Instruction Anchor: Dissecting the Mechanistic Dynamics of Modality Arbitration

Yu Zhang[1,2] **Mufan Xu**[3] **Xuefeng Bai**[1*] **Kehai Chen**[1,2] **Pengfei Zhang**[2]
Yang Xiang[2] **Min Zhang**[1,2]
[1]Harbin Institute of Technology, Shenzhen, China, [2]Peng Cheng Laboratory, Shenzhen, China
[3]Harbin Institute of Technology, Harbin, China
yuzhang2717@gmail.com, {baixuefeng,chenkehai}@hit.edu.cn

Abstract

Modality following is the ability to selectively leverage multimodal contexts based on user instructions. It is fundamental to the safety and reliability of multimodal large language models (MLLMs) in real-world deployments. However, the internal mechanisms governing this decision-making process remain largely under-explored. In this work, we investigate modality following from an information-flow perspective. Our findings identify instruction tokens as structural anchors that aggregate multimodal cues and mediate modality decisions: Shallow attention layers perform non-selective transfer of modality cues, forming a latent buffer, whereas deep attention layers selectively strengthen the instruction-compliant signal, with a sparse subset of attention heads driving this process. Targeted attention-head interventions further validate the functional specificity of these heads: blocking only 5% of the identified heads substantially degrades modality following while preserving general visual and language capabilities, whereas targeted amplification can restore failed modality-following samples by up to approximately 60%. Together, this work provides a mechanistic account of modality following and informs future efforts to improve how MLLMs integrate and utilize multimodal evidence under user instructions. [2]

1 Introduction

Multimodal instruction following (MIF) [5, 10, 48] has become a foundational capability for multimodal large language models (MLLMs) [1, 2, 58], enabling them to integrate information across different modalities to execute complex user directives [67, 60]. MIF is pivotal for real-world deployments, such as multi-turn dialogues [40], graphical user interface navigation [29], and embodied robotic control [59], where models are required to reliably follow user intent across heterogeneous modalities. Compared to conventional instruction following in large language models, which primarily focuses on complying with output-format constraints [5, 10, 37], MIF introduces an additional challenge: MLLMs are required to strictly use the modality evidence specified by the user instruction [17, 23], especially when different modalities provide conflicting information. We refer to this instruction-guided selection between competing modality signals as *modality arbitration*. Despite its significance, the internal decision-making process underlying this selective utilization remains a "black box", forming a major obstacle to diagnosing model failures and ensuring behavioral reliability.

In this work, we address this gap by dissecting the internal decision-making mechanisms of MIF through an information flow perspective. We begin by localizing the key attention pathways that transmit modality cues to the final prediction using Attention Knockout Analysis. To rigorously

*Corresponding author.
[2]All data and code are released at https://github.com/EchoDreamer/Modality-Following.

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

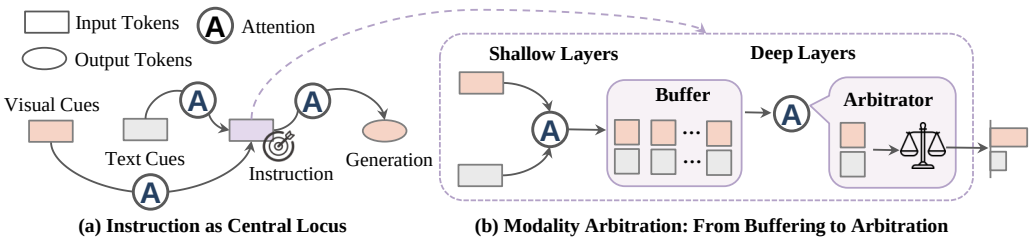


Figure 1: Information flow dissection for modality following. (a) Multimodal cues are preferentially aggregated at instruction tokens, which function as structural anchors. (b) Shallow attention accumulates both modality signals at instruction tokens with limited selectivity, forming a latent buffer. Deeper attention strengthens the instruction-compliant signal, resolving modality arbitration.

quantify the importance of each pathway for modality following while suppressing noise from task-irrelevant output variations, we propose Normalized Signed Structural Divergence ($\mathcal{I}_{\text{NSSD}}$), which measures probability shifts within the decision subspace spanned by the instruction-compliant and competing modalities. As illustrated in Fig. 1 (a), attention pathways that route modality cues into instruction tokens are far more prominent than those directly connecting cues to generated tokens, indicating that instruction tokens serve as the primary aggregation site for modality-relevant cues. Moreover, latent-decision tracking shows that the modality decision decoded at the instruction tokens aligns with the final prediction in over 95% of cases. This suggests that instruction tokens as a key mediation site for modality decisions and largely resolve modality arbitration.

Building on the observation that instruction tokens act as the primary aggregation site for multimodal cues, we further examine how modality arbitration unfolds inside the instruction tokens. Using a logit-difference attribution analysis, we find that, compared with MLP layers, attention exhibits a more stable positive shift in the arbitration margin toward the instruction-specified intent across both successful and failed modality-following samples. This suggests that attention is more closely associated with instruction-aligned modulation of the arbitration margin. Motivated by this observation, we further dissect attention across layers. As shown in Fig. 1 (b), shallow attention layers accumulates both modality signals at instruction tokens with limited selectivity, forming a latent buffer. Deeper attention more selectively strengthens the instruction-compliant signal, with a sparse subset of heads driving this arbitration. We perform targeted attention-head interventions to examine the functional specificity of the identified heads and their potential for effective modality following. As a result, blocking only 5% of the identified heads substantially degrades modality following while preserving general visual and language capabilities, supporting their functional specificity to modality-following behavior. Conversely, selectively amplifying these heads restores the modality-following ratio on failed cases by up to approximately 60%, demonstrating the role of these heads for targeted enhancement of modality following.

Our main contributions are summarized as follows:

- We investigate the under-explored internal mechanism of modality following from an information flow perspective.
- We identify instruction tokens as structural anchors for cue aggregation and decision mediation through attention-pathway analysis.
- We characterize a gradual transition from non-selective buffering to instruction-conditioned arbitration, driven by a sparse set of attention heads.
- This work proposes a new perspective for understanding modality arbitration and informs future efforts to improve multimodal evidence utilization in MLLMs.

2 Related Work

Multimodal language models (MLLMs) [2, 44, 22] have achieved remarkable success across a wide range of domains [47, 55, 46], demonstrating exceptional capabilities in integrating and reasoning over heterogeneous data [43, 53, 62, 32, 68]. Recent studies have highlighted the importance of multimodal instruction following (MIF) [5, 10, 48] for real-world deployments, where models are required to reliably follow user intent across heterogeneous modalities [40, 29, 65, 64, 24].

Multimodal Instruction Following (MIF). Research on MIF is fundamentally categorized into two dimensions: *1) Instruction-Driven Format Compliance* constrains the model's output format and has evolved from open-ended evaluation paradigms [5, 37] to rigorous benchmarks targeting complex, vision-dependent constraints [10, 18]. Correspondingly, enhancement efforts focus on scaling high-quality instruction-following data [7, 8] and applying preference-alignment strategies such as SFT and DPO to ensure strict adherence to structural output requirements [10, 18]; *2) Precise Context Utilization* focuses on the accurate synthesis of evidence from heterogeneous modalities guided by instructional intent [9, 17, 23]. Existing work has evolved from alignment-driven fine-tuning and rigorous evaluation [9, 17] to enhancing behavioral fidelity, which reveals challenges such as hallucinations in cross-modal understanding [23]. However, the internal mechanisms governing this decision-making process remain largely under-explored.

Interpretability in MLLMs. Existing literature on the mechanistic interpretability of MLLMs [3, 4, 11, 66, 20, 12] is predominantly anchored in a *perception-centric perspective*, focusing on the encoding, storage, and retrieval of visual information within the Transformer architecture [1, 41]. One trajectory focuses on pinpointing the specific neural topography responsible for multimodal processing, isolating modality-specific neurons [20, 36] or task-contingent sub-circuits [34] that disentangle cross-modal mechanisms. Another trajectory [3, 4, 52] interrogates the dynamic propagation of signals, employing causal interventions and attribution methods to trace the underlying information pathways. Concurrently, a burgeoning line of work seeks to decode semantic content by projecting activations onto human-understandable concepts through tools such as Sparse Autoencoders (SAEs) [28] or the Logit Lens [33, 50]. While perception-centric studies have advanced our understanding of MLLMs, the mechanisms governing cross-modal arbitration remain largely overlooked. This work elucidates the dynamics of modality-following by identifying instruction tokens as the critical structural locus for decision crystallization, providing a novel lens into multimodal information utilization.

3 Instruction Serves as Structural Anchor

We begin by investigating the pathways through which modality cues are aggregated and integrated to form decisions during modality following. Our findings identify instruction tokens as structural anchors that aggregate cross-modal cues (§3.2) and mediate the transmission of modality-decision information to generation (§3.3).

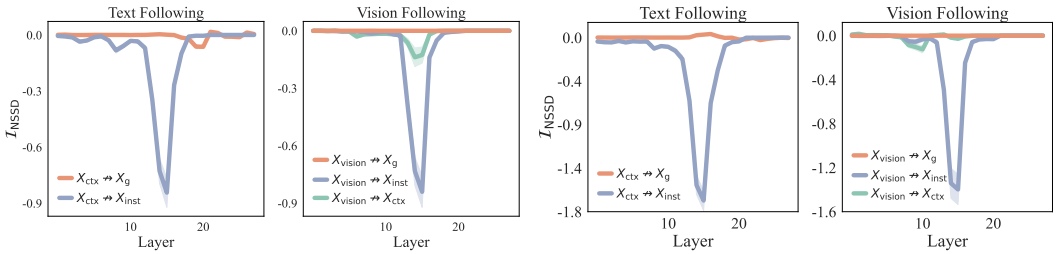
3.1 Diagnostic Setup

For mechanistic readout, we construct a controlled diagnostic setting based on [54], where visual and textual contexts support different answers. Each instance is abstracted as $S = \langle C_p, C_c, I, A_p, A_c, \mathcal{E}_p, \mathcal{E}_c \rangle$, where I denotes the active instruction specifying the target modality, C_p and C_c are the instruction-compliant and competing contexts, and A_p, A_c are the corresponding answers. The answer entity dictionaries $\mathcal{E}_p$ and $\mathcal{E}_c$ aggregate up to ten semantically equivalent surface forms to robustly track modality-specific signals.[3] The model produces an output Y conditioned on C_p, C_c and I, expected to align with C_p based on the intent I. We study decoder-only MLLMs built on the Transformer architecture. Given an input token sequence $X = [x_1, \ldots, x_N]$, we partition tokens into visual tokens (X_{vision}), textual context tokens (X_{ctx}), and instruction tokens (X_{inst}). Each token x_i passes through L Transformer residual layers:$\mathbf{h}_i^l = \mathbf{h}_i^{l-1} + \mathbf{A}_i^l + \mathbf{F}_i^l$, where $\mathbf{A}_i^l$ and $\mathbf{F}_i^l$ are the attention and MLP outputs, respectively. The answer dictionaries $\mathcal{E}_p$ and $\mathcal{E}_c$ define two sets of candidate-token unembedding directions, denoted by Y_p and Y_c. We use modality-arbitration subspace as operational terminology for these answer-defined sets. Let $\mathcal{V}_m$ be the token IDs corresponding to the verified surface forms in $\mathcal{E}_m$, using the first token for a multi-token expression. For a hidden state $\mathbf{h}_i^l$, we use Logit Lens [16] to read out the strongest activated answer direction:

$$\text{Logit}(Y_m \mid \mathbf{h}_i^l) = \max_{y \in \mathcal{V}_m} (\mathbf{E}\mathbf{h}_i^l)_y, \qquad m \in \{p, c\}, \tag{1}$$

where $\mathbf{E}$ is the unembedding matrix. This maximum-logit readout is used in the subsequent layer-wise and component-level analyses. Its comparison with an average-logit readout is reported in Apdx. E.1.

[3] Unless otherwise specified, our main analysis traces the internal routing and arbitration of successful modality-following instances, while failure cases are analyzed to localize the internal stage at which modality-following fails in §4.1.



