025.

ChameleonTeam. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

Lichang Chen, Hexiang Hu, Mingda Zhang, Yiwen Chen, Zifeng Wang, Yandong Li, Pranav Shyam, Tianyi Zhou, Heng Huang, Ming-Hsuan Yang, et al. Omnixr: Evaluating omni-modality language models on reasoning across modalities. In *International Conference on Learning Representations*, volume 2025, pp. 68100–68118, 2025b.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36:72842–72866, 2023.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Kanzhi Cheng, YanTao Li, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. Vision-language models can self-improve reasoning via reflection. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8876–8892, 2025.

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

Sanjoy Chowdhury, Mohamed Elmoghany, Yohan Abeysinghe, Junjie Fei, Sayan Nag, Salman Khan, Mohamed Elhoseiny, and Dinesh Manocha. Magnet: A multi-agent framework for finding audio-visual needles by reasoning over multi-video haystacks. *Advances in Neural Information Processing Systems*, 38:49255–49291, 2026.

Rhea Chowers, Oshri Naparstek, Udi Barzelay, and Yair Weiss. Is the modality gap a bug or a feature? a robustness perspective. *arXiv preprint arXiv:2603.29080*, 2026.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report, 2024.

Giordano Cicchetti, Eleonora Grassucci, and Danilo Comminiello. A triangle enables multimodal alignment beyond cosine similarity. *Advances in Neural Information Processing Systems*, 38: 112004–112022, 2026.

Yifan Dai, Zhenhua Wu, Bohan Zeng, Daili Hua, Jialing Liu, Bozhou Li, Yuran Wang, Chengzhuo Tong, Hao Liang, Xiaochen Ma, et al. Latentomni: Rethinking omni-modal understanding via unified audio-visual latent reasoning. *arXiv preprint arXiv:2605.22012*, 2026.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13142–13153, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740. IEEE, 2020.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in clip. In *International Conference on Learning Representations*, volume 2025, pp. 82070–82088, 2025.

Wanlong Fang, Tianle Zhang, and Alvin Chan. To align or not to align: Strategic multimodal representation alignment for optimal performance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21056–21064, 2026.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

Rui Fonseca, Bruno Martins, and Gil Rocha. Text-only training for image captioning with retrieval augmentation and modality gap correction. *arXiv preprint arXiv:2512.04309*, 2025.

Andreas Fürst, Elisabeth Rumetshofer, Johannes Lehner, Viet T Tran, Fei Tang, Hubert Ramsauer, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto, et al. Cloob: Modern hopfield networks with infoloob outperform clip. *Advances in neural information processing systems*, 35: 20450–20468, 2022.

GemmaTeam, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Juyeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico

Barbero, Joelle Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, Arthur Bražinskas, Ryan Burnell, Robert Busa-Fekete, Mu Cai, Glenn Cameron, Charlotte Caucheteux, Garima Chadha, Jetha Chan, Aditya Chawla, Blake Jianhang Chen, Jesse Chen, Lin Chen, Xu Chen, Derek Cheng, Tzu hsiang Chien, Nikolai Chinaev, Yi Chou, Zhaohui Chu, Benjamin Coleman, Pooja Consul, Sam Conway-Rahman, Scott Crowell, Dylan Cutler, Vivek Dani, Samira Daruki, Anil Das, Daniel Deutsch, Nishanth Dikkala, Li Ding, Qiuhan Ding, Shenil Dodhia, Konstantin Donhauser, Tulsee Doshi, Anca Dragan, Alex Druinsky, Sahil Dua, Zoltan Egyed, Danielle Eisenbud, Daniel Eppens, Cindy Fan, Bahare Fatemi, Yassir Fathullah, Vlad Feinberg, Milen Ferev, Takumi Fujimoto, Isaac Galatzer-Levy, João Gante, Simon Geisler, Soham Ghosal, Antonious M. Girgis, Alec Go, Alhaad Gokhale, Alex Grills, Yiming Gu, Pramod Gupta, Guru Guruganesh, Raia Hadsell, Hamza Harkous, Jitendra Harlalka, Demis Hassabis, Anja Hauth, Joe Heyward, Arian Hosseini, Chih-Yang Hsia, I-Hung Hsu, Xiaopeng Huang, Yangsibo Huang, Kevin Hui, Adrian Hutter, Te I, Fotis Iliopoulos, Advait Jain, Ganesh Jawahar, Ziwei Ji, Qilin Jin, Melvin Johnson, Kandarp Joshi, Arun Kandoor, Wang-Cheng Kang, Koray Kavukcuoglu, Mehran Kazemi, Kathleen Kenealy, Amr Khalifa, Phoebe Kirk, Suraj Kothawade, Vitaly Kovalev, Neel Kovelamudi, Adam Kraft, Ravin Kumar, Harish Kuppam, Justin Lannin, Chen-Yu Lee, Seungji Lee, Dmitry Lepikhin, Dongdong Li, Qiujia Li, Valentin Liévin, Ethan Lin, Ziqian Lin, Casper Liu, Tianlin Liu, Tianqi Liu, Xin Liu, Mayank Lunayach, Min Ma, Gagan Madan, Andrii Maksai, Eric Malmi, Michal Matuszak, Daniel McDuff, Gaurav Menghani, Daniil Mirylenka, Karolis Misiunas, Vedant Misra, Andreea Mitran, Kareem Mohamed, Maksim Mukha, Eric Noland, James O'Donnell, Kate Olszewska, Bernett Orlando, Wanqiong Pan, Rina Panigrahy, Unnati Parekh, Chunjong Park, Eric Paskie, Liqian Peng, Bryce Petrini, Slav Petrov, Jonas Pfeiffer, Bilal Piot, Martyna Plomecka, Siim Poder, Octavio Ponce, Arijit Pramanik, David Racz, Anish Rajan, Michelle Ramanovich, Anand Rao, Marvin Ritter, Vitor Rodrigues, Evan Visin, Alex Vitvitskyi, Kiran Vodrahalli, Weiyi Wang, Xin Wang, Tris Warkentin, Jan Wassenberg, John Wieting, Lechao Xiao, Hao Xu, Yuhui Xu, Fuzhao Xue, Arun Yadav, Jun Yan, Antoine Yang, Lin Yang, Ming-Hsuan Yang, Ziyu Ying, Jae Hyeon Yoo, Sajjad Zafar, Fred Zhang, Jiageng Zhang, Jianyi Zhang, Xiaofan Zhang, Chao Zhao, David Zhou, and Chen Zou. Gemma 4 technical report, 2026.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Advances in Neural Information Processing Systems*, 38:41819–41886, 2026.

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Xin Jin, Zhenguo Li, James Kwok, and Yu Zhang. Reasoning-aligned perception decoupling for scalable multi-modal reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 140359–140415, 2026.

Eleonora Grassucci, Giordano Cicchetti, Emanuele Frasca, Aurelio Uncini, and Danilo Comminiello. Closing the modality gap aligns group-wise semantics. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the platonic representation hypothesis: An aristotelian view. *arXiv preprint arXiv:2602.14486*, 2026.

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. In *The Fourteenth International Conference on Learning Representations*, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26584–26595, 2024.

Yinghui He, Simran Kaur, Adithya Bhaskar, Yongjin Yang, Jiarui Liu, Narutatsu Ri, Liam Fowl, Abhishek Panigrahi, Danqi Chen, and Sanjeev Arora. Self-distillation zero: Self-revision turns binary rewards into dense supervision. *arXiv preprint arXiv:2604.12002*, 2026.

Nouri Alexander Hilscher, Mateo de Mayo, Dominik Muhle, Christoph Otten genannt Hermes, and Daniel Cremers. Voca: Visual odometry with codec awareness. In *European Conference on Computer Vision*, pp. 594–614. Springer, 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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International conference on machine learning*, pp. 9226–9259. PMLR, 2022.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*, 2024.

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.

AI Inclusion, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chun-Xiang Jin, Chun-Jie Shen, Dan-Dan Zheng, Fudong Wang, Fu-Rong Xu, Guang-Ming Yao, Jun Zhou, Jingdong Chen, Jian-Xin Sun, Jiajia Liu, Jian-Jiang Zhu, Jun Peng, Kaixiang Ji, Kaiyou Song, Kaimeng Ren, Li-Bin Wang, Lixiang Ru, Lele Xie, Long-Hua Tan, Lyuxin Xue, Lan Wang, Mo-Chen Bai, Ning Gao, Pei Chen, Qingpei Guo, Qing-Long Zhang, Qiang Xu, Rui Liu, Rui-Jie Xiong, Si-Rui Gao, Ting Liu, Taisong Li, Weilong Chai, Xinyu Xiao, Xiaomei Wang, Xiaoxue Chen, Xiao Lu, Xiaoyu Li, Xingning Dong, Xuzheng Yu, Yi Yuan, Yuting Gao, Yun-Xiao Sun, Yipeng Chen, Yi fan Wu, Yongjie Lyu, Ziping Ma, Zipeng Feng, Zhijian Fang, Zhi-Hao Qiu, Ziyuan Huang, and Zhengyu He. Ming-omni: A unified multimodal model for perception and generation. *ArXiv*, abs/2506.09344, 2025.

Seongyu Kim, Seungwoo Lee, Hyeonggon Ryu, Joon Son Chung, and Arda Senocak. See & sniff: Learning visuo-olfactory representations. *arXiv preprint arXiv:2606.27307*, 2026.

Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

Stefanos Koutoupis, Michaela Areti Zervou, Konstantinos Kontras, Maarten De Vos, Panagiotis Tsakalides, and Grigorios Tsagkatakis. The more, the merrier: Contrastive fusion for higher-order multimodal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8825–8835, 2026.

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. *Advances in Neural Information Processing Systems*, 38, 2026.

Xiquan Li, Wenxi Chen, Ziyang Ma, Xuenan Xu, Yuzhe Liang, Zhisheng Zheng, Qiuqiang Kong, and Xie Chen. Drcap: Decoding clap latents with retrieval-augmented generation for zero-shot audio captioning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025a.

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*, 2025b.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Hangyu Guo, Yiming Liang, Jiaheng Liu, Noah Wang, Jian Yang, Siwei Wu, et al. Omnibench: Towards the future of universal omni-language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Le Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis philippe Morency. Multibench: Multiscale benchmarks for multimodal representation learning. *Advances in neural information processing systems*, 2021:1 – 20, 2021.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Chih-Ting Liao, Fei Shen, Xin Cao, and Tat-Seng Chua. Decodable is not grounded: A vision-ablation arbiter for vlm spatial reasoning. *arXiv preprint arXiv:2606.31257*, 2026.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Che Liu, Lichao Ma, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Xuerui Yang, and Fei Tian. Boosting omni-modal language models: Staged post-training with visually debiased evaluation. *arXiv preprint arXiv:2605.12034*, 2026.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Joy Zeng, Zhenwei Dai, Zhan Shi, Tian-Xin Wei, Benoît Dumoulin, and Hang-Hang Tong. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. *ArXiv*, abs/2510.17771, 2025a.

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*, 2025b.

Meituan LongCat LongcatTeam, Bin Xiao, Chao Wang, Chengjiang Li, Chi Zhang, Chong Peng, Hang Yu, Hao Yang, Haonan Yan, Haoze Sun, et al. Longcat-next: Lexicalizing modalities as discrete tokens. *arXiv preprint arXiv:2603.27538*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Chen-Yi Lu, Yueh-Shao Chen, and Somali Chaterji. What clip knows but cannot say: Recovering negation from frozen intermediate features. *arXiv preprint arXiv:2607.23271*, 2026.

Run Luo, Ting-En Lin, Haonan Zhang, Yuchuan Wu, Xiong Liu, Yongbin Li, Longze Chen, Jiaming Li, Lei Zhang, Xiaobo Xia, et al. Openomni: Advancing open-source omnimodal large language models with progressive multimodal alignment and real-time emotional speech synthesis. *Advances in Neural Information Processing Systems*, 38:158925–158953, 2026a.

Run Luo, Xiaobo Xia, Lu Wang, Longze Chen, Renke Shan, Jing Luo, Min Yang, and Tat-Seng Chua. Next-omni: Towards any-to-any omnimodal foundation models with discrete flow matching. In *International Conference on Learning Representations*, volume 2026, pp. 147298–147334, 2026b.

Ziyang Luo, Nian Liu, and Junwei Han. Chain of modality: From static fusion to dynamic orchestration in omni-mllms. *arXiv preprint arXiv:2604.14520*, 2026c.

MT McClure. Perception and thinking. *The Journal of Philosophy, Psychology and Scientific Methods*, 13(13):345–354, 1916.

NAVER Cloud HyperCLOVA X NAVERTeam. Hyperclova x 8b omni. *arXiv preprint arXiv:2601.01792*, 2026.

H Nguyen, K Tsui, A Radonjic, and C Schuhmann. Valid (video-audio large interleaved dataset), 2024.

Le Thien Phuc Nguyen, Zhuoran Yu, Samuel Low Yu Hang, Subin An, Jeongik Lee, Yohan Ban, SeungEun Chung, Thanh-Huy Nguyen, JuWan Maeng, Soochahn Lee, et al. See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2272–2283, 2026.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations. In *International Conference on Learning Representations*, volume 2025, pp. 66880–66913, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Yiming Qin, Bomin Wei, Jiaxin Ge, Konstantinos Kallidromitis, Stephanie Fu, Trevor Darrell, and XuDong Wang. Chain-of-visual-thought: Teaching vlms to see and think better with continuous visual tokens. *arXiv preprint arXiv:2511.19418*, 2025.

QwenTeam. Qwen3.5-omni technical report, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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

Dominick Reilly, Qiyu Wu, Hiromi Wakaki, Srijan Das, and Yuki Mistufuji. Mixture of probes: Learning from privileged modalities in multimodal llms through probing. *arXiv preprint arXiv:2607.08839*, 2026.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Dustin Stokes. *Thinking and Perceiving*. Routledge, 2021.

Kim Sung-Bin, Oh Hyun-Bin, Lee Jung-Mok, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. In *International Conference on Learning Representations*, volume 2025, pp. 24244–24271, 2025.

Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. Understanding the emergence of multimodal representation alignment. *arXiv preprint arXiv:2502.16282*, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Jiahang Tu, Fengyu Yang, Chenyang Ma, Xihang Yu, Ziyao Zeng, Shaokai Wu, Hanbin Zhao, Zhi Tao, Chao Zhang, Hui Qian, et al. Unitac: A unified multimodal model for cross-sensor tactile understanding and generation. *arXiv preprint arXiv:2606.31451*, 2026.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Kun Wang, Zherui Li, Zhenhong Zhou, Yitong Zhang, Yan Mi, Kun Yang, Yiming Zhang, Junhao Dong, Zhongxiang Sun, Qiankun Li, et al. Omni-safety under cross-modality conflict: Vulnerabilities, dynamics mechanisms and efficient alignment. *arXiv preprint arXiv:2602.10161*, 2026.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Xingrui Wang, Jiang Liu, Chao Huang, Xiaodong Yu, Ze Wang, Ximeng Sun, Jialian Wu, Alan Yuille, Emad Barsoum, and Zicheng Liu. Xmodbench: Benchmarking cross-modal capabilities and consistency in omni-language models. *arXiv preprint arXiv:2510.15148*, 2025.

Bajian Xiang, Shuaijiang Zhao, Tingwei Guo, and Wei Zou. Understanding the modality gap: An empirical study on the speech-text alignment mechanism of large speech language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5187–5202, 2025.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

Zhifei Xie and Changqiao Wu. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *arXiv preprint arXiv:2410.11190*, 2024.

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.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 1645–1653, 2017.

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

Mengyao Xu, Wenfei Zhou, Yauhen Babakhin, Gabriel Moreira, Ronay Ak, Radek Osmulski, Bo Liu, Even Oldridge, and Benedikt Schifferer. Omni-embed-nemotron: A unified multimodal retrieval model for text, image, audio, and video. *arXiv preprint arXiv:2510.03458*, 2025c.

Hanqi Yan, Xiangxiang Cui, Lu Yin, Jindong Gu, Paul Pu Liang, Yulan He, and Yifei Wang. Beyond cross-modal alignment: Measuring and leveraging modality gap in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 12123–12138, 2026a.

Xinru Yan, Boxi Cao, Yaojie Lu, Hongyu Lin, Weixiang Zhou, Le Sun, and Xianpei Han. Beyond text-dominance: Understanding modality preference of omni-modal large language models. *arXiv preprint arXiv:2604.16902*, 2026b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang

Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a.

Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. Humanomniv2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025b.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bo-Wen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Ke-Ming Lu, Ke-Qin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhen-Ru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024.

Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Audio-centric video understanding benchmark without text shortcut. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6580–6598, 2025c.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Can Yaras, Siyi Chen, Peng Wang, and Qing Qu. Explaining and mitigating the modality gap in contrastive multimodal learning. *arXiv preprint arXiv:2412.07909*, 2024.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*, 2025.

Yuanyang Yin, Yaqi Zhao, Yajie Zhang, Yuanxing Zhang, Ke Lin, Jiahao Wang, Xin Tao, Pengfei Wan, Wentao Zhang, and Feng Zhao. Sea: Supervised embedding alignment for token-level visual-textual integration in mllms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1058–1070, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STar: Bootstrapping reasoning with reasoning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yu-Gang Jiang, and Xipeng Qiu. Anygpt: Unified multimodal llm with discrete sequence modeling, 2025.

Xin Zhang, Qiqi Tao, Jiawei Du, Moyun Liu, and Joey Tianyi Zhou. Visual latents know more than they say: Unsilencing latent reasoning in mllms. *arXiv preprint arXiv:2605.02735*, 2026.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*, 2023.

Shanshan Zhao, Xinjie Zhang, Jintao Guo, Jiakui Hu, Lunhao Duan, Minghao Fu, Yong Xien Chng, Guo-Hua Wang, Qing-Guo Chen, Zhao Xu, et al. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025.

Vincent Zhihao Zheng, Étienne Marcotte, Arjun Ashok, Andrew Robert Williams, Lijun Sun, Alexandre Drouin, and Valentina Zantedeschi. Overcoming the modality gap in context-aided forecasting. *arXiv preprint arXiv:2603.12451*, 2026.

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. *Advances in Neural Information Processing Systems*, 38:45266–45297, 2026.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *International Conference on Learning Representations*, volume 2025, pp. 6446–6469, 2025.

Tinghui Zhu, Kai Zhang, Muhao Chen, and Yu Su. Is extending modality the right path towards omni-modality? *arXiv preprint arXiv:2506.01872*, 2025.

APPENDIX

A AI USE STATEMENT

We used AI tools to edit and style the manuscript, identify relevant literature, check the bibliography for potential hallucinations or inaccuracies, assist with writing proofs in the supplementary materials, reconstruct the single-modality-to-text evaluation dataset, and edit software code. The authors developed and verified all research ideas, conceptual development, methodological decisions, experimental design, analyses, and conclusions without AI assistance. The authors reviewed and verified all AI-assisted text, proofs, and code and take full responsibility for the final content of this work.

B ETHICS STATEMENT

By simply encouraging Omni-LLMs to describe non-textual modality inputs, our approach improves the balance of modality usage and mitigates over-reliance on a single dominant modality. We do not anticipate direct negative societal impacts beyond those inherent to the underlying pretrained models, whose biases our method neither introduces nor fully removes.

C REPRODUCIBILITY STATEMENT

We provide sufficient details to reproduce our experiments, including model configurations, datasets, evaluation protocols, implementation details, and training settings. We describe all prompts, hyperparameters, and experimental procedures in the paper and supplementary materials. We use publicly available datasets and pretrained models, and provide the code and configuration files necessary to reproduce our main results. For trained variants, we additionally report the training data, objectives, and optimization settings. We will release the code and other necessary resources upon publication.

D IMPLEMENTATION DETAILS

Datasets. We provide additional details on the datasets and evaluation settings used for the experiments that are omitted from the main paper. For single-modality-to-text evaluation, we use MS COCO (Chen et al., 2015) for image-text, MSVD QA (Xu et al., 2017) for video-text, Clotho (Drossos et al., 2020) for audio-text captioning, Ego4D (Grauman et al., 2022) for IMU-text, NSD (Allen et al., 2022) for fMRI-text, Objaverse (Deitke et al., 2023) for point-cloud-text tasks.

Because these datasets do not follow the standard Omni-LLM benchmark format, we reconstruct them into a common reasoning-oriented multiple-choice format. Table S1 summarizes the resulting task construction, including the input modality, source annotation, and target category. For each instance, we pool its ground-truth caption(s) or video QA text and use keyword-based scoring to map the source annotation to an abstract, modality-specific category (*e.g.*, scene/activity, sound source, action, or object type). The category is not required verbatim in the source text, so the mapping requires interpretation beyond surface-level vocabulary matching. We select the highest-scoring category as the correct answer, sample three non-matching categories as distractors, and discard instances when we cannot obtain sufficiently distinct distractors. Each resulting example therefore contains four text options, with correctness determined by exact option matching.

Modality	Ground-truth text (pooled)	Options		Answer
Image	"A child holding a flowered umbrella and petting a yak." "A young man holding an umbrella next to a herd of cattle." *(+3 more)*	A) A market, shop, or shopping scene C) An animal or pet is the main subject	B) None of the above D) People sharing a meal or eating food together	C (*animals_pets*)
Video	"who gives a wushu demonstration in front of a crowd?" → man "who demonstrated a sword routine for the crowd?" → man *(+34 more)*	A) An animal or pet is the main subject C) Someone dancing or performing on stage	B) A baby, infant, or child being cared for D) A person walking, standing, or sitting	C (*dancing_performing*)
IMU	"stops painting the block of wood"	A) Cooking or preparing food C) Painting, drawing, or writing	B) None of the above D) Picking up, placing, or putting down an object	C (*painting_drawing_writing*)
fMRI	"Three women are sitting at a table with pizza."	A) A market, shop, or shopping scene C) People sharing a meal or eating food together	B) None of the above D) Someone preparing or cooking food	C (*eating_dining*)
Point cloud	"a blue cartoon"	A) An electronic device C) A toy or cartoon/fictional character figure	B) None of the above D) A vehicle, such as a car, boat, train, or aircraft	C (*toy_or_character*)

Table S1: **Samples constructed for each single-modality domain.** The pooled ground-truth caption(s) or QA text never state the correct category verbatim; the model must infer it and select the matching option.

Metrics. We use task-appropriate metrics for the three evaluation settings. For the *multimodal* benchmark, we report reasoning accuracy. For the *single modality-conditioned* benchmark, we separately report perception accuracy (correct yeses) and hallucination accuracy (correct nos). For the *modality preference* benchmark (data along the x-axis in Figure 2), we quantify the balance of modality preference using the entropy of Normalized Modality Preference (NMP), where $\mathrm{NMP}_m = \frac{P(m|I,A,T,C)}{P(m|C)}$. Here, I, A, and T denote image, audio, and text inputs, respectively, and C denotes the set of modality options. Higher NMP indicates stronger reliance on modality

m after normalizing for the model's textual prior, while higher entropy indicates a more distributed preference across modalities. For audio captioning evaluation, we report standard automatic metrics, such as BLEU (Papineni et al., 2002) (averaged BLEU-1, BLEU-2, BLEU-3, and BLEU-4), METEOR (Banerjee & Lavie, 2005), ROUGE (Lin, 2004), and CIDER (Vedantam et al., 2015).

Model types and structures. To examine whether our findings extend beyond the models used in the main experiments, we additionally evaluate OMNI-CATALYST on `Qwen3-Omni-30B-A3B-Instruct` (Xu et al., 2025b), `OmniVinci` (Ye et al., 2025), and `MiniCPM-o-2_6` (Yao et al., 2024). Table S2 summarizes the architectures of all the tested Omni-LLMs, including their modality encoders and language-model backbones. Most of the evaluated Omni-LLMs adopt a Qwen-family language model as the LLM backbone, whereas OneLLM and the Gemma family use different language-model backbones.

Model	Audio Encoder	Visual Encoder	LLM
`OneLLM`	CLIP ViT-L (Radford et al., 2021) (*shared*)		LLaMA2-7B (Touvron et al., 2023)
`gemma-4-e2b-it`	USM-305M (Zhang et al., 2023)	ViT-150M (Dosovitskiy et al., 2020)	Gemma-4-2.3B (GemmaTeam et al., 2026)
`gemma-4-e4b-it`	USM-305M (Zhang et al., 2023)	ViT-150M (Dosovitskiy et al., 2020)	Gemma-4 4.5B (GemmaTeam et al., 2026)
`Qwen2.5-Omni-3B`	Whisper-large-v3 (Radford et al., 2023)	Qwen2.5-VL ViT (Bai et al., 2025)	Qwen2.5-3B (Yang et al., 2024)
`Qwen2.5-Omni-7B`	Whisper-large-v3 (Radford et al., 2023)	Qwen2.5-VL ViT (Bai et al., 2025)	Qwen2.5-7B (Yang et al., 2024)
`Qwen3-Omni-30B-A3B-Instruct`	AuT-650M (Xu et al., 2025b)	SigLIP2-So400M (Tschannen et al., 2025)	Qwen3-30B-A3B (Yang et al., 2025a)
`omnivinci`	Audio Flamingo 3 (Ghosh et al., 2026)	SigLIP-So400M (Zhai et al., 2023)	Qwen2.5-7B (Yang et al., 2024)
`HyperCLOVAX-SEED-Omni-8B`	Qwen2-Audio (Chu et al., 2024)	Qwen2.5-VL ViT (Bai et al., 2025)	HyperCLOVA X SEED-8B (NAVERTeam, 2026)
`MiniCPM-o-2_6`	Whisper-medium (Radford et al., 2023)	SigLIP-400M (Zhai et al., 2023)	Qwen2.5-7B (Yang et al., 2024)

Table S2: **Omni-LLM architectures.** The Omni-LLM structures evaluated in this paper consist of modality-specific encoders and LLMs.

Training details. For trained variants of OMNI-CATALYST, we implement LoRA (Hu et al., 2021) with rank 16 and $\alpha = 32$ and AdamW (Loshchilov & Hutter, 2019) using weight decay 0.01 and $\varepsilon = 10^{-7}$. We use a learning rate of 10^{-6} for Gemma and 3×10^{-6} for Qwen. STaR-based training (Zelikman et al., 2022) runs for four outer iterations with two epochs each and early stopping with patience 5 based on self-generated validation accuracy. Descriptions are greedily decoded with up to 48 new tokens, and images are capped at 100,352 pixels. We use seven VALID shards for training and two for validation, with at most four images and audio clips per sample and 2k MCQs per iteration. We evaluate 100 validation examples every 200 optimizer steps and train on two A100-SXM-80GB GPUs.

Self-training objectives. Our training variants use self-generated modality descriptions as supervision rather than relying on human-annotated reasoning traces. STaR (Zelikman et al., 2022), R^{3}V (Cheng et al., 2025), and SD-Zero (He et al., 2026) differ in how they generate, refine, filter, and learn from self-generated supervision. We adapt these strategies to generate descriptions of non-textual inputs, thereby providing a textual representation of modality-specific information that can be used during generation. Commonly, the model's own binary reward $r(\hat{\mathbf{d}}) = \mathbb{1}[\arg\max_{y'} p_\theta(y' \mid \mathbf{X}, \hat{\mathbf{d}}, q) = y]$ determines whether a generated description is useful for learning. Each outer iteration is initialized from the original checkpoint, following Zelikman et al. (2022).

Distinctively, StaR (Zelikman et al., 2022) iteratively generates and filters rationales before fine-tuning on successful reasoning trajectories, while subsequent multimodal extensions such as R^{3}V (Cheng et al., 2025) and SD-Zero (He et al., 2026) use self-generated supervision to improve reasoning without gold traces. The detailed objectives for the three self-training strategies are given below.

