Source-Modality Monitoring in Vision-Language Models

Etha Tianze Hua, Tian Yun, Ellie Pavlick
Department of Computer Science, Brown University
{tianze_hua,tian_yun,ellie_pavlick}@brown.edu

Abstract

We define and investigate *source-modality monitoring* – the ability of multimodal models to track and communicate the input source from which pieces of information originate. We consider source-modality monitoring as an instance of the more general *binding problem*, and evaluate the extent to which models exploit syntactic versus semantic signals in order to bind words like *image* in a user-provided prompt to specific components of their input and context (i.e., actual images). Across experiments spanning 11 vision-language models (VLMs) performing target-modality information retrieval tasks, we find that both syntactic and semantic signals play an important role, but that the latter tend to outweigh the former in cases when modalities are highly distinct distributionally. We discuss the implications of these findings for model robustness, and in the context of increasingly multimodal agentic systems.[1]

1 Introduction

We are moving toward a world of multimodal AI agents that must build and update world models from many sources of information and interaction (Li et al., 2024; Yang et al., 2025; Fung et al., 2025; Team et al., 2025a). In order for humans to interact effectively with such systems, AI models need to be able to track and communicate where (i.e., from which modality) information comes from. Borrowing terminology from cognitive psychology (Johnson et al., 1993; Henkel et al., 2000), we refer to this ability as *source-modality monitoring*.

We consider source-modality monitoring as an instance of the more general *binding problem* (Smolensky, 1990): a model must associate particular content with an abstract role, such as *image* or *text*, and retrieve the content corresponding to the queried role. Recent work has shown that language models solve related binding problems through abstract, symbolic-like mechanisms, for example by associating entities with their attributes (Feng & Steinhardt, 2024; Gur-Arieh et al., 2025) or binding values to variables (Wu et al., 2025). Given this, we might expect modern multimodal models to invoke a similar mechanism for binding inputs to modalities, especially since their inputs are often explicitly marked with modality-specific wrapper tokens such as `<image>` and `</image>`.

However, viewing source-modality monitoring as a case of purely symbolic binding misses many of the key properties. Unlike arbitrary role labels, words such as *image* and *text* are grounded in content with distinct distributional and semantic structure. Some kinds of inputs just *are* images, regardless of whether or not they are tagged as such. As a result, when a multimodal model is asked to report "what is in the image," it need not rely only on explicit modality markers; it should also be able to infer from the nature of the input itself whether or not a piece of information should bind to a word like *image*. Moreover, as AI becomes increasingly multimodal and agentic, it becomes unrealistic to assume that all the "modalities" to which a human might refer will be explicitly tagged. Humans might ask AI for information from "yesterday's meeting", or "the discussion about dinner plans". In such cases, AI will need to rely on contextual and distributional information about the content being bound, instead of or in addition to information available in explicit syntactic markers.

[1] All resources will be available at https://github.com/ethahtz/source-modality-monitoring.

In this paper, we ask if and how modern vision-language models (VLMs) accomplish source-modality monitoring. Source-modality monitoring can be studied at multiple levels, including mechanisms acquired over training (Khalifa et al., 2024) and mechanisms deployed in context at inference time. We focus on the latter, and investigate how it is accomplished in a variety of VLMs across model sizes and preprocessing choices. We first show that many of these models perform well on a task designed to test in-context source-modality monitoring, and that models with explicit marker tokens (e.g., `<image>&</image>`) generally outperform those without them. We then ask what kind of signal supports this performance. By structurally perturbing symbolic modality markers, we find that models' binding behavior is driven primarily by distributional semantic signals rather than purely symbolic ones, although explicit markers still improve selectivity. Finally, by applying optimized transformations to marker and content representations, we show that symbolic cues can nonetheless be exploited to reverse models' modality attribution. Together, these results provide a more nuanced picture of how syntactic and semantic signals combine to support binding in complex multimodal settings.

2 Related Work

Cross-Modal Conflict in VLMs Prior work has used inconsistent multimodal inputs to study modality bias in VLM predictions (Deng et al., 2025; Pezeshkpour et al., 2025; Zhang et al., 2025). While these studies show that VLMs can favor one modality over another depending on the model, dataset, or relative unimodal confidence, they typically use questions that remain ambiguously answerable from either modality, or focus mainly on the image side of the conflict. Our task builds on our prior work (Hua et al., 2025), which explicitly queries both image and caption information under inconsistent inputs. We broaden that evaluation to more VLMs across scales and model families, and study the underlying signals that enable VLMs to accomplish this task.

Interpretability Methods for Transformer-Based Models Analyses in this paper adapt three standard interpretability techniques to study how VLMs keep track of the source modality of a token. First, probing is a widely used technique to test whether a feature of interest is (linearly) decodable from a model's hidden states (Alain & Bengio, 2017; Conneau et al., 2018; Tenney et al., 2019), with control tasks separating genuine encoding from probe capacity (Hewitt & Liang, 2019; Belinkov, 2022). We use a linear probe with shuffled-label controls to establish that modality identity is linearly decodable from token representations. Next, activation patching establishes causal effects by copying activations from one forward pass into another and measuring the change in model output (Vig et al., 2020; Meng et al., 2022; Wang et al., 2023; Zhang & Nanda, 2024). Whereas patching is typically used to localize the components responsible for a behavior, we employ a freeze-remove variant – restoring contextualized content-token activations into marker-token-ablated runs – to test whether modality identity is written into content-token representations by the markers, rather than remaining localized at the marker positions. Finally, representation steering adds a vector to the residual stream at inference time, where vectors are either extracted from contrastive inputs (Li et al., 2023; Rimsky et al., 2024) or, as in our case, trained with gradient descent against an explicit objective with the model's native weights frozen (Subramani et al., 2022; Wu et al., 2024). While steering is typically a control method, we use trained vectors as a localization instrument, identifying the token positions and layer depths at which modality identity is most causally accessible.

3 How Well Do VLMs Bind Words to Their Inputs?

3.1 Target-Modality Retrieval Task

We first ask simply whether VLMs are capable of associating different parts of their input (i.e., images, text) with the words that refer to those inputs (*image*, *caption*)? We define the following task whose successful performance depends on a model's ability to monitor source modality. Each instance consists of a single in-context *inconsistent* image-caption

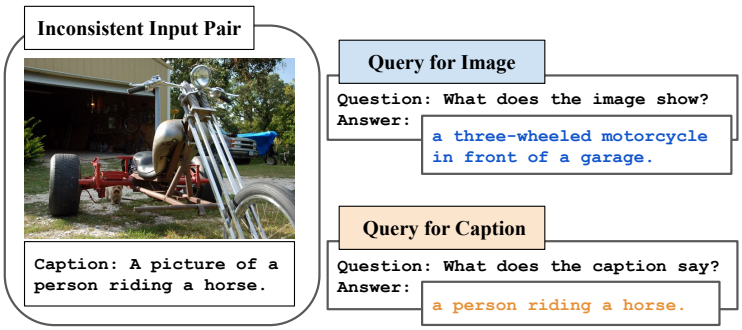


Figure 1: Example task instance.

pair, and the model is prompted to retrieve information from one specified input source (e.g., "what is in the image" or "what is in the caption"). We refer to the queried source as the *target modality* and the other source as the *non-target modality*. Figure 1 shows an example, illustrating the correct answers when a model is asked to retrieve image or caption information from its input pair. We exclude the possibility that VLMs rely on input order to retrieve information by evaluating them under both input orderings – image tokens first or text tokens first – and report the average result across the two conditions.[2]

Because the visual and textual inputs convey mutually exclusive information, answering correctly requires the model to retrieve content from the specified target modality. As the target modality is explicitly indicated, this task does not merely measure a model's overall modality bias, but instead targets at the model's ability to associate retrieved information with its input modality thus making it suitable for assessing source-modality monitoring.

3.2 Datasets and Models

To construct inconsistent image–caption pairs, we leverage existing image captioning datasets flickr30k (Young et al., 2014) and MSCOCO (Lin et al., 2015), and systematically sample captions that are inconsistent with the visual content of each image. We use a simple heuristic procedure to ensure that the caption paired with the image is genuinely mismatched at the semantic level. To do this, we require that the sampled caption have cosine similarity below a threshold of 0.2 with *any* of the five image's ground-truth associated captions, where similarity is computed using a BERT-style sentence embedding model (`sentence-transformers/all-mpnet-base-v2` (Song et al., 2020))[3].

We evaluate a range of 11 open-source VLMs on this task. Beyond differences in model sizes, these systems also vary along an architectural dimension that is especially relevant to our setting: whether the input preprocessing pipeline inserts explicit *image wrapper tokens* that mark the boundary of image tokens. Qwen-2.5-VL, InternVL3, and Gemma-3 models (Bai et al., 2025; Zhu et al., 2025; Team et al., 2025b) are trained with these wrapper tokens, while earlier models such as LLaVA-1.5 (Liu et al., 2023), InstructBLIP (Dai et al., 2023), and LLaVA-OneVision (Li et al., 2025) do not use such markers. See Table 5 for the image marker tokens used across model families. This variation allows us to ask whether the presence of modality-marking tokens is associated with stronger retrieval from the target modality.

3.3 Metrics and Baselines

Metrics To measure how reliably models retrieve information from the target modality, we first identify *valid predictions*, namely outputs that clearly report information from either the

[2]We adapt this approach to all models except InstructBLIP, which has a fixed order of having image tokens at the beginning of the input sequence.

[3]We find the threshold of 0.2 to be appropriate for filtering out semantically similar stimuli across modalities. For human validation of this threshold, see Appendix B.

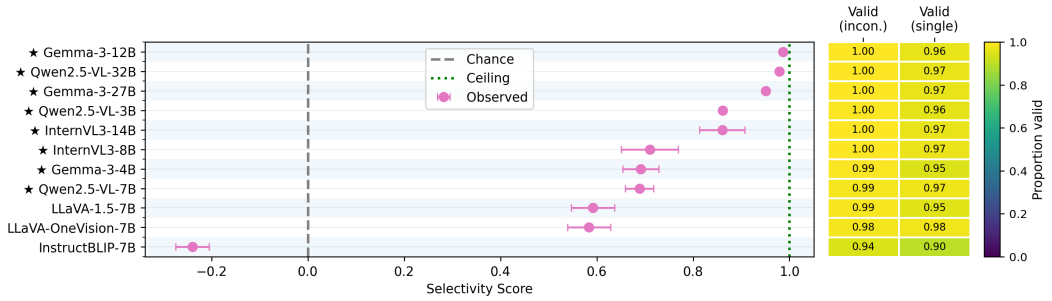


Figure 2: Left: Aggregated source-modality selectivity across VLMs and datasets. Error bars indicate standard deviation across datasets. ★ marks VLMs with special image wrapper tokens. Right: averaged rate of valid responses under inconsistent input pairs and under single modality inputs across datasets.

target modality or the non-target modality (i.e., as opposed to outputting nonsense, which models sometimes do). Then, we define *selectivity* as an aggregate metric which indicates the tendency of the model to return the modality that was requested in the prompt. To obtain selectivity, we compute the difference between the frequency with which the model reports information from the target modality and the frequency with which it reports information from the non-target modality, among these valid predictions.[4] This yields a score in the range $[-1, 1]$: a score of 1 indicates perfect retrieval from the target modality, a score of 0 indicates no preference between the two modalities, and a score of -1 indicates that the model always retrieves information from the non-target modality instead.

