LANGUAGE SHAPES VISION: BIDIRECTIONAL INFORMATION FLOW IN MULTIMODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

MLLMs have shown strong capabilities in jointly processing visual and linguistic information, making the underlying mechanism of multimodal integration a fundamental question. However, existing studies mainly examine autoregressive (AR) MLLMs, where causal attention has largely confined our understanding of this mechanism to a unidirectional flow from vision to language. In this work, we revisit multimodal integration mechanisms by studying diffusion MLLMs (DMLLMs) with full attention between vision and language. Surprisingly, we find that DMLLMs coordinate bidirectional information flow in a distinctly staged manner: language-to-vision flow dominates in shallow layers, and cooperates with vision-to-language flow in deeper layers. We further reveal that this language-to-vision flow prompts the alignment of question-relevant visual states with semantic concepts and facilitates subsequent vision-to-language transfer. Motivated by this finding, we introduce language-to-vision flow into the training of conventional AR MLLMs and find that it brings more question-targeted processing of visual information, improving vision-centric understanding capabilities. Our findings reveal bidirectional information flow as an important mechanism underlying multimodal integration and provide insights for designing next-generation MLLMs.

1 INTRODUCTION

Multimodal large language models (MLLMs) (Liu et al., 2023; Qwen Team, 2026; Wang et al., 2025) have achieved impressive performance across a wide range of multimodal tasks. Understanding mechanisms that govern integration of multimodal information within MLLMs is crucial for improving their interpretability and effectiveness. Existing mechanistic studies have largely examined autoregressive (AR) MLLMs: they characterize multimodal integration as an information flow from vision to language, which progressively integrates visual information into linguistic representations for answer prediction (Zhang et al., 2024; Kim et al., 2026; Suharitdamrong et al., 2026).

However, conventional AR MLLMs are structurally constrained by causal attention. As illustrated in Figure 1a, question tokens can utilize visual information through vision-to-language flow, while the language-to-vision flow is blocked, preventing visual representations from adapting to the question. Yet complex visual understanding tasks often require different visual information from the same image, which is highly dependent on the specific question. This motivates us to ask: *if models are also given a language-to-vision pathway during training, would they learn to exploit it alongside vision-to-language flow to facilitate multimodal integration?*

Recent diffusion-based MLLMs (DMLLMs) (You et al., 2025; Ye et al., 2025a; Li et al., 2025) provide a natural setting to investigate this problem. They employ bidirectional attention during multimodal training, enabling models to develop strategies for exploiting information flow in both vision-to-language and language-to-vision directions. To understand whether and how bidirectional information flow may contribute to multimodal integration, we focus on three aspects: (1) What cross-modal information-flow patterns emerge when both directions are available? (2) What role may the language-to-vision flow play in this integration process? (3) Can enabling this additional language-to-vision pathway during training also facilitate multimodal integration in AR MLLMs?

By tracing information flow (Geva et al., 2023) across multiple DMLLMs, we find that they exploit the two directions of information flow in a surprisingly staged manner (Figure 1b): *the language-to-vision flow dominates in shallow layers, and then cooperates with vision-to-language flow to support*

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

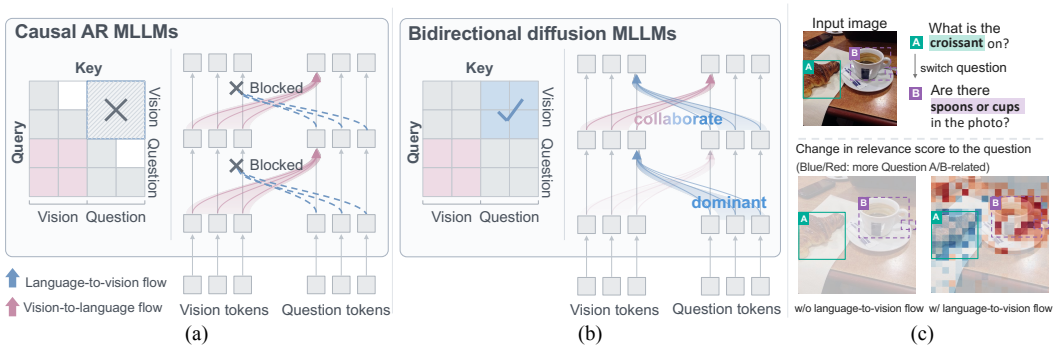


Figure 1: (a) Causal AR MLLMs permit only vision-to-language flow. (b) Bidirectional DMLLMs learn to exploit language-to-vision flow, which dominates in shallow layers before bidirectional cooperation in later layers. (c) Enabling language-to-vision flow in AR MLLMs also facilitates multimodal integration, allowing visual representations to adapt to different questions.

further multimodal integration. This reveals that, without the directional constraints imposed by causal attention, models naturally adopt a coordinated bidirectional information-flow strategy for multimodal integration, where the language-to-vision flow plays a crucial role.

Beyond identifying this surprising pattern, we further analyze how the language-to-vision pathway contributes to multimodal integration. Our analysis of cross-modal attention (Shukor & Cord, 2024) and correlation (Neo et al., 2025) reveals that language-to-vision flow strengthens the association between question tokens and relevant visual regions, helping the corresponding visual hidden states develop semantic information. As it cooperates with the vision-to-language flow in deeper layers, it also guides question tokens to focus more strongly on relevant visual evidence. These suggest that language-to-vision flow promotes the semantic alignment of question-relevant visual representations, and facilitates the subsequent transfer of relevant visual evidence into linguistic states.

We further investigate whether introducing language-to-vision flow can contribute to multimodal integration in AR MLLMs. During multimodal instruction tuning, we modify the causal mask within the multimodal prefix to permit image tokens to attend to question tokens. Using linear probing (Alain & Bengio, 2017; Raghu et al., 2021) to assess the relevance of the processed visual representations to the question, we find that *introducing language-to-vision flow enables visual representations to be organized according to the specific demands of the question* (Figure 1c). Extensive evaluations on downstream benchmarks further demonstrate improvements in vision-centric understanding capabilities. Thus, in AR paradigms, language-to-vision flow can also bring a more question-conditioned organization of visual representations for better multimodal integration.

To sum up, our study extends the previous understanding of single-pathway multimodal integration in MLLMs. **Firstly**, DMLLMs trained with bidirectional attention coordinate bidirectional pathways in a surprisingly staged manner. **Secondly**, language-to-vision flow in DMLLMs promotes task-relevant visual information processing and facilitates vision-to-language transfer. **Finally**, enabling this flow in AR MLLMs likewise induces more question-conditioned visual representations. We hope these findings inform the design of MLLMs with more effective multimodal integration.

2 Related Work

Autoregressive and diffusion MLLMs. MLLMs (Liu et al., 2023; Bai et al., 2023; Wang et al., 2025) extend language models with visual perception by connecting pre-trained visual encoders (Radford et al., 2021; Tschannen et al., 2025; Oquab et al., 2023) to language backbones (Touvron et al., 2023; Grattafiori et al., 2024; Yang et al., 2024) through multimodal alignment and instruction tuning, achieving strong performance on multimodal tasks (Yin et al., 2024). These models are trained under an autoregressive paradigm (Brown et al., 2020) with causal attention. DLLMs (Nie et al., 2025; Ye et al., 2025b) instead learn through token-level diffusion denoising with bidirectional attention during training, achieving performance comparable to similarly sized AR models on language tasks. Subsequent DMLLMs (You et al., 2025; Ye et al., 2025a; Li et al.,

2025) equip these diffusion backbones with multimodal capabilities. Our work investigates how DMLLMs exploit bidirectional cross-modal pathway enabled by multimodal training under bidirectional attention, and whether language-to-vision flow also benefits AR MLLMs.

Mechanistic Interpretability of MLLMs. Existing work investigates the internal mechanisms of multimodal understanding in MLLMs (Golovanevsky et al., 2025; Neo et al., 2025; Wu et al., 2026). Information-flow (Geva et al., 2023) analysis helps explain how information from different modalities is integrated within the model. By this approach, Zhang et al. (2024) trace how visual information is transferred to linguistic representations during multimodal integration in AR MLLMs. Subsequent studies extend this perspective to temporal reasoning in VideoLLMs (Kim et al., 2026) and the integration of auditory and visual information in AVLLMs (Selvakumar et al., 2026; Suharitdamrong et al., 2026). Recent studies in DMLLMs focus on visual attribution and generation failures (Zuo et al., 2026; Zhao et al., 2026; Hong et al., 2026), while the bidirectional information-flow patterns underlying multimodal integration remain underexplored. In contrast, we investigate multimodal integration mechanisms in DMLLMs through information-flow analysis and examine whether the benefits of language-to-vision flow extend across training paradigms.

3 PRELIMINARIES

3.1 MULTIMODAL LARGE LANGUAGE MODELS

The input sequence of MLLMs contains visual tokens followed by question tokens, with their positions indexed by $\mathcal{I}_{\mathrm{v}}$ and $\mathcal{I}_{\mathrm{q}}$, respectively. Let $\mathbf{H}^{\ell} \in \mathbb{R}^{N \times d}$ denote the token representations at layer ℓ, where N is the sequence length and $\mathbf{h}_i^{\ell}$ is the representation at position i. At the input layer, $\mathbf{H}^0_{\mathcal{I}_{\mathrm{v}}}$ and $\mathbf{H}^0_{\mathcal{I}_{\mathrm{q}}}$ denote the visual and question embeddings, respectively.