(a) Qwen2.5-VL-7B (b) InternVL3-8B

Figure 2: **Layer-wise $\mathcal{I}_{\text{NSSD}}$ under attention knockout.** Each curve corresponds to removing a specific attention pathway, where X_{vision}, X_{ctx}, X_{inst}, and X_{g} denote visual tokens, text context tokens, instruction tokens, and generated tokens. More negative values of $\mathcal{I}_{\text{NSSD}}$ indicate greater importance of the pathway for modality following. Pathways targeting X_{inst} exhibit larger negative shifts than those to X_{g}, highlighting instruction tokens as the central convergence site for modality cues.

3.2 Information Routing Reveals Instruction Anchors

3.2.1 Method: Attention Knockout Analysis

Attention serves as the primary conduit for cross-modal information flow [30, 52]. The attention output $\mathbf{A}_i^l$ for i-th token at layer l is computed via multi-head self-attention:

$$\mathbf{A}_i^l = \sum_{j=1}^{J} \text{Head}_i^{l,j}, \quad \text{Head}_i^{l,j} = \text{Softmax}\Big(\frac{\mathbf{q}_i^{l,j}(\mathbf{K}^{l,j})^\top}{\sqrt{d_k}} + \mathbf{M}_i^l\Big)\mathbf{V}^{l,j} \cdot \mathbf{W}_O^{l,j}, \tag{2}$$

Here, $\mathbf{q}_i^{l,j}$ is the query vector of token i, $\mathbf{K}^{l,j}$ and $\mathbf{V}^{l,j}$ are the key and value projections from the previous layer, $\mathbf{W}_O^{l,j}$ is the output projection, and $\mathbf{M}^l \in \mathbb{R}^{N\times N}$ is the causal mask ($M_{s,t}^l = 0$ for $t \le s$, $-\infty$ otherwise). We employ Attention Knockout Analysis [15] to selectively remove edges in the attention graph and probe critical attention pathways. A *Target Pathway* $\mathcal{P}_{\text{src}\to\text{dst}}$ comprises edges from source token set X_{src} to destination set X_{dst}. Following [52], we block the target pathway by modifying the causal mask across all heads within a window of k layers centered at layer l. We set $k = 3$ unless otherwise noted. Sensitivity analyses in Apdx. C.1 confirm robustness across window sizes. The modified mask is:

$$\tilde{M}_{s,t}^l = e\begin{cases} -\infty, & (s,t) \in \mathcal{P}_{\text{src}\to\text{dst}}, \\ M_{s,t}^l, & \text{otherwise.} \end{cases} \tag{3}$$

3.2.2 Metric: Normalized Signed Structural Divergence

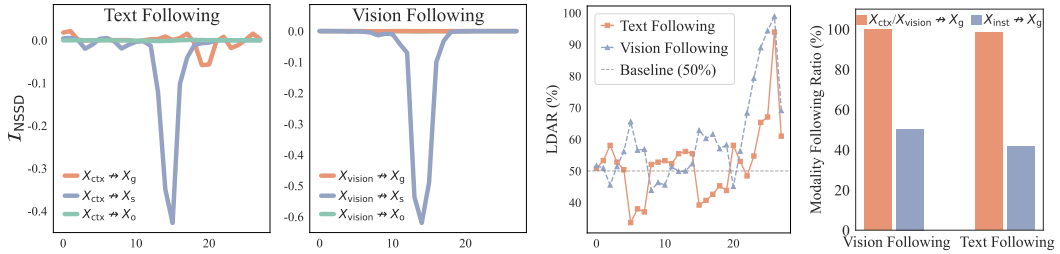
We quantify the effect of attention knockout through the signed change in the model's output distribution over the two answer-defined subspaces Let P and $\widetilde{P}$ denote the output token distributions before and after intervention. We denote by $P_{\mathcal{U}}$ and $\widetilde{P}_{\mathcal{U}}$ the corresponding subspace-normalized distributions. This restriction isolates the prediction components directly involved in modality following, allowing the metric to focus on intervention-induced changes in the arbitration-relevant decision structure. We define the Normalized Signed Structural Divergence as:

$$\mathcal{I}_{\text{NSSD}} := \text{sign}\Big[\widetilde{P}_{\mathcal{U}}(Y_p) - P_{\mathcal{U}}(Y_p)\Big]\, D_{\text{KL}}\Big(P_{\mathcal{U}} \,\Big\|\, \widetilde{P}_{\mathcal{U}}\Big), \tag{4}$$

where the sign term specifies the direction of the knockout effect on the instruction-compliant interpretation. The KL divergence term measures how strongly the intervention reshapes the prediction distribution within the modality-arbitration subspace. Intuitively, a negative sign means that removing the pathway shifts probability mass away from the instruction-compliant region, indicating that the blocked pathway originally supported modality following.

3.2.3 Results

In our experiments, we examine the information flow among visual tokens (X_{vision}), text context tokens (X_{ctx}), instruction tokens (X_{inst}), and generated tokens (X_{g}) to characterize how modal signals



(a) Comparison with semantic and format instruction (b) Layer-wise LDAR (c) Effect of Pathway

Figure 3: **(a)** Layer-wise $\mathcal{I}_{\text{NSSD}}$ under attention knockout, where X_{vision}, X_{ctx}, X_{g}, X_{s} and X_{o} denote visual, text context, generation, semantic instruction for modality following and output-format instruction tokens. **(b)** Layer-wise LDAR of instruction tokens on successful modality-following examples. Higher LDAR indicates greater alignment between latent state of instruction tokens and the final output. **(c)** Modality following ratio after cutting attention pathways from instruction X_{inst} or instruction-compliant modal context ($X_{\text{vision}}/X_{\text{ctx}}$) to generated tokens ($X_{\text{g}}$). The lower the value, the more important the pathway.

are routed during modality following. For each layer, we compute the mean and standard deviation of $\mathcal{I}_{\text{NSSD}}$ across all evaluation samples. Several key findings emerge:

Instruction Tokens as a Central Locus. Fig. 2 reports the layer-wise $\mathcal{I}_{\text{NSSD}}$ profiles for Qwen2.5-VL-7B and InternVL3-8B. Several observations can be drawn from these results. **First**, cutting pathways from the instruction-compliant modal context to the generated tokens produces only marginal changes in $\mathcal{I}_{\text{NSSD}}$. This indicates that the generated tokens do not appear to rely primarily on direct attention to the modal cues when producing modality-following responses. **Second**, cutting the pathways from X_{vision} or X_{ctx} to the instruction tokens X_{inst} results in a pronounced negative shift in $\mathcal{I}_{\text{NSSD}}$. This suggests that the pathway from modal cues to instruction is the most sensitive routing component for maintaining the correct instruction following. **Third**, in vision-following cases, cutting the attention pathway $X_{\text{vision}} \rightarrow X_{\text{ctx}}$ produces a markedly smaller effect than $X_{\text{vision}} \rightarrow X_{\text{inst}}$. This contrast indicates that the main structural sensitivity in vision-following is not distributed uniformly over textual tokens. Instead, it is more concentrated around pathways involving instruction tokens. Taken together, these results reveal a cross-modal relay pattern: instruction tokens form a prominent convergence site that aggregates multimodal cues during modality following.

Instruction Convergence Is Not a Positional Artifact. To test whether convergence on instruction tokens is due to their proximity to generated tokens, we analyze two types of instruction tokens: modality-specifying semantic tokens (X_s) and output-format constraint tokens (X_o), e.g., "Answer the question with a single word," both located near the generated tokens. We selectively cut the attention pathway from modality cues to X_s or X_o in Fig. 3 (a). Cutting the pathway to X_o has minimal effect, while X_s substantially reduces modality following, indicating that convergence reflects semantic, not positional, factors. Additional experiments—including swapping X_s and X_o positions and inserting unrelated instructions (Apdx. C.2)—further support this conclusion.

Generalization and Robustness of Attention Routing. In Apdx. C.3, the routing analysis extends to larger dense models, an MoE model, and an additional diagnostic dataset, as well as a cross-attention architecture. These evaluations recover the instruction-centered routing pattern while revealing architecture-dependent differences. Captioning, subset, and instruction-style analyses further examine the scope and robustness of the routing results. Besides, the same routing pattern persists for longer contexts.

Takeaways. Instruction tokens serve as the primary aggregation site for modality-relevant cues.

3.3 Decoding Modality Arbitration in Instruction Tokens

The information routing analysis in §3.2 identifies instruction tokens as a structural locus where cross-modal cues converge. We next ask how the modality states represented at these positions evolve across layers and relate to the model's final modality decision.

3.3.1 Method: Internal Belief Tracking

To answer this, we track layer-wise evolution of the internal belief using Logit Lens [16] for instruction-compliant subspace Y_p and competing subspace Y_c. Formally, we define a readout function S_m to quantify the intensity of a specific subspace within instruction tokens:

$$S_m(\mathbf{H}^l_{\text{inst}}) := \frac{1}{K} \sum_{i \in \mathcal{T}^l_m} \text{Logit}(Y_m \mid \mathbf{h}^l_i), \quad m \in \{p, c\}, \tag{5}$$

where $\mathbf{H}^l_{\text{inst}}$ denotes the hidden states of the instruction tokens at layer l, $\mathcal{T}^l_m$ is the set of indices of the top-K instruction tokens with the highest logit activations in subspace m. This follows the *semantic sparsity* hypothesis, focusing on the peak activations that represent the crystallized intent [27]. In practice, we set $K = 1$ and robustness analyses for varying K are provided in Apdx. E.

3.3.2 Metric: Latent Decision Alignment Rate

We use the Latent Decision Alignment Rate (LDAR) to measure alignment between the instruction-token readout and the final modality decision on successful modality-following examples. Given a sample, we consider instruction tokens at layer l to be aligned with final output if the signal strength of the instruction-compliant subspace dominates its competitor. Let $\mathcal{D}_{\text{succ}}$ denote this successful subset. We define the LDAR at layer l using the readout function in Eq. 5:

$$\text{LDAR}(l) := \frac{1}{|\mathcal{D}_{\text{succ}}|} \sum \mathbb{I}\left[S_p(\mathbf{H}^l_{\text{inst}}) > S_c(\mathbf{H}^l_{\text{inst}})\right], \tag{6}$$

where the summation is taken over all samples in $\mathcal{D}_{\text{succ}}$ and $\mathbb{I}[\cdot]$ is the indicator function. LDAR $= 1.0$ means instruction tokens' internal states can correctly predict the final decision for all samples, while 0.5 indicates chance-level alignment. Because the final answer follows the requested modality on this subset, LDAR measures how often the intermediate readout favors the requested modality.

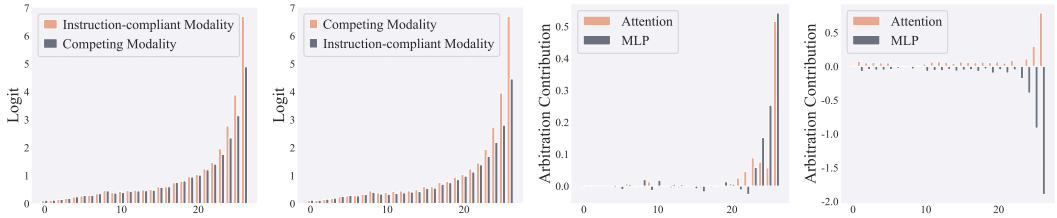
3.3.3 Result

As shown in Fig. 3 (b), LDAR remains around 0.5 in shallow layers, indicating the instruction-compliant subspace has not yet dominated. In deeper layers, LDAR rises sharply to over 95%, showing that modality arbitration is effectively finalized within the instruction tokens. We further examine the functional role through following three additional analyses:

Critical Role of Instruction-Mediated Information Flow. We assess the role of instruction-mediated information flow by cutting the pathways from instruction tokens (X_{inst}) or instruction-compliant modality cues to generated tokens (X_{g}) in the critical layers highlighted by the LDAR analysis. As shown in Fig. 3 (c), cutting the attention pathway from X_{inst} to X_{g} leads to a dramatic drop in modality-following performance, whereas cutting the attention pathway from the modality cues to X_{g} has little effect. This contrast supports a causal mediating role for the instruction-to-generation pathway in transmitting decision-relevant information.