Since we construct our input pairs from captioning datasets, model outputs are often free-form and cannot be evaluated with exact string matching. We therefore use `GPT5.4-mini` (OpenAI, 2026) as an LLM judge to evaluate whether model outputs are grounded in the image, the caption, or neither. In Appendix D, we conduct a human validation study of the LLM judge, showing that it achieves a high alignment rate with human judgment. We also include examples of model outputs and judged outcomes in Appendix I.

Single-Modality Baseline To contextualize the rate of valid responses under the inconsistent two-modality input pairs, we include a single-modality baseline in which a model receives input from only one modality and is prompted to report information from that modality. We use the same GPT model as the LLM Judge to evaluate whether the response sufficiently reports what is in its input.

3.4 Results

Valid Responses Figure 2 shows the valid-response rates and selectivity scores for all 11 models in our evaluation. With the exception of InstructBLIP-7B, most models maintain a near-perfect valid-response rate under inconsistent inputs, in some cases even exceeding their single-modality baselines. One possible explanation is that inconsistent inputs encourage models to give more specific responses about the queried target modality, yielding slightly higher valid-response rates than in the single-modality setting. Still, the difference is small, suggesting that these models can report information from at least one modality with a level of specificity comparable to that seen when only a single modality is provided.

Modality Selectivity All models, except for InstructBLIP-7B, outperform the random baseline in selectivity, indicating that they retrieve information from the target modality more often than from the non-target modality. That said, there is substantial variation between models, and many smaller models struggle to reliably return the requested information. InstructBLIP-7B shows an interesting negative selectivity, indicating it more frequently

[4]See Appendix C for the formal definitions of valid prediction rate and selectivity.

reports information from the non-target modality than the target modality. Conceptually, the worst performing model would be at chance level – we suspect InstructBLIP is below chance level because of its fixed-order design, which puts the image tokens always before the textual tokens, that lead to a constant bias which leads to the negative selectivity. On the other end, models such as Gemma-3-12B and Qwen2.5-VL-32B perform nearly perfectly, almost achieving the ceiling performance. In general, larger models tend to perform better than smaller models; but we also observe that models with special image wrapper tokens (marked with ★ in the figure) outperform models without such tokens. We explore this mechanism in detail in subsequent sections.

4 Role of Symbolic vs. Distributional Signals in Word-Input Binding

The strong performance of VLMs which have special image wrapper tokens on the target-modality retrieval task might suggest that source-modality monitoring is driven primarily by these symbolic marker tokens. This would resemble the largely symbolic binding mechanisms previously studied in language models, such as binding entities to their attributes (Feng & Steinhardt, 2024; Gur-Arieh et al., 2025). In this section, however, we show a more nuanced picture. First, we show that VLMs are indeed capable of solving a purely symbolic binding problem when there is no distributional signal to exploit. Second, we confirm that image and text content tokens differ markedly in their representations' distributions, providing an additional/alternative cue for modality identity. Third, we show that source-modality monitoring is a hybrid syntactic-semantic mechanism: models rely in part on distributional content cues to guide retrieval, but symbolic markers help contextualize content tokens, leading to a more robust encoding of modality identity. We focus our analysis on three best-performing models from different model families from the previous section.

4.1 VLMs Are Capable of Symbolic Binding

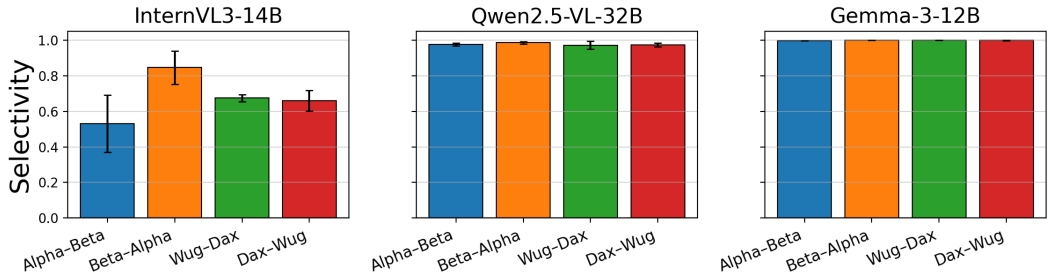


Figure 3: Performance on the purely symbolic retrieval task across four arbitrary label assignments. Each bar shows selectivity when image and caption content are associated with arbitrary labels rather than modality names. LabelA–LabelB indicates that LabelA is associated with image content and LabelB with caption content.

We begin by asking whether models can use symbols as pure indexing devices. To isolate this capacity, we test VLMs with a simpler symbolic binding task in which the two input sources are labeled by arbitrary labels such as *Alpha* and *Beta* or *Dax* and *Wug*. Row 2 of Table 1 shows an example in which *Dax* is associated with image content and *Wug* with caption content. In this setting, successful performance requires the model to bind each content span to its assigned label and later retrieve the queried one, but it does not require any notion of modality. In Figure 3, Qwen2.5-VL-32B and Gemma-3-12B perform nearly perfectly under this task. Although InternVL3-14B performs worse than the other two models, it still succeeds well above chance. These results show that VLMs are capable of using arbitrary symbols as indexing devices, just like their language model counterparts.

Condition	Input format
Unperturbed	<image> IM_1 ··· IM_n </image> Caption: a person riding a horse .
Arbitrary	Dax: IM_1 ··· IM_n Wug: a person riding a horse .
Remove	IM_1 ··· IM_n a person riding a horse
Swap	Caption: IM_1 ··· IM_n . <image> a person riding a horse </image>

Table 1: Image–caption span of the prompt under the 4 marker conditions used in our experiments. In **Unperturbed**, original modality markers are preserved; in **Arbitrary**, original modality markers are replaced with arbitrary labels; in **Remove**, both image and caption markers are deleted; and in **Swap**, the image and caption markers are exchanged.

4.2 VLMs Encode Tokens of Different Modalities in a Linearly Separable Manner

A second potential source of distinction comes from the distributional properties of image and text token representations. Here, we ask whether these representations are already separable before any contextual processing by the language model. Therefore, we examine them at the embedding layer of the backbone language model. To do so, we sample 200 instances from a dataset and, for each instance, sample the embedding representation of one token from each modality. We then measure cosine similarity between embedding pairs within and across modalities, and train a linear probe to distinguish image-token representations from text-token representations. To evaluate linear separability, we use stratified 3-fold cross-validation.

Across all model-dataset pairs, we find that the averaged within-modality cosine similarity (0.21 ± 0.06) is substantially higher than cross-modality cosine similarity (0.02 ± 0.02). This suggests that tokens from the same modality are more similar to one another than to tokens from the other modality. Consistent with this, a linear probe can separate image and caption tokens with perfect cross-validation accuracy across all models and datasets we test, indicating that image and text representations are linearly separable to a very high degree.[5]

4.3 Structural Perturbation on Symbolic Markers

Next, we study how symbolic and distributional signals each or together affect VLMs' ability to monitor source modalities, and retrieve from the target modality. To approach this, we systematically perturb the symbolic marker tokens by removing them or swapping them across modalities (see row 3 and 4 in Table 1). If the model uses a purely symbolic mechanism, we would expect to see a performance near chance when symbolic markers are removed, and models' predictions completely reversed when they are swapped; on the other hand, if the model relies solely on the distributional signals, we would expect these perturbations to have minimal effects on models' performance.

Figure 4 shows the result of swapping and removing symbolic marker tokens on models' performance in the target modality retrieval task. The aggregated results (subplots 3-5) shows that while removing marker tokens significantly reduces models' selectivity, it does not drop towards near-zero level, which indicates VLMs are able to use the residual signals within the distributions of image/text tokens to tell a difference between modalities. Moreover, swapping the marker tokens does not reverse models' predictions, rather it remains positive well above chance, which emphasizes that when symbolic and distributional signals conflict with each other, the model relies on the distributional signals more.

We also include results broken down by target modality in subplots 6–8 of Figure 4. We find that accuracy for reporting image content remains high across perturbation conditions, whereas accuracy for reporting caption content falls to near chance level (and even below chance for InternVL3). This sharp asymmetry suggests that VLMs learn a particularly robust

[5]See detailed results in Table 6 in the appendix.

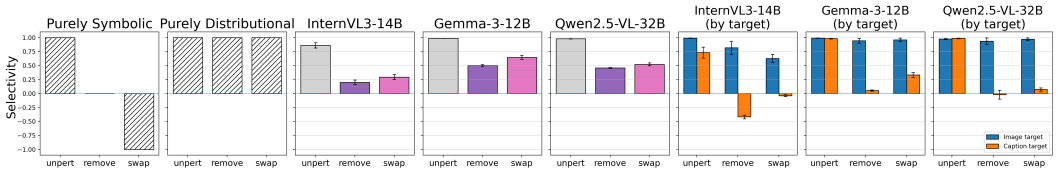


Figure 4: Selectivity under conditions where we remove or swap symbolic marker tokens, alongside purely symbolic and purely distributional expectations. Across all three VLMs, disrupting markers lowers selectivity but does not eliminate it, showing that source-modality retrieval is not reducible to symbolic cues alone. Modality-specific breakdowns show that image-target performance remains relatively unaffected, whereas caption-target performance drops sharply under perturbation, indicating that symbolic markers play a larger role in supporting caption retrieval.