Within each transformer layer, self-attention enables information exchange among tokens. Each attention head projects $\mathbf{h}_i^{\ell}$ into a query $\mathbf{q}_i^{\ell}$, a key $\mathbf{k}_i^{\ell}$, and a value $\mathbf{u}_i^{\ell}$. The attention weights determine how each query aggregates information from the value tokens:

$$A_{ij}^{\ell} = \mathrm{softmax}_j\left(\frac{(\mathbf{q}_i^{\ell})^{\top}\mathbf{k}_j^{\ell}}{\sqrt{d_h}} + M_{ij}^{\ell}\right). \tag{1}$$

The mask entry $M_{ij}^{\ell} \in \{0, -\infty\}$ allows or blocks information flow from the key and value at position j to the query at position i, respectively. AR MLLMs learn to generate a response $\mathbf{r} = (r_1, \ldots, r_{N_R})$ by minimizing the next-token prediction loss

$$\mathcal{L}_{\mathrm{AR}}(\theta) = -\sum_{i=1}^{N_R} \log p_\theta(r_i \mid \mathbf{H}^0_{\mathcal{I}_{\mathrm{v}}}, \mathbf{H}^0_{\mathcal{I}_{\mathrm{q}}}, \mathbf{r}_{<i}), \tag{2}$$

where $\mathbf{r}_{<i}$ is the response prefix and p_θ is the model's token distribution. To implement left-to-right prediction without access to future response tokens, conventional AR MLLMs apply a causal attention mask with $M_{ij}^{\ell,\mathrm{causal}} = 0$ when $j \le i$ and $M_{ij}^{\ell,\mathrm{causal}} = -\infty$ otherwise. DMLLMs are instead trained through masked denoising. This training objective relies on the available context to reconstruct masked tokens, allowing the DMLLMs studied here to use full bidirectional attention: $M_{ij}^{\ell,\mathrm{full}} = 0$ for all input tokens.

3.2 ATTENTION KNOCKOUT

Attention Knockout disables selected attention connections during inference, allowing us to assess how information flow between token groups contributes to answer prediction (Geva et al., 2023; Zhang et al., 2024; Kim et al., 2026). For source positions $\mathcal{S}$, target positions $\mathcal{T}$, and intervened layers $\mathcal{L}$, we set $\widetilde{M}_{ij}^{\ell} = -\infty$ when $i \in \mathcal{T}$, $j \in \mathcal{S}$, and $\ell \in \mathcal{L}$, retaining $\widetilde{M}_{ij}^{\ell} = M_{ij}^{\ell}$ otherwise. Following prior work (Zhang et al., 2024; Kim et al., 2026), we apply attention knockout within sliding windows to trace the contribution of information flow across layers. Each window is defined as $\mathcal{L}(c, k) = \{\ell : |\ell - c| \le \lfloor k/2 \rfloor\}$, where c is the center layer and k is the number of layers in the window. The relative change in correct-answer probability from p_0 before knockout to p_{ko} after knockout, $\Delta p = (p_{\mathrm{ko}} - p_0)/p_0$, reflects the importance of the blocked information flow at the stage covered by the window.

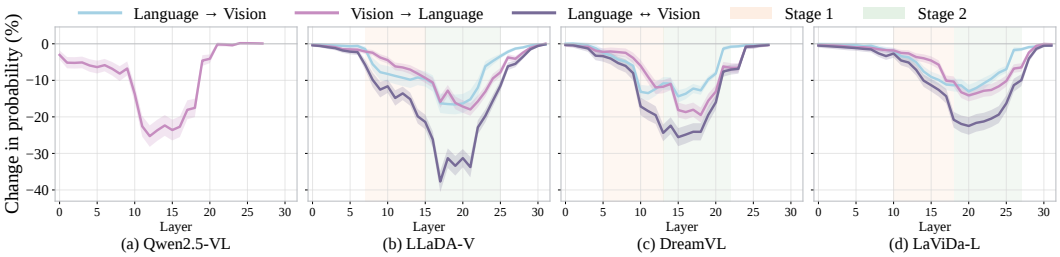


Figure 2: **Comparison of cross-modal information flow between AR and diffusion MLLMs.** AR MLLMs with causal attention, such as Qwen2.5-VL in (a), integrate multimodal information through unidirectional vision-to-language flow. In contrast, DMLLMs with bidirectional attention in (b–d) coordinate both directions of cross-modal information flow in a two-stage process.

4 How Does Multimodal Information Flow Differ under Causal and Bidirectional Attention?

In this section, we examine how MLLMs integrate visual and linguistic information for answer prediction when both directions of cross-modal information pathways are permitted. To investigate this, we apply Attention Knockout (Geva et al., 2023) in a sliding-window manner to DMLLMs trained with bidirectional attention, which allows us to identify influential directions of cross-modal information flow and trace how their contributions evolve across layers.

4.1 Experimental Setup

Models. We study DMLLMs trained with full bidirectional attention during both multimodal alignment pre-training and instruction tuning, which are allowed to learn multimodal integration mechanisms with both directions of cross-modal information flow available. We study three models that meet this requirement: LLaDA-V (You et al., 2025), Dream-VL (Ye et al., 2025a), and LaViDa-L (Li et al., 2025).[1] And we also include Qwen2.5-VL (Bai et al., 2025) as a reference in AR MLLM with causal attention.

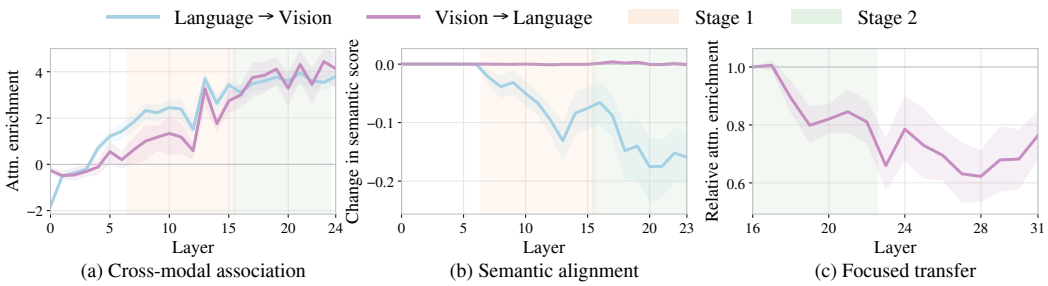
Data. Following Zhang et al. (2024), we collect our analysis data from six tasks in the GQA validation set, covering object and attribute recognition, relational understanding, and logical reasoning. To examine effective cross-modal information-flow patterns, we first run inference with each model and retain only the examples it answers correctly. Details of the data composition and task-specific results are provided in Appendices A.2 and A.6, and validation on other datasets in Appendix A.5.

Protocol. We apply Attention Knockout within sliding windows of 9 layers to examine cross-modal information flow. Blocking attention from question tokens to vision tokens interrupts vision-to-language flow, and vice versa. We perform these interventions separately and jointly within each window. For the AR reference, we only block question tokens from attending to vision tokens, since its causal mask blocks language-to-vision flow architecturally. Larger probability drops indicate a stronger contribution of the blocked information flow to multimodal integration for answer prediction. Comparisons of diffusion decoding strategies and knockout window sizes are provided in Appendices A.3 and A.4, respectively.

4.2 From Unidirectional to Bidirectional Information Flow

We first examine how cross-modal information-flow patterns differ under models trained with causal and bidirectional attention, asking whether models use both directions jointly for multimodal integration when permitted or continue to favor conventional vision-to-language flow. Prior analyses of AR MLLMs observe only vision-to-language flow, as causal attention prohibits the language-to-vision pathway (Zhang et al., 2024; Kim et al., 2026; Suharitdamrong et al., 2026). Consistent

[1] Some DMLLMs, such as DiffusionVL and Dimple, introduce bidirectional attention only during selected training stages or within particular decoding segments, leaving bidirectional information-flow directions constrained during multimodal learning. See Appendix A.1 for a detailed analysis.



(a) Cross-modal association (b) Semantic alignment (c) Focused transfer

Figure 3: **Functions of language-to-vision flow in Bidirectional DMLLMs.** (a) Cross-modal attention enrichment between keywords and their corresponding visual regions. (b) Change in semantic scores of relevant visual tokens after blocking information flow in each direction. (c) Vision-to-language attention enrichment after blocking language-to-vision flow.

with these observations, blocking vision-to-language flow in Qwen2.5-VL reduces correct-answer probability in shallow and middle layers as shown in Figure 2a.

We find that blocking either vision-to-language or language-to-vision flow substantially reduces correct-answer probability across LLaDA-V, Dream-VL, and LaViDa-L, all of which underwent multimodal alignment pre-training and instruction tuning with full bidirectional attention, as depicted in Figure 2b–d. This shows that the models spontaneously utilize the reverse language-to-vision pathway when it is available during training, and blocking both directions within the same window produces larger drops than blocking either direction alone, further supporting their joint contribution to multimodal integration.

4.3 A Two-Stage Pattern of Bidirectional Integration

Having identified bidirectional cross-modal information flow, we next analyze how the two directions collaborate across backbone layers. The models exhibit a two-stage pattern of multimodal integration, with early language-to-vision dominance followed by stronger vision-to-language flow alongside the persistent reverse pathway in middle layers.

As illustrated in Figure 2b, LLaDA-V relies more strongly on language-to-vision flow than on vision-to-language flow in layers 7–15. In the following layers 15–25, vision-to-language flow becomes more influential and cooperates with the persistent language-to-vision flow to support multimodal integration. The earlier contribution of language-to-vision flow suggests that it may help shape visual representations (analyzed in Section 5) before stronger vision-to-language transfer.

Takeaway 1: *Models trained with bidirectional attention spontaneously coordinate both directions of cross-modal information flow in a distinctly staged manner. In particular, the language-to-vision pathway, which is blocked by causal attention in conventional AR MLLMs, dominates in early layers and remains active as vision-to-language flow strengthens in later layers.*

5 How Does Language-to-Vision Flow Contribute to Multimodal Integration?

Having demonstrated that reverse language-to-vision flow supports multimodal integration, we next investigate the mechanisms underlying this contribution. We present results of LLaDA-V on data collected in Section 4.1. Results for other models are provided in Appendix B.3.