- **STaR** (Zelikman et al., 2022) retains greedily generated descriptions that yield a correct answer, $\mathcal{D}^{(t)} = \{(\mathbf{X}, q, \hat{\mathbf{d}}) \mid r(\hat{\mathbf{d}}) = 1\}$, optionally recovering failures by re-generating with the ground-truth caption as a hint:

$$\mathcal{L}_{\mathrm{STaR}} = \mathbb{E}_{\mathcal{D}^{(t)}}\big[\mathcal{L}_{\mathrm{U}} + \mathcal{L}_{\mathrm{G}}(\hat{\mathbf{d}})\big]. \tag{S1}$$

- **R3V** (Cheng et al., 2025) samples K descriptions per item and splits them into $\mathcal{D}^+$ and $\mathcal{D}^-$ by reward. The same model additionally acts as a refiner and a selector, where $\mathcal{C} = \{(d^{(k)}, \hat{y}^{(k)})\}_{k=1}^{K}$:

$$\mathcal{L}_{\mathrm{R3V}} = \mathcal{L}_{\mathrm{STaR}} - \underbrace{\log \pi_\theta^{\mathrm{Ref}}(d^+ \mid d^-)}_{\text{refiner}} - \underbrace{\log \pi_\theta^{\mathrm{Sel}}(y \mid \mathcal{C})}_{\text{selector}}. \tag{S2}$$

- **SD-Zero** (He et al., 2026) uses the binary outcome reward $r(\cdot)$ not only to filter samples (like STaR) or pair samples for sequence-level SFT (like R3V) but as context for a self-reviser. One model acts as a generator $\pi_\theta^{\text{Gen}}(\cdot) = p_\theta(\cdot \mid \mathbf{X}, q_{\text{G}})$ and a reviser $\pi_\theta^{\text{Rev}}(\cdot \mid \mathbf{d}, P) = p_\theta(\cdot \mid \mathbf{X}, q_{\text{rev}}, \mathbf{d}, P)$. $r(\mathbf{d}_{\text{init}})$ selects a control phrase P_r (rephrase if correct, restart if incorrect), which is appended to the context before the model generates a revision $\mathbf{d}_{\text{rev}}$; the resulting trace is kept only if $r(\mathbf{d}_{\text{rev}}) = 1$,

$$\mathcal{L}_{\text{SRT}} = \mathcal{L}_{\text{U}} - \underbrace{\log \pi_\theta^{\text{Gen}}(\mathbf{d}_{\text{init}}, P_r, \mathbf{d}_{\text{rev}})}_{\text{generator}} - \underbrace{\log \pi_\theta^{\text{Rev}}(\mathbf{d}_{\text{rev}} \mid \mathbf{d}_{\text{init}}, P_r)}_{\text{reviser}}, \tag{S3}$$

after which the reviser, frozen at θ_{SRT}, serves as the teacher, and its token-level distribution is distilled into the generator on on-policy samples $\mathbf{d} \sim \pi_\theta^{\text{Gen}}$:

$$\mathcal{L}_{\text{SD}} = \mathbb{E}_{\mathbf{d}\sim\pi_\theta^{\text{Gen}}}\left[\sum_{t=1}^{T_{\mathbf{d}}} \text{KL}\big(\underbrace{\pi_\theta^{\text{Gen}}(\cdot \mid d_{<t})}_{\text{generator (student)}} \,\|\, \underbrace{\pi_{\theta_{\text{SRT}}}^{\text{Rev}}(\cdot \mid \mathbf{d}, P_{r(\mathbf{d})}, d_{<t})}_{\text{frozen reviser (teacher)}}\big)\right] + \lambda \mathcal{L}_{\text{U}}. \tag{S4}$$

Prompt details and sensitivity. The default prompt used to generate modality descriptions during training OMNI-CATALYST is ("*Please listen carefully to the audio and observe the image. Describe what you see and hear.*"). Figure S2 compares the default training-free OMNI-CATALYST (Prompt 1) with alternative formulations that vary the wording and emphasis of the description instruction. As quantified in Table S3, OMNI-CATALYST performance remains within a similar range across these prompt variants. Thus, the observed gains are not tied to a particular phrasing of the description instruction.

Model	Baseline	Prompt 1	Prompt 2	Prompt 3
`gemma-4-e2b-it`	22.3	30.7$_{+8.4}$	33.0$_{+10.7}$	29.3$_{+7.0}$
`gemma-4-e4b-it`	32.0	42.0$_{+10.0}$	44.0$_{+12.0}$	46.3$_{+14.3}$
`Qwen2.5-Omni-3B`	39.3	41.7$_{+2.4}$	42.3$_{+3.0}$	40.0$_{+0.7}$
`Qwen2.5-Omni-7B`	42.3	48.0$_{+5.7}$	51.7$_{+9.4}$	51.7$_{+9.4}$
Avg.	34.0	40.6$_{+6.6}$	42.8$_{+8.8}$	41.8$_{+7.9}$

Table S3: **Prompt sensitivity.** Training-free OMNI-CATALYST consistently improves over the baseline across prompt variants, with comparable gains across models, indicating limited sensitivity to prompt wording.

Figure S1 illustrates the two description insertion positions considered in the prompt ablation: immediately before the corresponding modality and after the system prompt. Table S4 compares these alternative settings on both the multimodal (OmniBench (Li et al., 2026)) and single modality-conditioned (CMM (Leng et al., 2026)) benchmarks. Complementing the original inputs with modality descriptions (TV, TA, TC, and TE) generally performs better than replacing the corresponding modalities with descriptions (MV, MA, and ME). On the modality-conditioned benchmark, conditionally complementing the description associated with the queried modality (TC) performs particularly well, suggesting that descriptions are most useful when they provide information relevant to the required modality in single-modality-conditioned tasks.

Benchmark	Placement	Model	Baseline		Modality masking			OMNI-CATALYST (two-pass)				Oracle description		
			Orig	CoT	Visual	Audio	All	Visual	Audio	Cond.	All	Visual	Audio	All
OmniBench	Prepended	Qwen	45.2	46.1	39.2	42.6	34.3	40.2	45.7*	40.7	42.2	39.9	**53.1*****	50.7***
		Gemma	24.3	28.0***	15.3	29.7***	20.7	23.8	27.7***	24.0	23.7	26.0	**40.8*****	38.3***
	System-level	Qwen	45.2	46.2	38.8	40.3	35.6	41.0	46.9*	41.0	42.9	40.9	**54.0*****	51.1***
		Gemma	24.3	28.0***	20.5	29.0***	23.2	24.1	27.0***	24.0	24.8	27.2*	**42.9*****	38.5***
CMM	Prepended	Qwen	84.0	75.0	79.0	80.8	71.7	83.8	83.8	**84.2**	81.0	83.8	**84.2**	80.4
		Gemma	70.2	70.7	65.5	69.0	56.5	71.1	71.1	**72.5**	71.0	69.9	71.7	71.2
	System-level	Qwen	84.1	75.8	78.0	83.0	74.6	84.4	84.3	**84.7**	82.9	83.8	84.2	80.4
		Gemma	72.0	70.3	67.0	71.9	60.6	71.8	71.5	**72.4**	71.8	69.9	71.7	71.2

Table S4: **Effect of description position.** Complementing the original modalities with descriptions generally outperforms replacing modalities with their descriptions. On the modality-conditional benchmark, conditionally complementing the query-relevant description (**Cond.**) performs best, while oracle audio descriptions provide the largest gains on OmniBench. Results are consistent across both description placements (Figure S1).

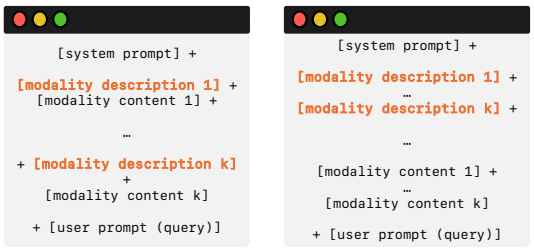


Figure S1: **Effect of description position.** We compare two insertion positions for self-generated modality descriptions: before each corresponding modality (*left*) and after the system prompt (*right*).

Prompt 1: Transcribe the audio.
- SPEECH: Begin each speaker's line with '[the <gender> <role>] says' (e.g. '[the man] says', '[the boy] says', '[the woman] says', '[the girl] says'). Infer gender from voice. Use role if clear (e.g. '[the male officer] says'). Transcribe their words verbatim after that. Note tone only if strongly marked (shouting, crying, whispering).
- NON-SPEECH: Describe literally only if relevant to the question (e.g. 'loud thud', 'dog barking'). Do not infer cause or meaning.
Be concise (2–4 sentences). Do not paraphrase or interpret.

Prompt 2: Write down what is said or heard in the audio.
- DIALOGUE: For every person who speaks, lead with '[the <gender> <role>] says' (e.g. '[the man] says', '[the boy] says', '[the woman] says', '[the girl] says'), guessing gender from the voice and adding a role when it's obvious (e.g. '[the male officer] says'). Give their exact words right after. Mention tone only when it's unmistakable (yelling, sobbing, whispering).
- OTHER SOUNDS: Note them plainly, only when they matter for the question (e.g. 'loud thud', 'dog barking'). Don't guess at their cause or meaning.
Keep it short (2–4 sentences), and stick to what's there — no paraphrasing or interpretation.

Prompt 3: Listen to the audio clip and describe what happens in a couple of sentences. Do not answer the question.

Prompt 1: First, write '[Audio Transcription]' on its own line, followed by: Transcribe the audio.
- SPEECH: Begin each speaker's line with '[the <gender> <role>] says' (e.g. '[the man] says', '[the boy] says', '[the woman] says', '[the girl] says'). Infer gender from voice. Use role if clear (e.g. '[the male officer] says'). Transcribe their words verbatim after that. Note tone only if strongly marked (shouting, crying, whispering).
- NON-SPEECH: Describe literally only if relevant to the question (e.g. 'loud thud', 'dog barking'). Do not infer cause or meaning.
Be concise (2–4 sentences). Do not paraphrase or interpret. Do not mention any answer options in the transcription.

Then, write '[Answer]' on its own line, followed by your answer to the question below. Write each section only once.

Prompt 2: First, write '[Audio Transcription]' on its own line, followed by: Write down what is said or heard in the audio.
- DIALOGUE: For every person who speaks, lead with '[the <gender> <role>] says' (e.g. '[the man] says', '[the boy] says', '[the woman] says', '[the girl] says'), guessing gender from the voice and adding a role when it's obvious (e.g. '[the male officer] says'). Give their exact words right after. Mention tone only when it's unmistakable (yelling, sobbing, whispering).
- OTHER SOUNDS: Note them plainly, only when they matter for the question (e.g. 'loud thud', 'dog barking'). Don't guess at their cause or meaning.
Keep it short (2–4 sentences), and stick to what's there — no paraphrasing or interpretation.
Do not mention any answer options in the transcription.

Then, write '[Answer]' on its own line, followed by your answer to the question below. Write each section only once.

Prompt 3: First, write '[Audio Transcription]' on its own line, followed by: Listen to the audio clip and describe what happens in a couple of sentences. Do not mention any answer options in the transcription.

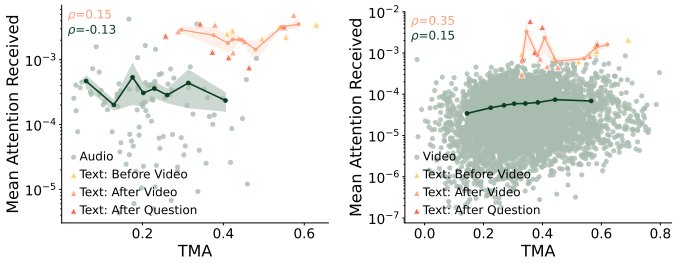
Then, write '[Answer]' on its own line, followed by your answer to the question below. Write each section only once.

Figure S2: **Prompt variations.** Alternative prompts for generating audio descriptions. Prompt 1 is used throughout the main experiments, which achieves performance within the range of the alternative prompts (Figure S3).

E VISUALIZATION DETAILS

TMA and attention. We first provide additional evidence for the relationship between TMA and attention. Figure 3a extends the analysis in Figure S3) with additional audio and video examples. Higher-TMA tokens tend to receive more generation-time attention, but the association is modest and absolute attention remains low. TMA and attention thus capture related but distinct aspects of generation accessibility: alignment with response representations and attention allocated during decoding, respectively.

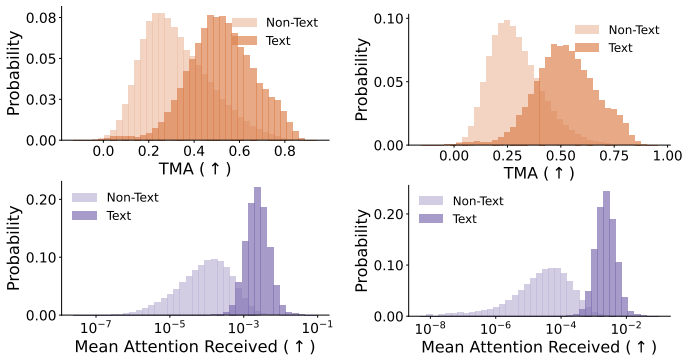
We further examine this relationship across token distributions and model families. Text tokens tend to exhibit higher TMA and attention than non-text tokens for both modalities. Figures S4 and S5 for each modality and different Omni-LLMs (corresponding to Figure 3b).



(a) Audio sample (b) Video sample

Figure S3: **Relation between TMA and attention received at the sample level.** Within samples, input modality (e.g., audio and video) tokens with higher TMA tend to receive higher attention weights from generated-token queries relatively, but exhibit low absolute TMA values. In contrast, text tokens yield higher attention weights and TMA values than non-text tokens, regardless of insertion position. This figure corresponds to audio and video from the two samples in Figure 3a.