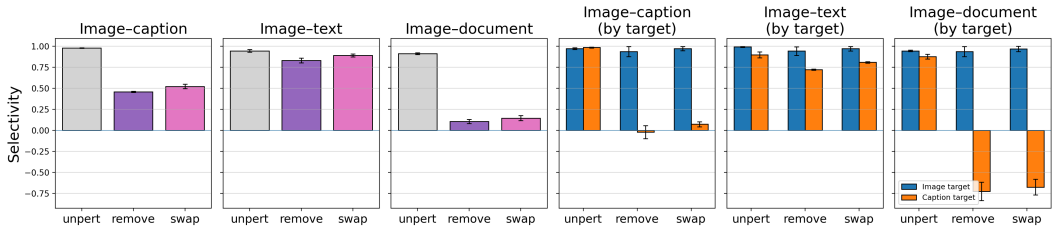


Figure 5: Selectivity across image–caption, image–text, and image–document settings under unperturbed, marker-removal, and marker-swap conditions in Qwen2.5-VL-32B. Marker perturbations have little effect in image–text, but substantially reduce performance in image–caption and image–document, where models rely heavily on markers.

way of associating the concept of "image" with image tokens, while relying more heavily on symbolic cues when retrieving caption information.

One plausible reason for the asymmetry noted above is that caption tokens are embedded among other textual spans in the input, so there is nothing about them that makes them inherently "caption" in the way image tokens are inherently "image."

We follow up on this intuition by varying the word used to refer to the caption, and observing the effect on the above trends. In this experiment, the text source is labeled as either *caption*, *text*, or *document*, and the query correspondingly asks what is in the image, caption, text, or document.

The results are shown in Figure 5. Models achieve high selectivity in the unperturbed condition, showing that they can reliably retrieve information from the queried source when the source markers are present. The key difference emerges under marker perturbations, which reveal that the model relies heavily on symbolic marker tokens when the text is referred to as the *document*, and relies on them very little when it is referred to as *text*. This suggests an interesting and intuitive mechanism, in which models encode a notion of the type of input that is typically referred to by specific words. That is, the model is happy to bind the word *text* to text in the prompt, whether or not that text is prefixed with a marker such as "Text:". In contrast, the model is resistant to considering an arbitrary short piece of text to be a *document* unless it is explicitly labeled as such.[6]

[6]One plausible reason is that the category *document* is intrinsically less distributionally distinct from images than *text*: documents can contain heterogeneous content, including embedded images, and therefore may provide a weaker standalone signal of source identity.

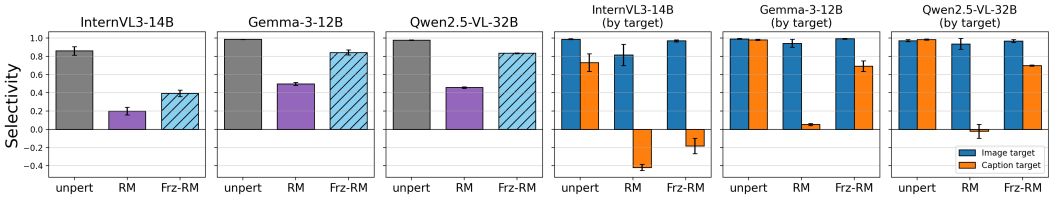


Figure 6: Selectivity under the freeze-remove intervention. Restoring contextualized content-token activations partially recovers performance relative to removing markers from the beginning, indicating that symbolic marker information is propagated into downstream content-token representations during contextualization.

4.4 Propagation of Symbolic Signals onto Content Tokens

In this subsection, we investigate how symbolic and distributional signals interact, focusing on whether symbolic marker tokens help write modality identity information into content-token representations. Our hypothesis is that this happens through contextualization across layers: marker tokens influence the hidden states of downstream content tokens, causing modality identity to become embedded in those representations rather than remaining localized at the marker positions.

To isolate this effect, we first run the model on the normal input with intact markers and collect the hidden activations at the content-token positions after they have been contextualized in the presence of those markers. We then perform a second run in which the marker tokens are removed, but the contextualized content-token activations from the clean run are patched in and frozen (Figure 11). We refer to this intervention as the freeze-remove condition (Frz-RM).

Figure 6 shows that symbolic information does not remain confined to the marker positions. In the Frz-RM condition, selectivity is partially restored relative to conditions in which markers are removed from the outset. This suggests that, during contextualization, information carried by the marker tokens is propagated into downstream content-token representations. Once contextualized in this way, content tokens encode not only their native distributional signals, but also the signal of modality identity inherited from the symbolic markers.

The modality-specific breakdown further shows that both image and caption content benefit from this contextualization, though the gain is much larger for caption content. For Gemma-3-12B and Qwen2.5-VL-32B, caption-target selectivity rises from near chance to around 0.6 under Frz-RM. For InternVL3-14B, the effect is weaker, but still mitigates the negative selectivity observed when markers are removed without using the contextualized activations. These results suggest that symbolic markers become more effective not only as explicit tokens, but also by shaping the representations of downstream content tokens.

5 Controllability of Symbolic vs Distributional Signals of Modality

5.1 Experiment Setup

We have shown that VLMs rely on both symbolic signals from marker tokens and distributional signals from content tokens to monitor input modalities, we now ask whether these signals can be manipulated to induce misattribution. Specifically, we study a setting in which we add two learned vectors, δ_1 and δ_2, to the image span and caption span, respectively (Figure 7), either at the marker positions or at the content-token positions at a specific layer. We then prompt the model to report, for example, the image content, but optimize these vectors to make the model instead produce the caption content, using the model's own caption-side prediction from a natural run as the target; symmetrically, we also optimize in the reverse direction, asking for caption content while driving the model toward its natural image-side prediction. If these marker and content representations indeed support source-modality monitoring, then there should exist transformations of them that

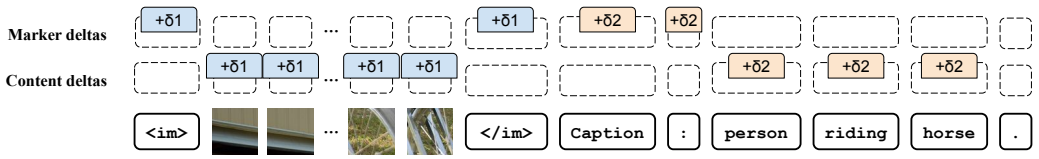


Figure 7: Illustration of the learned-vector intervention used to induce source misattribution. We add two trainable vectors, δ_1 and δ_2, either to the symbolic marker tokens (*marker deltas*) or to the modality-specific content tokens (*content deltas*) at a specific layer. The vectors are applied respectively to the image span and caption span, and are optimized to drive the model to report content from the non-queried modality.

systematically induce source misattribution. We initialize both δ vectors to zero and train them on 4,000 samples from the training split of MSCOCO. For each model, we optimize them at various relative layer depths and we repeat the procedure across 3 random seeds.

5.2 Results

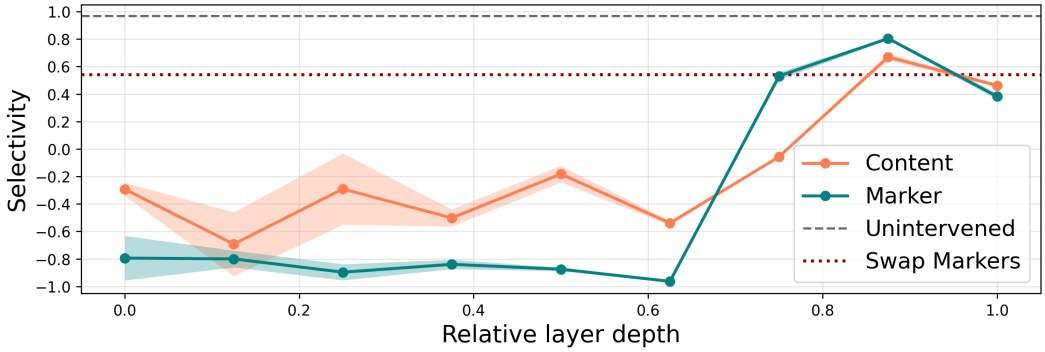


Figure 8: Selectivity of Qwen2.5-VL-32B after learned interventions at different layer depths. In early and intermediate layers, both intervention types can cause selectivity below zero, indicating that the model retrieves information from the non-queried modality more often than from the queried one. Marker-token interventions are especially effective, pushing selectivity close to -1 across a broad range of layers. Dashed and dotted lines show the selectivity of the unintervened model and under the token-level marker-swap condition.

Figure 8 shows the selectivity of Qwen2.5-VL-32B after applying learned δ vectors to either content spans or marker spans at different relative layer depths, evaluated on 100 validation examples unseen during training. In early and middle layers, interventions on both content and marker representations can drive the model to systematically report the opposite modality of the target modality. This effect is especially strong for marker-token interventions, which achieve selectivities close to -1 across a broad range of layers.

Notably, these learned interventions are substantially more effective than the marker-swapping manipulation presented earlier in the paper. This suggests that learned transformation vectors exploit modality-relevant information in marker representations more effectively than a simple symbolic swap. While both can influence how modality identity is propagated into downstream content representations, the learned vectors are optimized for this purpose and thus more successfully override native distributional cues. We observe similar trends in Gemma-3-12B and InternVL3-14B (see Figures 14 and 15, respectively).

The intervention effect weakens in later layers for both marker and content locations. This suggests that the representations supporting source-modality monitoring are most causally relevant in early and middle layers, and become harder to exploit after later-stage processing has already built on them. Taken together, these results show that both marker and content

representations encode source modality signals as directions that can be steered, but that the marker-tokens provide a stronger causal handle for accessing this direction.[7]

6 Discussion and Conclusion

We investigate the mechanism by which modern VLMs bind words in their input prompts (e.g., words like *image* or *text*) to specific components of their input/context (e.g., the image or the text). In experiments spanning 11 different vision-language models (VLMs), we find evidence that this type of binding involves both syntactic and semantic signals. That is, models exploit explicit marker tokens (such as `<image>` & `</image>`) when available, but also exploit distributional semantic cues. Indeed, we find that when the two signals are in conflict, models rely more on the distributional semantic cues than on the syntactic ones, though this varies as a function of the strength of the distributional signal.

Our experiments focus specifically on VLMs, but offer insights that are likely to extend to multimodal and agentic models in general. In particular, as models become increasingly complex, multimodal, and agentic, we expect that the importance of semantic signals for binding will become increasingly pronounced. This is due to the fact that models appear to have an innate preference for distributional semantic signals (when both signals are available), and because it will become increasingly impractical to provide explicit markers as the complexity of the input space grows. Indeed, our results comparing the difference between binding the words *text* vs. *caption* vs. *document* reveal the challenge of grounding references to "micromodalities" – segments of the input space that are different enough to humans to warrant different words, but which might be hard to distinguish to the model based on distributional signals alone. These point to an important direction for future work – to endow models with rich enough world models that they can not only predict future states, but refer back to aspects of those states and their origin via natural language.