5.1 Earlier Associations Establishment between Question Keywords and Relevant Visual Regions

We first analyze the attention patterns underlying language-to-vision flow. In this flow, visual tokens act as queries and attend to question tokens as keys and values. We quantify cross-modal



Figure 4: **Visualization of semantic alignment effect of language-to-vision flow.** Top: object boxes. Bottom: keyword logit-lens scores before minus after blocking this flow, with highlighted question keywords below. Warm/cool colors denote positive/negative differences.

attention enrichment between corresponding visual and question tokens by measuring the extent to which attention from relevant visual tokens is concentrated on words referring to them (details in Appendix B.1). Higher enrichment indicates more focused attention between visual regions and their corresponding question words in language-to-vision flow. And for vision-to-language flow, this enrichment is reflected in the concentration of attention from these question words on the corresponding visual regions.

As shown in Figure 3a, language-to-vision flow exhibits higher attention enrichment between corresponding visual and question tokens than vision-to-language flow in shallow layers, and the two directions reach comparable levels in middle layers. Thus, the language-to-vision pathway exhibits stronger concentration between question keywords and the corresponding visual regions during the early stages of multimodal integration.

5.2 Semantic Alignment of Question-Relevant Visual Representations

We further investigate how the cross-modal associations established by language-to-vision flow affect visual representations. Following the logit lens (nostalgebraist, 2020; Wang et al., 2026; Neo et al., 2025), we use the target-word probability decoded from question-relevant visual representations as the semantic score. We then measure its relative change after blocking each direction of information flow (Appendix B.2). Values close to zero indicate little effect, whereas larger drops suggest a stronger contribution of the blocked flow to concept decodability of visual hidden states.

Figure 3b shows that retaining language-to-vision flow into the relevant visual region during shallow and middle layers strengthens semantic alignment, while retaining vision-to-language attention to the same visual region has little effect. Figure 4 further visualizes this alignment effect by comparing logit-lens scores for question keywords of each visual hidden state before and after blocking language-to-vision flow. The score increases brought by language-to-vision flow are concentrated primarily within the corresponding object bounding boxes. These results indicate that language-to-vision flow helps align representations of question-relevant visual regions with semantic concepts.

5.3 Guiding the Transfer of Relevant Visual Evidence to Language

As vision-to-language flow strengthens in middle layers, we ask whether continued language-to-vision flow supports the subsequent transfer of visual evidence into language representations. To investigate this, we block the language-to-vision flow to all visual tokens at middle layers. Sub-

sequent attention enrichment between corresponding visual and question tokens is compared using the ratio of values after and before the intervention. Figure 3c shows that blocking the language-to-vision pathway decreases the concentration of vision-to-language attention in relevant visual regions. These results suggest that continued language-to-vision flow helps focus vision-to-language transfer on task-relevant visual evidence.

Takeaway 2: *language-to-vision flow in models trained with bidirectional attention establishes earlier associations between question keywords and relevant visual regions, promotes semantic alignment of question-relevant visual representations, and facilitates more focused propagation of visual evidence to question tokens.*

6 Can AR MLLMs Benefit from Language-to-Vision Flow?

After discovering that language-to-vision flow in dMLLMs supports multimodal integration, we next investigate whether AR MLLMs can also benefit when this pathway is introduced. By enabling this reverse flow during multimodal instruction tuning, we explore whether question information helps organize visual representations and improves performance on downstream tasks.

6.1 Enabling Language-to-Vision Flow under Autoregressive Training

In conventional AR MLLMs, image tokens cannot receive question information under the standard causal attention mask. We relax this constraint during instruction tuning by introducing *reverse attention*, which allows image tokens to attend to question tokens and thereby enables language-to-vision information flow.

Enabling reverse attention. We augment standard causal attention with connections from image queries to question tokens. Let $\mathcal{I}$ and $\mathcal{Q}$ denote the image and question positions, respectively. For query position i and key position j, the modified mask is

$$M_{ij} = \begin{cases} 0, & j \le i \text{ or } (i \in \mathcal{I} \text{ and } j \in \mathcal{Q}), \\ -\infty, & \text{otherwise.} \end{cases} \tag{3}$$

We apply this mask in all decoder blocks, allowing image tokens to attend to subsequent question tokens, excluding system tokens. Attention within the image and text token sequences remains causal. This design isolates the effect of reverse attention from gains associated with bidirectional attention among image tokens (Wu et al., 2026; Beyer et al., 2024), while preserving autoregressive training and inference. More detailed comparisons are provided in Appendix C.2.

Training setup. We use the LLaVA-1.5 framework with Vicuna-7B-v1.5 and Qwen2.5-7B as language backbones and CLIP, SigLIP2, and DINOv2 as visual encoders. During multimodal alignment pre-training, we train a two-layer MLP projector on image–caption pairs under causal attention. Then we perform instruction tuning on visual question answering datasets under different attention mask settings. Additional training details are provided in Appendix C.1.

6.2 Question-Conditioned Organization of Visual Representations

To assess how language-to-vision enabled by reverse attention shapes visual representations, we train linear probes(Alain & Bengio, 2017; Raghu et al., 2021; Wu et al., 2026) to predict each visual representation's relevance to the question. The evaluation covers $1,180$ GQA questions. Following Liu et al. (2025), we use AUROC to measure discrimination between target and background patches and NDCG@all to assess how highly target patches rank across the full image. To further assess whether visual representations adapt to changes in the question, these questions are organized into 590 pairs, each asking about different target regions in the same image. Target-shift AUROC is then adopted to assess whether switching the question changes relevance scores in favor of the newly queried region over the previously queried one. Implementation details and metric definitions are provided in Appendices C.3 and C.4, respectively.

For a given question, reverse attention makes relevant visual regions more distinguishable from the background. As shown in Table 1, reverse attention improves both AUROC and NDCG@all across

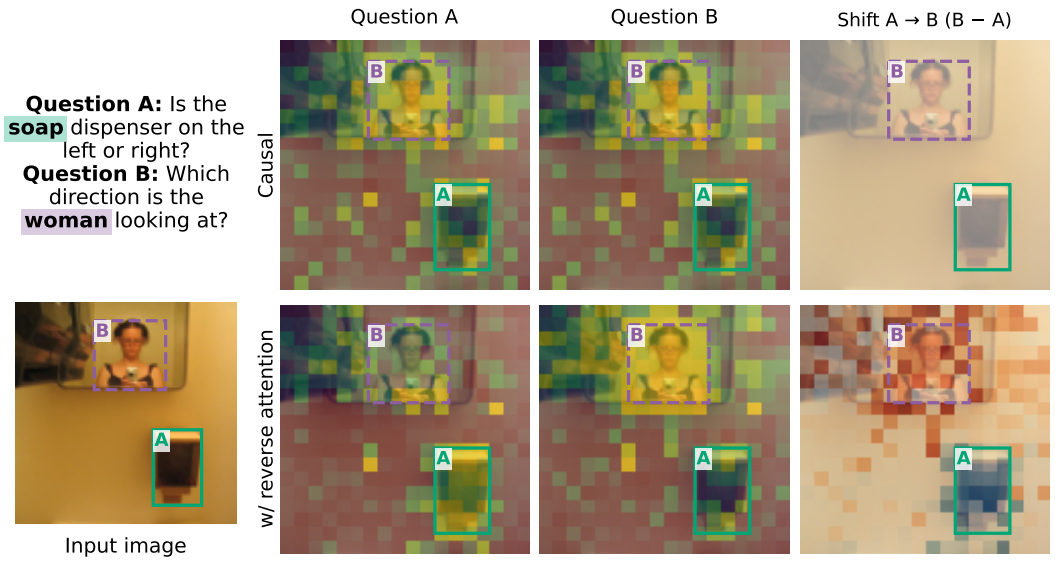


Figure 5: **With reverse attention, regions receiving higher relevance scores shift with the question's target.** Linear probe score maps show the question relevance of visual representations for the same image under two questions, comparing causal attention with and without reverse attention. Boxes A and B mark the corresponding target regions. The rightmost column shows the signed relevance-score change from question A to question B when reverse attention is enabled.

Table 1: **Question relevance of visual representations and their adaptation to changing question targets.** Higher scores indicate better discrimination of question-relevant regions (AUROC, NDCG) and stronger adaptation to target changes (Target-shift AUROC). Scores are scaled to 0–100; Target-shift AUROC has chance level 50. Off/On denotes instruction tuning without/with reverse attention.

Metric	Vicuna-7B-v1.5						Qwen2.5-7B					
	CLIP		SigLIP2		DINOv2		CLIP		SigLIP2		DINOv2	
	Off	On	Off	On	Off	On	Off	On	Off	On	Off	On
Relevance discrimination to a single question												
AUROC	77.60	**82.21**	75.44	**76.66**	76.00	**80.85**	76.13	**80.64**	75.51	**80.34**	75.76	**81.81**
NDCG@all	57.30	**63.90**	58.68	**61.27**	57.84	**64.92**	55.86	**64.36**	57.90	**68.87**	57.16	**65.03**
Adaptation to question changes												
Target-shift AUROC	49.51	**79.28**	49.46	**69.06**	48.45	**80.02**	50.19	**77.63**	50.33	**77.58**	49.28	**80.25**

all evaluated combinations of language backbone and visual encoder. For example, with Vicuna-7B-v1.5 and CLIP, AUROC increases from 0.78 to 0.82 and NDCG@all from 0.57 to 0.64.