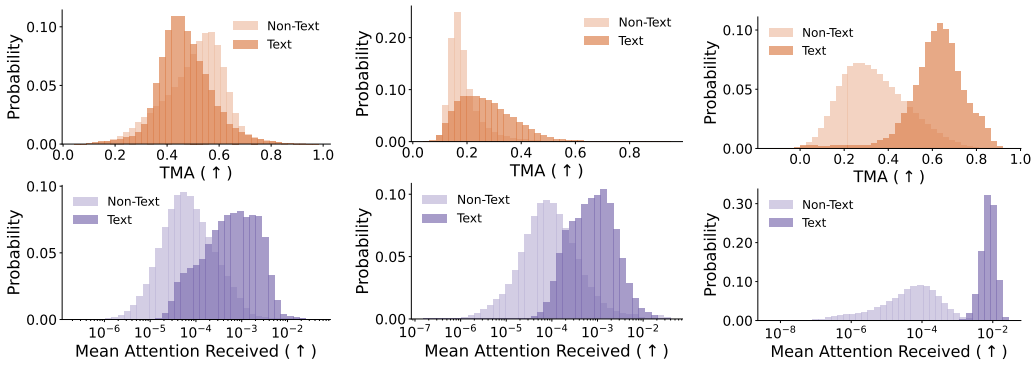
Now, we provide conditions under which the empirical relationship hypothesized in § 3 can arise. Let $k^*(j)$ denote the response token maximizing the TMA of modality token j. Suppose that the generated token representation contains a positively aligned component: $\mathbf{h}^r_{i,t} = \gamma_{i,t,j}\mathbf{h}^r_{i,k^*(j)} + \boldsymbol{\epsilon}_{i,t,j}, \gamma_{i,t,j} > 0$, where the residual has a negligible contribution to its similarity with the modality token: $\boldsymbol{\epsilon}^\top_{i,t,j}\mathbf{h}^m_{i,j} \approx 0$. The attention logit from generated token t to modality token j is $\ell_{i,t,j} = \frac{(W_Q\mathbf{h}^r_{i,t})^\top(W_K\mathbf{h}^m_{i,j})}{\sqrt{d_k}} = \frac{(\mathbf{h}^r_{i,t})^\top W_Q^\top W_K\mathbf{h}^m_{i,j}}{\sqrt{d_k}}$. As a special case, if the query-key projections preserve the ordering of similarities between the relevant response and modality representations, then increasing $s^{m\to r}_{i,j}$ also increases the corresponding attention logit: $s^{m\to r}_{i,j} \uparrow \implies \ell_{i,t,j} \uparrow$. Under the stated approximation, $\ell_{i,t,j} \approx (\mathbf{h}^{(r)}_{i,k^*(j)})^\top\mathbf{h}^m_{i,j} \approx s^{m\to r}_{i,j}$. Since attention weights are obtained by softmax normalization, $\alpha_{i,t,j} = \mathrm{softmax}_j(\ell_{i,t,j})$, gives $s^{m\to r}_{i,j} \uparrow \implies \alpha_{i,t,j} \uparrow \implies \bar{a}^m_{i,j} \uparrow$. Thus, the observed positive association between TMA and attention follows under a specific representation and projection condition; this analysis does not imply that TMA and attention are equivalent.



(a) Audio distribution (b) Video distribution

Figure S4: **TMA and attention distribution for each modality.** Text tokens exhibit higher TMA and attention than non-text tokens for both modalities, averaged in Figure 3b.

Attention map. To visualize the spatial and temporal distribution of attention within each non-textual modality, we project the token-level attention values back to their corresponding input positions. For each sample i and modality m, we retain tokens in the upper $\gamma\%$

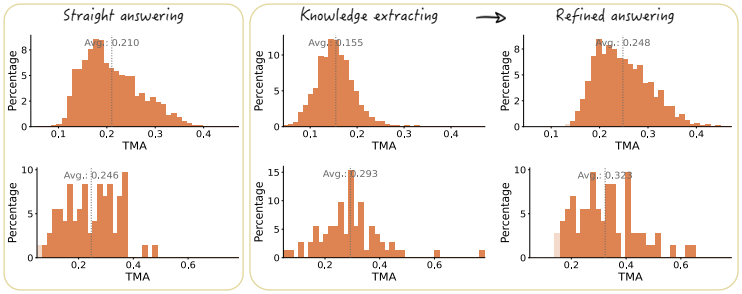


(a) `gemma-2b` (b) `gemma-4b` (c) `Qwen-3B`

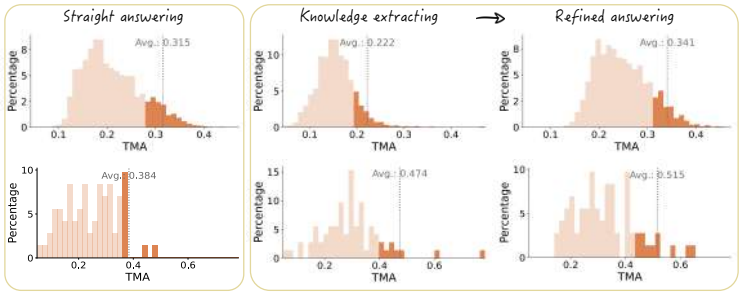
Figure S5: **TMA and attention distribution for remaining Omni-LLMs.** Text tokens generally exhibit higher TMA and attention than non-text tokens in both modalities (`Qwen-7b` in Figure 3b).

of the within-sample TMA distribution: $\mathcal{G}_i^m = \left\{ j \in M_i^m ; \middle| ; \mathrm{TMA}_{i,j} \geq q_\gamma^{i,m} \right\}$ where $q_\gamma^{i,m} = \mathrm{Quantile}_\gamma \left(\mathrm{TMA}_{i,k} : k \in M_i^m \right)$. We visualize only tokens in $\mathcal{G}_i^m$, leaving the remaining positions transparent. Figure S7 visualizes how progressively restricting attention to higher-TMA tokens changes the resulting spatial attention map (γ ranging from 100% to 15%). For video-only, we apply Gaussian smoothing to the attention, and TMA maps over the spatial patch grid before computing the normalization and selection threshold: $\tilde{x}_{i,j}(G_\sigma * x_i)_j$, where x_i denotes either the attention or TMA map ($\sigma = 0.8$). As the TMA threshold becomes more selective, the retained regions remain concentrated around prompt-relevant content, such as the upper-left region containing the cat's paws, rather than uniformly distributed over the input. Figure S8 additionally illustrates the relation between attention and TMA for each modality across stages.

Knowledge and reasoning performance. We next examine the distinction between perceptual knowledge and downstream reasoning under different modality conditions. For each task, we compute reasoning performance directly from the benchmark: binary accuracy for AVHBench and multiple-choice accuracy for OmniBench. We evaluate perceptual knowledge using the corresponding open-ended modality-description task. An independent text-only judge extracts a small set of key content elements from the ground-truth description and checks whether each element appears in the model response verbatim or with semantic equivalence. We report the fraction of examples with non-zero keyword recall. For multimodal conditions, we evaluate only content associated with the modalities actually provided. Figure S6a shows the raw attention distribution used during knowledge extraction, while S6b shows the corresponding distribution after TMA filtering. The two views illustrate how the extraction stage can preferentially attend to audio content before the subsequent answering stage.



(a) Raw attention weights distribution for video (top) and audio (bottom)



(b) Attention + TMA filter distribution for video (top) and audio (bottom)

Figure S6: **TMA and attention distribution for each modality.** The knowledge extraction stage shows higher (a) raw and (b) filtered attention weights for the audio modality, affecting the subsequent refined answering step.

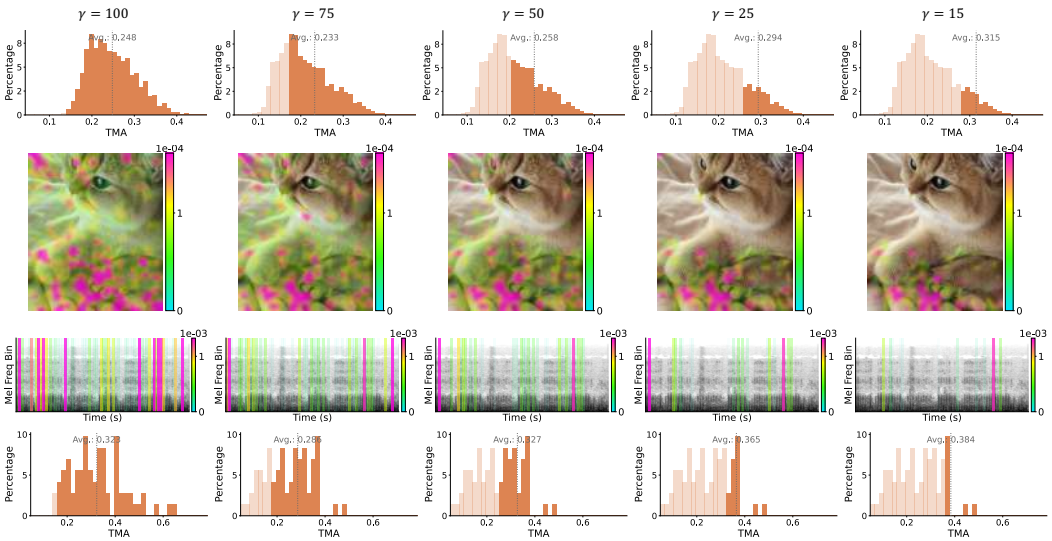


Figure S7: **Probe of how Omni-LLM attention, capped with TMA, changes.** The highlighted blots remain in the prompt-relevant area (*e.g.*, upper-left, near where the cat's paws are; less in the upper-left background regions) as γ decreases.

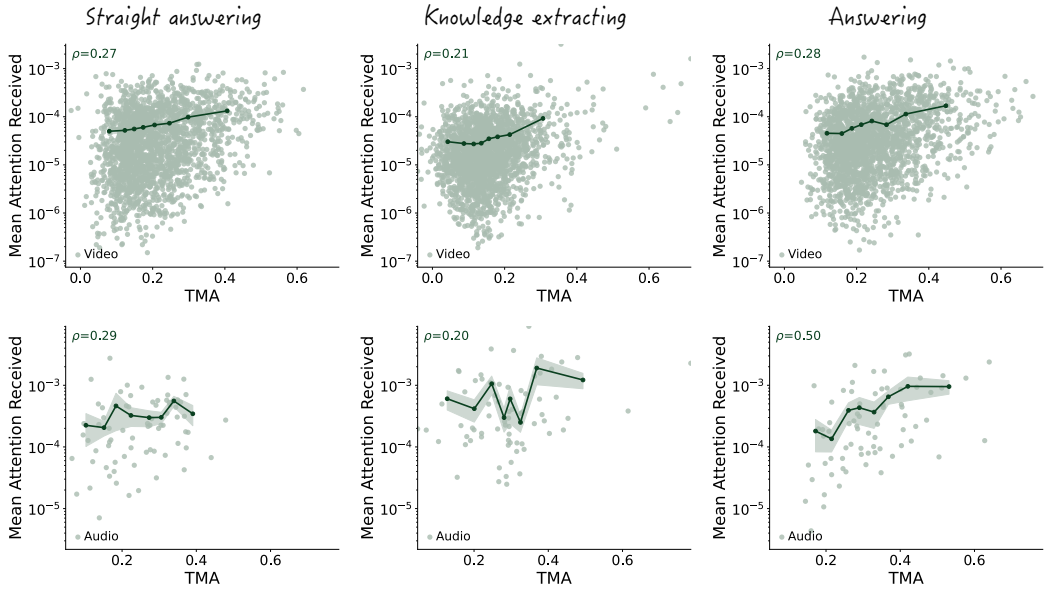
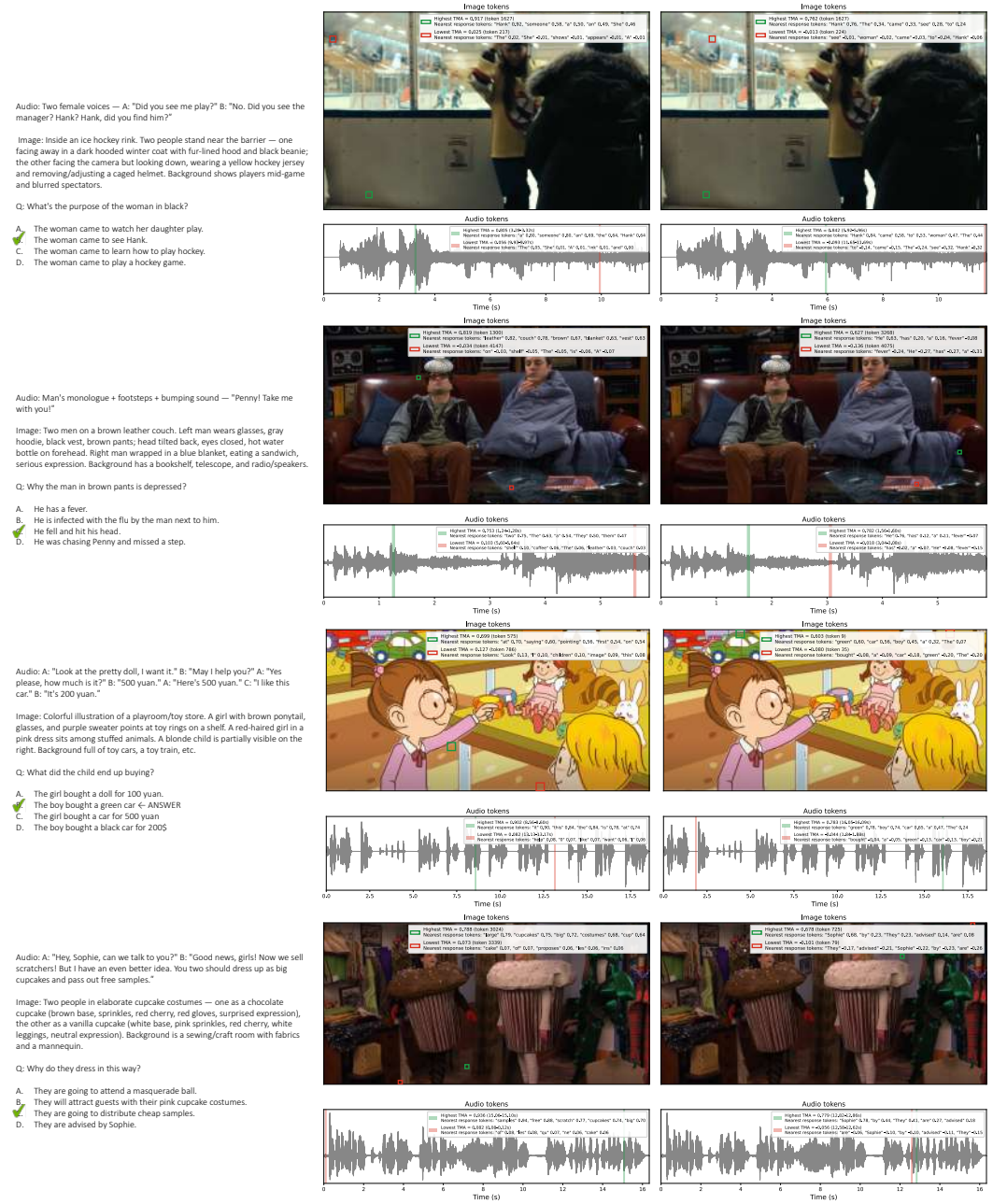


Figure S8: **TMA reflects accessibility to the generation pathway.** Within samples, input modality tokens with higher TMA tend to receive greater attention from generated-token queries across steps.