Our results also provide a valuable case study for how the abstract "binding problem" in AI becomes more complicated in realistic contexts. Binding has frequently been used as an example of an abstract symbolic problem that challenges the innate tendencies of neural networks. The type of binding we consider, that necessary for source-modality monitoring, represents a more realistic type of binding which involves both abstract symbolic and associative semantic signals. Our experiments provide an initial picture of how these signals are traded off within modern AI models, and provide a framework for investigating similar mechanisms in general, for example considering how models balance logical and heuristic inferences (Dasgupta et al., 2024).

Acknowledgments

We are very grateful to the members of the Language Understanding and Representation (LUNAR) Lab at Brown University – especially Zhuonan Yang and Jennifer Meng Lu – as well as other members of the Brown Superlab for their valuable feedback on this paper. Etha Tianze Hua would like to thank Selina Liang for her support during this work. This project was in part supported by Schmidt Sciences Grant #GR5300958 and the NSF AI Research Institute on Interaction for AI Assistants Grant #GR5300593. Ellie Pavlick is a paid consultant for Google DeepMind. The content of this article does not necessarily reflect the views of the US Government or of Google, and no official endorsement of this work should be inferred.

[7] To provide additional baselines to the effects of intervening modality-related tokens, we include results of intervening non-modality-related token positions in Appendix L.2. See qualitative examples of VLMs' predictions being flipped after intervention in Appendix L.3.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

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.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL https://aclanthology.org/P18-1198/.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. Language models show human-like content effects on reasoning tasks, 2024. URL https://arxiv.org/abs/2207.07051.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3867–3876, 2025. doi: 10.1109/CVPR52734.2025.00366.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zb3b6oKO77.

Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, et al. Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355*, 2025.

Yoav Gur-Arieh, Mor Geva, and Atticus Geiger. Mixing mechanisms: How language models retrieve bound entities in-context, 2025.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=QYvFUlF19n.

Linda Henkel, N. Franklin, and {M. K.} Johnson. Cross-modal confusions between perceived and imagined events. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 26, January 2000. doi: 10.1037/0278-7393.26.2.321.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Tianze Hua, Tian Yun, and Ellie Pavlick. How do vision-language models process conflicting information across modalities? *arXiv preprint arXiv:2507.01790*, 2025.

Marcia K Johnson, Shahin Hashtroudi, and D Stephen Lindsay. Source monitoring. *Psychological bulletin*, 114(1):3, 1993.

Muhammad Khalifa, David Wadden, Emma Strubell, Honglak Lee, Lu Wang, Iz Beltagy, and Hao Peng. Source-aware training enables knowledge attribution in language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=UPyWLwciYz.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=zKv8qULV6n.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aLLuYpn83y.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, 2021.

Zaijing Li, Yuquan Xie, Rui Shao, Gongwei Chen, Dongmei Jiang, and Liqiang Nie. Optimus-1: Hybrid multimodal memory empowered agents excel in long-horizon tasks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=XXOMCwZ6by.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36, 2022. arXiv:2202.05262.

OpenAI. Introducing gpt-5.4 mini and nano, March 2026. URL https://openai.com/index/introducing-gpt-5-4-mini-and-nano/. Accessed March 31, 2026.

Pouya Pezeshkpour, Moin Aminnaseri, and Estevam Hruschka. Mixed signals: Decoding VLMs' reasoning and underlying bias in vision-language conflict. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24833–24848, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1351. URL https://aclanthology.org/2025.findings-emnlp.1351/.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1):159–216, 1990. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(90)90007-M. URL https://www.sciencedirect.com/science/article/pii/000437029090007M.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting latent steering vectors from pretrained language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.48. URL https://aclanthology.org/2022.findings-acl.48/.

Gemini Robotics Team, Abbas Abdolmaleki, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Ashwin Balakrishna, Nathan Batchelor, Alex Bewley, Jeff Bingham, Michael Bloesch, Konstantinos Bousmalis, Philemon Brakel, Anthony Brohan, Thomas Buschmann, Arunkumar Byravan, Serkan Cabi, Ken Caluwaerts, Federico Casarini, Christine Chan, Oscar Chang, London Chappellet-Volpini, Jose Enrique Chen, Xi Chen, Hao-Tien Lewis Chiang, Krzysztof Choromanski, Adrian Collister, David B. D'Ambrosio, Sudeep Dasari, Todor Davchev, Meet Kirankumar Dave, Coline Devin, Norman Di Palo, Tianli Ding, Carl Doersch, Adil Dostmohamed, Yilun Du, Debidatta Dwibedi, Sathish Thoppay Egambaram, Michael Elabd, Tom Erez, Xiaolin Fang, Claudio Fantacci, Cody Fong, Erik Frey, Chuyuan Fu, Ruiqi Gao, Marissa Giustina, Keerthana Gopalakrishnan, Laura Graesser, Oliver Groth, Agrim Gupta, Roland Hafner, Steven Hansen, Leonard Hasenclever, Sam Haves, Nicolas Heess, Brandon Hernaez, Alex Hofer, Jasmine Hsu, Lu Huang, Sandy H. Huang, Atil Iscen, Mithun George Jacob, Deepali Jain, Sally Jesmonth, Abhishek Jindal, Ryan Julian, Dmitry Kalashnikov, M. Emre Karagozler, Stefani Karp, Matija Kecman, J. Chase Kew, Donnie Kim, Frank Kim, Junkyung Kim, Thomas Kipf, Sean Kirmani, Ksenia Konyushkova, Li Yang Ku, Yuheng Kuang, Thomas Lampe, Antoine Laurens, Tuan Anh Le, Isabel Leal, Alex X. Lee, Tsang-Wei Edward Lee, Guy Lever, Jacky Liang, Li-Heng Lin, Fangchen Liu, Shangbang Long, Caden Lu, Sharath Maddineni, Anirudha Majumdar, Kevis-Kokitsi Maninis, Andrew Marmon, Sergio Martinez, Assaf Hurwitz Michaely, Niko Milonopoulos, Joss Moore, Robert Moreno, Michael Neunert, Francesco Nori, Joy Ortiz, Kenneth Oslund, Carolina Parada, Emilio Parisotto, Amaris Paryag, Acorn Pooley, Thomas Power, Alessio Quaglino, Haroon Qureshi, Rajkumar Vasudeva Raju, Helen Ran, Dushyant Rao, Kanishka Rao, Isaac Reid, David Rendleman, Krista Reymann, Miguel Rivas, Francesco Romano, Yulia Rubanova, Peter Pastor Sampedro, Pannag R Sanketi, Dhruv Shah, Mohit Sharma, Kathryn Shea, Mohit Shridhar, Charles Shu, Vikas Sindhwani, Sumeet Singh, Radu Soricut, Rachel Sterneck, Ian Storz, Razvan Surdulescu, Jie Tan, Jonathan Tompson, Saran Tunyasuvunakool, Jake Varley, Grace Vesom, Giulia Vezzani, Maria Bauza Villalonga, Oriol Vinyals, René Wagner, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Chengda Wu, Markus Wulfmeier, Fei Xia, Ted Xiao, Annie Xie, Jinyu Xie, Peng Xu, Sichun Xu, Ying Xu, Zhuo Xu, Jimmy Yan, Sherry Yang, Skye Yang, Yuxiang Yang, Hiu Hong Yu, Wenhao Yu, Wentao Yuan, Yuan Yuan, Jingwei Zhang, Tingnan Zhang, Zhiyuan Zhang, Allan Zhou, Guangyao Zhou, and Yuxiang Zhou. Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer, 2025a. URL https://arxiv.org/abs/2510.03342.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025b.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In *International conference on learning representations*, volume 2024, pp. 17282–17333, 2024.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Yiwei Wu, Atticus Geiger, and Raphaël Millière. How do transformers learn variable binding in symbolic programs? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=kVtyv7bpnw.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=fykjplMc0V.

Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, Yuquan Deng, Lars Liden, and Jianfeng Gao. Magma: A foundation model for multimodal ai agents, 2025. URL https://arxiv.org/abs/2502.13130.

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. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006/.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Zhuoran Zhang, Tengyue Wang, Xilin Gong, Yang Shi, Haotian Wang, Di Wang, and Lijie Hu. When modalities conflict: How unimodal reasoning uncertainty governs preference dynamics in mllms, 2025. URL https://arxiv.org/abs/2511.02243.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Dataset Details

We use two image–captioning datasets: MSCOCO 2017 captions and Flickr30k.

Flickr30k. For Flickr30k, we use the `test` split from the `lmms-lab/flickr30k` release on HuggingFace. We then randomly permute the examples using a fixed seed and subsample 4,000 examples for train, 2,000 for validation, and 2,000 for test.

MSCOCO. For MSCOCO, we use the 2017 captioning data. We sample the training set from the official `train2017` split, and the validation and test sets from the official `val2017` split. Specifically, after randomly permuting examples with a fixed seed, we sample 4,000 examples from `train2017` for train, and 2,000 examples each from `val2017` for validation and test.

Preprocessing. We lightly normalize captions before use by collapsing repeated whitespace, stripping surrounding quotation marks, and ensuring sentence-final punctuation when appropriate. These steps reduce superficial formatting variation while preserving semantic content.

Use in our experiments. Both datasets are used to construct the image–text inputs in our source-modality monitoring task. Because each image is paired with textual descriptions, these datasets allow us to build inconsistent image–text examples, as well as image-only and text-only controls.

B Validation of Cosine-Similarity Threshold Used for Dataset Construction

The function of the threshold is to ensure that the semantic content of the image and its pairing caption is sufficiently dissimilar. To verify that this threshold reasonably enforces the desired dissimilarity constraint between an image and a caption, we conduct the following analysis.

For each sample in the test split of the MSCOCO dataset (which contains 1 image and 5 ground-truth captions per sample), we randomly select one of the 5 ground-truth captions and embed it into a vector using the same BERT-style sentence embedding model employed in the main paper. Then, for each interval of width 0.2 between 0 and 1, we sample 20 pairs of vectors whose cosine similarity falls within that interval. For each pair of vectors, we retrieve the corresponding original image-caption pairs and submit them for human evaluation to determine whether the image and caption are semantically aligned. In total, we collect 100 such pairs, with results summarized in Table 2.

Caption Cosine-Similarity Bin	Aligned / Total	Alignment Rate
0.0–0.2	0/20	0.00
0.2–0.4	0/20	0.00
0.4–0.6	2/20	0.10
0.6–0.8	8/20	0.40
0.8–1.0	16/20	0.80

Table 2: Human-judged alignment rate as a function of caption cosine-similarity bin.

We observe that as cosine similarity increases, the likelihood that a human evaluator judges the image-caption pair as aligned also increases. To ensure that our inputs are consistently misaligned, we adopt a threshold of 0.2: within this bin, none of the 20 sampled pairs were judged as aligned. This result supports the validity of using caption cosine similarity < 0.2 as a proxy for filtering out inconsistent image-caption pairs.