This effect is further supported by the observation that visual representations adapt to different questions about the same image. As shown in Figure 5, with reverse attention, regions receiving higher relevance scores shift toward the newly queried region when the question changes, while the causal maps remain largely unchanged. This pattern is supported by the quantitative results in Table 1: Target-shift AUROC is consistently above chance with reverse attention, but remains near chance under causal attention. These results suggest that language-to-vision information flow enables visual representations to reflect which regions matter for the current question.

6.3 Evaluation on Downstream Tasks

We next investigate whether enabling language-to-vision information flow improves downstream task performance. Since we observe that language-to-vision information flow improves the organization of visual representations, we evaluate on a range of vision-centric benchmarks, including

Table 2: **Downstream evaluations on vision-centric tasks.** Off/On indicates instruction tuning without/with reverse attention. Green and orange indicate performance increases and decreases, respectively, while gray indicates no change.

Model	Reverse Attention	RWQA	CV-Bench^{2D}	MMVP	MM*	BLINK			
						Object local.	Semantic corr.	Spatial relation	Visual corr.
Backbone: Vicuna-7B-v1.5									
CLIP	Off	48.63	50.76	27.33	33.27	50.00	23.02	60.84	30.81
	On	50.07$_{+1.44}$	52.16$_{+1.40}$	27.67$_{+0.34}$	32.21$_{-1.06}$	52.46$_{+2.46}$	24.37$_{+1.35}$	64.34$_{+3.50}$	31.40$_{+0.59}$
SigLIP2	Off	50.72	57.51	28.00	33.40	51.64	24.46	70.63	30.81
	On	51.37$_{+0.65}$	57.72$_{+0.21}$	29.67$_{+1.67}$	33.80$_{+0.40}$	51.82$_{+0.18}$	24.89$_{+0.43}$	70.34$_{-0.29}$	31.98$_{+1.17}$
DINOv2	Off	45.23	52.92	26.00	32.80	50.00	23.74	66.43	26.74
	On	45.10$_{-0.13}$	54.31$_{+1.39}$	28.67$_{+2.67}$	33.13$_{+0.33}$	59.02$_{+9.02}$	23.74$_{+0.00}$	68.53$_{+2.10}$	29.65$_{+2.91}$
Backbone: Qwen2.5-7B									
CLIP	Off	53.07	55.70	31.33	39.20	61.48	28.78	69.93	26.74
	On	54.50$_{+1.43}$	57.09$_{+1.39}$	30.67$_{-0.66}$	40.40$_{+1.20}$	61.52$_{+0.04}$	30.94$_{+2.16}$	74.83$_{+4.90}$	28.49$_{+1.75}$
SigLIP2	Off	57.39	64.19	44.00	42.87	57.38	29.50	79.02	31.40
	On	57.78$_{+0.39}$	63.77$_{-0.42}$	49.33$_{+5.33}$	45.53$_{+2.66}$	57.64$_{+0.26}$	29.62$_{+0.12}$	78.65$_{-0.37}$	32.74$_{+1.34}$
DINOv2	Off	49.02	54.45	30.67	36.67	54.10	33.81	74.83	29.65
	On	49.93$_{+0.91}$	56.47$_{+2.02}$	34.67$_{+4.00}$	35.07$_{-1.60}$	54.36$_{+0.26}$	34.53$_{+0.72}$	75.52$_{+0.69}$	29.65$_{+0.00}$

RWQA (xAI, 2024), CV-Bench^{2D} (Tong et al., 2024a), and MMVP (Tong et al., 2024b), together with MMStar (Chen et al., 2024), which requires visual evidence for multimodal reasoning. For BLINK (Fu et al., 2024), we focus on object localization, spatial relations, and semantic and visual correspondence, which directly assess the ability to locate visual targets and distinguish relationships within and across images.

As shown in Table 2, augmenting causal attention with reverse attention brings performance gains across different language backbones and visual encoders. For example, Vicuna with DINOv2 gains 9.02 percentage points on BLINK object localization, and Qwen2.5 with SigLIP2 improves MMVP pair accuracy by 5.33 points. Qwen2.5 with CLIP also gains 4.90 points on BLINK spatial relations. These gains span tasks that require different forms of visual evidence: identifying the queried object, distinguishing fine-grained visual details, and resolving spatial relationships. This suggests that the more question-conditioned visual representations enabled by language-to-vision flow (Section 6.2) facilitate the extraction and use of cues from question-relevant visual regions, supporting downstream tasks that require precise visual evidence.

Takeaway 3: *language-to-vision information flow remains beneficial under autoregressive training, promoting better question-conditioned organization of visual representations and enhancing models' vision-centric understanding capabilities.*

7 Discussion and Limitations

In this work, we explore the possibility of bidirectional cross-modal information flow during multimodal integration in MLLMs. We find that models trained with bidirectional attention spontaneously leverage bidirectional information flow for multimodal integration. Specifically, the reverse language-to-vision flow enhances the processing of visual representations across different training paradigms. These findings highlight bidirectional flow as an important mechanism for multimodal integration in MLLMs and provide insights for designing future large multimodal models. Our study primarily analyzes multimodal understanding tasks in LLaVA-style architectures. Future work could extend to unified models that combine understanding and image generation to explore how bidirectional information flow is coordinated when both multimodal tasks may cooperate and compete.

AI USE STATEMENT

AI-assisted tools were used to polish the language and presentation of the manuscript. All AI-assisted revisions were carefully reviewed and revised by the authors to ensure accuracy and fidelity to the intended meaning. We independently formulated the research questions, designed the methodology, conducted the experiments, and drew the scientific conclusions. We take full responsibility for the content of this paper.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding Intermediate Layers Using Linear Classifier Probes. In *International Conference on Learning Representations*, 2017.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer. *arXiv:2407.07726*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, 2024.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023.

Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. What do VLMs NOTICE? a mechanistic interpretability pipeline for Gaussian-noise-free text-image corruption and evaluation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11462–11482, April 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6325–6334, 2017.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Sujung Hong, Chanyong Yoon, and Seongjae Hwang. Mitigating mask prior drift and positional attention collapse in large diffusion vision-language models. *arXiv preprint arXiv:2605.14530*, 2026.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702, 2019.

Minji Kim, Taekyung Kim, and Bohyung Han. Map the flow: Revealing hidden pathways of information in videollms. In *International Conference on Learning Representations (ICLR)*, 2026.

Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. Lavida: A large diffusion model for vision-language understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2023.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Joy Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Benoit Dumoulin, and Hanghang Tong. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. *arXiv preprint arXiv:2510.17771*, 2025.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3190–3199, 2019.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 57172–57189, 2025.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

nostalgebraist. Interpreting GPT: The logit lens. AI Alignment Forum, August 2020. URL https://www.alignmentforum.org/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023.

Xiaohuan Pei, Tao Huang, Yanxiang Ma, and Chang Xu. Rethinking causal mask attention for vision-language inference. In *International Conference on Learning Representations (ICLR)*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 12116–12128, 2021.

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

Mustafa Shukor and Matthieu Cord. Implicit multimodal alignment: On the generalization of frozen llms to multimodal inputs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130848–130886, 2024.

Wish Suharitdamrong, Muhammad Awais, Xiatian Zhu, and Sara Atito. From senses to decisions: The information flow of auditory and visual perception in multimodal llms. *arXiv preprint arXiv:2606.10147*, 2026.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems*, volume 37, pp. 87310–87356, 2024a.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*, 2024b.

Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

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

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Zichuan Wang, Songlin Yang, Bo Peng, Zhenchen Tang, Yang Li, Beibei Dong, and Jing Dong. Same attention, different truths: Put logit-lens over visual attention to detect and mitigate lvlm object hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25315–25325, 2026.

Boyong Wu, Sanghwan Kim, and Zeynep Akata. From drop-off to recovery: A mechanistic analysis of segmentation in MLLMs. *arXiv preprint arXiv:2603.17228*, 2026.

xAI. Grok-1.5 vision preview, 2024. URL `https://x.ai/news/grok-1.5v`. Introduces the RealWorldQA benchmark.

An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Jiacheng Ye, Shansan Gong, Jiahui Gao, Junming Fan, Shuang Wu, Wei Bi, Haoli Bai, Lifeng Shang, and Lingpeng Kong. Dream-vl & dream-vla: Open vision-language and vision-language-action models with diffusion language model backbone. *arXiv preprint arXiv:2512.22615*, 2025a.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025b.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12), November 2024. ISSN 2053-714X. doi: 10.1093/nsr/nwae403.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*, 2025.

Runpeng Yu, Xinyin Ma, and Xinchao Wang. Dimple: Discrete diffusion multimodal large language model with parallel decoding. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2026.

Lunbin Zeng, Jingfeng Yao, Bencheng Liao, Hongyuan Tao, Wenyu Liu, and Xinggang Wang. Diffusionvl: Translating any autoregressive models into diffusion vision language models. *arXiv preprint arXiv:2512.15713*, 2025.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. *arXiv preprint arXiv:2411.18620*, 2024.

Qiyan Zhao, Xiaofeng Zhang, Shuochen Chang, Qianyu Chen, Xiaosong Yuan, Xuhang Chen, LUOQI LIU, Jiajun Zhang, Xu-yao Zhang, and Da-Han Wang. Context tokens are anchors: Understanding the repeat curse in dmllms from an information flow perspective. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 152550–152578, 2026.

Haomin Zuo, Yidi Li, Luoxiao Yang, and Xiaofeng Zhang. Diffusion-cam: Faithful visual explanations for dmllms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

A Implementation Details and additional results of Information Flow Analysis

A.1 Analysis of Diffusion Multimodal Large Language Models

Diffusion MLLMs differ in when and how bidirectional attention is introduced during multimodal training. Table 3 compares LLaDA-V (You et al., 2025), Dream-VL (Ye et al., 2025a), LaViDa-L (Li et al., 2025), Dimple (Yu et al., 2026), and DiffusionVL (Zeng et al., 2025).