Decision Synchronization between Instruction and Generated Tokens. To characterize how instruction-token states relate to generation, we measure the sample-wise agreement between the modality decisions decoded at X_{inst} and X_{g} on successful modality-following examples. Unlike LDAR, which measures alignment between the modality decisions decoded at instruction-token positions and the final correct modality decisions, this agreement measures alignment between the modality decisions decoded at instruction-token positions and generation-token positions, where the decoded decisions does not necessarily align with the final correct modality decision. We find that the agreement exceeds 90% across the critical layers in Fig. 3(b), even where LDAR remains below 70%. This shows that instruction tokens and generated tokens follow tightly synchronized modality-decision trajectories even before their intermediate readouts align with the final answer.

Activation-Patching Validation. To assess the functional relevance of the instruction-token readouts, we perform activation patching between paired successful runs with opposite modality instructions (See Apdx. E.3 for details). As shown in Table 11, patching instruction-token states at the LDAR-peak layer produces a larger mean logit-margin shift toward the donor modality (+1.08) than patching before (+0.28) or after (+0.41) the peak. At the peak layer, patching the top-five instruction positions selected by the Logit-Lens-based margin also yields a stronger shift than patching five random



(a) (b) (c) (d)

Figure 4: **Layer-wise component analysis.** (a,b) Layer-wise evolution of the instruction-compliant and competing subspace readouts for successful and failed modality-following samples, respectively. (c,d) Layer-wise contributions of attention and MLP to the modality arbitration margin for successful and failed samples, respectively. A positive contribution indicates that the corresponding component shifts the arbitration margin toward the instruction-specified intent.

positions (+0.87 versus +0.38). These results provide complementary intervention evidence that the readout identifies functionally relevant instruction-token states.

Takeaways. The convergence of modality cues, synchronization of decoded states, and sensitivity to instruction-to-generation pathway interventions identify instruction tokens as a key aggregation and mediation site for modality decisions. This role allows decision-relevant information to be integrated and relayed to generation.

4 Mechanistic Dissection of Modality Arbitration

We have identified instruction tokens as a key aggregation and mediation site for modality arbitration (§3). In this section, we investigate how instruction tokens mediate the arbitration process. The component analysis (§4.1) characterizes the transition from non-selective buffering to instruction-conditioned arbitration. At a finer granularity, we find the arbitration process is largely driven by a sparse subset of specialized attention heads (§4.1). Finally, the attention intervention experiments confirm that the identified attention heads are functionally specialized and can effectively modulate modality following (§4.2).

4.1 Component Analysis for Modality Arbitration

We quantify the behavior of attention and MLP for modality arbitration based on the readout function $S_m(\mathbf{H}^l_{\text{inst}})$ in Eq. (5), which measures the layer-wise logit intensity of subspace $m \in \{p, c\}$ at instruction tokens. Following the residual update, $\mathbf{H}^l_{\text{inst}} = \mathbf{H}^{l-1}_{\text{inst}} + \mathbf{A}^l_{\text{inst}} + \mathbf{F}^l_{\text{inst}}$ we also apply this readout to the intermediate states as $S_m(\mathbf{H}^{l-1}_{\text{inst}} + \mathbf{A}^l_{\text{inst}})$. Besides, we define the layer-wise modality arbitration margin as $\Delta S = S_p - S_c$, which provides a direct measure of the residual-stream signals relevant for modality arbitration within instruction tokens. A larger positive ΔS indicates that the residual-stream signal at instruction tokens is more aligned with the instruction-compliant modality than with its competitor. For convenience, let $\mathcal{M} \in \{S_p, S_c, \Delta S\}$ denote a generic readout. To this end, we track how attention and MLP modulate the logit intensity of *each subspace* or the *modality arbitration margin* with attribution of logit difference [31, 35]:

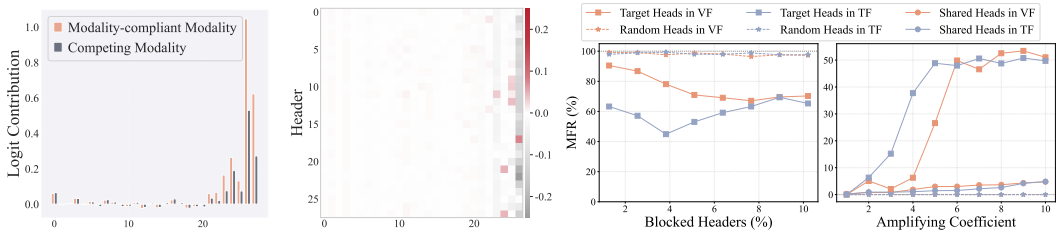
$$\delta^l_A(\mathcal{M}) := \mathcal{M}(\mathbf{H}^{l-1}_{\text{inst}} + \mathbf{A}^l_{\text{inst}}) - \mathcal{M}(\mathbf{H}^{l-1}_{\text{inst}}), \tag{7}$$

$$\delta^l_F(\mathcal{M}) := \mathcal{M}(\mathbf{H}^l_{\text{inst}}) - \mathcal{M}(\mathbf{H}^{l-1}_{\text{inst}} + \mathbf{A}^l_{\text{inst}}). \tag{8}$$

For robustness, we average these metrics across all samples. In the following, we focus on text-following results, as vision-following shows similar trends (see Apdx. D.1).

Emergence of Modality Decisions in Deep Layers. We first track the layer-wise subspace signals $S_m(\mathbf{H}^l_{\text{inst}})$ in Fig. 4(a,b). The instruction-compliant and competing signals are similar in shallow layers for both successful and failed examples. In deeper layers, their trajectories diverge: the instruction-compliant signal gradually dominates on successful examples, whereas the competing signal prevails on failed examples. This separation is consistent with the increasing LDAR on the successful subset in §3.3.3 and motivates the component-level analysis below.

Attention Modulates Modality Arbitration Toward Instruction Intent. Then we examine how attention and MLP modulate the modality arbitration margin using $\delta^l_A(\Delta S)$ and $\delta^l_F(\Delta S)$. As shown



(a) (b) (c)

Figure 5: **Attention analysis and validation of functional specialization.** (a) Layer-wise attention roles to the instruction-compliant and competing subspaces for successful text-following samples. (b) Layer-wise roles of individual attention heads to the modality arbitration margin for successful text-following samples. (c) Attention intervention analysis through attention head blocking (left) and amplification (right). We compare the Modality Following Ratio (MFR) under targeted, random, and shared-head interventions for Text Following (TF) and Vision Following (VF).

in Fig. 4 (c,d), attention consistently exhibits positive modulation of the modality arbitration margin, regardless of successful or failed modality-following samples, while MLP does not show such consistent behavior. This demonstrates that attention tends to help resolve modality arbitration based on instruction intent. These observations align with prior findings that attention in LLMs primarily orchestrates external context [13], whereas MLPs mainly encode internal parametric knowledge, which can occasionally counteract external signals [31, 57].

Attention: From Buffering to Arbitration. Given the role of attention in resolving modality arbitration toward instruction intent, we further track its modulation of the individual instruction-compliant and competing subspaces using $\delta_A^l(\mathcal{S}_p)$ and $\delta_A^l(\mathcal{S}_c)$ for successful modality-following samples in Fig. 5 (a). We find that in shallow layers, attention modulates both modalities to a similar extent. Correspondingly, its effect on the modality arbitration margin is near zero, indicating that shallow attention primarily performs undifferentiated information transfer, acting as a latent buffer. In deep layers, attention selectively amplifies the instruction-compliant modality relative to the competing modality, driving the resolution of modality arbitration, consistent with Fig. 4 (c).

Sparse Attention Head Arbitrators. For finer analysis, we further decompose $\delta_A^l(\Delta S)$ into individual attention heads $\text{Head}^{l,j}$ in Eq. (2):

$$\delta_A^{l,j} = \Delta\mathcal{S}(\mathbf{H}_{\text{inst}}^{l-1} + \mathbf{Head}_{\text{inst}}^{l,j}) - \Delta\mathcal{S}(\mathbf{H}_{\text{inst}}^{l-1}). \tag{9}$$

As shown in Fig. 5 (b), only a small subset of deep-layer heads exhibit strong modulation of the arbitration margin. While vision-following and text-following tasks activate largely distinct heads, a few top-ranking heads overlap, indicating a set of modality-shared arbitrators.

Takeaways. The component analysis characterizes a gradual shift from non-selective buffering of modality cues to instruction-conditioned arbitration. A sparse subset of attention heads makes strong positive contributions to the arbitration margin, motivating the targeted interventions in §4.2. Corresponding analyses in other MLLMs are provided in Apdx. D.2.

4.2 Validation of Functional Specialization of Attention Heads

Having identified a sparse set of attention heads as potential modality arbitrators, we evaluate their functional specialization to confirm the reliability of our analysis framework and demonstrate their effect on modality-following behavior.

Experimental Setup. For each attention head j at layer l, we manipulate its output at instruction tokens ($\text{Head}_{\text{inst}}^{l,j}$) in two ways. 1) *Attention Blocking*: The head's output is zeroed ($\text{Head}_{\text{inst}}^{l,j} = 0$) to test its necessity for modality-following behavior, while monitoring its impact on general language and vision capabilities. 2) *Attention Amplifying*: The head's output is scaled by a factor $\alpha > 1$ to intensify its signal, demonstrating the potential to effectively modulate modality-following behavior.

For each following direction, we rank attention heads by their mean attribution scores on successful modality-following examples and select the Top-G heads. The rankings correspond to the text-following analysis in Fig. 5(b) and the vision-following analysis in Fig. 12(b) in Apdx. D.1. The selected head set is fixed across evaluation prompts. We use the Modality Following Ratio (MFR) to

Table 1: Impact of blocking identified attention heads on general capabilities of Qwen2.5-VL-7B for vision understanding with TextVQA (Acc) and Flickr30k (METEOR/ CIDEr/ SPICE/ PPL); language understanding with MMLU (Acc) and IFEval (Acc). Higher is better except for PPL.

Method	TextVQA	Flickr30k	MMLU	IFEval
Qwen2.5-VL-7B	63.10	21.4/ 28.4/ 14.9/ 19.6	68.3	66.4
Block-#10	62.95	21.5/ 28.0/ 14.9/ 19.1	68.5	66.0
Block-#20	63.26	21.4/ 26.9/ 14.9/ 18.4	67.8	65.7
Block-#30	63.32	21.4/ 28.5/ 14.8/ 20.3	66.9	65.3

quantify the effects of these interventions. We compare four configurations as follows. 1) Original: Standard inference without intervention. 2) Targeted Heads: Interventions applied to the Top-G heads identified by our framework. 3) Random Heads: Interventions applied to an equal number of randomly selected heads. 4) Shared Heads: Interventions applied only to modality-shared heads. Blocking is evaluated on the successful examples used for head selection, with an original MFR of 100%, and measures the necessity of the selected heads on this subset. Amplification is evaluated on failed examples, with an original MFR of 0%. This complementary evaluation tests whether enhancing the same heads can recover modality following beyond the selection subset.

General capabilities are assessed on both visual and text understanding benchmarks. For vision understanding tasks, the performance is measured on VQA task, TextVQA [39] measured by Accuracy and image captioning task, Flickr30k [49] measured by METEOR, CIDEr, SPICE, and PPL. For language capabilities: we evaluate pure text understanding task, MMLU [19] measured by Accuracy and instruction following task, IFEval [61] measured by Accuracy.

Blocking Targeted Heads Suppresses Modality Following with Minimal Impact on General Capabilities. As shown in the left panel of Fig. 5 (c), progressively blocking targeted heads leads to a clear decline in MFR, whereas blocking the same number of randomly selected heads has minimal effect. Notably, blocking only the Top-40 heads (around 5% of total heads) causes a dramatic MFR drop (around 60% for text following), supporting the view that these heads play an important role in modulating modality-following behavior. Gradually increasing the number of blocked heads further highlights their coordinated effect. Blocking only the shared heads produces negligible changes, indicating that effective modality-following requires the modality-specific heads. Since around 30 blocked heads are sufficient to substantially reduce modality-following performance, we restrict interventions to this number. Table 1 shows that blocking these heads has a negligible impact on general vision or language tasks, including the instruction following task, IFEval. This demonstrates the functional specificity of the identified attention heads for modality following.