C Measuring Valid Response Rate and Selectivity

We evaluate source-modality monitoring with two separate metrics: *valid-response rate* (p_{valid}) and *selectivity*. Intuitively, p_{valid} measures whether the model gives an on-task response that can be attributed to one of the two candidate modalities, while selectivity measures, among such valid responses, whether the model reports information from the requested modality rather than the non-requested one.

Formally, let $\hat{y}_i$ denote the model output for example i, and let the queried modality be $m \in \{\text{image}, \text{caption}\}$, with the other modality denoted $\neg m$. For each modality, define a modality-assignment indicator

$$M_i^{(m)}(\hat{y}_i) \in \{0, 1\},$$

which equals 1 if the output is judged to be grounded in modality m, and 0 otherwise. Likewise, $M_i^{(\neg m)}(\hat{y}_i)$ indicates whether the output is judged to be grounded in the non-queried modality. We use `GPT5.4-mini` as an LLM judge to decide whether a piece of response is grounded in image, in caption, or in neither.

We define a response to be *valid* if it is an on-task answer that the judge assigns to any one of the two modalities. This gives the binary indicator

$$\text{valid}(\hat{y}_i) = M_i^{(m)}(\hat{y}_i) + M_i^{(\neg m)}(\hat{y}_i).$$

The valid-response rate is then

$$p_{\text{valid}} = \mathbb{E}[\text{valid}(\hat{y})].$$

We define selectivity as the tendency, among valid responses, to report information from the queried modality rather than the non-queried modality:

$$S = \mathbb{E}\Big[M^{(m)}(\hat{y}) - M^{(\neg m)}(\hat{y}) \;\Big|\; \text{valid}(\hat{y}) = 1\Big].$$

This score ranges from -1 to 1: a value of 1 indicates that all valid responses are grounded in the queried modality, a value of 0 indicates no preference between the two modalities, and a value of -1 indicates that all valid responses are grounded in the non-queried modality.

D LLM Judge Validation

To validate our use of GPT-5.4-mini as the LLM judge, we randomly sample 100 VLM predictions from Section 3.4, along with their corresponding classification labels produced by the GPT LLM judge. We then present the exact same instructions and content given to the GPT LLM judge to human evaluators, and compare the classification labels produced by the GPT LLM judge against those produced by the human evaluators. We obtain an agreement rate of 88% across the 100 classifications. Table 3 provides a breakdown of the classifications from GPT and human evaluators.

Human ↓ / GPT →	Image	Text	Neither	Row Total
Image	**50**	0	0	50
Text	1	**36**	0	37
Neither	9	2	**2**	13
Col. Total	60	38	2	**100**

Table 3: Agreement between GPT-judge and human classifications of VLM predictions. Diagonal entries (bold) indicate agreement: $88/100 = 0.88$.

We observe that the overall agreement rate (0.88) is very high, though GPT judgments are not perfectly aligned with human judgments. In particular, many VLM predictions that humans classify as grounded in neither the image nor the caption are instead classified by

Dimension	Values	Description
Source order	image first; caption first	Whether the image span appears before or after the caption span in the prompt.
Image query wording	*What is in the image?*; *Describe the image.*	Alternative phrasings for querying image content.
Caption query wording	*What is in the caption?*; *What does the caption say?*	Alternative phrasings for querying caption content.

Table 4: Controlled prompt variations used in the target-modality retrieval task.

```
<bos><bos><start_of_turn>user

<start_of_image><image_soft_token><end_of_image>

Caption: A city view of an E-train on a track.
Question: What is in the image?
Answer the question with a short sentence.
<end_of_turn>
<start_of_turn>model
Answer:
```

Figure 9: Representative raw prompt template for one model family, shown with the original chat formatting and special image tokens.

GPT as grounded in one of the input modalities. Upon inspecting individual instances, we find that the GPT judge tends to be more permissive in these cases: some predictions from earlier-generation models (LLaVA, InstructBLIP) are topically related but very brief, and the GPT judge classifies these as grounded in one of the input modalities, whereas human evaluators tend not to. Despite this asymmetry, the strong overall agreement indicates that GPT judgments serve as a reasonable proxy for human evaluation at the scale required by our study.

E Input and Query Prompt Templates for VLMs

Our prompt templates vary along two controlled dimensions: (1) the order in which the image and caption inputs appear, and (2) the surface form used to query the target modality. For source order, we evaluate both image-first and caption-first variants. For query wording, we use two formulations for each modality: for image queries, *What is in the image?* and *Describe the image.*; for caption queries, *What is in the caption?* and *What does the caption say?* Table 4 summarizes these variations, and Figure 9 shows one representative raw prompt including the model-specific chat template and special image tokens.

F Evaluation Prompt Templates for LLM Judge

See Figure 10 for prompt templates that we use for evaluating VLM model outputs with an LLM Judge.

G Examples of Marker and Content Tokens

Table 5 summarizes representative marker and content tokens for the image and caption modalities across the model families studied in this paper.

Prompt template: Inconsistent input

System
You are a careful evaluator for vision-language model outputs. Use only the user message. Reply with JSON only, matching the requested shape.

User
The model saw an image and a written caption that describes a *different* scene (inconsistent). It was told to report one modality (image vs. text).

Reference A (what is in the image):
`[image caption set]`

Reference B (caption text next to the image; may disagree with the image):
`[paired caption]`

Model answer:
`[prediction]`

Which reference is the answer primarily grounded in? Output only the JSON object with `closer_to` one of: `image`, `text`, `neither`.

Prompt template: Image-only input

System
You are a careful evaluator for vision-language model outputs. Use only the user message. Reply with JSON only, matching the requested shape.

User
The model saw only an image.

Reference descriptions of the image:
`[image caption set]`

Model answer:
`[prediction]`

Decide whether the answer is broadly about the same image content. Be permissive: the answer does not need to match the reference exactly. Output only the JSON object with boolean field `aligns_with_image`.

Prompt template: Text-only input

System
You are a careful evaluator for vision-language model outputs. Use only the user message. Reply with JSON only, matching the requested shape.

User
The model saw only this text:

`[paired caption]`

Model answer:
`[prediction]`

Decide whether the answer is broadly about the same text content. Output only the JSON object with boolean field `aligns_with_text`.

Figure 10: Prompt templates used for GPT-based evaluation in the inconsistent, image-only, and text-only settings.

Modality	Model Family	Start Marker	Content	End Marker
Image	LLaVA-1.5	–	VISUAL TOKENS	–
	Qwen2.5-VL	<vision_start>	VISUAL TOKENS	<vision_end>
	InternVL3	<im>	VISUAL TOKENS	</im>
	Gemma3	<start_of_image>	VISUAL TOKENS	<end_of_image>
Text (Caption)	All models	"Caption:"	TEXT TOKENS	–

Table 5: Marker tokens for image and caption inputs across different VLM model families.

H Additional Results on distributional separation between visual and textual content-token embeddings

Table 6 shows the results of cosine similarity (within- and across- modalities) and 3-fold cross validation probe accuracy on classifying image vs caption tokens.

Metric	Mean	Std
Within-image cosine similarity	0.325	0.093
Within-caption cosine similarity	0.101	0.039
Average Within-modality cosine similarity	0.213	0.064
Cross-modality cosine similarity	0.020	0.018
Linear probe accuracy (3-fold CV)	0.999	0.001
Control: shuffled labels – accuracy (3-fold CV)	0.491	0.018

Table 6: Distributional separation between visual and textual content-token embeddings across 16 model-dataset pairs. Cosine rows: mean and sample standard deviation across model-dataset pairs (within-modality is the mean of within-image and within-caption for each pair). Linear probe rows: average of mean 3-fold cross-validation scores and average of cross-validation standard errors across pairs. Control task rows use the same cross-validation splits with labels randomly permuted (Hewitt & Liang, 2019).

I Qualitative Examples of Model Outputs

Here, we provide qualitative examples showing model behaviors under inconsistent inputs.

GPT predicted label: `image` (closer to gold image captions)

Model: `Salesforce/instructblip-vicuna-7b` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`image`
GPT label	`image` (*outcome*: `correct`)
Gold image caption	*A very aged woman wearing blue is standing in front of a pile of things many in black trash bags.*
Paired caption	*a girl in a green uniform and a girl in a white uniform are holding lacrosse sticks.*
Model answer	a girl in a blue uniform is holding a banana

GPT predicted label: `text` (closer to paired caption)

Model: `Salesforce/instructblip-vicuna-7b` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`image`
GPT label	`text` (*outcome*: `misled`)
Gold image caption	*A man in a leather coat plays a saxophone into a microphone.*
Paired caption	*the man is wearing a backpack and sunglasses while walking through an outdoor market.*
Model answer	the man is wearing a backpack and sunglasses while walking through an outdoor market

GPT predicted label: `neither`

Model: `Salesforce/instructblip-vicuna-7b` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`image`
GPT label	`neither` (*outcome*: `neither`)
Gold image caption	*Two young women are embracing at the side of a road and looking at an mp3 player.*
Paired caption	*A small car decorated with flames and other paint is driving amongst a crowd with a tub and flags on the roof.*
Model answer	carnival

GPT predicted label: `image` (closer to gold image captions)

Model: `Salesforce/instructblip-vicuna-7b` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`text`
GPT label	`image` (*outcome*: `misled`)
Gold image caption	*Girl in black pants and an orange top jumping with one leg outstreached and foot pointed other in background dressed the same.*
Paired caption	*Three people are having a conversation in a living room.*
Model answer	people

GPT predicted label: `text` (closer to paired caption)

Model: `Salesforce/instructblip-vicuna-7b` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`text`
GPT label	`text` (*outcome*: `correct`)
Gold image caption	*Brown horse standing in the background while man in a white cowboy hat and blue long-sleeve shirt is walking off.*
Paired caption	*A brown and white dog jumps in the air to catch a snowball.*
Model answer	dog

GPT predicted label: `neither`

Model: `Salesforce/instructblip-vicuna-7b` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`text`
GPT label	`neither` (*outcome*: `neither`)
Gold image caption	*A man wearing a blue t-shirt and shorts is holding a newborn baby who is yawning.*
Paired caption	*A dog playing with a purple toy.*
Model answer	nothing

GPT predicted label: `image` (closer to gold image captions)

Model: `llava-hf/llava-1.5-7b-hf` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`image`
GPT label	`image` (*outcome*: `correct`)
Gold image caption	*A young man wearing a bandanna on his head stands leaning on a short brick wall with flowers behind him.*
Paired caption	*A group of people in mostly casual dress meet at an L-shaped table with laptops and notebooks.*
Model answer	a man wearing a white tank top is sitting on a brick wall