Models trained with full bidirectional attention. LLaDA-V, Dream-VL, and LaViDa-L use full bidirectional attention during both multimodal alignment and instruction tuning, allowing visual and question tokens to attend to each other throughout these stages. The language backbone of both LLaDA-V and LaViDa-L is LLaDA (Nie et al., 2025), while Dream-VL is built on Dream (Ye et al., 2025b). Note that LLaDA and Dream obtain their bidirectional language backbones in different ways: LLaDA is trained from scratch with bidirectional attention under the diffusion paradigm, whereas Dream is converted from the AR Qwen2.5-7B model to bidirectional diffusion training, potentially retaining some left-to-right inductive bias from its AR initialization.

Others. Dimple instead performs multimodal alignment and the first instruction-tuning stage with causal attention under an autoregressive objective, and switches to full attention in the subsequent diffusion-based stage. DiffusionVL directly converts pretrained AR-VLMs such as Qwen2.5-VL (Bai et al., 2025) into diffusion VLMs through end-to-end block-diffusion finetuning. Its attention is bidirectional within each block but remains causal across blocks. Thus, neither Dimple nor DiffusionVL provides unrestricted bidirectional cross-modal information flow throughout the full multimodal training process.

We therefore focus on LLaDA-V, Dream-VL, and LaViDa-L because both cross-modal pathways are available throughout their multimodal training. This shared property lets us examine how models learn to use bidirectional flow to achieve multimodal integration.

A.2 Data Collection

Following Zhang et al. (2024), we select six tasks from the GQA validation set (Hudson & Manning, 2019): ChooseAttr, ChooseCat, ChooseRel, CompareAttr, LogicalObj, and QueryAttr. These tasks span object recognition, attributes, relations, and logical reasoning. For each model, we retain examples correctly answered by the original model to examine the information-flow mechanisms underlying its successful predictions. Table 4 summarizes the final subsets used for each model's information-flow analysis. Within each model, we use the same subset for all knockout conditions to ensure consistent comparisons across flow directions and layer windows.

Table 3: **Attention patterns of different DMLLMs during multimodal training.** Base type describes the training strategy of the backbone model. Full denotes bidirectional attention; Block denotes bidirectional attention within blocks and causal attention between them. The arrow denotes successive training stages. † indicates that the model skips the alignment stage and directly inherits the pretrained backbone's visual capabilities.

Model	Base model	Base type	Alignment	Later training
Models used in the main analysis				
LLaDA-V	LLaDA-8B-Instruct	Full (from scratch)	Full	Full
Dream-VL	Dream-v0-Instruct-7B	Full (from AR)	Full	Full
LaViDa-L	LLaDA-8B-Instruct	Full (from scratch)	Full	Full
Others				
Dimple	Dream-v0-Instruct-7B	Full (from AR)	Causal	Causal → Full
DiffusionVL	Qwen2.5-VL-7B-Instruct	Causal AR	Causal†	Block

Table 4: **GQA subsets for information-flow analyses in different models.** Attention knockout interventions are performed only on questions each model answers correctly without intervention, allowing us to examine the information-flow mechanisms supporting successful predictions.

Model / pool	ChooseAttr	ChooseCat	ChooseRel	CompareAttr	LogicalObj	QueryAttr	Total
Candidates	1,000	1,000	964	570	991	1,000	**5,525**
LLaDA-V	755	514	850	551	844	933	**4,447**
Dream-VL	781	511	902	548	928	937	**4,607**
LaViDa	715	499	740	523	892	878	**4,247**

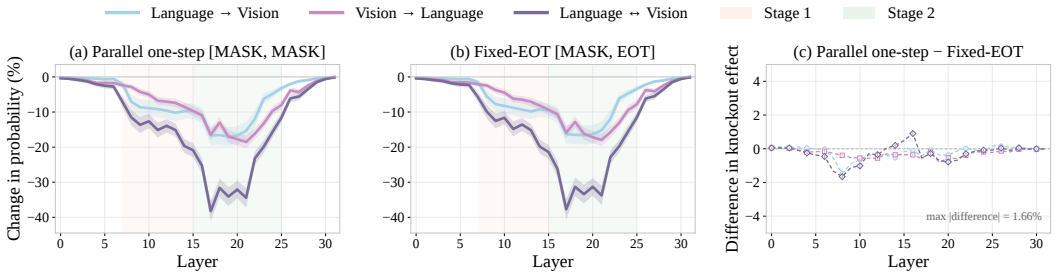


Figure 6: **Effect of diffusion decoding strategies on cross-modal information flow.** Attention Knockout effects under (a) parallel one-step and (b) fixed-EOT decoding, and (c) their difference in percentage points. Both strategies preserve the two-stage flow pattern, with only minor differences in knockout effects. Results are based on LLaDA-V.

A.3 Comparison of Diffusion Decoding Strategies

We find that the observed information-flow pattern is robust across different diffusion decoding strategies. We use a two-position generation block for the answer token and the end-of-turn (EOT) token, comparing three strategies: *Parallel one-step* decodes both positions in parallel from a fully masked block in a single forward pass; *Two-step* decodes one position at a time over two forward passes; *Fixed-EOT*, the setting used in the main analysis, directly conditions on EOT while keeping the answer position masked, requiring only one forward pass.

We find that *two-step* decoding consistently generates EOT before the answer, yielding the same answer probability as *fixed-EOT*. Figure 6 further shows that *parallel one-step* and *fixed-EOT* decoding preserve the overall layerwise pattern and the relative strength of the two information-flow directions, with a maximum difference of 1.66 percentage points in knockout effects.

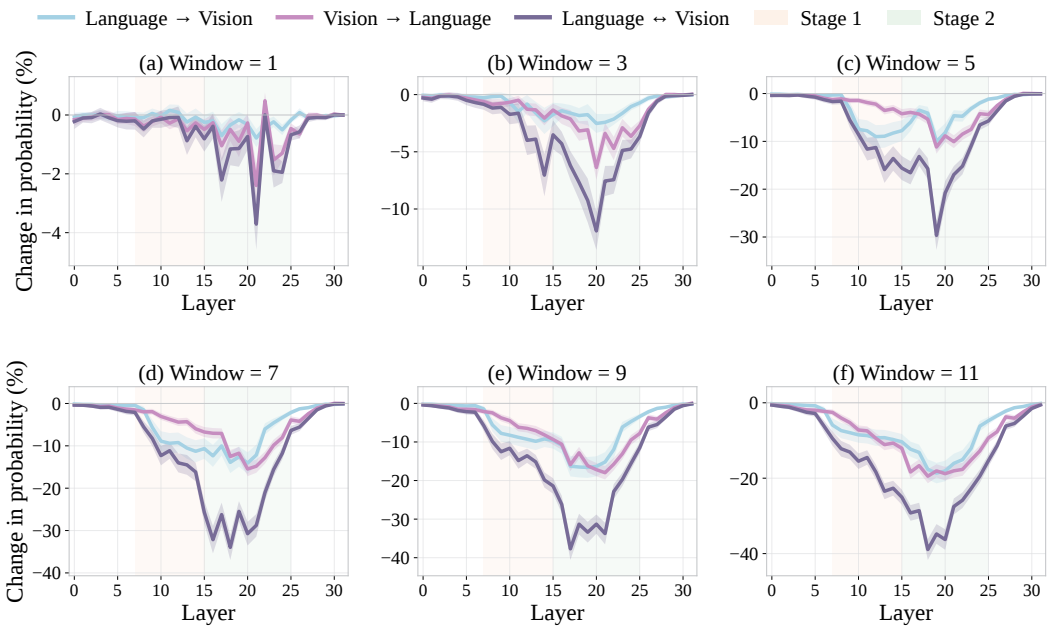


Figure 7: **Effect of knockout window size on cross-modal information flow.** Small windows yield weak effects, whereas widths of 5 to 11 layers reveal a consistent two-stage pattern. Results are based on LLaDA-V.

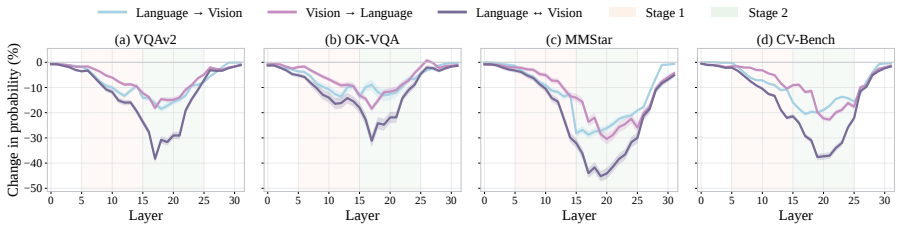


Figure 8: **Cross-modal information flow on additional datasets.** The two-stage pattern broadly holds across datasets, with an earlier emphasis on language-to-vision flow followed by stronger cooperation between both directions. Results are based on LLaDA-V.

A.4 Effect of Knockout Window Size

We vary the knockout window width in LLaDA-V over $k \in \{1, 3, 5, 7, 9, 11\}$, keeping the GQA subset and other settings fixed. Windows slide with a stride of one layer and are clipped at network boundaries. As shown in Figure 7, windows of one or three layers produce small probability drops, making the information-flow pattern difficult to discern. Widths of five to nine layers reveal clearer effects and a consistent two-stage pattern across similar layer ranges. Increasing the width from 9 to 11 yields only a modest increase in the peak joint effect, while making it harder to locate the layer ranges responsible for each stage. We therefore use nine-layer windows to balance intervention strength with the precision of identifying where each direction contributes.

A.5 Results on Additional Datasets

We extend the information-flow analysis to VQAv2 (Goyal et al., 2017) and OK-VQA (Marino et al., 2019), as well as the multiple-choice visual understanding benchmarks MMStar (Chen et al., 2024) and CV-Bench (Tong et al., 2024a), using the same protocol.