We further inspect the modality token that is most strongly aligned with the generated response. For each sample, we identify the single highest-TMA token and decode it through its nearest response token. The accompanying benchmark questions and task-specific visualizations are included in Figures S9 and S10.

Across $N = 120$ samples per task pooled over four Omni-LLMs, the selected token exhibits a comparatively balanced modality distribution for knowledge evaluation (Shannon entropy $H = 0.82$ bits out of a binary maximum of 1 bit), with $H \in [0.47, 1.00]$ across individual models. For reasoning, the corresponding pooled entropy is substantially lower ($H = 0.07$ bits, with $H \in [0, 0.21]$ across models). These examples are consistent with the view that successful reasoning may depend on a relatively localized piece of evidence that is explicitly verbalized, whereas perceptual knowledge can be distributed across multiple modalities and tokens. In this sense, the model appears to *know* more, diffusely, across modalities than its most response-aligned token alone would reveal during reasoning.



(a) Benchmark question (b) Knowledge task (c) Reasoning task

Figure S9: **Sample tokens extracted during knowledge and reasoning tasks for `Qwen-7B`.** Tokens with the highest TMA generally highlight key parts of the image/audio, while tokens with the lowest TMA point to semantically less meaningful regions/frames.

Modality-count analysis. We illustrate the knowledge and reasoning performance gap by modality count (OmniBench in Figure 1b; AVHBench in Figure S11). We consider one real modality (averaged across video/image-only and audio-only), two real modalities (video/image+audio), and two real modalities plus task-irrelevant text. Together, these conditions distinguish retaining perceptual content from using that content to obtain the correct answer.

Using the same setting as above, we also examine how TMA changes with additional modalities (Figure 4b). For each example, we compute TMA for every available input modality and take the

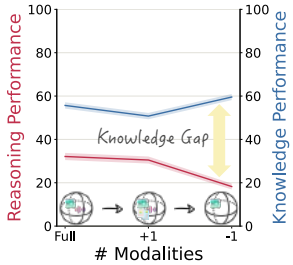
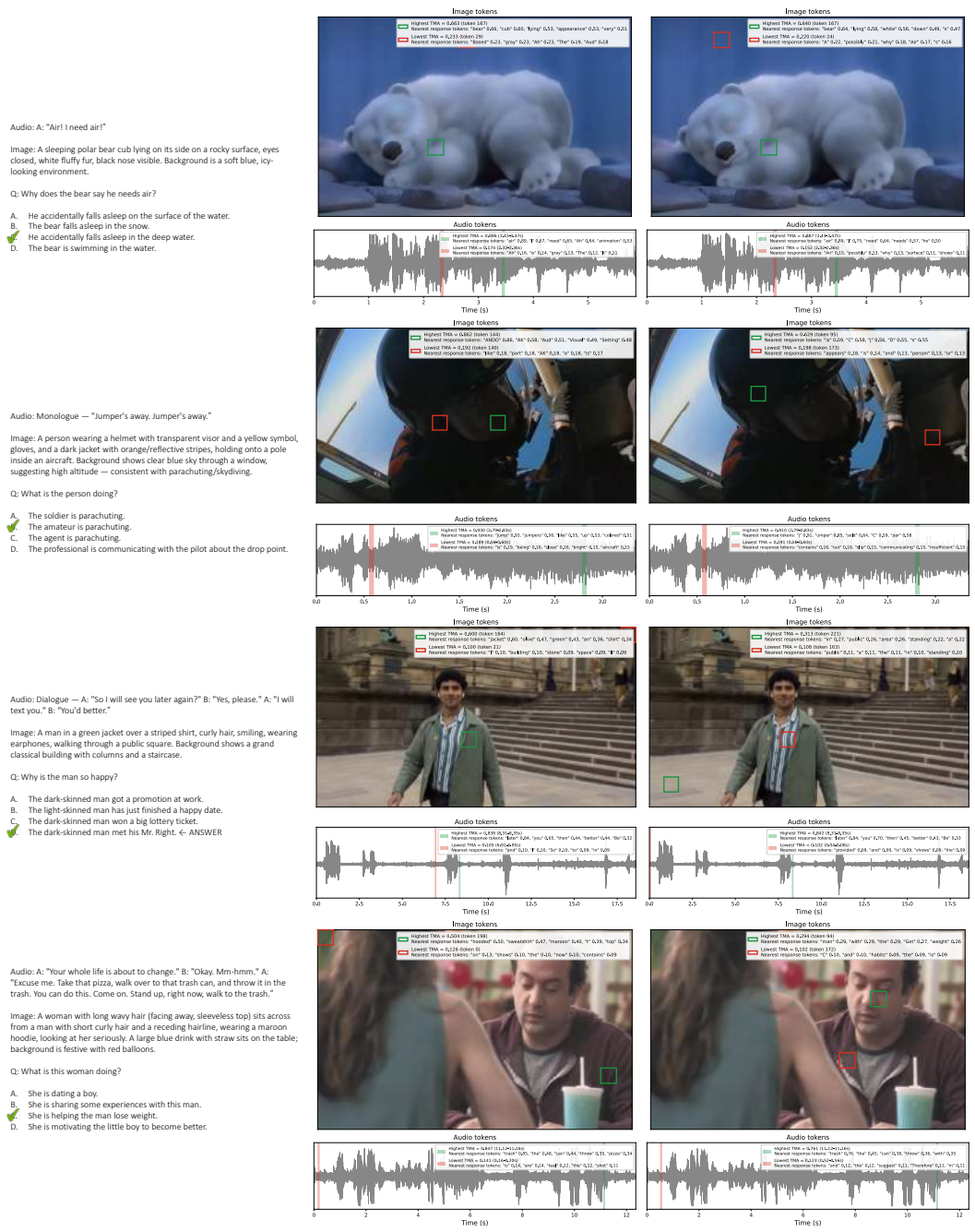


Figure S11: **Modality sensitivity.** The gap increases with the # of missing modalities.



(a) Benchmark question (b) Knowledge task (c) Reasoning task

Figure S10: **Sample tokens extracted during knowledge and reasoning tasks for `gemma-4B`.** Tokens with the highest TMA generally highlight key parts of the image/audio, while tokens with the lowest TMA point to semantically less meaningful regions/frames.

maximum value across modalities. We then average the scores across examples and models. The resulting curves therefore characterize how the strongest modality-to-generation alignment changes as additional inputs become available. The effect is most pronounced when the added modality is semantically relevant to the query. Because TMA measures alignment with the generated response rather than modality identity, a high score for an irrelevant modality can reflect a generic modality preference and should not be interpreted as evidence of semantic grounding.

Correctness of TMA. We validate whether TMA identifies modality tokens that are functionally relevant to generation through tiered masking (Figure 4a). For each sample, let $V_i = M_i^{\text{Visual}} \cup$

M_i^{Audio}, and partition its modality tokens into low-, mid-, and high-TMA tiers according to within-sample quantiles:

$$V_i^{\text{low}} = j : \text{TMA}_{i,j} \le q_\alpha^{(i)}, \quad V_i^{\text{mid}} = j : q_\alpha^{(i)} < \text{TMA}_{i,j} \le q_\beta^{(i)}, \quad V_i^{\text{high}} = j : \text{TMA}_{i,j} > q_\beta^{(i)}. \quad \text{(S5)}$$

For each tier, we zero out the selected representations at every decoder layer during prefill and regenerate the response. We then evaluate both downstream correctness and perceptual knowledge across Omni-LLMs. The mean TMA of the masked tokens $\overline{\text{TMA}}_i^{\text{tier}} = \frac{1}{|V_i^{\text{tier}}|} \sum_{j \in V_i^{\text{tier}}} \text{TMA}_{i,j}$, is used as the corresponding ablated performance. Masking high-TMA tokens causes a larger degradation in downstream performance, whereas perceptual knowledge remains comparatively stable. This separation is consistent with TMA capturing accessibility to the generation pathway rather than simply the presence of information in the input representation, capturing the "knows but cannot say" phenomenon.

Robustness of TMA. To assess robustness to response variation, we resample each model's response K times from the same fixed multimodal prompt and recompute TMA for each prompt token. The prefill pass is identical across samples, so modality-token representations remain fixed; only the decoded response varies. We quantify the stability of the continuous TMA values using the two-way random-effects, single-measure, absolute-agreement Intraclass Correlation Coefficient–ICC(2, 1) (Koo & Li, 2016):

$$\text{ICC}(2,1) = \frac{MS_R - MS_E}{MS_R + (K-1)MS_E + \frac{K}{N}(MS_C - MS_E)}, \quad \text{(S6)}$$

where MS_R, MS_C, and MS_E denote between-modality token, between-response token, and residual mean squares, respectively. A high $\text{ICC}(2,1)$ indicates that token-level differences in TMA are stable across sampled responses.

We assign each visual or audio token to the same low/mid/high TMG tiers used for masking. We then quantify the stability of their low/mid/high tier assignments across the K independent resamples using Fleiss' κ (Fleiss, 1971), $\kappa = \frac{\bar{P} - P_e}{1 - P_e}$, where $\bar{P}$ is the mean pairwise tier agreement per token and P_e is the chance agreement implied by the marginal tier frequencies.

F Additional Results

TMA robustness. Table S5 reports $\text{ICC}(2,1)$ for visual, audio, and random-text tokens and pooled κ for visual+audio tokens across $K = 5$ independently sampled responses and up to $N = 300$ OmniBench samples. TMA shows high consistency across response samples, with $\text{ICC}(2,1)$ of 0.76, 0.73, and 0.67 for visual, audio, and random-text tokens, respectively, in `gemma-4-e2b-it`, accompanied by $\kappa = 0.61$ for low/mid/high TMA tier agreement. The consistency is further improved for larger models, reaching $\text{ICC}(2,1) = 0.99$, 0.98, and 0.92 and $\kappa = 0.93$ for `Qwen3-Omni-30B`. Although k^* in Eq. 1 depends on the sampled response under the argmax formulation, these results indicate that the resulting TMA scores are largely stable across independently sampled responses rather than being driven by response-specific variation. In contrast, random-text tokens exhibit lower consistency, particularly for `Qwen2.5-Omni-7B` ($\text{ICC}(2,1) = 0.54$), as their TMA is unrelated to multimodal content and therefore depends more strongly on incidental lexical overlap with the sampled response.

Modality gap. Figure S12 compares three measures of modality alignment, expressed as modality gaps: centroid distance (CD), local 1-CKNNA, and 1-TMA. The comparison shows that existing mitigation methods do not consistently reduce all three measures simultaneously. In particular, randomly inserted text can reduce centroid distance and 1-TMA despite being semantically unrelated to the response, illustrating that representation proximity alone does not guarantee semantic correspondence. In contrast, its higher 1-CKNNA reflects the absence of semantically aligned neighboring samples.

Model	Vis.	Aud.	Txt.	$\kappa \uparrow$
`gemma-4-e2b-it`	76.4	72.7	67.3	60.8
`Qwen2.5-Omni-7B`	81.9	75.8	53.8	59.9
`Qwen3-Omni-30B`	98.9	97.9	92.4	93.0

Table S5: **Robustness of TMA to response sampling.** Omni-LLMs overall show low variation across K sampled responses, indicated by Vis./Aud./Txt. ICC(2,1) (↑) and κ, or agreement of low/mid/high TMA tier across the K responses (all scaled by 100).

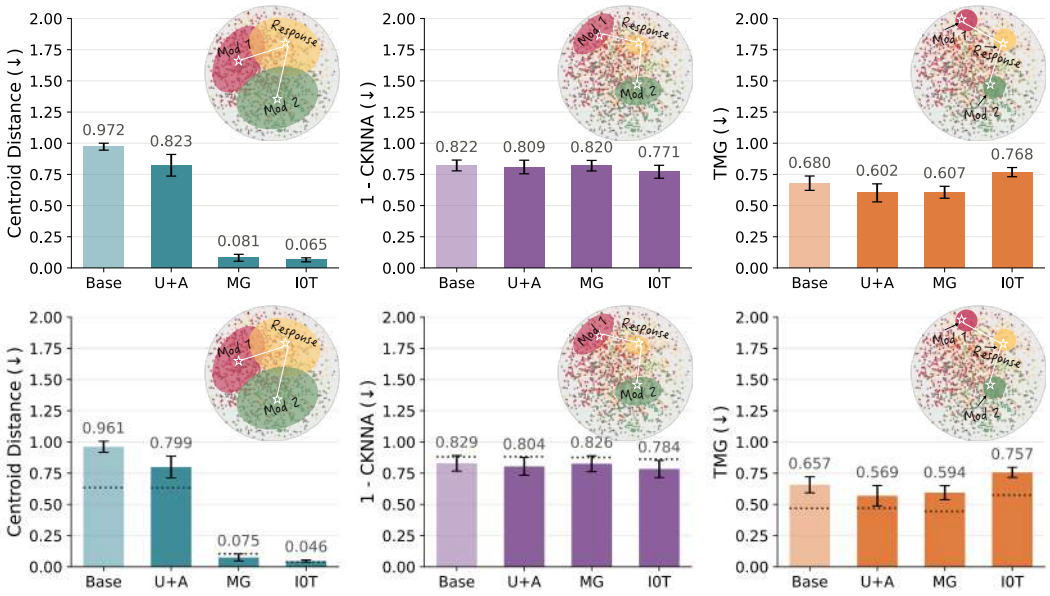


Figure S12: **Modality gap across mitigation methods.** We compare global CD, local 1-CKNNA, and 1-TMA without (top) and with (bottom) random descriptions. Dotted lines show the corresponding gap for randomly inserted text inputs. Prior modality-mitigation methods do not consistently reduce the gap across all three measures, highlighting that improved representational alignment under one measure does not necessarily imply alignment under others.