Targeted Amplification Enhances Modality Following. As shown in the right panel of Fig. 5 (c), amplifying the top 30 targeted heads substantially improves MFR, yielding an absolute gain of around 60%. By contrast, applying the same amplification to random or shared heads leads to smaller changes. Besides, MFR rises progressively with the amplification coefficient before reaching a plateau. The experiments in Table 9 in Apdx. E.4, indicate that excessive amplification (e.g., $\alpha = 18$) can impair modality-following by disrupting general language capabilities.

Targeted Attention-Head Fine-Tuning. A complementary adaptation experiment fine-tunes the fixed set of selected heads while keeping the rest of the model frozen. Evaluation on the held-out portion of a 7:3 train/test split shows higher modality-following accuracy than direct amplification or random-head fine-tuning (Apdx. D.3).

Takeaways. Taken together, these experiments confirm that the identified attention heads are functionally specialized and can effectively modulate modality-following behavior. This provides empirical validation for our mechanistic analysis of modality arbitration. We further verify that these patterns can generalize to other open-ended scenarios, such as modality interference tasks in Apdx. D.2. Finally, we also perform additional ablation studies on key design components S_m in Eq. (5) to assess the robustness of our diagnostic framework in Apdx. E.1.

5 Discussion

In this work, we adopt a cross-modal relay perspective to study how external multimodal contexts are integrated and leveraged, in contrast to prior studies that focus solely on information changes at

the generated tokens. Based on these findings, we highlight two potential directions: 1) **Efficient computation**: Insights from instruction-anchor–mediated context integration could inform context-compression strategies to achieve more efficient and effective context learning. 2) **Architectural design for omni-modal models**: Current architectures involve substantial redundant computation, particularly in attention pathways. Alternative designs that better align computation with context usage may improve capability ceilings under constrained resources.

6 Conclusion

This paper investigates the underlying mechanisms of modality following in MLLMs through the lens of information flow. We identify instruction tokens as structural anchors that aggregate competing multimodal cues and mediate their influence on generation. Our analysis reveals a functional stratification within the transformer architecture: shallow attention layers act as latent buffers, while deep attention layers resolve modality arbitration based on instruction intent, with a sparse subset of attention heads driving this process. Targeted blocking and amplification provide intervention-based functional validation of these findings, while the contrast between instruction-token-only and global blocking highlights the importance of intervention position. Together, this work provides an empirical mechanistic account of modality following and informs future efforts to improve how MLLMs integrate and utilize multimodal evidence under user instructions.

Acknowledgments and Disclosure of Funding

This work was supported in part by the New Generation Artificial Intelligence-National Science and Technology Major Project (2026ZD0127802), in part by National Key R&D Program of China (2024YFE0215300), in part by the Science Fund for Creative Research Groups of the National Natural Science Foundation of China (62521006), in part by the National Natural Science Foundation of China (62406091, 62276077, U23B2055, 62350710797), in part by Guangdong S&T Program (2024B0101050003), in part by the Guangdong Basic and Applied Basic Research Foundation (2026A1515011718, 2024A1515011205), in part by Guangdong Provincial Key Laboratory of Artificial General Intelligence and Large Models (2026B1212010005), in part by Major Key Project of PCL (PCL2025A12), and in part by the Shenzhen Science and Technology Program (KQTD20240729102154066).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 3, 18

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 2, 16, 19

[3] Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. *Advances in Neural Information Processing Systems*, 37:7400–7426, 2024. 3

[4] Gabriela Ben Melech Stan, Estelle Aflalo, Raanan Yehezkel Rohekar, Anahita Bhiwandiwalla, Shao-Yen Tseng, Matthew Lyle Olson, Yaniv Gurwicz, Chenfei Wu, Nan Duan, and Vasudev Lal. Lvlm-intrepret: An interpretability tool for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8182–8187, 2024. 3

[5] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. VisIT-Bench: A benchmark for vision-language instruction following inspired by real-world use. In *NeurIPS, Datasets and Benchmarks*, 2023. 1, 2, 3

[6] Rui Cai, Bangzheng Li, Xiaofei Wen, Muhao Chen, and Zhe Zhao. Diagnosing and mitigating modality interference in multimodal large language models. *arXiv preprint arXiv:2505.19616*, 2025. 21, 25

[7] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*, 2024. 3

[8] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions, 2023. 3

[9] Tianle Chen, Chaitanya Chakka, Arjun Reddy Akula, Xavier Thomas, and Deepti Ghadiyaram. Some modalities are more equal than others: Decoding and architecting multimodal integration in mllms. *arXiv preprint arXiv:2511.22826*, 2025. 3

[10] Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Mm-ifengine: Towards multimodal instruction following. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1099–1109, October 2025. 1, 2, 3

[11] 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, pages 21056–21064, 2026. 3

[12] Wanlong Fang, Tianle Zhang, Wen Tao, and Alvin Chan. Towards understanding modality interaction in multimodal language models via partial information decomposition. In *Forty-third International Conference on Machine Learning*, 2026. 3

[13] Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. Measuring the mixing of contextual information in the transformer. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. 8

[14] Xingyu Fu, Siyi Liu, Yinuo Xu, Pan Lu, Guangqiuse Hu, Tianbo Yang, Taran Anantasagar, Christopher Shen, Yikai Mao, Yuanzhe Liu, et al. Learning human-perceived fakeness in ai-generated videos via multimodal llms. *arXiv preprint arXiv:2509.22646*, 2025. 16

[15] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore, December 2023. Association for Computational Linguistics. 4

[16] Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 30–45, 2022. 3, 6

[17] Yuxin Guo, Shuailei Ma, Shijie Ma, Xiaoyi Bao, Chen-Wei Xie, Kecheng Zheng, Tingyu Weng, Siyang Sun, Yun Zheng, and Wei Zou. Aligned better, listen better for audio-visual large language models. *arXiv preprint arXiv:2504.02061*, 2025. 1, 3

[18] Weilei He, Feng Ju, Zhiyuan Fan, Rui Min, Minhao Cheng, and Yi R Fung. Empowering reliable visual-centric instruction following in mllms. *arXiv preprint arXiv:2601.03198*, 2026. 3

[19] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. 9

[20] Kaichen Huang, Jiahao Huo, Yibo Yan, Kun Wang, Yutao Yue, and Xuming Hu. Miner: Mining the underlying pattern of modality-specific neurons in multimodal large language models. *arXiv preprint arXiv:2410.04819*, 2024. 3

[21] Lingjie Jiang, Shaohan Huang, Yixia Li, Guanhua Chen, Dongdong Zhang, Furu Wei, et al. Viscodex: Unified multimodal code generation via merging vision and coding models. In *International Conference on Learning Representations*, volume 2026, pages 41476–41513, 2026. 16

[22] Yiming Lei, Guozhen Peng, Zeming Liu, Hui Qiu, Haitao Leng, Shaoguo Liu, Tingting Gao, Qingjie Liu, Annan Li, and Yunhong Wang. A3bench: an audience-aligned multilingual benchmark for video audience insights understanding. *Frontiers of Computer Science*, 20(9):2009377, 2026. 2

[23] Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *arXiv preprint arXiv:2410.12787*, 2024. 1, 3

[24] Boyi Li, Yifan Shen, Yuanzhe Liu, Yifan Xu, Jiateng Liu, Xinzhuo Li, Zhengyuan Li, Jingyuan Zhu, Yunhan Zhong, Fangzhou Lan, et al. Toward cognitive supersensing in multimodal large language model. *arXiv preprint arXiv:2602.01541*, 2026. 2

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 16

[26] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 16, 17

[27] Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Re, et al. Deja vu: Contextual sparsity for efficient llms at inference time. In *International Conference on Machine Learning*, pages 22137–22176. PMLR, 2023. 6

[28] Hantao Lou, Changye Li, Jiaming Ji, and Yaodong Yang. Sae-v: Interpreting multimodal models for enhanced alignment. *arXiv preprint arXiv:2502.17514*, 2025. 3

[29] Quanfeng Lu, Wenqi Shao, Zitao Liu, Lingxiao Du, Fanqing Meng, Boxuan Li, Botong Chen, Siyuan Huang, Kaipeng Zhang, and Ping Luo. Guiodyssey: A comprehensive dataset for cross-app gui navigation on mobile devices. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22404–22414, 2025. 1, 2

[30] Siyu Lu, Mingzhe Liu, Lirong Yin, Zhengtong Yin, Xuan Liu, and Wenfeng Zheng. The multi-modal fusion in visual question answering: a review of attention mechanisms. *PeerJ Computer Science*, 9:e1400, 2023. 4

[31] Ang Lv, Yuhan Chen, Kaiyi Zhang, Yulong Wang, Lifeng Liu, Ji-Rong Wen, Jian Xie, and Rui Yan. Interpreting key mechanisms of factual recall in transformer-based language models. *arXiv preprint arXiv:2403.19521*, 2024. 7, 8

[32] Jinlong Ma, Yu Zhang, Xuefeng Bai, Kehai Chen, Yuwei Wang, Zeming Liu, Jun Yu, and Min Zhang. Beyond unimodal shortcuts: Mllms as cross-modal reasoners for grounded named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 43518–43539, 2026. 2

[33] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*, 2024. 3

[34] Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. Same task, different circuits: Disentangling modality-specific mechanisms in vlms. *arXiv preprint arXiv:2506.09047*, 2025. 3

[35] Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8420–8436, 2024. 7

[36] Haowen Pan, Yixin Cao, Xiaozhi Wang, Xun Yang, and Meng Wang. Finding and editing multi-modal neurons in pre-trained transformers. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1012–1037, 2024. 3

[37] Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. MIA-Bench: Towards better instruction following evaluation of multimodal llms. In *ICLR*, 2025. 1, 3

[38] Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in vlms. *Advances in Neural Information Processing Systems*, 38:17929–17960, 2026. 16

[39] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 9

[40] Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. Multimodal dialogue response generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2854–2866, Dublin, Ireland, May 2022. Association for Computational Linguistics. 1, 2

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[42] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 16, 19

[43] Xi Xiao, Xingjian Li, Yunbei Zhang, Cheng Han, Tianming Liu, Tianyang Wang, Runmin Jiang, Jihun Hamm, Xiao Wang, and Min Xu. Layer-specific prompt fusion discovery via differentiable search in vision foundation models. In *European Conference on Computer Vision*, pages 391–409. Springer, 2026. 2

[44] Xi Xiao, Chen Liu, Chih-Ting Liao, Yunbei Zhang, Qizhen Lan, Yuxiang Wei, Lin Zhao, Janet Wang, Jianyang Gu, Muchao Ye, et al. Staying vigil ant: Mitigating visual laziness via counterfactual visual alignment in mllms. In *European Conference on Computer Vision*, pages 287–305. Springer, 2026. 2

[45] Xi Xiao, Chenrui Ma, Yunbei Zhang, Chen Liu, Zhuxuanzi Wang, Yanshu Li, Lin Zhao, Guosheng Hu, Tianyang Wang, and Hao Xu. Not all directions matter: Towards structured and task-aware low-rank model adaptation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2132–2154, 2026. 16

[46] Xi Xiao, Yunbei Zhang, Xingjian Li, Tianyang Wang, Xiao Wang, Yuxiang Wei, Jihun Hamm, and Min Xu. Visual instance-aware prompt tuning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2880–2889, 2025. 2

[47] Xi Xiao, Tianchen Zhao, Youngeun Kim, Zhuowei Li, Linghan Xu, Jiaye Wu, Zheng Zhang, Xiang Xu, Xuanbai Chen, Farhan Tejani, et al. Rethinking latent visual reasoning: Grounding latent reasoning in visual evidence. *arXiv preprint arXiv:2609.34563*, 2026. 2

[48] Zhiyang Xu, Ying Shen, and Lifu Huang. MultiInstruct: Improving multi-modal zero-shot learning via instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11445–11465, Toronto, Canada, July 2023. Association for Computational Linguistics. 1, 2

[49] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 9

[50] Zeping Yu and Sophia Ananiadou. Understanding multimodal llms: the mechanistic interpretability of llava in visual question answering. *arXiv preprint arXiv:2411.10950*, 2024. 3