Figure 8 shows that the two-stage pattern broadly holds across these datasets: language-to-vision flow contributes more strongly in earlier layers, followed by stronger vision-to-language transfer alongside the persistent reverse pathway. The larger drops under joint knockout in middle layers

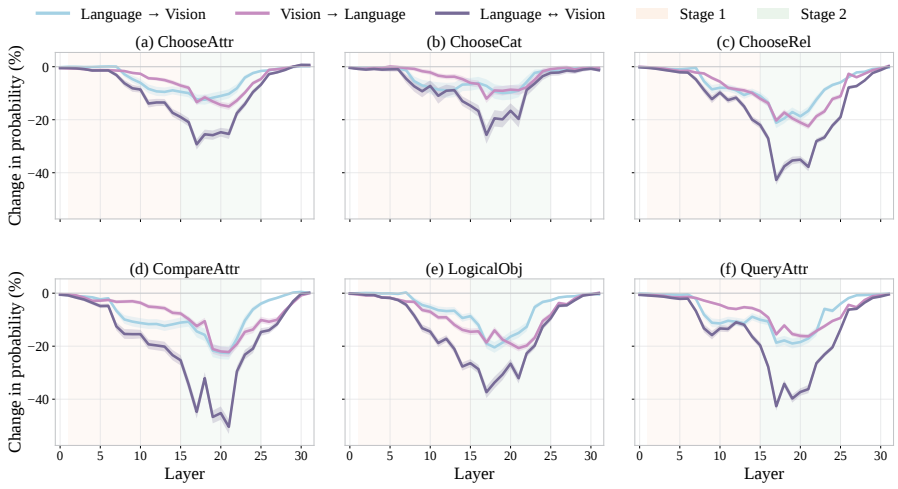


Figure 9: **Cross-modal information flow in different tasks.** Relative changes in answer probability after blocking either or both directions within nine-layer windows across six GQA tasks. Shaded backgrounds mark the two stages identified in the main analysis. Results are based on LLaDA-V.

further support this later cooperation. Although the relative strengths of the two pathways vary across tasks, these results support the staged bidirectional cross-modal interaction beyond GQA.

A.6 Information Flow Analysis in different tasks

Figure 9 shows cross-modal information-flow patterns across six GQA subtasks, providing task-level support for the two-stage cooperation identified in the main analysis. In ChooseAttr, ChooseCat, and CompareAttr, language-to-vision knockout produces larger probability drops in earlier windows, whereas vision-to-language effects strengthen to comparable or greater levels in later windows while the reverse pathway remains influential. In ChooseRel and QueryAttr, joint knockout also produces substantially larger drops than either direction alone in the middle layers, supporting the contribution of both pathways during the cooperative stage. LogicalObj shows a relatively weaker contribution from reverse flow in early layers, possibly because it requires less fine-grained visual perception.

B Metrics and Additional Results of Language-to-Vision Flow in DMLLMs

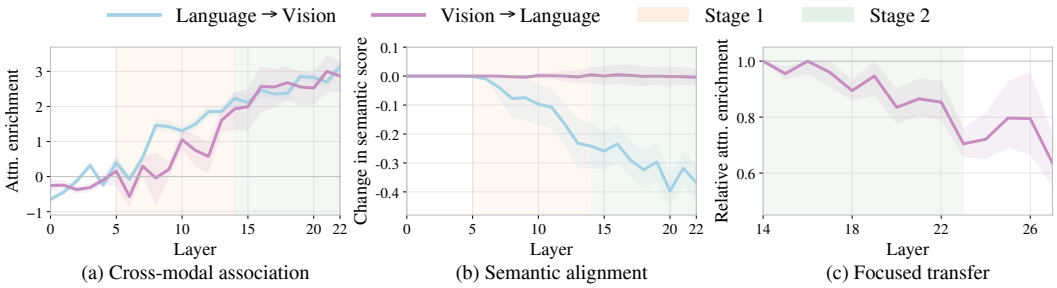
B.1 Metric of Cross-Modal Attention Enrichment

In Section 5, we use attention enrichment to assess whether cross-modal attention preferentially connects question-relevant visual regions with the words referring to them, relative to the image-wide attention pattern. The same metric is applied to both information-flow directions.

Let $\mathcal{V}_{\text{rel}}$ denote question-relevant visual patches, and let $\mathcal{T}_{\text{rel}}$ contain the question tokens referring to them. Here, $\mathcal{I}_{\text{v}}$ and $\mathcal{I}_{\text{t}}$ contain all visual and text tokens, respectively, and h indexes attention heads. For language-to-vision flow, we aggregate attention as $a^{\ell}(\mathcal{V},\mathcal{T}) = \sum_h \sum_{i\in\mathcal{V}} \sum_{t\in\mathcal{T}} A_{i,t}^{\ell,h}$. For vision-to-language flow, we use the same definition with $A_{i,t}^{\ell,h}$ replaced by $A_{t,i}^{\ell,h}$. For either direction, enrichment is

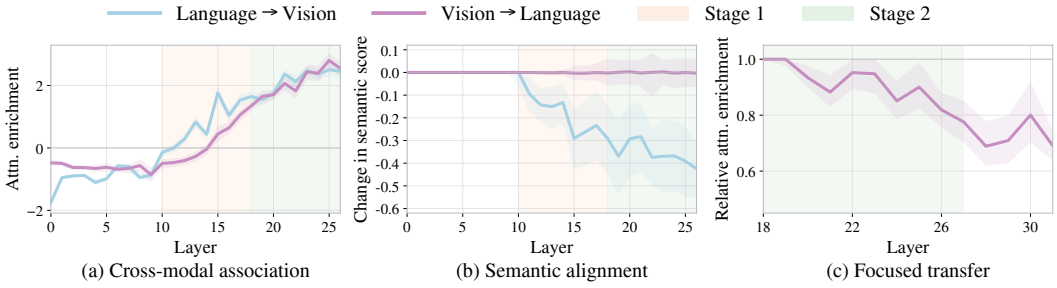
$$E^{\ell} = \frac{a^{\ell}(\mathcal{V}_{\text{rel}},\mathcal{T}_{\text{rel}})/a^{\ell}(\mathcal{V}_{\text{rel}},\mathcal{I}_{\text{t}})}{a^{\ell}(\mathcal{I}_{\text{v}},\mathcal{T}_{\text{rel}})/a^{\ell}(\mathcal{I}_{\text{v}},\mathcal{I}_{\text{t}})}. \tag{4}$$

Taking language-to-vision flow as an example, each inner ratio measures the fraction of cross-modal attention assigned to the corresponding question tokens relative to all text tokens. The two ratios are computed over the relevant visual regions and the full image, respectively. Attention weights are summed over visual tokens and heads before normalization, and enrichment is defined as the ratio of the regional fraction to the corresponding fraction over the full image. We report the base-2



(a) Cross-modal association (b) Semantic alignment (c) Focused transfer

Figure 10: **Functions of language-to-vision flow in Dream-VL.** (a) Cross-modal attention enrichment. (b) Semantic-score changes after blocking information flow in each direction. (c) Relative vision-to-language enrichment after blocking reverse flow. The roles of reverse flow in visual representation processing and subsequent information transfer generalize to Dream-VL.



(a) Cross-modal association (b) Semantic alignment (c) Focused transfer

Figure 11: **Functions of language-to-vision flow in LaViDa.** The roles of reverse flow in visual representation processing and subsequent information transfer generalize to LaViDa.

logarithm of the mean enrichment across valid questions. Positive values indicate more focused attention between question-relevant visual regions and the corresponding question tokens, while zero indicates no enrichment.

B.2 Semantic Score of Visual Representations

In Section 5.2, we measure changes in semantic score to assess how blocking each cross-modal information-flow direction affects target-word decodability from question-relevant visual representations.

Let R denote visual tokens within the annotated target region and w the corresponding target word. We apply the model's final normalization and language-model head to each visual hidden state h_v^ℓ and average the target-word probability over the region:

$$P^\ell(w) = \frac{1}{|R|} \sum_{v \in R} \left[\mathrm{softmax}\big(W_{\mathrm{LM}}\,\mathrm{Norm}(h_v^\ell)\big)\right]_w . \tag{5}$$

The softmax is computed over the full vocabulary. For each flow direction, the relative change between the knockout (KO) and unmodified (clean) runs is

$$\Delta^\ell = \frac{P^\ell_{\mathrm{KO}}(w) - P^\ell_{\mathrm{clean}}(w)}{P^\ell_{\mathrm{clean}}(w)}. \tag{6}$$

Values near zero indicate little effect, whereas larger negative changes suggest a stronger contribution of the blocked flow to the semantic alignment of the question-relevant visual representations.

B.3 Language-to-Vision Flow in Dream-VL and LaViDa

We extend the three analyses in Section 5 to Dream-VL and LaViDa. Figures 10 and 11 show that both models exhibit patterns consistent with the role of reverse flow identified in LLaDA-V. These demonstrate that language-to-vision flow supports the semantic processing of relevant visual representations and helps guide their subsequent transfer to question tokens across studied DMLLMs.

Table 5: **Discrimination of question-relevant visual regions across question types.** We report AUROC and NDCG for each GQA semantic type. Higher scores indicate that linear probes better distinguish relevant regions from background and rank them more highly. Results use visual representations at Layer 13 and are reported on a 0–100 scale. Off/On indicates whether reverse attention is enabled.