Figure S13 further relates modality gap to downstream performance. Across existing modality-gap mitigation methods, lower 1-CKNNA is associated with incremental performance improvements. Together, Figures S12 and S13 highlight the distinction between representational alignment and accessibility to generation.

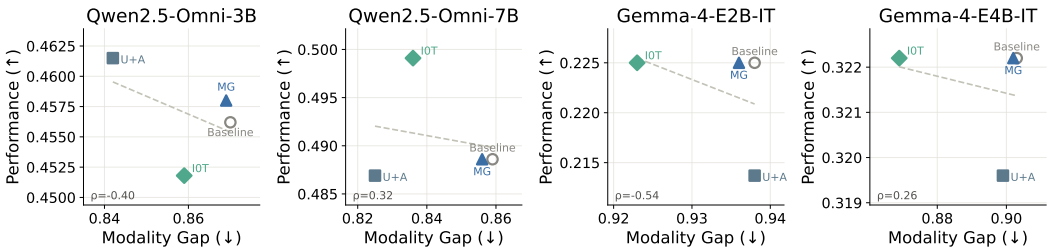


Figure S13: **Modality gap and performance.** Existing modality gap mitigation methods yield incremental performance gains as the modality gap, measured by (1-CKNNA), decreases.

Modality preference. Table S6 reports the detailed modality preference distributions for each model. Consistent with the main results, OMNI-CATALYST generally produces a more distributed modality preference than prior modality-gap mitigation methods, as reflected by higher normalized preference entropy.

Additional Omni-LLMs. Table S7 extends the main evaluation to additional Omni-LLMs. OMNI-CATALYST consistently improves performance across these models, with particularly large gains for `OmniVinci` (+11.5 percentage points on average). The larger `Qwen3-Omni-30B-A3B-Instruct` also benefits from OMNI-CATALYST, although the gain is smaller than that observed for the smaller models. These results indicate that the method is not limited to a particular model scale or architecture.

Training ablation. Table S8 reports per-model results for the different training strategies. Applying training-free OMNI-CATALYST at inference time improves trained models' performance, while incorporating modality-description generation into training provides additional gains. This comparison separates the benefit of the inference-time mechanism from that of learning the description capability.

Condition	Qwen2.5-Omni-3B				Qwen2.5-Omni-7B				gemma-4-e2b-it				gemma-4-e4b-it				*Avg.*			
	T	I	A	$\hat{H}$	T	I	A	$\hat{H}$	T	I	A	$\hat{H}$	T	I	A	$\hat{H}$	T	I	A	$\hat{H}$
Ideal	33.3	33.3	33.3	1.000	33.3	33.3	33.3	1.000	33.3	33.3	33.3	1.000	33.3	33.3	33.3	1.000	33.3	33.3	33.3	1.000
Baselines																				
No augmentation	23.4	56.1	20.5	0.900	8.6	59.1	32.3	0.807	24.8	51.1	24.1	0.939	21.3	61.3	17.4	0.849	19.5	56.9	23.5	0.892
MG	10.1	60.6	29.3	0.814	1.7	58.9	39.4	0.681	12.0	70.2	17.8	0.737	11.4	64.9	23.7	0.791	8.8	63.7	27.5	0.780
IoT	72.8	22.1	5.1	0.652	78.3	16.2	5.5	0.587	41.6	19.9	38.5	**0.959**	68.3	19.6	12.1	0.761	65.3	19.4	15.3	0.805
Self-CoT	22.5	45.1	32.5	0.965	9.2	58.8	32.0	0.816	20.4	53.8	25.8	0.917	20.1	62.5	17.3	0.837	18.1	55.0	26.9	0.902
Single-modality masking with its description																				
Mask image	1.1	87.7	11.2	0.372	0.5	92.1	7.4	0.269	1.8	85.0	13.2	0.436	3.4	80.7	15.9	0.528	1.7	86.4	11.9	0.409
Mask audio	1.9	7.7	90.5	0.329	3.0	9.4	87.6	0.405	7.5	39.7	52.8	0.818	12.6	31.1	56.2	0.863	6.3	22.0	71.8	0.678
Oracle descriptions																				
GT text only	33.3	48.4	18.3	0.936	11.7	59.6	28.7	0.835	38.8	39.7	21.5	0.969	29.2	58.8	12.0	0.844	28.2	51.6	20.1	0.930
GT image only	5.2	86.8	8.0	0.436	2.1	86.3	11.6	0.417	3.7	92.5	3.8	0.290	4.7	92.4	3.0	0.292	3.9	89.5	6.6	0.369
GT audio only	12.0	27.1	60.8	0.829	1.5	26.0	72.5	0.589	13.5	38.4	48.1	0.902	12.7	35.6	51.8	0.883	9.9	31.8	58.3	0.827
GT all	4.2	57.3	38.4	0.747	0.4	53.0	46.6	0.651	5.1	72.3	22.6	0.658	5.6	65.1	29.3	0.728	3.8	61.9	34.2	0.718
OMNI-CATALYST *(two-pass)*																				
Text only	32.4	49.6	18.0	0.930	13.6	56.9	29.5	0.867	44.6	35.5	19.9	0.955	32.0	54.7	13.3	0.877	30.6	49.2	20.2	0.942
Image only	15.7	66.0	18.3	0.797	7.5	64.1	28.4	0.761	12.1	64.0	23.9	0.804	13.0	73.1	13.9	0.700	12.1	66.8	21.1	0.777
Audio only	25.4	43.6	30.9	0.977	4.5	45.9	49.6	0.768	30.6	47.1	22.3	0.957	23.5	56.8	19.7	0.894	21.0	48.4	30.7	0.948
All	13.9	52.4	33.7	0.892	2.1	54.6	43.3	0.705	14.9	59.3	25.8	0.858	13.7	65.6	20.7	0.796	11.1	58.0	30.9	0.841
OMNI-CATALYST *(single-pass)*																				
Text only	27.0	41.6	31.4	**0.985**	10.9	46.2	42.9	**0.875**	20.6	49.4	30.1	0.942	25.2	55.9	18.9	0.898	20.9	48.3	30.8	0.948
Image only	23.9	47.9	28.1	0.957	8.2	61.7	30.1	0.787	11.3	46.0	42.7	0.881	30.5	51.9	17.5	0.917	18.5	51.9	29.6	0.922
Audio only	25.6	32.2	42.2	0.981	7.4	35.4	57.2	0.802	10.5	44.2	45.4	0.870	36.0	41.2	22.8	**0.974**	19.9	38.2	41.9	**0.959**
All	26.1	42.7	31.2	0.981	9.6	63.7	26.7	0.786	13.2	48.1	38.7	0.899	32.6	48.0	19.4	0.943	20.4	50.6	29.0	0.936

Table S6: **Modality preference distributions and balance.** For each model, we report the softmax-normalized preference mass over Text (T), Image (I), and Audio (A), together with the normalized entropy $\hat{H} = H/\log_2 3$, where $\hat{H} = 1$ denotes perfectly balanced preference. **Bold** and underline mark the highest and second-highest $\hat{H}$ per column among non-oracle conditions.

Method	Music	Sound Event	Speech	*Avg.*
`Qwen3-Omni-30B-A3B-Instruct`	53.8	60.0	48.4	51.6
OMNI-CATALYST	53.8$_{+0.0}$	58.9$_{-1.1}$	50.3$_{+1.9}$	52.6$_{+1.0}$
`omnivinci`	44.3	38.1	35.3	36.8
OMNI-CATALYST	57.6$_{+13.3}$	48.3$_{+10.2}$	47.0$_{+11.7}$	48.3$_{+11.5}$
`MiniCPM-o-2_6`	50.0	40.0	37.6	39.1
OMNI-CATALYST	48.1$_{-1.9}$	45.7$_{+5.7}$	39.8$_{+2.2}$	42.0$_{+2.9}$

Table S7: **OmniBench performance on additional Omni-LLMs.** OMNI-CATALYST improves performance across additional Omni-LLMs, with particularly large gains for `omnivinci`.

Method	gemma-2B	gemma-4B	Qwen-3B	Qwen-7B	*Avg.*
Baseline	21.4	27.1	43.8	46.5	34.7
OMNI-CATALYST	32.6$_{+11.2}$	40.5$_{+13.4}$	49.6$_{+5.8}$	51.7$_{+5.2}$	43.6$_{+8.9}$
Infer-only	22.0	32.1	44.7	48.1	36.7
OMNI-CATALYST	32.0$_{+10.0}$	39.8$_{+7.7}$	46.2$_{+1.5}$	52.8$_{+4.7}$	42.7$_{+6.0}$
Infer+Perceive	22.8	32.6	45.1	48.4	37.2
OMNI-CATALYST	32.1$_{+9.3}$	40.3$_{+7.7}$	44.0$_{-1.1}$	53.0$_{+4.6}$	42.4$_{+5.1}$
STaR (Zelikman et al., 2022)	22.6	31.7	44.9	48.9	37.0
OMNI-CATALYST	32.2$_{+9.6}$	40.1$_{+8.4}$	48.2$_{+3.3}$	52.1$_{+3.2}$	43.2$_{+6.1}$
R3V (Cheng et al., 2025)	22.6	32.0	45.5	48.2	37.1
OMNI-CATALYST	32.1$_{+9.5}$	40.2$_{+8.2}$	47.0$_{+1.5}$	52.2$_{+4.0}$	42.9$_{+5.8}$
SD-Zero (He et al., 2026)	22.6	32.2	45.2	49.1	37.3
OMNI-CATALYST	31.5$_{+8.9}$	40.4$_{+8.2}$	46.9$_{+1.7}$	52.4$_{+3.3}$	42.8$_{+5.5}$
Perceive-only	22.9	31.5	46.2	49.3	37.5
OMNI-CATALYST	31.6$_{+8.7}$	40.2$_{+8.7}$	48.5$_{+2.3}$	53.2$_{+3.9}$	43.4$_{+5.9}$
Perceive + STaR	22.2	33.0	45.9	49.1	37.6
OMNI-CATALYST	31.6$_{+9.4}$	40.4$_{+7.4}$	44.8$_{-1.1}$	51.3$_{+2.2}$	42.0$_{+4.5}$

Table S8: **Adaptability across training strategies.** Training-free OMNI-CATALYST consistently improves models trained with different objectives applied with our *perceive* → *infer* paradigm.

Table S9 further examines modality-description configurations for Omni-LLMs trained with the single-pass formulation across target modalities. Single-pass training generally improves the effectiveness of inference-time descriptions across trained modalities and description types. Among the two-turn configurations, audio descriptions provide the most consistent improvement, regardless of whether the model was trained with audio, visual, or all-modality descriptions. Visual or all-modality descriptions produce smaller or occasionally negative gains. This pattern is consistent with the main observation that audio is often the less accessible modality and therefore benefits particularly from an explicit textual description.

Transfer to audio captioning. Table S10 evaluates trained models on an out-of-distribution audio-captioning task. Perceive + STaR improves captioning performance for both Gemma and Qwen compared with both the baseline and perceive-only training. The improvement suggests that the learned ability to generate modality descriptions can transfer beyond the multiple-choice reasoning tasks used for training.