GPT predicted label: `text` (closer to paired caption)

Model: `llava-hf/llava-1.5-7b-hf** **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`image`
GPT label	`text` (*outcome*: `misled`)
Gold image caption	*The lady in the black dress is standing near the edge of the cliff taking pictures.*
Paired caption	*Two girls with brown hair are taking a picture of themselves.*
Model answer	two girls are taking a picture of themselves

GPT predicted label: `neither`

Model: `llava-hf/llava-1.5-7b-hf` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`image`
GPT label	`neither` (*outcome*: `neither`)
Gold image caption	*Two young girls are arm wrestling in their hotel room while another girl watches.*
Paired caption	*A young female wearing sandals is in a courtyard and is wearing a colorful backpack and green pants.*
Model answer	a girl

GPT predicted label: `image` (closer to gold image captions)

Model: `llava-hf/llava-1.5-7b-hf` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`text`
GPT label	`image` (*outcome*: `misled`)
Gold image caption	*Man with blue plaid shirt and white pants skateboards through a building that has graffiti art.*
Paired caption	*A group of elderly people stand around a booth at a food co-op.*
Model answer	a man riding a skateboard

GPT predicted label: `text` (closer to paired caption)

Model: `llava-hf/llava-1.5-7b-hf` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`text`
GPT label	`text` (*outcome*: `correct`)
Gold image caption	*A group of people in blue Arabian robes and white headdresses is on yellow bicycles in the street in front of a tall building.*
Paired caption	*Three men in work-related uniforms standing next to a building.*
Model answer	three men in work-related uniforms standing next to a building

GPT predicted label: `neither`

Model: `llava-hf/llava-1.5-7b-hf` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`text`
GPT label	`neither` (*outcome*: `neither`)
Gold image caption	*A man in a dark suit talks with a middle-aged man and woman on a stage in front of a large logo of the letter " D " with the numeral five within it.*
Paired caption	*One man is helping another man with red shorts do a back flip in the yard.*
Model answer	nothing

GPT predicted label: `image` (closer to gold image captions)

Model: `Qwen/Qwen2.5-VL-32B-Instruct` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`image`
GPT label	`image` (*outcome*: `correct`)
Gold image caption	*Racers run for the finish line in an indoor meet and one man has a big lead.*
Paired caption	*A young woman in a pink shirt and jeans is walking past some trees.*
Model answer	the image shows athletes competing in an indoor track event

GPT predicted label: `text` (closer to paired caption)

Model: `Qwen/Qwen2.5-VL-32B-Instruct` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`image`
GPT label	`text` (*outcome*: `misled`)
Gold image caption	*Female football fan wearing brown tank top and blue jean shorts attends a football game in a stadium.*
Paired caption	*Two blond women are smiling while one is holding a camera to capture their image.*
Model answer	two blond women are smiling while one is holding a camera to capture their image

GPT predicted label: `neither`

Model: `Qwen/Qwen2.5-VL-32B-Instruct` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`image`
GPT label	`neither` (*outcome*: `neither`)
Gold image caption	*A woman in a blue dress and rain boots mops the cobblestone street.*
Paired caption	*People are walking in front of a tourist information building displaying the following on its facade : pain is temporary quitting lasts forever.*
Model answer	a person wearing blue is sweeping the street while people walk by a tourist information building

GPT predicted label: `image` (closer to gold image captions)

Model: `Qwen/Qwen2.5-VL-32B-Instruct` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`text`
GPT label	`image` (*outcome*: `misled`)
Gold image caption	*People are walking in front of a tourist information building displaying the following on its facade : pain is temporary quitting lasts forever.*
Paired caption	*A woman with blond-hair and a white shirt and white apron is cooking.*
Model answer	the caption says, "pain is temporary, quitting lasts forever."

GPT predicted label: `text` (closer to paired caption)

Model: `Qwen/Qwen2.5-VL-32B-Instruct` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`text`
GPT label	`text` (*outcome*: `correct`)
Gold image caption	*Two gentle of asian descent sitting at a table or desk seemingly to be concentrating on a project the two are doing together.*
Paired caption	*There is a gray-haired man wearing a black and orange jacket sitting on chair on the sidewalk playing an accordion.*
Model answer	a gray-haired man wearing a black and orange jacket sits on a chair on the sidewalk playing an accordion

GPT predicted label: `neither`

Model: `Qwen/Qwen2.5-VL-32B-Instruct` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`text`
GPT label	`neither` (*outcome*: `neither`)
Gold image caption	*Two women with white and purple gowns while posing for the camera.*
Paired caption	*An Asian man sells a hot confection from his stand in a crowded market.*
Model answer	a caption is a brief description of an image or scene

GPT predicted label: `image` (closer to gold image captions)

Model: `google/gemma-3-12b-it` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`image`
GPT label	`image` (*outcome*: `correct`)
Gold image caption	*A man wearing a light blue shirt and gray pants is crouched in a field holding nuts in his hand.*
Paired caption	*Skateboarder performs a stunt on an uphill residential street.*
Model answer	the image shows a man crouched in a field, tending to plants and holding a root vegetable in his hand

GPT predicted label: `text` (closer to paired caption)

Model: `google/gemma-3-12b-it` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`image`
GPT label	`text` (*outcome*: `misled`)
Gold image caption	*A woman wearing a blue shirt and high heels stands on the sidewalk next to a man.*
Paired caption	*A man in jeans and a leather jacket walking by a poster advertising Miller Lite.*
Model answer	the image shows a woman in a blue dress and a man with headphones standing on a sidewalk next to a store window displaying lingerie

GPT predicted label: `image` (closer to gold image captions)

Model: `google/gemma-3-12b-it` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`text`
GPT label	`image` (*outcome*: `misled`)
Gold image caption	*A man uses two sticks to control two tan goats in a pen while a black and white dog watches.*
Paired caption	*Two females sitting on large backed stools in front of a counter inside of a house with one female turning her head back.*
Model answer	the caption describes a scene of a sheep herding demonstration with a handler, a dog, and sheep

GPT predicted label: `text` (closer to paired caption)

Model: `google/gemma-3-12b-it` **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`text`
GPT label	`text` (*outcome*: `correct`)
Gold image caption	*A hockey player in black and white collides with another player wearing blue.*
Paired caption	*A man in a gray shirt sits with several pairs of shoes around him.*
Model answer	the caption says "a man in a gray shirt sits with several pairs of shoes around him."

GPT predicted label: `neither`

Model: `google/gemma-3-12b-it** **Dataset:** `flickr30k`
GPT judge: `openai/gpt-5.4-mini`

Target modality	`text`
GPT label	`neither` (*outcome*: `neither`)
Gold image caption	*People viewing a sculpture with a man without a face but with a suit tie glasses and a hat holding a multi color umbrella.*
Paired caption	*Three men are working on some kind of equipment but I 'm not sure whether it 's a ship or an oil rig.*
Model answer	the caption says "guts and capture."

J Experimental Setup for the Freeze-Remove Condition

Figure 11 shows an illustration of the freeze-remove experiment condition. We collect activations from a clean run (with the marker tokens intact), and then patch these collected activations into a corrupt run where the marker tokens are removed – we then allow any downstream computation to continue based on the patched-in activations. This allows us to study the effect of contextualization of content tokens in source-modality monitoring.

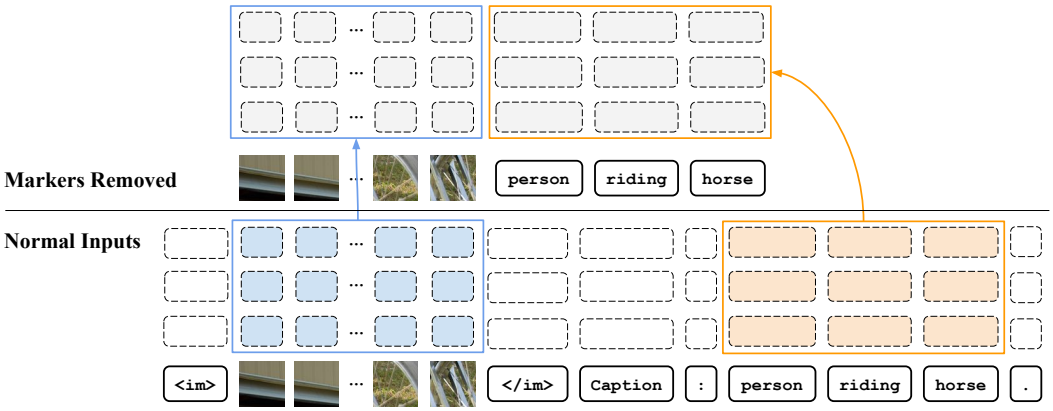


Figure 11: Illustration of the freeze-remove condition. We first collect hidden activations at content-token positions from a normal run with intact markers, then patch these activations into a run where the marker tokens are removed. This tests whether symbolic marker information has been contextualized into the content-token representations.

K Additional Results for Comparing Alternative Source-Modality Pairs

In the main paper, we show result for Qwen2.5-VL-32B on comparing alternative source-modality pairs of image–caption, image–text, and image–document. Here, we include results for Gemma-3-12B and InvernVL-3-14B (Figure 12 and Figure 13). The overall trend is similar across all three models.

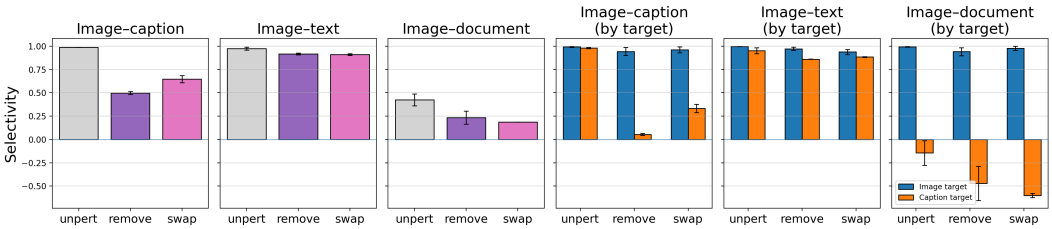


Figure 12: Selectivity across image–caption, image–text, and image–document settings under unperturbed, marker-removal, and marker-swap conditions in Gemma-3-12B. Marker perturbations have little effect in image–text, but substantially reduce performance in image–caption and image–document, where models rely heavily on markers.