[51] Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pages 1651–1678, 2024. 27

[52] Pingrui Zhang, Yifei Su, Pengyuan Wu, Dong An, Li Zhang, Zhigang Wang, Dong Wang, Yan Ding, Bin Zhao, and Xuelong Li. Cross from left to right brain: Adaptive text dreamer for vision-and-language navigation. *arXiv preprint arXiv:2505.20897*, 2025. 3, 4

[53] Yu Zhang, Kehai Chen, Xuefeng Bai, Zhao Kang, Quanjiang Guo, and Min Zhang. Question-guided knowledge graph re-scoring and injection for knowledge graph question answering. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 8972–8985, 2024. 2

[54] Yu Zhang, Jinlong Ma, Yongshuai Hou, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. Evaluating and steering modality preferences in multimodal large language model. *arXiv preprint arXiv:2505.20977*, 2025. 3, 16

[55] Yu Zhang, Chuyang Sun, Kehai Chen, Xuefeng Bai, Yang Xiang, and Min Zhang. Mitigating multimodal hallucination via phase-wise self-reward. *arXiv preprint arXiv:2604.17982*, 2026. 2

[56] Zhuoran Zhang, Tengyue Wang, Xilin Gong, Yang Shi, Haotian Wang, Di Wang, and Lijie Hu. Beyond fixed biases: Decoding the role of reasoning uncertainty in mllm modality conflicts. In *Forty-third International Conference on Machine Learning*. 20

[57] Jun Zhao, Yongzhuo Yang, Xiang Hu, Jingqi Tong, Yi Lu, Wei Wu, Tao Gui, Qi Zhang, and Xuanjing Huang. Understanding parametric and contextual knowledge reconciliation within large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 8

[58] Chen Zhe, Wu Jiannan, Wang Wenhai, Su Weijie, Chen Guo, Xing Sen, Zhong Muyan, Zhang Qinglong, Zhu Xizhou, Lu Lewei, Li Bin, Luo Ping, Lu Tong, Qiao Yu, and Dai Jifeng. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1, 16, 19

[59] Jinliang Zheng, Jianxiong Li, Dongxiu Liu, Yinan Zheng, Zhihao Wang, Zhonghong Ou, Yu Liu, Jingjing Liu, Ya-Qin Zhang, and Xianyuan Zhan. Universal actions for enhanced embodied foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22508–22519, June 2025. 1

[60] Xiangqing Zheng, Chengyue Wu, Kehai Chen, and Min Zhang. Locot2v-bench: A benchmark for long-form and complex text-to-video generation. *arXiv preprint arXiv:2510.26412*, 2025. 1

[61] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. 9

[62] Chunzheng Zhu, Jianxin Lin, Feng Wang, Cheng Jiang, Guanghua Tan, Zhenyu Zhou, Shengli Li, and Kenli Li. Subspace-guided semantic and topological invariant registration for annotation-free ultrasound plane quality control. *arXiv preprint arXiv:2605.25396*, 2026. 2

[63] Chunzheng Zhu, Lei Tian, Bohan Tan, Ziqi Zhou, Yuxuan Sun, Yijun Wang, Chengchao Lv, Yilin Wen, Yijun He, Jinghao Lin, et al. The path to self-evolving clinical systems: Scaling medical agents from assistance to autonomy. *arXiv preprint arXiv:2607.11175*, 2026. 16

[64] Chunzheng Zhu, Yijun Wang, Jianxin Lin, Feng Wang, Hongwei Wang, Lei Zhao, Shengli Li, and Kenli Li. Anatomy-anchored self-supervision: Distilling vision foundation models for invariant ultrasound representation. *MICCAI 2026 (Early Accept)*, 2026. 2

[65] Chunzheng Zhu, Yijun Wang, Jiaqi Zeng, Junyu Jiang, and Jianxin Lin. Medsynapse-v: Bridging visual perception and clinical intuition via latent memory evolution. In *European Conference on Computer Vision*, pages 559–577. Springer, 2026. 2

[66] Chunzheng Zhu, Jiaqi Zeng, Hongbo Zhao, Yihang Chen, Yijun Wang, and Jianxin Lin. Craft: Causal responsibility and failure tracing in medical vision language models, 2026. arXiv:2609.38810. 3

[67] Yingjie Zhu, Xuefeng Bai, Kehai Chen, Yang Xiang, Youcheng Pan, Xiaoqiang Zhou, and Min Zhang. Decoupling skeleton and flesh: Efficient multimodal table reasoning with disentangled alignment and structure-aware guidance. *arXiv preprint arXiv:2602.03491*, 2026. 1

[68] Yingjie Zhu, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. Benchmarking and improving large vision-language models for fundamental visual graph understanding and reasoning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30678–30701, Vienna, Austria, July 2025. Association for Computational Linguistics. 2

[69] Fei Zuo, Kehai Chen, Yu Zhang, Zhengshan Xue, and Min Zhang. Inimagetrans: Multimodal llm-based text image machine translation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20256–20277, 2025. 16

Appendices

Our supplementary materials are summarized as follows:

- Appendix A: Limitations, Use of LLM, Impact Statement and License of Assets.
- Appendix B: Detailed Diagnostic Setup.
- Appendix C: More Results for Attention Knockout Analysis.
- Appendix D: More Results for Mechanistic Dissection of Modality Arbitration.
- Appendix E: Ablation Studies and Robustness Analysis.

A Discussion

A.1 Limitations

This paper investigates the underlying mechanisms of modality following in MLLMs through the lens of information flow. Our analysis reveals a functional stratification within the transformer architecture: shallow attention layers act as latent buffers, while deep attention layers resolve modality arbitration based on instruction intent. However, we recognize that a more microscopic investigation into neuron-level activation patterns could potentially uncover even more fundamental principles of cross-modal arbitration. We leave this fine-grained circuit decomposition—transitioning from functional heads to atomic neurons—for future research to further refine the theoretical boundaries of multimodal integration.

A.2 Use of LLM

In this work, we leveraged DeepSeek-V3 [26] to curate and generate an answer entity dictionary, which served as the foundation for constructing our analysis dataset. Furthermore, we conducted a mechanistic interrogation of modality-following behaviors in several state-of-the-art MLLMs, including Qwen2.5-VL-7B [2], InternVL3-8B [58], InternVL3-14B [58] and DeepSeek-VL2-Tiny [42]. Additionally, large language models were employed to assist with grammatical refinement and linguistic polishing of the manuscript.

A.3 Impact Statement

This study elucidates the mechanisms of modality following in MLLMs, revealing how instruction-driven arbitration governs the resolution and prioritization of multimodal inputs. These findings may also inform the design of future MLLM-based applications that require reliable, instruction-guided integration of multimodal information [38, 14, 69, 63, 45, 21]. While these insights highlight potential vulnerabilities where safety filters might be bypassed, they primarily establish a structural foundation for developing more robust and transparent AI safeguards.

A.4 License of Assets

All images used are publicly available from COCO [25]. We release our analysis under a Creative Commons Attribution 4.0 License (CC BY 4.0) to enhance global accessibility and foster innovation and collaboration in research.

B Detailed Diagnostic Setup

We construct our analysis dataset based on MC^2 [54], which includes various visual and textual evidences corresponding to different answer candidates For each sample, we instantiate a modality-following sample by attaching an explicit modality following instruction that specifies which source the model should follow, e.g., "You should follow the textual context rather than the visual content." Under the abstraction used in the main text, this yields a modality-following sample:

$$S = \langle C_p, C_c, I, A_p, A_c, \mathcal{E}_p, \mathcal{E}_c \rangle,$$

where I denotes the active instruction specifying the target modality, C_p and C_c are the instruction-compliant and competing contexts, and A_p, A_c are the corresponding answers. The answer entity dictionaries $\mathcal{E}_p$ and $\mathcal{E}_c$ aggregate up to ten semantically equivalent surface forms to robustly track modality-specific signals. To support robust latent-signal tracing under surface-form variation, we construct the Answer Entity Dictionaries $\mathcal{E}_p$ and $\mathcal{E}_c$ from the canonical answers A_p and A_c based on lexical resources and LLMs [26]. The construction process consists of two stages:

1. Candidate Retrieval. We first use WordNet via NLTK to retrieve semantically related candidate entities for each canonical answer.

2. Semantic Verification. We then filter these candidates using DeepSeek-V3 [26] to ensure strict semantic alignment with the canonical answer in the sample-specific context. The semantic verification step uses the following prompt template:

Validate candidate answers using DeepSeek-V3

Instruction:
You are an expert English linguist helping to judge if a candidate word can be a synonym of a label in a specific context.
Your task: 1. You are given: - a natural language question (the context), - an original label (the target meaning), - a single candidate word to evaluate. 2. Decide if the candidate word can express approximately the same meaning as the label in this question's context. 3. The candidate word may differ in tense or part of speech (noun/verb/adjective/etc.) but should still preserve the core meaning. 4. Answer "yes" if the candidate word is a valid synonym/near-synonym in this context, "no" otherwise.
Output format: You MUST respond using the following tags: <answer>yes </answer> or <answer>no </answer>
<reason>brief English explanation of your judgment </reason>
Do NOT output anything other than these tags.
Question: {**question**}
Original label: {**label**}
Candidate word: {**candidate**}
Instructions: In the context of the Question, can the candidate word be considered a valid synonym or near-synonym of the Original label? Answer yes or no using the required tags.

Preliminary experiments further showed that intermediate hidden states are often decoded into Chinese tokens. To improve coverage in latent-space readout, we therefore augment each dictionary with Chinese candidate expressions generated by DeepSeek-V3 using the following prompt template:

Generate Chinese candidate answers using DeepSeek-V3

Instruction:
You are a Chinese linguist expert assisting in generating Chinese synonyms or related terms based on English vocabulary and specific contexts.
Your task:
You are given:
An English word (label),
An English question (the context).
Understand the specific meaning of the English word within the context of the question.
Generate a list of Chinese synonyms or related terms, ensuring that:
The Chinese terms accurately convey the core meaning of the English word in the given context.
They may include synonyms, near-synonyms, or related expressions.
Return only individual words or phrases; do not provide full sentences.
The quantity should be controlled between 5 and 15 terms.
Output format: You MUST respond using the following tag format: <chinese words >word1, word2, word3, ... </chinese words >
<explanation >A brief explanation in Chinese describing how these terms correspond to the English word's meaning </explanation >
Do NOT output anything other than these tags.
Question: {**question**}
Original label: {**label**}
Instructions: Generate a list of corresponding Chinese synonyms or related terms based on the meaning of this English word within the context of the question. Output the results using the required tag format.

To ensure data quality, we employ both LLM verification and manual verification methods: 1) Verification of LLM: We utilize ChatGPT-4o-mini [1] to verify whether each candidate word is appropriate for the given question; 2) Human Verification: We engaged three students in the field of artificial intelligence to conduct cross-validation of the candidate words. Only when all three agreed that the candidate word matched the question, did we consider it as a valid candidate.

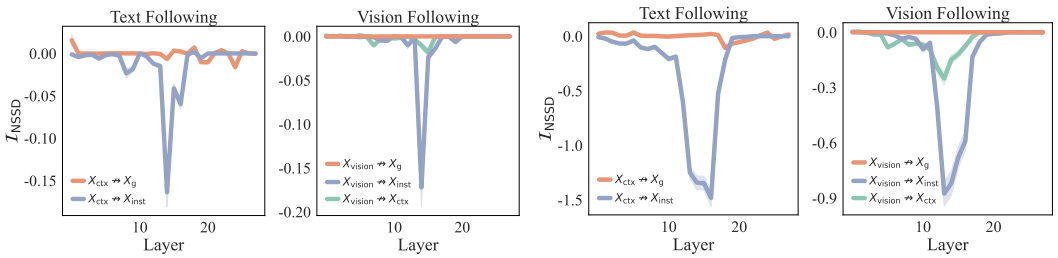
Finally, each canonical answer is ultimately associated with up to 10 bilingual synonym candidates. After filtering out samples with insufficient answer-set coverage, the final dataset contains 2,000 instances for analysis. We also verify that the analysis dataset is sufficient for robust modality-following analysis, the results in Fig. 9 in Apdx. C.3 demonstrate that consistent results are obtained when the analysis is conducted on smaller subsets and different instruction styles. All experiments are implemented on a single H20-80G GPU.