Type	Reverse	Vicuna-7B-v1.5						Qwen2.5-7B					
		CLIP		SigLIP2		DINOv2		CLIP		SigLIP2		DINOv2	
		AUROC	NDCG	AUROC	NDCG	AUROC	NDCG	AUROC	NDCG	AUROC	NDCG	AUROC	NDCG
Attribute	Off	76.36	56.78	73.82	57.41	74.37	55.57	74.41	54.54	74.21	56.31	75.26	56.13
	On	**82.30**	**62.55**	**75.74**	**59.19**	**80.80**	**64.80**	**80.11**	**64.43**	**80.35**	**69.82**	**82.87**	**66.69**
	Δ	+5.94	+5.77	+1.92	+1.78	+6.42	+9.23	+5.70	+9.89	+6.15	+13.51	+7.61	+10.56
Object	Off	77.12	55.51	76.93	57.49	75.44	55.43	77.14	55.07	76.02	56.95	74.63	55.43
	On	**81.40**	**60.43**	**77.16**	**59.15**	**79.41**	**62.21**	**81.76**	**63.53**	**81.62**	**69.85**	**80.02**	**62.66**
	Δ	+4.28	+4.92	+0.23	+1.66	+3.97	+6.78	+4.62	+8.47	+5.60	+12.90	+5.39	+7.23
Relation	Off	78.71	59.27	**76.53**	59.41	77.41	59.46	77.31	56.44	76.38	58.76	76.41	57.66
	On	**82.38**	**62.15**	76.35	**59.99**	**81.28**	**65.03**	**80.86**	**63.91**	**80.10**	**67.51**	**81.42**	**63.73**
	Δ	+3.67	+2.88	−0.18	+0.58	+3.87	+5.58	+3.55	+7.47	+3.72	+8.76	+5.00	+6.07
Category	Off	76.78	76.35	**73.80**	73.37	76.19	76.27	75.01	72.42	76.60	73.69	75.73	75.46
	On	**80.92**	**77.20**	73.61	**73.59**	**78.96**	**80.45**	**79.21**	**79.44**	**78.97**	**78.78**	**80.58**	**78.46**
	Δ	+4.13	+0.85	−0.19	+0.22	+2.77	+4.18	+4.20	+7.01	+2.37	+5.09	+4.85	+3.00

C Implementation Details and additional results of Reverse Attention in AR MLLMs

C.1 Training Details

We combine Vicuna-7B-v1.5 or Qwen2.5-7B-Instruct with a frozen CLIP ViT-L/14, SigLIP2 SO400M/14, or DINOv2 ViT-L/14 encoder through a two-layer MLP projector with GELU activation. Following LLaVA-1.5, training has two stages: causal alignment on 558K LLaVA image–caption pairs updates only the projector with a batch size of 256 and a learning rate of 10^{-3}; instruction tuning on the LLaVA-1.5 665K updates both the projector and language model with a batch size of 128 and a learning rate of 2×10^{-5}. We enable reverse attention only during instruction tuning for comparison with the causal-attention baseline.

C.2 Comparison with Prior works introducing Bidirectional Attention in AR

PaliGemma (Beyer et al., 2024) trains a 3B vision-language model with bidirectional attention over the image–text prefix and autoregressive answer generation, and evaluates alternative training masks through downstream task performance. Its prefix mask enables both within-modality bidirectionality and language-to-vision flow, whereas ours adds only the latter while preserving causal attention within each modality to isolate its effect on visual representations.

Wu et al. (2026) enable bidirectional attention among image tokens to improve access to visual context for semantic segmentation. In our design, attention within each modality remains causal, and only visual tokens gain access to question tokens. This isolates the contribution of language-to-vision flow from that of bidirectional attention among image tokens.

Pei et al. (2026) discuss whether allowing visual tokens to attend to subsequent text tokens at inference improves the performance of models trained with causal attention. In contrast, we introduce this pathway during multimodal instruction tuning, allowing models to adapt to this new pattern of bidirectional information flow. Through in-depth probing of visual representations, we further investigate whether the newly introduced language-to-vision flow improves multimodal integration.

C.3 Implementation Details of Linear Probe

Probe training. We train linear probes on $1,597$ GQA questions from 676 images. Visual patches overlapping the question's target boxes are labeled positive, and the remaining patches negative.

Table 6: **Adaptation of visual representations to changing question targets within the same image.** We report Target-shift AUROC for question pairs grouped by their task types (attr: attribute; obj: object; rel: relation). Higher scores indicate that relevance changes favor the newly queried region over the previous target. Results use visual representations at Layer 13 and are reported on a 0–100 scale, with 50 indicating chance. Off/On indicates whether reverse attention is enabled.

Semantic pair	Reverse	Vicuna-7B-v1.5			Qwen2.5-7B		
		CLIP	SigLIP2	DINOv2	CLIP	SigLIP2	DINOv2
attr+rel	Off	50.86	50.57	47.43	50.64	51.01	49.23
	On	**80.11**	**68.26**	**78.56**	**75.92**	**74.67**	**79.00**
	Δ	+29.25	+17.69	+31.13	+25.28	+23.65	+29.76
rel+rel	Off	49.89	48.28	47.61	50.57	50.27	48.66
	On	**77.67**	**67.64**	**79.41**	**77.56**	**77.90**	**79.59**
	Δ	+27.78	+19.36	+31.80	+26.98	+27.64	+30.92
attr+attr	Off	49.59	48.91	51.41	49.72	50.99	49.04
	On	**83.72**	**72.63**	**83.19**	**79.52**	**82.19**	**82.40**
	Δ	+34.13	+23.73	+31.78	+29.80	+31.20	+33.36
attr+obj	Off	49.35	46.75	50.02	51.36	48.88	50.66
	On	**82.53**	**71.55**	**81.51**	**82.19**	**81.08**	**81.76**
	Δ	+33.18	+24.80	+31.49	+30.83	+32.20	+31.10
obj+rel	Off	52.22	52.99	47.13	47.87	48.62	49.51
	On	**82.22**	**68.07**	**81.14**	**76.22**	**77.23**	**81.56**
	Δ	+30.00	+15.08	+34.01	+28.35	+28.61	+32.06

For each backbone–encoder configuration and attention condition, we independently train a two-class linear probe at the projector output and after every decoder layer. Each probe receives only the frozen visual hidden state of an individual patch. Features are standardized using training-set statistics, and each question contributes up to 16 positive and 16 negative patches.

Evaluation data. We evaluate on $1,180$ GQA questions organized into 590 same-image question pairs, with no image overlap with probe training. Each pair targets different regions of the same image, with limited overlap between the two target regions. The probe's positive-class probability serves as the patch relevance score. We evaluate all visual patches and keep the probe fixed when switching between the two questions, so score changes reflect changes in the visual representations. Evaluation covers all probed layers. The main analysis reports results at Layer 13, while Appendix C.5 presents results across all layers.

C.4 Details of Evaluation Metrics

In Section 6.2, we use three probe-based metrics to evaluate how visual representations are organized around the question. AUROC and NDCG@all measure how well the probe separates and ranks question-relevant patches, respectively; Target-shift AUROC measures whether relevance scores shift toward the newly queried region when the question changes for the same image.

Let s_i^q be the probe's relevance score for visual patch i under question q, and let $y_i^q \in \{0, 1\}$ indicate whether the patch belongs to the target region.

AUROC. For target patches P_q and background patches N_q, define

$$\mathrm{AUROC}_q = \frac{1}{|P_q||N_q|} \sum_{i \in P_q} \sum_{j \in N_q} \left[\mathbf{1}(s_i^q > s_j^q) + \tfrac{1}{2}\mathbf{1}(s_i^q = s_j^q) \right]. \tag{7}$$

This measures how often a target patch scores above a background patch, with half credit for ties. Higher values indicate that question-relevant and irrelevant visual states are more separable by the linear probe; 0.5 denotes chance and 1 denotes perfect separation.

NDCG@all. For all n_q patches ranked by decreasing score as π_q,

$$\mathrm{DCG}_q = \sum_{k=1}^{n_q} \frac{y^q_{\pi_q(k)}}{\log_2(k+1)}, \qquad \mathrm{NDCG@all}_q = \frac{\mathrm{DCG}_q}{\mathrm{IDCG}_q}, \tag{8}$$

where IDCG uses the ideal ordering of the same binary labels. Ties are handled by averaging over all possible orderings of equally scored patches. NDCG assigns greater weight to target patches near the top of the ranking. Higher values indicate that the probe ranks question-relevant patches closer to the top; 1 denotes an ideal ranking with all target patches preceding background patches.

Target-shift AUROC. For a question change from A to B, let $\Delta_i = s_i^B - s_i^A$, with positive set $P = P_B \setminus P_A$ and negative set $N = P_A \setminus P_B$. We compute

$$\text{Target-shift AUROC}_{A,B} = \frac{1}{|P||N|} \sum_{i \in P} \sum_{j \in N} \left[\mathbf{1}(\Delta_i > \Delta_j) + \tfrac{1}{2}\mathbf{1}(\Delta_i = \Delta_j) \right]. \tag{9}$$

Shared target patches and patches outside both targets are excluded. Higher values indicate that score changes more consistently favor the newly queried region over the previous target. A value of 1 means every B-exclusive patch has a larger score change than every A-exclusive patch; 0.5 indicates no systematic target-directed shift, and values below 0.5 indicate a shift in the opposite direction.

C.5 More Results on Question-Dependent Visual Representations analysis

We extend the representation analysis across layers, individual question types, and question-pair types to examine whether the effects of reverse attention persist beyond a single layer or task category. These analyses show improvements in question relevance and target-directed adaptation across tasks and model layers.

Layer-wise analysis. Figure 12 extends the analysis from the projector output through all 32 decoder layers of Vicuna and all 28 decoder layers of Qwen2.5. Reverse attention improves linearly decodable relevance across broad layer ranges, with larger differences generally emerging in intermediate layers. Causal Target-shift AUROC remains near chance, whereas reverse attention yields target-directed changes across all six configurations. Under reverse attention, scores generally rise toward intermediate layers and then decline slightly in deeper layers. Vicuna with SigLIP2 shows somewhat smaller gains in AUROC and NDCG, which may reflect differences in how the language backbone incorporates features from each visual encoder. Its clear improvement in Target-shift AUROC nevertheless supports the broader finding that reverse attention enables question-dependent visual adaptation.