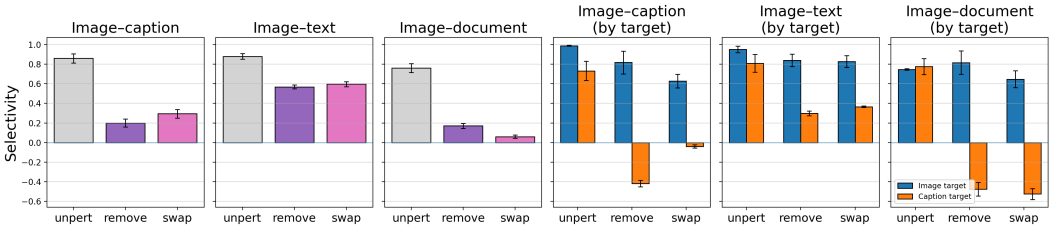


Figure 13: Selectivity across image–caption, image–text, and image–document settings under unperturbed, marker-removal, and marker-swap conditions in InternVL3-14B. Marker perturbations have somewhat effect in image–text, but substantially reduce performance in image–caption and image–document, where models rely heavily on markers.

L Additional Results for Controllability of Marker vs Content Token Representation

L.1 Quantitative Results for Gemma-3-12B and InternVL3-14B

In the main paper, we present results for Qwen2.5-VL-32B on the exploitability of marker versus content tokens. Here, we include results for Gemma-3-12B and InternVL3-14B (Figures 14 and 15). The general trend holds in both models: in early and middle layers, learned vectors induce a much higher rate of misattribution, with a weaker effect in later layers. Intervening on the marker-token representations also in general produces a stronger misattribution effect.

L.2 Baseline Comparisons: Intervening at Non-Marker/Content Positions

The experiments in Section 5 aim to locate *where* causal manipulation of source-modality attribution is most effective, rather than to isolate a clean, disentangled modality-specific direction or subspace. In our setting, compactly isolating a single direction or low-dimensional subspace that encodes input-modality identity is difficult, in part because image inputs span hundreds of tokens—unlike the few-token entity/attribute setups of prior activation-intervention work such as Feng & Steinhardt (2024). We therefore frame these results as identifying the positions at which intervention is most effective, and we find that intervening on both marker and content tokens is effective, more so on the marker tokens.

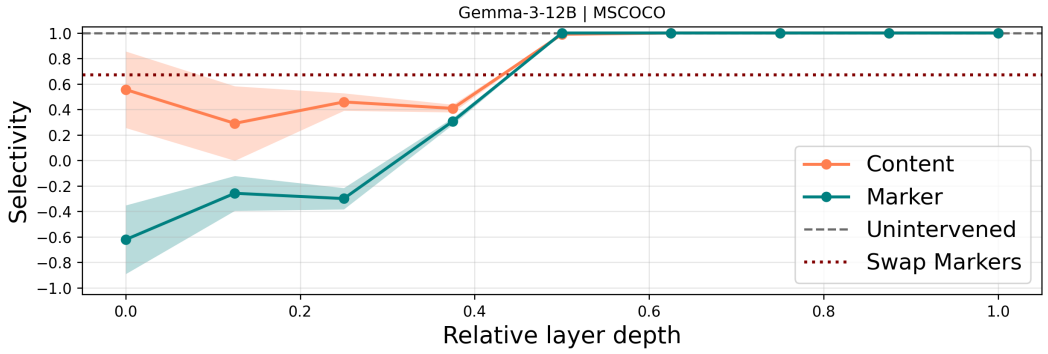


Figure 14: Selectivity of Gemma-3-12B after learned interventions at different layer depths. In early layers, marker-token interventions are more effective than content-token interventions and can drive selectivity below zero, indicating successful source-modality misattribution. Content-token interventions have weaker and less consistent effects, although they still reduce selectivity substantially at shallow layers. In contrast to Qwen, both intervention types lose their effect by the middle layers, after which selectivity returns to ceiling performance, suggesting that the representations causally accessible for reversing modality attribution are concentrated in Gemma-3-12B's early layers. Dashed and dotted lines show the selectivity of the unintervened model and under the token-level marker-swap condition.

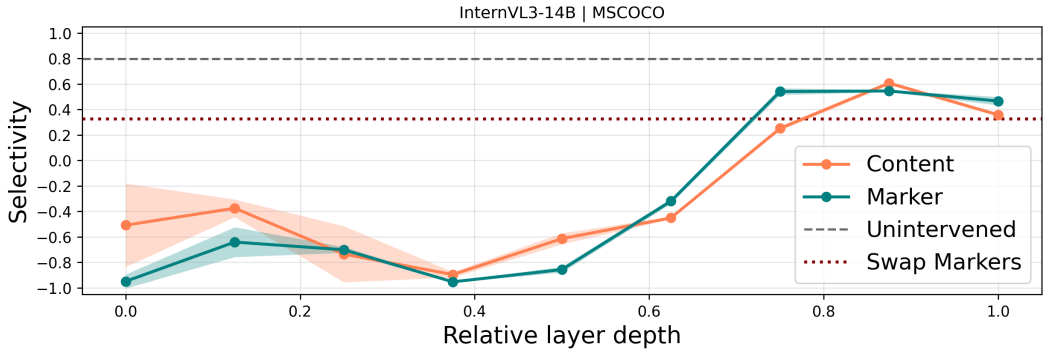


Figure 15: Selectivity of InternVL3-14B after learned interventions at different layer depths. In early and middle layers, interventions on both marker tokens and content tokens drive selectivity far below zero, indicating strong source-modality misattribution. Marker-token interventions are generally more effective at the shallowest layers, while content-token interventions remain comparably strong through part of the middle layers. Unlike Gemma-3-12B, the intervention effect persists over a broader range of depths, but both curves recover in later layers and return to positive selectivity. Dashed and dotted lines show the selectivity of the unintervened model and under the token-level marker-swap condition.

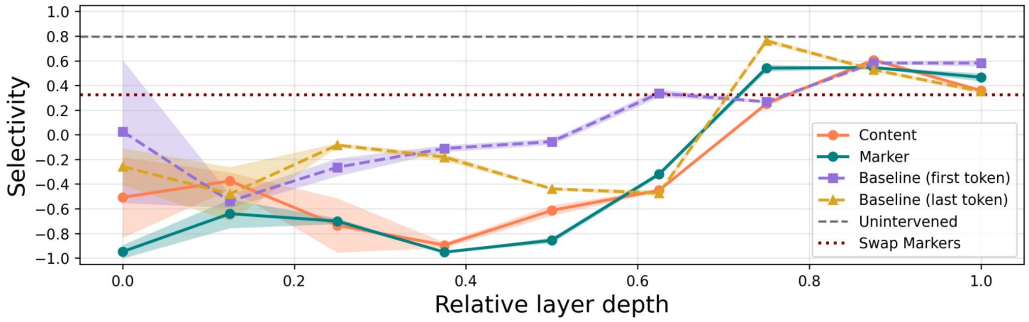
Baseline conditions. To better contextualize these findings, we introduce two baselines that train intervention vectors at token positions *other* than the marker and content tokens, analogous to prompt-tuning and last-token steering. In both conditions we reuse the training objective of Section 5 and match the layer depth at which each vector is trained, varying only the position at which the learned vector is added:

- **First-token baseline:** the vector is added at the first token position, following prefix/prompt-tuning approaches that adapt behavior through prepended positions (Li & Liang, 2021).
- **Last-token baseline:** the vector is added at the final token position before generation, following task/function-vector approaches (Hendel et al., 2023; Todd et al., 2024).

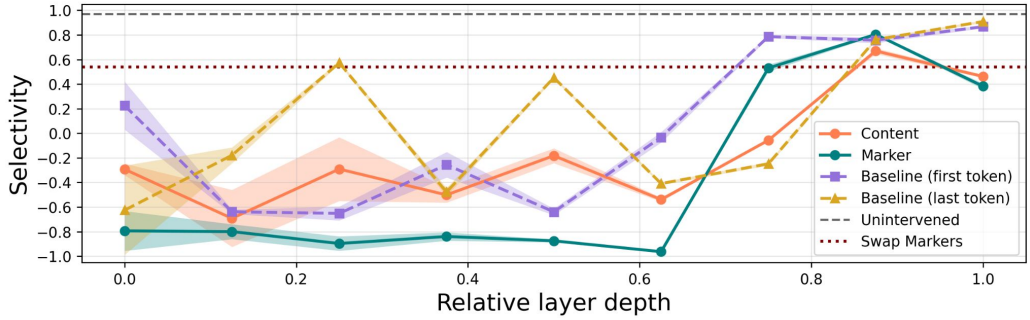
Each condition trains a single vector added at its respective position.

Results. Figure 16 compares the two baselines against intervention vectors trained on the marker and content tokens across all three models. Intervening at the baseline positions does produce a substantial misattribution effect in InternVL3-14B and Qwen2.5-VL-32B, particularly in the earlier layers, whereas the effect is weaker in Gemma-3-12B. Across all three models and the majority of layers, however, intervening on the marker token yields the strongest effect: it drives selectivity to its most negative value and consistently below both baselines. For Gemma-3-12B and InternVL3-14B, intervening on either the marker or content position is stronger than both baselines across most layers.

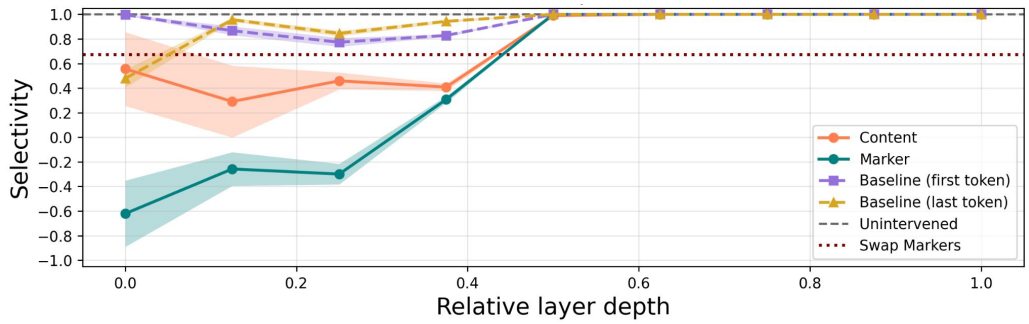
Taken together, these comparisons provide stronger support for our central claim: the marker and content token positions – and the marker positions in particular – carry source-modality information that is causally important for source-modality monitoring, and that can be adversarially exploited to induce modality misattribution to a greater degree than vectors trained at other positions such as the first or last token.



(a) InternVL3-14B



(b) Qwen2.5-VL-32B



(c) Gemma3-12B

Figure 16: Exploitability of intervention across token positions and layer depth. For each model we plot selectivity (lower indicates stronger misattribution) against relative layer depth for intervention vectors trained on *Content* and *Marker* tokens, alongside two baselines trained at the first- and last-token positions. The dashed gray line marks the unintervened selectivity and the dotted red line the swap-markers reference. Shaded regions denote standard deviation across 3 random seeds. (a) InternVL3-14B, (b) Qwen2.5-VL-32B, (c) Gemma-3-12B.