C More Results for Causal Attention Analysis

In this section, we provide more results for causal attention analysis including the results for sensitivity analysis of attention knockout windows in Apdx. C.1, additional evidence that instruction convergence is not a positional artifact in Apdx. C.2, more results for generalization and robustness of attention routing in Apdx.C.3 including more results for larger dense MLLM and MoE-based MLLM, the results of generalization to open-ended, multi-token generation tasks and the sensitivity analysis for dataset scale and instruction styles.

C.1 Sensitivity Analysis of Attention Knockout Windows

In §3.2.1, our primary analysis utilized a default attention knockout window 3. To ensure the robustness of our mechanistic findings and confirm that the observed trends are not artifacts of a specific window configuration, we conducted a sensitivity analysis across various window sizes. Specifically, maintaining the experimental setup described in Fig. 2, we report the $\mathcal{I}_{\text{NSSD}}$ results of Qwen2.5VL-7B for alternative window sizes in Fig. 6. We observe that while the absolute values of $\mathcal{I}_{\text{NSSD}}$ fluctuate slightly as the window expands, the fundamental conclusions regarding attention patterns remain consistent, further validating the stability of our method.



(a) Window Size 1 (b) Window Size 5.

Figure 6: **Layer-wise $\mathcal{I}_{\text{NSSD}}$ under attention-knockout for different window sizes.** Each curve represents cutting a attention pathway, where X_{vision}, X_{ctx}, X_{inst}, and X_{g} denote visual tokens, text context tokens, instruction tokens, and generated tokens. Cutting pathways to X_{inst} cause larger negative shifts than those to X_{g}, highlighting instruction tokens as the central convergence site for modality cues.

Table 2: Attention-knockout results for Qwen2.5-VL-32B. Txt and Vis denote textual and visual cues; Inst and G denote instruction and generated tokens. Blocking pathways to Inst produces larger negative NSSD shifts than blocking direct pathways to G.

Following	Pathway	L0–24	L28	L32	L36	L40	L44	L48–64
Text	Txt→G	0	0	-3	0	-1	0	0
Text	Txt→Inst	0	-18	-34	-46	-30	-9	0
Vision	Vis→G	0	0	-3	-1	-1	0	0
Vision	Vis→Inst	0	-25	-48	-65	-37	-11	0

C.2 Additional Evidence that Instruction Convergence Is Not a Positional Artifact

Beyond the evidence in Fig. 3 (a) of the main text, we present four experiments in a more compact form to validate that instruction tokens function as genuine semantic anchors rather than positional artifacts.

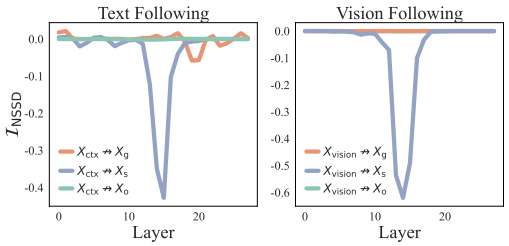
First, we examined the effect of swapping the positions of modality-specifying semantic tokens (X_s) and output-format constraint tokens (X_o) within the input sequence. knockout attention analysis was performed to measure the impact of cutting attention from modality cues to X_s or X_o. The results, shown in Fig. 7 (a,b), indicate that cutting attention to X_s substantially reduces modality-following performance, whereas cutting attention to X_o has only a marginal effect.

Second, to further rule out positional confounds, we introduced a longer general system prompt ("You are a helpful and honest assistant who strictly follows all user instructions and provides accurate responses ..."), placed either before or between X_s and X_o. knockout attention analysis. Here, the newly introduced system prompt and the original constraint tokens are both treated as X_o. As shown in Fig. 7 (c,d), exhibits the same pattern: cutting attention from modality cues to X_s significantly impairs modality following, whereas cutting attention to X_o has negligible effect. Collectively, These results confirm that the convergence on X_s cannot be explained by positional proximity alone.

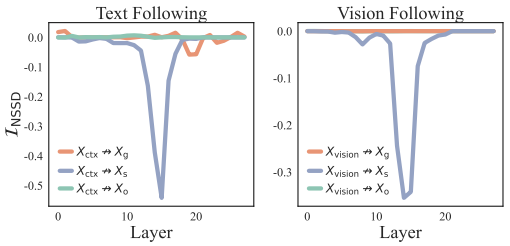
C.3 Results for Generalization and Robustness of Attention Routing

Attention-knockout analysis across model scales and architectures. We extend the attention-knockout analysis to larger dense models (InternVL3-14B [58] and Qwen2.5-VL-32B [2]), an MoE-based model (DeepSeek-VL2-Tiny [42]), and a cross-attention model (Llama-3.2-11B-Vision-Instruct).[4] For InternVL3-14B and DeepSeek-VL2-Tiny, Fig. 8 shows that blocking evidence-to-instruction pathways produces substantially larger negative $\mathcal{I}_{\text{NSSD}}$ shifts than blocking direct evidence-to-generation pathways, consistent with the smaller models in Fig. 2. Qwen2.5-VL-32B exhibits the same pattern in both following directions, with the largest contrasts in the middle-to-late layers (Table 2).

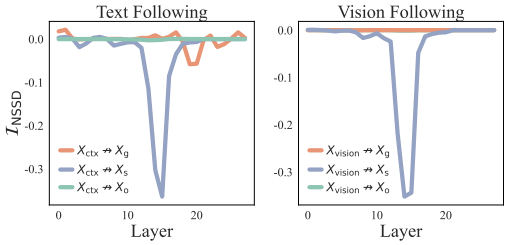
[4] https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct



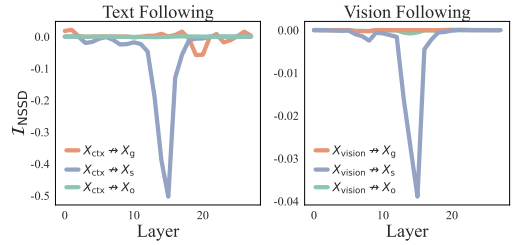
(a) Semantic instruction tokens (X_s) are placed before the output-format constraint instruction tokens (X_o).



(b) Semantic instruction tokens (X_s) are placed after the output-format constraint instruction tokens (X_o).

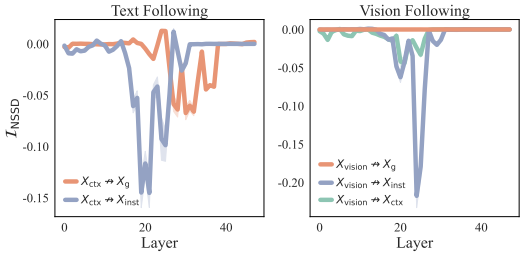


(c) A general system prompt is placed before the semantic and output-format instruction tokens. The system prompt and output-format tokens are grouped into X_o.

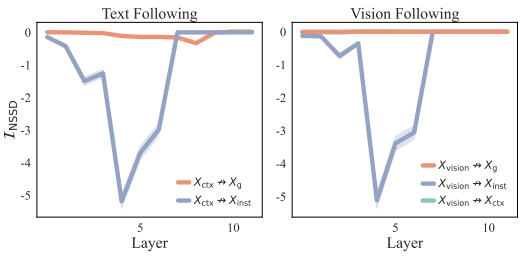


(d) A general system prompt is placed between the semantic and output-format instruction tokens. The system prompt and output-format tokens are grouped into X_o.

Figure 7: Additional knockout attention analyses verifying that modality-specifying semantic tokens serve as true semantic anchors rather than positional artifacts. Here, the newly introduced system prompt and the original constraint tokens are both treated as X_o.



(a) InternVL3-14B



(b) DeepSeek-VL2-Tiny

Figure 8: **Layer-wise $\mathcal{I}_{\text{NSSD}}$ profiles under different attention-knockout pathways for larger dense model and MoE-based model.** Each curve denotes cutting a pathway from a source token set to a target token set, where X_{vision}, X_{ctx}, X_{inst}, and X_{g} denote visual tokens, text context tokens, instruction tokens, and generated tokens, respectively. Across both models, pathways targeting X_{inst} produce substantially larger negative shifts than direct pathways targeting X_{g}, indicating that instruction tokens serve as a central convergence site for modality-relevant cues.

For Llama-3.2-11B-Vision-Instruct, visual information enters the language decoder through cross-attention. We therefore examine self-attention pathways from textual cues for text following and all eight cross-attention layers for vision following, separately blocking pathways to instruction tokens and generated tokens. Table 3 shows a similar instruction-centered routing pattern, particularly in the middle layers. However, the direct visual-to-generation pathway has a more noticeable effect than in the decoder-only models, indicating architecture-specific differences in routing. Together, these results extend the instruction-centered routing pattern to the evaluated larger dense, MoE-based, and cross-attention models.

Routing on an additional diagnostic dataset. To examine the dependence of the routing pattern on the diagnostic data, we perform attention-knockout analysis on a dataset drawn from [56]. As shown

Table 3: Attention-knockout results for Llama-3.2-11B-Vision-Instruct. Text-following interventions target self-attention pathways, whereas vision-following interventions target the eight cross-attention layers. Txt, Vis, Inst, and G follow Table 2. Instruction-directed pathways produce larger negative NSSD shifts, while direct Vis→G effects remain non-negligible in some cross-attention layers.

(a) Text following: self-attention

Pathway	L0–12	L16	L20	L24	L28	L32	L36–39
Txt→G	0	-1	-2	-1	-2	0	0
Txt→Inst	0	-14	-31	-38	-33	-17	0

(b) Vision following: cross-attention

Pathway	C3	C8	C13	C18	C23	C28	C33	C38
Vis→G	0	-1	-2	-10	-7	-8	-4	-2
Vis→Inst	0	-5	-23	-44	-37	-20	-7	-3

Table 4: Attention-knockout results with a neutral passage inserted before or after the instruction. F denotes filler tokens. Evidence-to-instruction pathways remain more sensitive to knockout than pathways to the filler or generated tokens in both conditions.

Position	Following	Pathway	L0–10	L12	L14	L16	L18	L20	L22–28
Before	Text	Txt→G	0	0	0	-1	0	0	0
		Txt→Inst	0	-18	-34	-42	-26	-9	0
		Txt→F	0	2	-4	-3	-2	0	0
	Vision	Vis→G	0	0	0	0	0	0	0
		Vis→Inst	0	-24	-52	-60	-34	-7	0
		Vis→F	0	-3	1	-2	-3	-3	0
After	Text	Txt→G	0	0	0	0	0	0	0
		Txt→Inst	0	-14	-29	-36	-20	-6	0
		Txt→F	0	1	0	-2	-3	0	0
	Vision	Vis→G	0	0	0	-1	0	0	0
		Vis→Inst	0	-20	-44	-52	-29	-8	0
		Vis→F	0	-5	1	-6	-2	-3	0

in Table 5, for both text and vision following, blocking evidence-to-instruction pathways produces pronounced negative NSSD shifts, whereas blocking direct evidence-to-generation pathways has little effect. This contrast is consistent with the instruction-centered routing observed on the primary diagnostic dataset.

Additional-context and recency controls. We examine the instruction-anchor effect under controlled context-length and recency perturbations by inserting a neutral passage of approximately 200 words before or after the instruction. The Before condition adds preceding context, whereas the After condition places unrelated text closer to generation than the instruction. Passages are sampled from PG-19/C4 validation data after filtering out answer entities, modality- or task-related terms, and instruction-like phrases.

As shown in Table 5, cutting pathways from modality evidence to instruction tokens produces substantially negative $\mathcal{I}_{\text{NSSD}}$, whereas cutting pathways to the filler or generated tokens has little effect. The persistence of this contrast in the After condition suggests that recency alone does not explain the instruction-anchor pattern.