External description models. Table S11 evaluates descriptions generated by larger external Omni-LLMs. Using these external descriptions increases average performance from 34.2 to 39.8, indi-

Method	Trained Modality Desc	+Modality Desc	gemma-2B	gemma-4B	Qwen-3B	Qwen-7B	*Avg.*
ST OMNI-CATALYST	Audio	-	15.1	**27.5**	44.8	49.8	34.3
ST OMNI-CATALYST	Audio	Audio	29.3$_{+14.2}$	36.1$_{+8.6}$	**47.7**$_{+2.9}$	**55.9**$_{+6.1}$	**42.3**$_{+8.0}$
ST OMNI-CATALYST	Audio	Visual	30.1$_{+15.0}$	**38.9**$_{+11.4}$	42.2$_{-2.6}$	52.9$_{+3.1}$	**41.0**$_{+6.7}$
ST OMNI-CATALYST	Audio	All	**29.3**$_{+14.2}$	**35.7**$_{+8.2}$	**43.7**$_{-1.1}$	**53.9**$_{+4.1}$	**40.7**$_{+6.4}$
ST OMNI-CATALYST	Audio	Audio (two turns)	19.3$_{+4.2}$	26.4$_{-1.1}$	**48.1**$_{+3.3}$	51.0$_{+1.2}$	36.2$_{+1.9}$
ST OMNI-CATALYST	Audio	Visual (two turns)	15.8$_{+0.7}$	23.8$_{-3.7}$	41.7$_{-3.1}$	47.3$_{-2.5}$	32.2$_{-2.1}$
ST OMNI-CATALYST	Audio	All (two turns)	18.4$_{+3.3}$	22.9$_{-4.6}$	43.4$_{-1.4}$	48.7$_{-1.1}$	33.4$_{-0.9}$
ST OMNI-CATALYST	Perceive→Audio	-	2.2	3.9	45.4	50.3	25.5
ST OMNI-CATALYST	Perceive→Audio	Audio	18.0$_{+15.8}$	21.1$_{+17.2}$	42.3$_{-3.1}$	54.0$_{+3.7}$	33.9$_{+8.4}$
ST OMNI-CATALYST	Perceive→Audio	Visual	18.5$_{+16.3}$	26.3$_{+22.4}$	38.2$_{-7.2}$	**53.7**$_{+3.4}$	34.2$_{+8.7}$
ST OMNI-CATALYST	Perceive→Audio	All	12.9$_{+10.7}$	25.0$_{+21.1}$	39.6$_{-5.8}$	53.1$_{+2.8}$	32.7$_{+7.2}$
ST OMNI-CATALYST	Perceive→Audio	Audio (two turns)	8.5$_{+6.3}$	11.0$_{+7.1}$	47.0$_{+1.6}$	**51.8**$_{+1.5}$	29.6$_{+4.1}$
ST OMNI-CATALYST	Perceive→Audio	Visual (two turns)	1.5$_{-0.7}$	3.2$_{-0.7}$	41.9$_{-3.5}$	**49.4**$_{-0.9}$	24.0$_{-1.5}$
ST OMNI-CATALYST	Perceive→Audio	All (two turns)	6.3$_{+4.1}$	7.6$_{+3.7}$	43.6$_{-1.8}$	**49.7**$_{-0.6}$	26.8$_{+1.3}$
ST OMNI-CATALYST	Visual	-	18.6	27.4	44.5	48.2	34.7
ST OMNI-CATALYST	Visual	Audio	25.6$_{+7.0}$	**36.3**$_{+8.9}$	47.0$_{+2.5}$	53.6$_{+5.4}$	40.6$_{+5.9}$
ST OMNI-CATALYST	Visual	Visual	26.8$_{+8.2}$	37.7$_{+10.3}$	41.4$_{-3.1}$	50.2$_{+2.0}$	39.0$_{+4.3}$
ST OMNI-CATALYST	Visual	All	20.5$_{+1.9}$	34.8$_{+7.4}$	42.3$_{-2.2}$	50.4$_{+2.2}$	37.0$_{+2.3}$
ST OMNI-CATALYST	Visual	Audio (two turns)	20.0$_{+1.4}$	27.3$_{-0.1}$	46.2$_{+1.7}$	51.1$_{+2.9}$	36.2$_{+1.5}$
ST OMNI-CATALYST	Visual	Visual (two turns)	16.0$_{-2.6}$	22.6$_{-4.8}$	43.2$_{-1.3}$	45.6$_{-2.6}$	31.9$_{-2.8}$
ST OMNI-CATALYST	Visual	All (two turns)	18.3$_{-0.3}$	23.7$_{-3.7}$	44.8$_{+0.3}$	48.4$_{+0.2}$	33.8$_{-0.9}$
ST OMNI-CATALYST	Perceive→Visual	-	19.9	21.1	44.5	49.6	33.8
ST OMNI-CATALYST	Perceive→Visual	Audio	21.1$_{+1.2}$	32.1$_{+11.0}$	44.8$_{+0.3}$	54.0$_{+4.4}$	38.0$_{+4.2}$
ST OMNI-CATALYST	Perceive→Visual	Visual	24.5$_{+4.6}$	36.0$_{+14.9}$	39.5$_{-5.0}$	50.9$_{+1.3}$	37.7$_{+3.9}$
ST OMNI-CATALYST	Perceive→Visual	All	17.2$_{-2.7}$	33.3$_{+12.2}$	42.5$_{-2.0}$	51.2$_{+1.6}$	36.1$_{+2.3}$
ST OMNI-CATALYST	Perceive→Visual	Audio (two turns)	17.7$_{-2.2}$	22.9$_{+1.8}$	46.9$_{+2.4}$	50.5$_{+0.9}$	34.5$_{+0.7}$
ST OMNI-CATALYST	Perceive→Visual	Visual (two turns)	9.5$_{-10.4}$	18.0$_{-3.1}$	41.9$_{-2.6}$	47.5$_{-2.1}$	29.2$_{-4.6}$
ST OMNI-CATALYST	Perceive→Visual	All (two turns)	13.7$_{-6.2}$	15.7$_{-5.4}$	44.3$_{-0.2}$	46.8$_{-2.8}$	30.1$_{-3.7}$
ST OMNI-CATALYST	All	-	**22.9**	23.2	**47.6**	49.7	**35.9**
ST OMNI-CATALYST	All	Audio	**30.5**$_{+7.6}$	30.3$_{+7.1}$	45.7$_{-1.9}$	52.1$_{+2.4}$	39.7$_{+3.8}$
ST OMNI-CATALYST	All	Visual	**30.5**$_{+7.6}$	30.2$_{+7.0}$	40.0$_{-7.6}$	48.4$_{-1.3}$	37.3$_{+1.4}$
ST OMNI-CATALYST	All	All	28.7$_{+5.8}$	26.1$_{+2.9}$	42.7$_{-4.9}$	46.9$_{-2.8}$	36.1$_{+0.2}$
ST OMNI-CATALYST	All	Audio (two turns)	22.1$_{-0.8}$	**27.4**$_{+4.2}$	47.6$_{+0.0}$	49.8$_{+0.1}$	**36.7**$_{+0.8}$
ST OMNI-CATALYST	All	Visual (two turns)	**18.5**$_{-4.4}$	**26.2**$_{+3.0}$	45.2$_{-2.4}$	47.1$_{-2.6}$	**34.3**$_{-1.6}$
ST OMNI-CATALYST	All	All (two turns)	21.3$_{-1.6}$	**26.4**$_{+3.2}$	45.9$_{-1.7}$	47.5$_{-2.2}$	**35.3**$_{-0.6}$
ST OMNI-CATALYST	Perceive→All	-	18.6	21.0	46.6	**50.5**	34.2
ST OMNI-CATALYST	Perceive→All	Audio	24.3$_{+5.7}$	34.0$_{+13.0}$	43.3$_{-3.3}$	53.2$_{+2.7}$	38.7$_{+4.5}$
ST OMNI-CATALYST	Perceive→All	Visual	29.1$_{+10.5}$	29.9$_{+8.9}$	**44.8**$_{-1.8}$	49.0$_{-1.5}$	38.2$_{+4.0}$
ST OMNI-CATALYST	Perceive→All	All	21.5$_{+2.9}$	35.1$_{+14.1}$	42.8$_{-3.8}$	48.3$_{-2.2}$	36.9$_{+2.7}$
ST OMNI-CATALYST	Perceive→All	Audio (two turns)	**26.4**$_{+7.8}$	23.7$_{+2.7}$	47.8$_{+1.2}$	49.0$_{-1.5}$	**36.7**$_{+2.5}$
ST OMNI-CATALYST	Perceive→All	Visual (two turns)	16.9$_{-1.7}$	23.5$_{+2.5}$	**46.0**$_{-0.6}$	45.8$_{-4.7}$	33.1$_{-1.1}$
ST OMNI-CATALYST	Perceive→All	All (two turns)	**21.5**$_{+2.9}$	25.3$_{+4.3}$	**46.2**$_{-0.4}$	46.1$_{-4.4}$	34.8$_{+0.6}$

Table S9: **Self-trained (ST) OMNI-CATALYST ablation results.** Performance of self-trained OMNI-CATALYST under different trained and inference-time modality-description configurations. **Bold** and underline denote the best and second-best results, respectively, for each +Modality Desc setting across trained modalities.

Model	Method	BLEU	METEOR	ROUGE-L	CIDEr
gemma	Baseline	12.8	8.1	23.0	4.4
gemma	Perceive-only	12.7$_{-0.1}$	8.1$_{-0.0}$	23.0$_{-0.0}$	4.4$_{-0.0}$
gemma	Perceive + STaR	13.0$_{+0.3}$	8.2$_{+0.1}$	23.1$_{+0.1}$	4.6$_{+0.2}$
Qwen	Baseline	28.2	15.8	34.9	35.6
Qwen	Perceive-only	28.7$_{+0.5}$	16.0$_{+0.2}$	35.7$_{+0.8}$	36.5$_{+0.9}$
Qwen	Perceive + STaR	31.4$_{+3.2}$	16.3$_{+0.5}$	36.0$_{+1.1}$	37.1$_{+1.5}$

Table S10: **Out-of-distribution audio captioning.** Audio-captioning performance of models trained with different description objectives. Perceive + STaR improves over both the baseline and perceive-only training for both Gemma and Qwen.

cating that OMNI-CATALYST benefits from higher-quality modality descriptions. This experiment also separates the utility of the textual modality representation from the identity of the model that generates it.

Model	No Description	Qwen2.5-Omni-7B	Qwen3-Omni-30B
gemma-4-e2b-it	23.0	29.0$_{+6.0}$	29.3$_{+6.3}$
gemma-4-e4b-it	32.0	35.0$_{+3.0}$	37.7$_{+5.7}$
Qwen2.5-Omni-3B	39.3	45.3$_{+6.0}$	44.3$_{+5.0}$
Qwen2.5-Omni-7B	42.3	44.0$_{+1.7}$	48.0$_{+5.7}$
Avg.	34.2	38.3$_{+4.2}$	39.8$_{+5.7}$

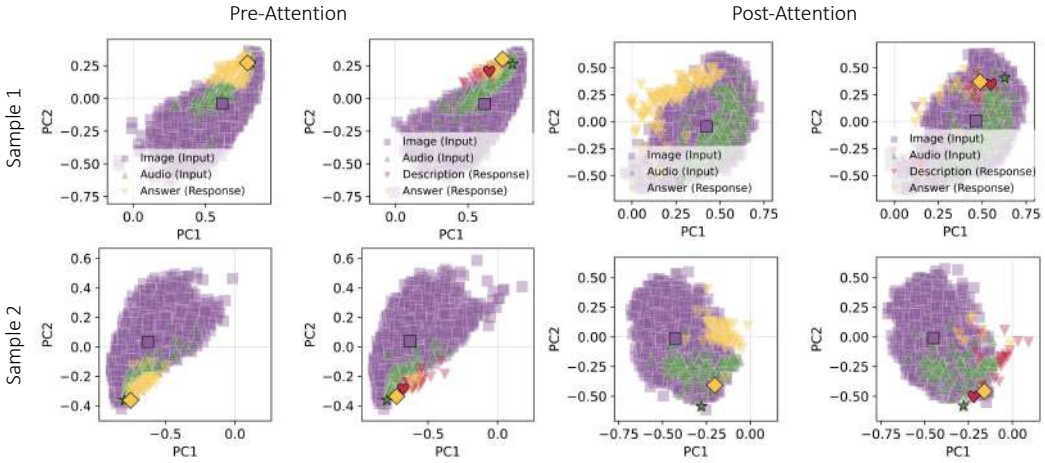
Table S11: **External model usage for modality description generation in OMNI-CATALYST.** The final omni-modal reasoning performance does not significantly differ when using modality descriptions generated by external models (spanned across columns) for the two-call scheme.

Thinker mode. Table S12 evaluates OMNI-CATALYST with thinker mode enabled for Gemma. Although the main Gemma experiments use non-thinker mode for consistency with Qwen, OMNI-CATALYST improves performance in both settings. Thus, the benefit of modality descriptions complements the model's internal thinking mode rather than depending on a particular decoding configuration.

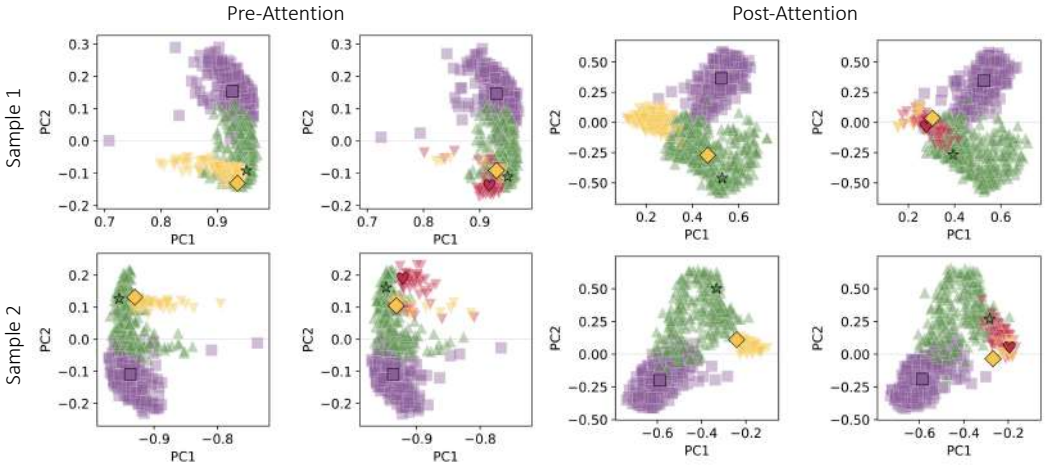
Model	No Thinker Mode		Thinker Mode	
	Baseline	OMNI-CATALYST	Baseline	OMNI-CATALYST
gemma-4-e2b-it	23.7	28.3$_{+4.6}$	36.1	37.3$_{+1.2}$
gemma-4-e4b-it	34.0	42.0$_{+8.0}$	42.0	42.3$_{+2.3}$

Table S12: **Effect of thinker mode on the Gemma family.** OMNI-CATALYST improves performance in both thinker and non-thinker modes, indicating that its benefit complements the model's thinking mode.

Hidden-state visualization. Figures S14a and S14b visualize PCA projections of token representations before and after attention in the final layer. For both `Qwen-7B` and `Gemma-4B`, the self-generated audio-description representations occupy a representation space closer to the audio and answer representations than to the image representations. This qualitative pattern is consistent with the description's role as a textual medium that makes audio information more accessible to the language-generation pathway.