Relevance across question types. We group questions by their task types: attribute (attr), object (obj), relation (rel), and category (cat). Table 5 reports within-group averages using the fixed probes at Layer 13. Reverse attention improves NDCG and AUROC in most groups, suggesting that it helps visual representations capture question-relevant information across different semantic types.

Target shifts across question-pair types. We also group question pairs according to the task types of the two questions. Table 6 reports results for five question-pair types. Reverse attention increases Target-shift AUROC in every task group across all six configurations, while causal scores remain near chance. These results suggest that reverse attention helps visual representations adapt to shifts in the question's target across a range of semantic demands.

C.6 Additional Visualizations of Question-Dependent Relevance Shifts

Figure 13 shows relevance shifting toward newly queried regions with reverse attention, while causal maps remain largely unchanged. For example, the question changes from the clock's location to the presence of donuts and a napkin; relevance decreases around the clock and increases over the tabletop target. Similarly, when the question changes from whether a fire hydrant is blue to the color of a car, relevance decreases around the hydrant and increases over the car. These examples suggest that language-to-vision flow enabled by reverse attention allows the processing of visual representations to adapt to changes in the queried objects.

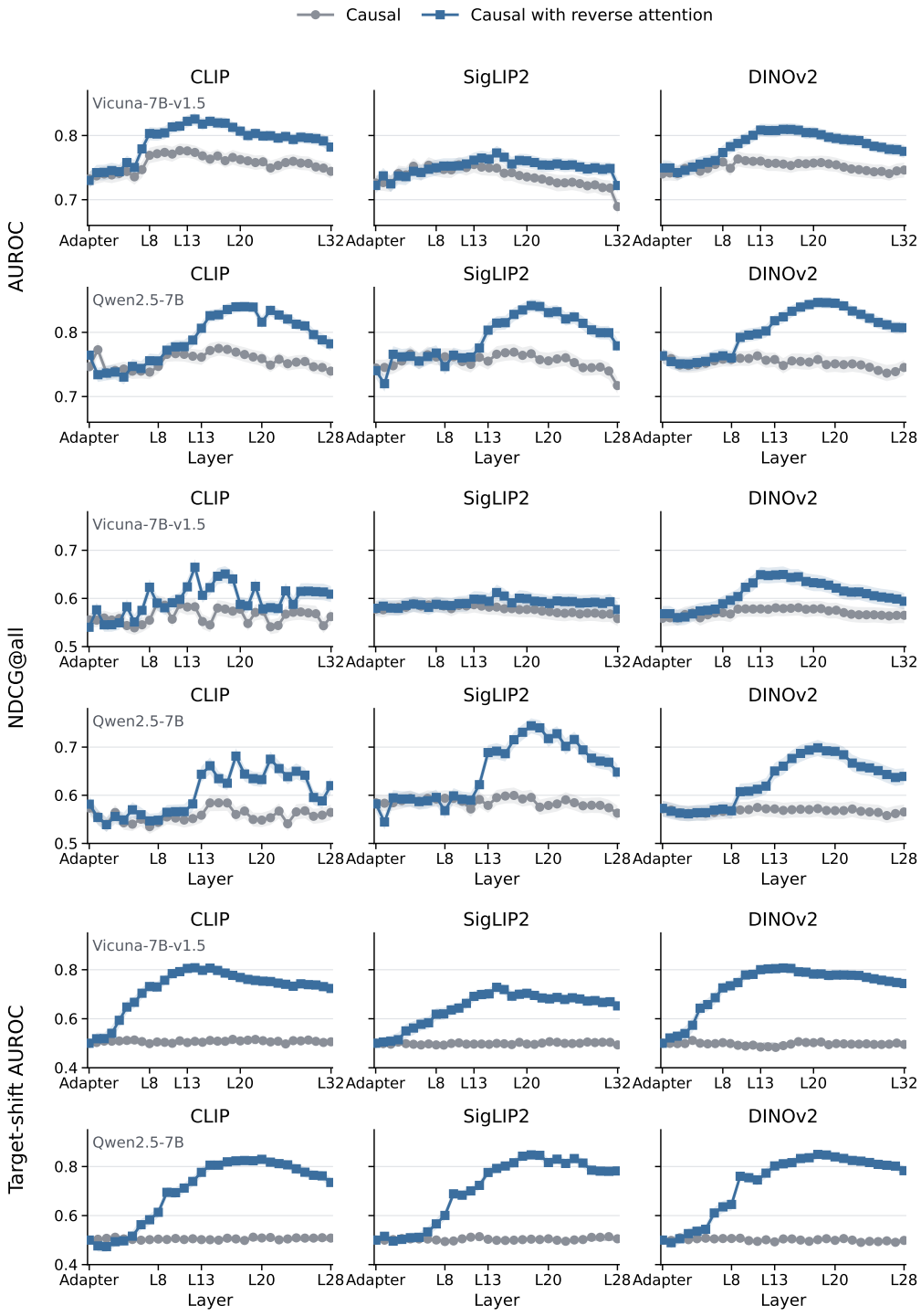


Figure 12: **Layer-wise question relevance and adaptation to question changes.** From top to bottom: AUROC, NDCG@all, and Target-shift AUROC. Each metric compares the two language backbones and three visual encoders, using an independent probe per layer and condition. Target-shift AUROC has a chance level of 0.5.

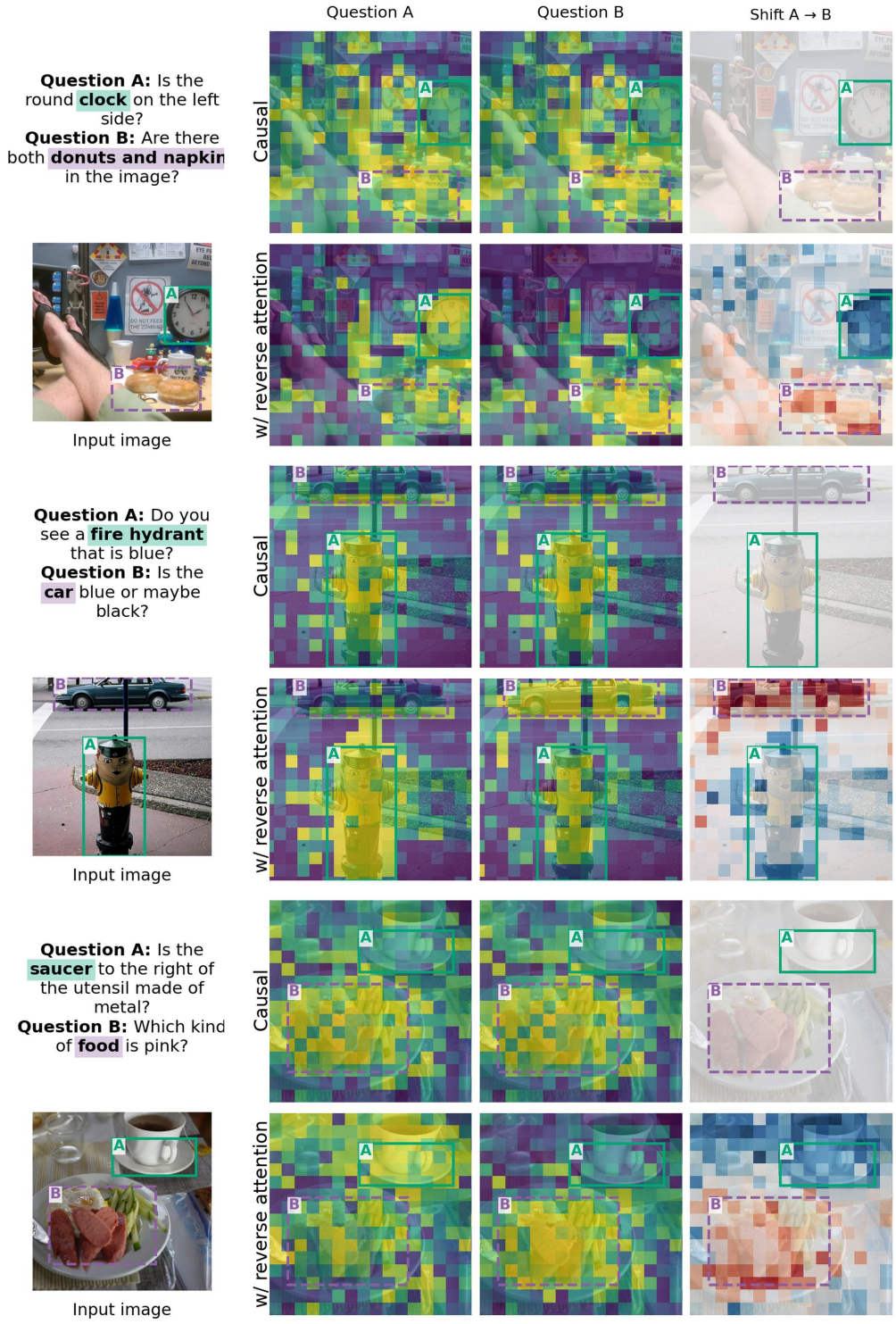


Figure 13: **Additional examples of question-dependent relevance shifts.** Rows compare causal and reverse attention; columns show relevance maps for questions A and B and their difference. Boxes mark the question targets. In difference maps, red/blue indicates increased/decreased relevance. Results are from Layer 13 of Vicuna-7B-v1.5 paired with the CLIP ViT-L/14 visual encoder.