Generalization to open-ended, multi-token generation scenarios In this section, we verify that the attention patterns observed at instruction anchors in §3.2.3 generalize to open-ended, multi-token generation tasks. To this end, we investigate their role in modality interference tasks [6]. Specifically, we conduct experiments on an image captioning task paired with a misleading text context. We selectively cut attention pathways from the vision context to the misleading text context, the instruction tokens, or the generated tokens, and measure the impact on generation quality using

Table 5: **Additional-context and recency controls.** A neutral passage is inserted before or after the instruction. Txt and Vis denote textual and visual cues; Inst, F, and G denote instruction, filler, and generated tokens. Evidence-to-instruction pathways exhibit larger negative NSSD shifts than pathways to filler or generated tokens in both conditions.

Position	Following	Pathway	L0–10	L12	L14	L16	L18	L20	L22–28
Before	Text	Txt→G	0	0	0	-1	0	0	0
		Txt→Inst	0	-18	-34	-42	-26	-9	0
		Txt→F	0	2	-4	-3	-2	0	0
	Vision	Vis→G	0	0	0	0	0	0	0
		Vis→Inst	0	-24	-52	-60	-34	-7	0
		Vis→F	0	-3	1	-2	-3	-3	0
After	Text	Txt→G	0	0	0	0	0	0	0
		Txt→Inst	0	-14	-29	-36	-20	-6	0
		Txt→F	0	1	0	-2	-3	0	0
	Vision	Vis→G	0	0	0	-1	0	0	0
		Vis→Inst	0	-20	-44	-52	-29	-8	0
		Vis→F	0	-5	1	-6	-2	-3	0

Table 6: Results on Flickr30k under misleading text interference by cutting the specific attention pathway. The performance is measured by METEOR, CIDEr and SPICE (Higher is better).

Method	METEOR	CIDEr	SPICE
Qwen2.5-VL-7B	20.0	18.6	13.5
Vision Context $\nrightarrow$ Text Context	19.6	18.5	13.3
Vision Context $\nrightarrow$ Generated Tokens	20.0	18.4	13.6
Vision Context $\nrightarrow$ Instruction	2.3	0.8	0.6

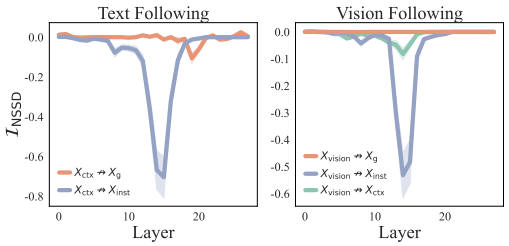
METEOR, CIDEr and SPICE. The interventions target layers 14–16, which were identified as critical for instruction-mediated information flow in Fig. 2 (a) for Qwen2.5-VL-7B.

As summarized in Table 6, cutting the attention pathways from the vision context to either the misleading text context or the generated tokens produces negligible changes across all metrics. In contrast, severing the pathways from the vision context to the instruction tokens consistently degrades performance across all measures. These results demonstrate that the attention patterns centered on instruction tokens effectively generalize to open-ended, multi-token generation tasks, highlighting the structural role of instruction anchors in guiding cross-modal information integration.

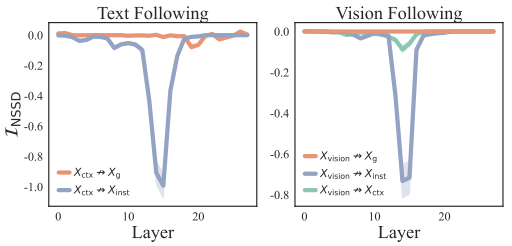
Sensitivity analysis for dataset scale. To verify that the current dataset scale is sufficient to robustly support our causal attention analysis, we repeat the main experiment in Fig. 2 using only 20%, 40%, 60% and 80% of the evaluation data, and compare the results with those obtained on the full dataset.

As shown in Fig. 9, the core attention patterns remain highly consistent across different data ratios. In particular, the overall layer-wise patterns are preserved even when only a fraction of the data is used, indicating that our findings do not depend on a specific sample scale. At the same time, as the amount of evaluation data increases, the curves exhibit lower variance, suggesting improved statistical reliability. These results provide empirical evidence that the current dataset size is sufficiently robust to capture the attention patterns.

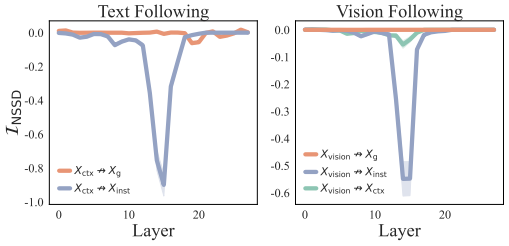
Sensitivity analysis for instruction styles. To verify that the our causal attention analysis is robust for different instruction styles, we repeat the main experiment in Fig. 2 using different instruction styles: "Prioritize the information in the text over any visual elements" and "Focus on the textual context; ignore visual content.". As shown in Fig. 10, the core mechanistic conclusions remain highly consistent across different instruction styles, which confirms that our causal attention analysis is robust for different instruction styles.



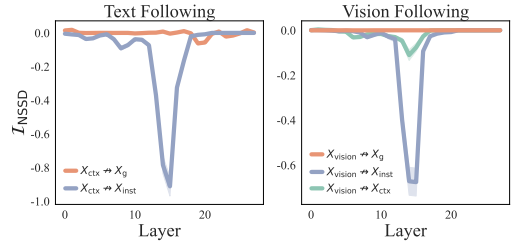
(a) 20% of the evaluation data.



(b) 40% of the evaluation data.

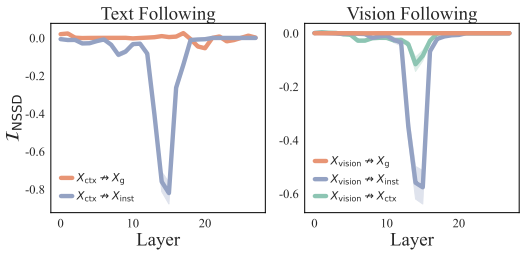


(c) 60% of the evaluation data.

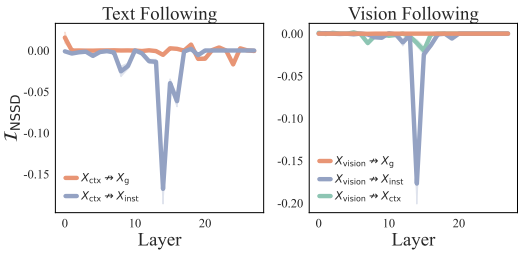


(d) 80% of the evaluation data.

Figure 9: Sensitivity analysis for dataset scale. We compare the results for layer-wise causal attention analysis under different ratios: 20%, 40%, 60% and 80% of the evaluation dataset. The overall attention patterns remain consistent across different scales, while larger sample sizes produce smoother curves with lower variance, indicating improved stability of the estimates.



(a) New Instruction Style 1.



(b) New Instruction Style 2.

Figure 10: **Layer-wise $\mathcal{I}_{\text{NSSD}}$ under attention-knockout for different instruction styles.** Each curve represents cutting a attention pathway, where X_{vision}, X_{ctx}, X_{inst}, and X_{g} denote visual tokens, text context tokens, instruction tokens, and generated tokens. The overall attention patterns remain consistent across different instruction styles.

D More Results for Mechanistic Dissection of Modality Arbitration

In this section, we include more results for mechanistic dissection of modality arbitration: the results for successful vision following in Apdx. D.1.1 and the results for generalization across other MLLM and open-ended, multi-token generation task in Apdx. D.2.

D.1 Detailed Results for Mechanistic Dissection of Modality Arbitration for Qwen2.5VL-7B

D.1.1 Results for Successful Vision Following

Due to the similar trends, we provide the results of mechanistic dissection of modality arbitration for text following in Fig. 4 and Fig. 5 (a,b) in §4.1. In this section, we report the results for vision following. As shown in Fig. 11, these results show similar trend with Fig. 4.

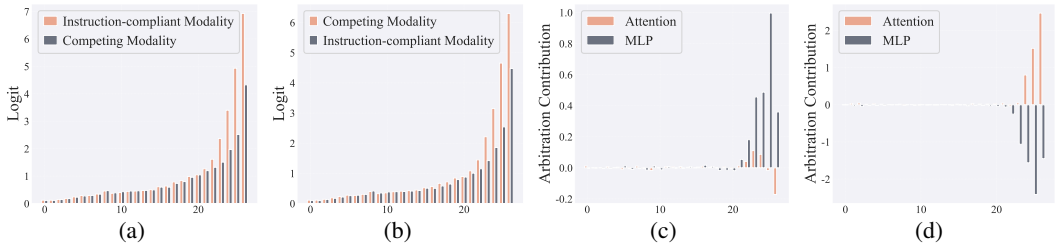


Figure 11: **Layer-wise component analysis.** (a,b) Layer-wise evolution of the instruction-compliant and competing subspace readouts for successful and failed vision-following samples, respectively. (c,d) Layer-wise contributions of attention and MLP to the modality arbitration margin for successful and failed vision-following samples, respectively. A positive contribution indicates that the corresponding component shifts the arbitration margin toward the instruction-specified intent.

Besides, we also observe that while vision-following and text-following tasks exhibit largely distinct head activation patterns, a small subset of high-contribution heads overlaps across tasks in Fig. 5 (b) and Fig. 12 (b). Specifically, among the Top-40 heads ranked by their contribution to the Modality Arbitration Margin, five heads overlap between vision-following and text-following, all within the Top-10. These constitute the modality-shared heads. The remaining heads are largely modality-specific, indicating that most attention heads specialize for a single modality.

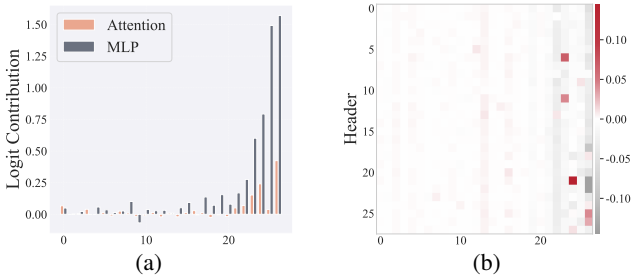


Figure 12: **Functional analysis of attention and validation of functional specialization.** (a) Layer-wise attention contributions to the instruction-compliant and competing subspaces for successful vision-following samples. (b) Layer-wise contributions of individual attention heads to the modality arbitration margin for successful samples.

D.2 Generalization for Mechanistic Dissection of Modality Arbitration

In this section, we verify that the findings for mechanistic dissection of modality arbitration observed in §4.1 can generalize to other MLLM and open-ended, multi-token generation tasks.

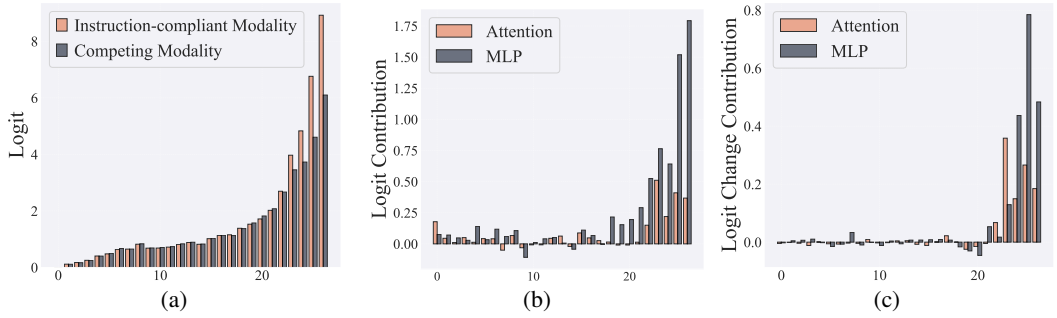


Figure 13: **Layer-wise diagnostic analysis for text following for InternVL3-8B.** (a) Layer-wise evolution of the instruction-compliant and competing subspace readouts for successful and failed text-following samples, respectively. (b) Layer-wise attention contributions to the instruction-compliant and competing subspaces for successful text-following samples. (c) Layer-wise contributions of attention and MLP to the modality arbitration margin for successful and failed text-following samples, respectively. A positive contribution indicates that the corresponding component shifts the arbitration margin toward the instruction-specified intent.