(a) `Qwen-7b`



(b) `gemma-4b`

Figure S14: **PCA projections of token representations before and after attention in the last layer for (a) `Qwen-7b` and (b) `gemma-4b`.** Each point per plot corresponds to a token of the image, audio, answer, and self-generated audio description. In both models, the audio description representations share the same PCA coordinate space as the audio and answer representations, unlike the image.

Qualitative results. Figure S15 shows an example from trained OMNI-CATALYST. The knowledge-extraction stage places greater emphasis on audio evidence, which subsequently affects the refined answering stage. The resulting attention pattern is more balanced between visual and auditory evidence, illustrating how learned modality descriptions can alter the information available to downstream generation.

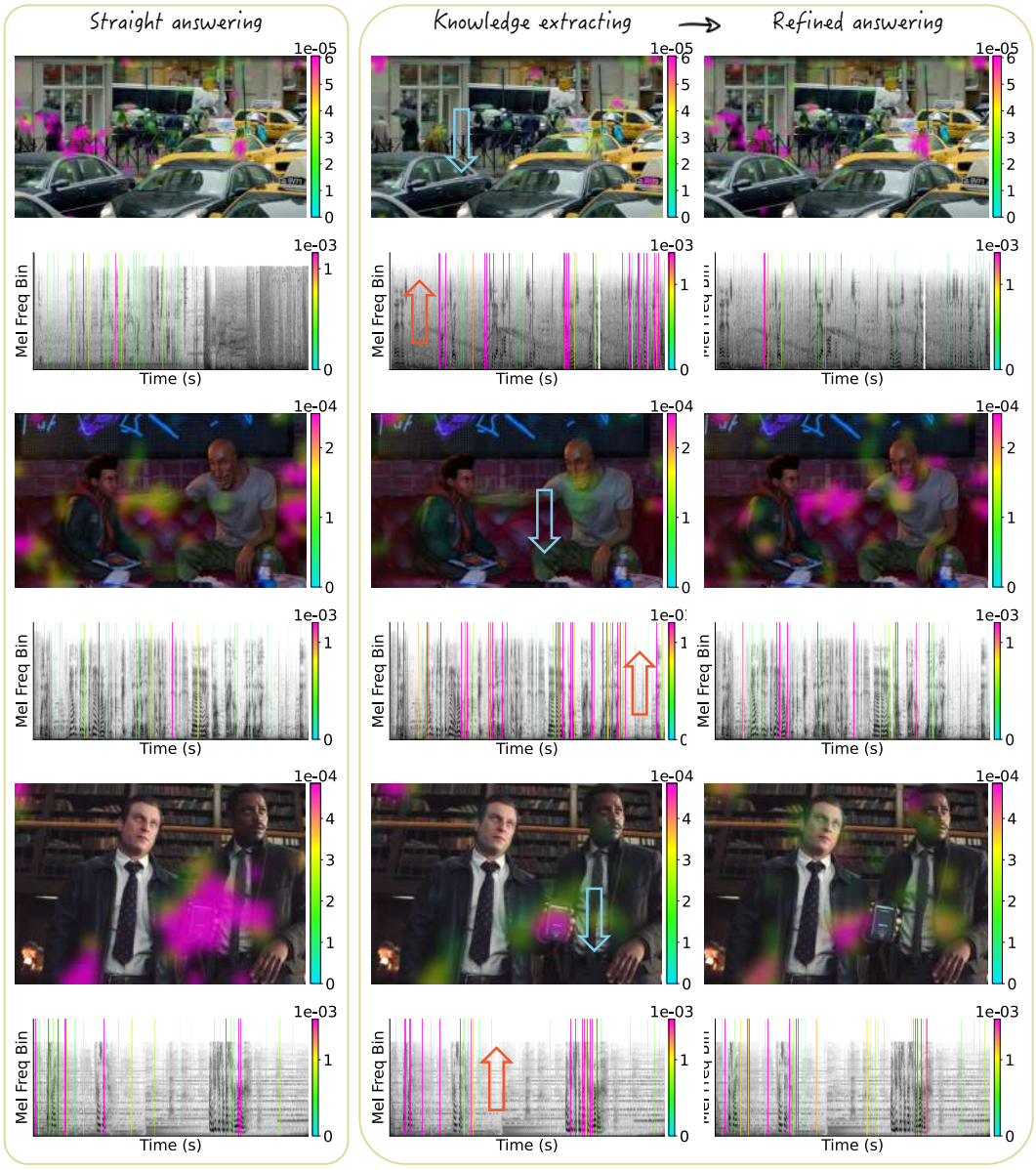
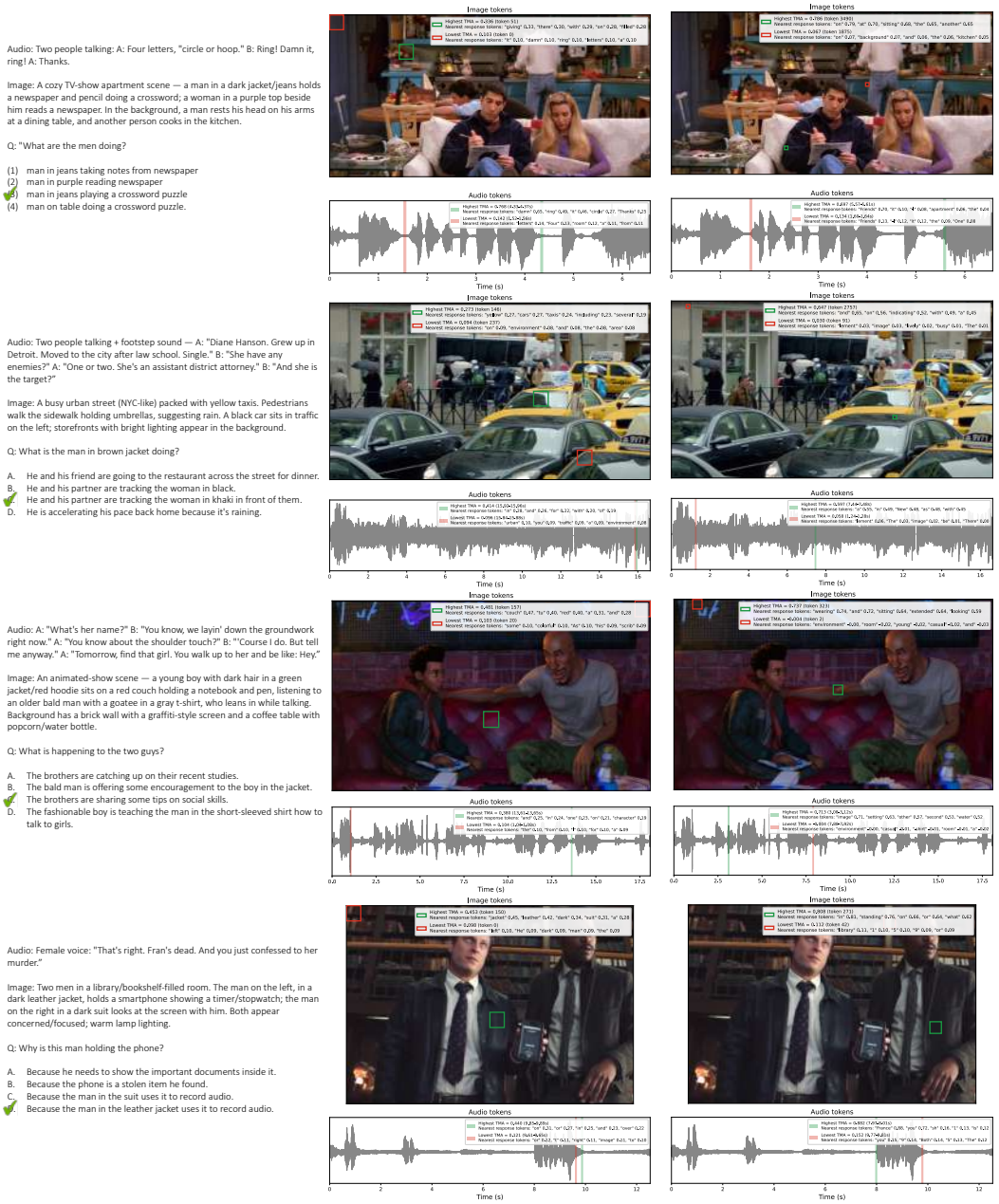


Figure S15: **Probe of how Omni-LLM attention changes using our TMA filter.** The knowledge-extraction step in *trained* OMNI-CATALYST enables the model to "hear" better, which in turn affects the refined answering stage, showing more balanced "seeing" and "hearing."

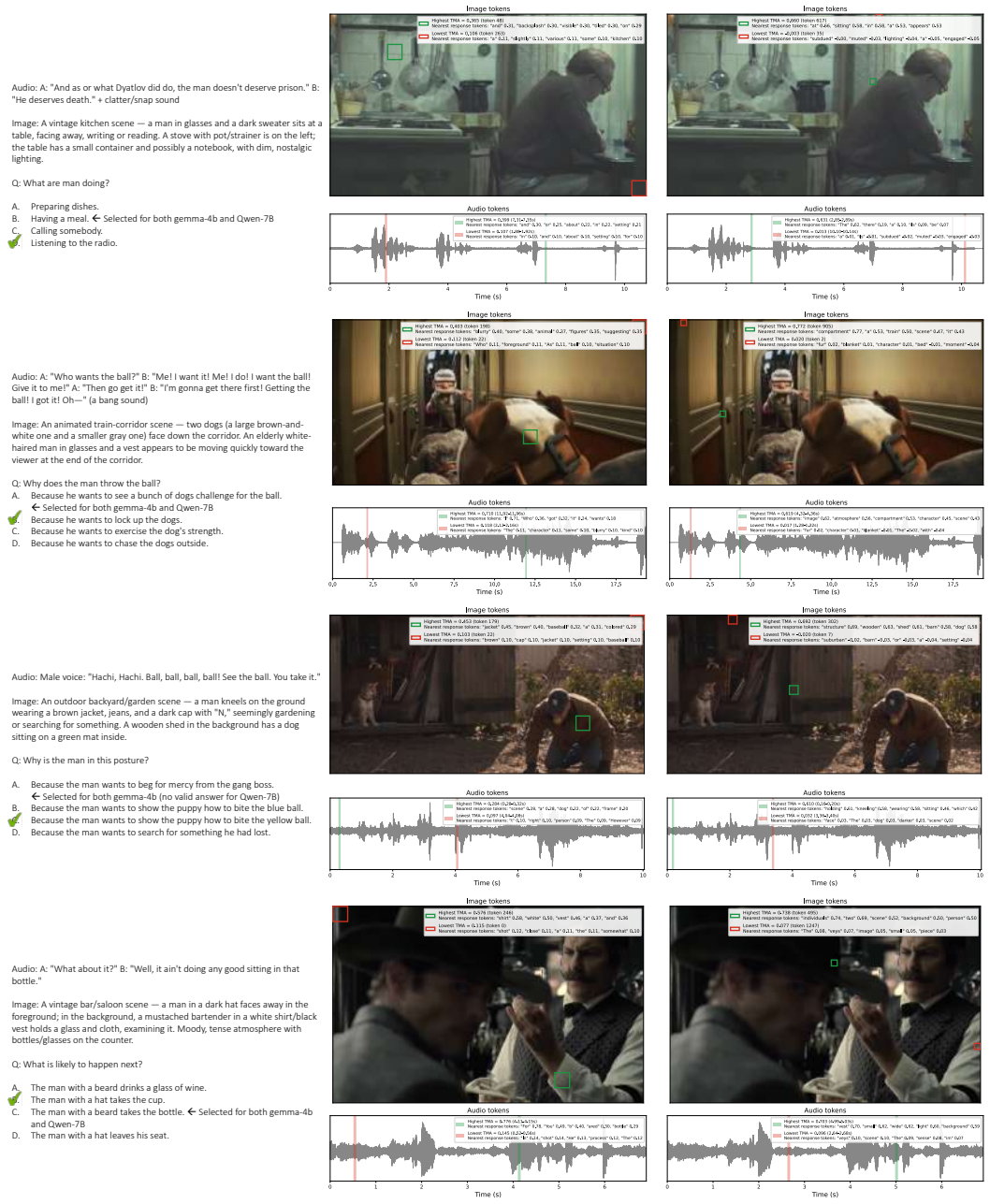
Figures **??** and **??** present additional successful and unsuccessful examples of training-free OMNI-CATALYST, together with the corresponding decoded TMA. In each panel, we highlight the image and audio tokens with the highest TMA values in red and those with the lowest values in green, using bounding boxes on video frames and vertical markers on audio waveforms. We also report their TMA values and the response tokens closest to them in embedding space. In the successful examples (Figure **??**), the lowest-TMA tokens consistently correspond to background image regions or silent audio segments, whereas the highest-TMA tokens correspond to visual regions or speech segments relevant to the correct answer. These include the newspaper and crossword-related speech in the first example, and the pedestrian walking behind the others and the dialogue identifying the target in the second. This spatial and temporal alignment suggests that token-level TMA reflects the relevance of the input evidence to the generated response. In the unsuccessful examples (Figure **??**), this correspondence is less apparent. When a question requires reasoning beyond directly observable evidence, such as inferring off-screen sounds, intentions, or future events, the highest-TMA token

corresponds to a region or temporal segment with limited relevance to the answer. Examples include the wall or shirt rather than the man, and the bartender's sleeves or the man's hat rather than the cup. These cases illustrate the limitations of relying on a single maximally aligned token as sufficient evidence, as some questions require integrating information across multiple regions or modalities.



(a) Benchmark question (b) `gemma-4b` (c) `Qwen-7b`

Figure S16: **Successful qualitative samples.** While the lowest-TMA tokens fall on background or silence, the highest-TMA image and audio tokens land on the visual regions and speech segments the answer depends on.



(a) Benchmark question (b) `gemma-4b` (c) `Qwen-7b`

Figure S17: **Unsuccessful qualitative samples.** When the answer requires reasoning beyond direct perception, such as off-screen sounds, intent, or future events, the highest-TMA tokens spread onto weakly related regions and segments. In other words, the evidence is too diffuse to capture with a single most-aligned token.