Generalization to other MLLMs. In §4.1, we provide the results of mechanistic dissection of modality arbitration for Qwen2.5-VL-7B. In this section, we demonstrate that these findings can generalize to InternVL3-8B. As shown in Fig. 13 and Fig. 14, we can observe that similar patterns

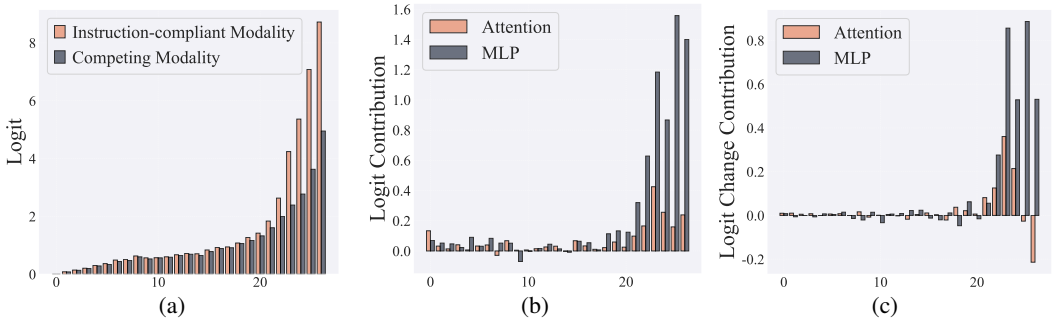


Figure 14: **Layer-wise diagnostic analysis for vision following for InternVL3-8B.** (a) Layer-wise evolution of the instruction-compliant and competing subspace readouts for successful and failed text-following samples, respectively. (b) Layer-wise attention contributions to the instruction-compliant and competing subspaces for successful text-following samples. (c) Layer-wise contributions of attention and MLP to the modality arbitration margin for successful and failed text-following samples, respectively. A positive contribution indicates that the corresponding component shifts the arbitration margin toward the instruction-specified intent.

for the evolution of subspace, and the contribution of attention and MLP to the logit intensity of the instruction-compliant modality and the modality arbitration margin with Qwen2.5VL-7B. These results show that the findings for mechanistic dissection of modality arbitration can generalize to other MLLM.

Table 7: Results on Flickr30k under text interference by amplifying the identified attention heads or random heads. The performance is measured by METEOR, CIDEr and SPICE (Higher is better).

Type	Method	METEOR	CIDEr	SPICE
-	Qwen2.5-VL-7B	20.0	18.6	13.5
Random	Amplify#10	19.8	18.6	13.6
	Amplify-#20	19.8	18.7	13.5
	Amplify-#30	19.7	18.8	13.3
Ours	Amplify-#10	18.4	19.9	13.3
	Amplify-#20	21.1	24.3	13.5
	Amplify-#30	20.3	23.5	13.8

Generalization to open-ended, multi-token scenarios. In this section, we verify that the mechanistic dissection of modality arbitration in §4.1 can generalize to open-ended, multi-token generation tasks. To this end, we investigate the roles of these findings in modality interference tasks [6]. Specifically, we conduct experiments on an image captioning task paired with a misleading text context. Following the attention amplifying methods in §4.2, we validate the functional roles of the identified attention heads through vision following. The performance is measured by METEOR, CIDEr and SPICE.

As summarized in Table 7, progressively amplifying the targeted attention heads improves the performances, especially for METEOR and CIDEr, whereas amplifying on an equal number of randomly selected heads has negligible effect. These results demonstrate that the mechanistic dissection of modality arbitration maybe extend to open-ended, multi-token generation tasks.

D.3 Targeted Component Fine-Tuning

We examine whether the fixed head sets identified by the component analysis support adaptation through targeted fine-tuning. For each following direction, only parameters associated with the selected heads are updated, while the rest of the model remains frozen. The analysis dataset is divided into training and test sets at a 7:3 ratio, and modality-following accuracy is evaluated on the held-out test portion. We compare fine-tuning 30 or 40 targeted heads with equal-sized random-head controls and direct amplification.

Table 8: Modality-following accuracy on the held-out test portion of a 7:3 train/test split. Targeted- and random-head fine-tuning update the indicated number of heads while keeping the remaining parameters frozen.

Method	Text-following accuracy (%)	Vision-following accuracy (%)
Original	21.5	23.6
Direct amplification	52.1	57.4
Random-head fine-tuning, 30 heads	28.8	29.1
Random-head fine-tuning, 40 heads	27.6	28.9
Targeted-head fine-tuning, 30 heads	63.1	64.3
Targeted-head fine-tuning, 40 heads	64.2	65.5

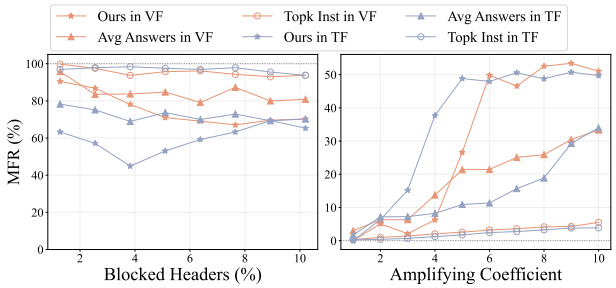


Figure 15: Interventional validation of alternative readout strategies. We compare the Modality Following Ratio (MFR) for Text Following (TF) and Vision Following (VF) across three settings: (1) TopK Inst: utilizing Top-K instruction aggregation ($K > 1$); we report results for $K = 2$ as a representative instance, given the similar performance trends observed across $K > 1$. (2) Avg Answer: employing an averaging strategy that aggregates logits across all semantically equivalent answer tokens; (3) Ours: the default max-pooling strategy used in the main text. Left: Impact of the number of blocked heads on MFR. Right: Impact of the amplification coefficient α on MFR.

As shown in Table 8, fine-tuning 40 targeted heads achieves 64.2% text-following accuracy and 65.5% vision-following accuracy, compared with 27.6% and 28.9% for 40 random heads and 52.1% and 57.4% for direct amplification. These results show that the identified heads support a fixed-component adaptation strategy in addition to inference-time amplification.

E Ablation Studies and Robustness Analysis

To verify the validity and robustness of our diagnostic framework, we conduct a series of ablation experiments on the key readout function S_m in Eq. (5) in Apdx. E.1 and the amplification coefficient in Apdx. E.4.

Maximum logit strategy. S_m leverages the layer-wise subspace probing $\mathrm{Logit}(Y_m \mid \mathbf{h}_i^l)$ with the

E.1 Ablation Studies on the Readout Function

We analyse different probing strategies to identify the critical attention heads. We then compare the effects on modality following by blocking or amplifying attention using different attention heads identified via various strategies, in order to evaluate the efficacy of the design choices.

Maximum-logit strategy. We compare the maximum-logit readout in Eq. 1 with an alternative that averages logits over candidate answer tokens. Each readout is used to identify attention heads, whose functional relevance is then evaluated through blocking and amplification. As shown in Fig. 15, heads selected using the maximum-logit readout produce larger changes in MFR than those selected using the average-logit alternative, supporting the default readout in this setting.

Instruction aggregation strategy. S_m utilizes a Top-K strategy to quantify the intensity of a specific subspace within instruction tokens in Eq. (5), with $K = 1$ as the default. We vary K to evaluate this

Table 9: Modality Following Ratio (MFR, %) for Text Following and Vision Following under progressively larger amplification coefficients. Performance remains stable within a moderate range, but collapses sharply once the intervention becomes excessively strong.

Coefficient	8	10	12	14	16	18
Text Following	48.8	49.7	47.5	34.1	21.0	2.0
Vision Following	52.5	51.0	48.5	30.6	21.9	3.3

Table 10: LDAR (%) with English-only candidate dictionaries on Qwen2.5-VL-7B. Chinese candidate expressions are removed to evaluate the effect of candidate coverage on the readout.

Following	L0–18	L20	L22	L24	L26	L28
Text	≈ 50	53	62	75	83	55
Vision	≈ 50	67	81	88	90	51

design choice and observe that $K > 1$ results in only minor variations in the modality following ratio (MFR) for both attention blocking and amplification interventions, as shown in Fig. 15. This observation provides empirical validation for the strategy adopted in our main text.

E.2 Sensitivity to Bilingual Candidate Coverage

We assess the sensitivity of the diagnostic readouts to candidate coverage by removing Chinese expressions from the answer dictionaries and recomputing $\mathcal{I}_{\mathrm{NSSD}}$ and LDAR on Qwen2.5-VL-7B. The NSSD routing pattern remains largely stable, likely because English candidates already receive higher output probabilities in this setting. LDAR, however, is lower than with the bilingual dictionaries in Fig. 3(b), reaching 83% for text following and 90% for vision following at layer 26.

As shown in Table 10, removing Chinese candidates excludes answer directions that may be activated in intermediate states. The resulting decrease in LDAR illustrates the importance of candidate coverage for Logit-Lens-based readouts and supports the use of bilingual answer dictionaries in our analysis.

E.3 Activation-Patching Validation of Instruction-Token Readouts

In this section, we provide additional details on the activation-patching validation of instruction-token readouts, as discussed in §3.3.3 of the main text. We assess the functional relevance of instruction-token readouts using activation patching [51] between paired successful text-following and vision-following runs. The donor instruction-token hidden states are taken from a successful run under one modality instruction and inserted into the paired recipient run whose instruction specifies the opposite modality. We compare patching at three layers around the LDAR peak: a pre-peak layer, the peak layer, and a post-peak layer.

The intervention effect is measured by the mean logit-margin change toward the donor modality, $\Delta M_{\mathrm{patch}}$, and the proportion of examples with a margin shift in that direction. As shown in Table 11, patching at the LDAR peak layer produces a mean shift of +1.08 and a directional shift rate of 86%, compared with +0.28 and 43% before the peak and +0.41 and 56% after it. To examine whether this effect is localized to the positions identified by the readout, we compare the top-five instruction positions ranked by the Logit-Lens-based margin with five random instruction positions at the peak layer. The selected positions yield a larger mean shift (+0.87 versus +0.38) and a higher directional shift rate (71% versus 55%). Together, the layer-wise and position-wise comparisons support the functional relevance of the instruction-token states identified by the readout.

E.4 Sensitivity to the Amplification Coefficient

To examine the stability boundary of amplification intervention, we further evaluate the effect of progressively increasing the amplification coefficient and analyze the resulting collapse behavior.

Table 11: Activation-patching validation of instruction-token readouts. Mean ΔM_{patch} measures the logit-margin change toward the donor modality. The directional shift rate is the percentage of samples whose logit margin shifts toward that modality, not the percentage whose generated answers switch modalities.

Intervention	Mean ΔM_{patch}	Directional shift rate
(a) Layer-wise comparison		
Pre-peak layer	+0.28	43%
LDAR-peak layer	+1.08	86%
Post-peak layer	+0.41	56%
(b) Position-wise comparison at the LDAR peak		
Random five instruction-token positions	+0.38	55%
Top-five instruction-token positions	+0.87	71%

As shown in Table 9, the model collapses once the amplification coefficient becomes too large. At a threshold of 18, modality-following performance drops to nearly zero, indicating that excessive intervention destabilizes the arbitration process.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: **Section 1**

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: **Apdx. A.1**

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

Justification: This paper does not include theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: **Section 3.1 and Apdx. B**

   Guidelines:

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

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

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

Answer: [Yes]

Justification: In Apdx. B, we provide the detail of used data and used MLLMs.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 3, Section 4 and Apdx. B

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The results in the paper are all presented as the mean values of all test samples, and some also report the standard deviation results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Apdx. B.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Confirmed

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Apdx. A.3

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: We do not release new pretrained models or scraped datasets with high misuse risk; only training code and instructions are planned for release.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: Apdx. A.4

    Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification:

    Guidelines:

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

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

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

    Answer: [No]

    Justification: Three students manually verified the candidate answers, as described in Apdx. B, but full annotation instructions and compensation details are not provided.

    Guidelines:

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

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

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

    Answer: [No]

    Justification: The paper does not report IRB approval or an equivalent institutional review for the annotation procedure.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [Yes]

Justification: Apdx. A.2 and Apdx. B describe the use of LLMs for candidate-answer generation and verification, together with subsequent human checks. LLMs were also used for language editing.

Guidelines:

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