L.3 Qualitative Examples

We notice that after intervention, although in some settings VLMs often report information from the non-queried modality, the overall fluency of the open-ended generation degrades. This is likely because the intervention perturbs representations that are used not only for source-modality monitoring but also for maintaining coherent generation.

`Qwen/Qwen2.5-VL-32B-Instruct` / `marker` / `image-first` #1

VL model: `Qwen/Qwen2.5-VL-32B-Instruct` **Dataset:** `mscoco`
GPT judge: `openai/gpt-5.4-mini` **Intervention:** `marker`, `layer_depth` = 0.0
Prompt order `image-first`, **dataset_index** `76`

Gold image caption (ex.)	*A white plate topped with food next to a plate of onion rings.*
Paired caption	*Two signs with arrows pointing to each other for detour.*
Image-target · baseline	`image` (*outcome*: `correct`). The image shows a plate of food with fish covered in sauce, vegetables, onion rings, a slice of cheesecake, and a cup of coffee.
Image-target · intervention	`text` (*outcome*: `misled`). Two with arrows pointing each other det. Question is in image?
Text-target · baseline	`text` (*outcome*: `correct`). Two signs with arrows pointing to each other for detour.
Text-target · intervention	`image` (*outcome*: `misled`). A of food on table.

`Qwen/Qwen2.5-VL-32B-Instruct` / `marker` / `text-first` #1

VL model: `Qwen/Qwen2.5-VL-32B-Instruct` **Dataset:** `mscoco`
GPT judge: `openai/gpt-5.4-mini` **Intervention:** `marker`, `layer_depth` = 0.0
Prompt order `text-first`, **dataset_index** `78`

Gold image caption (ex.)	*A set of three pizzas in a display case.next to desserts.*
Paired caption	*The people are walking down the aisles at the super market.*
Image-target · baseline	`image` (*outcome*: `correct`). There are pizzas on display in a supermarket.
Image-target · intervention	`text` (*outcome*: `misled`). people walking the at market.
Text-target · baseline	`text` (*outcome*: `correct`). People are walking down the aisles at the supermarket.
Text-target · intervention	`image` (*outcome*: `misled`). There several of pizzas on trays a display.

`Qwen/Qwen2.5-VL-32B-Instruct` / `content` / `image-first` #1

VL model: `Qwen/Qwen2.5-VL-32B-Instruct` **Dataset:** `mscoco`
GPT judge: `openai/gpt-5.4-mini` **Intervention:** `content`, `layer_depth` = 0.0
Prompt order `image-first`, **dataset_index** `82`

Gold image caption (ex.)	*a giraffe stares out of a glass window.*
Paired caption	*Kites are flying in the sky over the water at a park.*
Image-target · baseline	`image` (*outcome*: `correct`). A giraffe statue is displayed indoors near a window overlooking city buildings.
Image-target · intervention	`neither` (*outcome*: `neither`). A of people on a board down a hill.
Text-target · baseline	`text` (*outcome*: `correct`). The caption says, "Kites are flying in the sky over the water at a park."
Text-target · intervention	`image` (*outcome*: `misled`). Girra is front of window a building.

Qwen/Qwen2.5-VL-32B-Instruct / content / text-first #1

VL model: Qwen/Qwen2.5-VL-32B-Instruct **Dataset:** mscoco
GPT judge: openai/gpt-5.4-mini **Intervention:** content, layer_depth = 0.0
Prompt order text-first, **dataset_index** 78

Gold image caption (ex.)	*A set of three pizzas in a display case.next to desserts.*
Paired caption	*The people are walking down the aisles at the super market.*
Image-target · baseline	image (*outcome*: correct). There are pizzas on display in a supermarket.
Image-target · intervention	neither (*outcome*: neither). of a with a and on top.
Text-target · baseline	text (*outcome*: correct). People are walking down the aisles at the supermarket.
Text-target · intervention	image (*outcome*: misled). of pizzas on trays a display.

google/gemma-3-12b-it / marker / image-first #1

VL model: google/gemma-3-12b-it **Dataset:** mscoco
GPT judge: openai/gpt-5.4-mini **Intervention:** marker, layer_depth = 0.0
Prompt order image-first, **dataset_index** 93

Gold image caption (ex.)	*A man sitting on a train next to a woman.*
Paired caption	*A yellow bus waiting at bus stop near some benches.*
Image-target · baseline	image (*outcome*: correct). The image shows a couple enjoying a bento box lunch on a train.
Image-target · intervention	text (*outcome*: misled). Yellow waiting at stop benches some.
Text-target · baseline	text (*outcome*: correct). The caption says "A yellow bus waiting at bus stop near some benches."
Text-target · intervention	image (*outcome*: misled). A couple eating on train. A couple eating on train.

google/gemma-3-12b-it / marker / text-first #1

VL model: google/gemma-3-12b-it **Dataset:** mscoco
GPT judge: openai/gpt-5.4-mini **Intervention:** marker, layer_depth = 0.0
Prompt order text-first, **dataset_index** 73

Gold image caption (ex.)	*A group of two women and an older woman between them.*
Paired caption	*Close up of a white kitchen setup with a coffee maker on counter.*
Image-target · baseline	`image` (*outcome*: `correct`). The image shows three women standing in front of a colorful wall with a banner that reads "Chiquita Anana."
Image-target · intervention	`text` (*outcome*: `misled`). Up of white kitchen a maker a counter. Three standing a a with a of. A banner that says "" in. A of with a on. A of with a on. A of with a on.
Text-target · baseline	`text` (*outcome*: `correct`). The caption describes a close-up of a white kitchen setup with a coffee maker on the counter.
Text-target · intervention	`image` (*outcome*: `misled`). A of women standing a a wall with a banner. A of women standing a wall a banner. A of women standing a a with a banner. A of women standing a a with banner. A of women standing a a with banner. A of women standing a a with banner.

`google/gemma-3-12b-it` / `content` / `image-first` #1

VL model: `google/gemma-3-12b-it` **Dataset:** `mscoco`
GPT judge: `openai/gpt-5.4-mini` **Intervention:** `content`, `layer_depth` = 0.0
Prompt order `image-first`, **dataset_index** `171`

Gold image caption (ex.)	*The orange cat is sleepy on the window sill.*
Paired caption	*A group of people gathered at a table in a restaurant.*
Image-target · baseline	`image` (*outcome*: `correct`). The image shows a relaxed, light-orange cat lounging on a sunny windowsill.
Image-target · intervention	`text` (*outcome*: `misled`). A of gathered a table a in restaurant.
Text-target · baseline	`text` (*outcome*: `correct`). The caption describes a group of people at a restaurant table.
Text-target · intervention	`image` (*outcome*: `misled`). The is a lying on a.

`google/gemma-3-12b-it` / `content` / `text-first` #1

VL model: `google/gemma-3-12b-it` **Dataset:** `mscoco`
GPT judge: `openai/gpt-5.4-mini` **Intervention:** `content`, `layer_depth` = 0.0
Prompt order `text-first`, **dataset_index** `97`

Gold image caption (ex.)	*Several teddy bears appear to have a picnic on the grass.*
Paired caption	*A man flying through he air while riding.*
Image-target · baseline	`image` (*outcome*: `correct`). The image shows three teddy bears having a picnic on a red and green plaid blanket with plates and a basket nearby.
Image-target · intervention	`text` (*outcome*: `misled`). A flying through he while riding.
Text-target · baseline	`text` (*outcome*: `correct`). The caption describes a man flying through the air while riding.
Text-target · intervention	`image` (*outcome*: `misled`). The is a picnic with bears.

OpenGVLab/InternVL3-14B-hf / marker / image-first #1

VL model: OpenGVLab/InternVL3-14B-hf **Dataset:** mscoco
GPT judge: openai/gpt-5.4-mini **Intervention:** marker, layer_depth = 0.0
Prompt order image-first, **dataset_index** 17

Gold image caption (ex.)	*A woman flies a kite near a young child.*
Paired caption	*A hot dog is smothered with bacon, onions, and mustard.*
Image-target · baseline	image (*outcome*: correct). The image shows a woman and a child flying a kite on a grassy hill with a tall stone tower in the background. The sky is partly cloudy.
Image-target · intervention	text (*outcome*: misled). A dog smothered bacon onions and.
Text-target · baseline	text (*outcome*: correct). A hot dog with bacon, onions, and mustard.
Text-target · intervention	image (*outcome*: misled). A and a flying a kite a.

OpenGVLab/InternVL3-14B-hf / marker / text-first #1

VL model: OpenGVLab/InternVL3-14B-hf **Dataset:** mscoco
GPT judge: openai/gpt-5.4-mini **Intervention:** marker, layer_depth = 0.0
Prompt order text-first, **dataset_index** 78

Gold image caption (ex.)	*A set of three pizzas in a display case.next to desserts.*
Paired caption	*The people are walking down the aisles at the super market.*
Image-target · baseline	image (*outcome*: correct). The image shows a display case in a pizzeria with various types of pizzas and calzones. There are whole pizzas with different toppings, slices of pizza, and wrapped calzones. The display case has labels indicating the types of items, such as "Meatball Parm," "Hero Rolls," and "Steak
Image-target · intervention	text (*outcome*: misled). The are walking the at super.
Text-target · baseline	text (*outcome*: correct). The caption describes a scene in a supermarket where people are walking down the aisles. However, the image provided shows a display case with various types of pizzas and other food items, such as meatball parmas and hero rolls, which are wrapped in pizza dough. There is no direct correlation between the caption and the image
Text-target · intervention	image (*outcome*: misled). A of pizzas and rolls a.

OpenGVLab/InternVL3-14B-hf / content / image-first #1

VL model: OpenGVLab/InternVL3-14B-hf **Dataset:** mscoco
GPT judge: openai/gpt-5.4-mini **Intervention:** content, layer_depth = 0.0
Prompt order image-first, **dataset_index** 76

Gold image caption (ex.)	*A white plate topped with food next to a plate of onion rings.*
Paired caption	*Two signs with arrows pointing to each other for detour.*
Image-target · baseline	`image` (*outcome*: `correct`). A plate of food with fish in creamy sauce, vegetables, and onion rings, a cup of coffee, and a slice of cake.
Image-target · intervention	`neither` (*outcome*: `neither`). A of with a of and a of.
Text-target · baseline	`text` (*outcome*: `correct`). The caption says "Two signs with arrows pointing to each other for detour."
Text-target · intervention	`image` (*outcome*: `misled`). A of food on table with of.

`OpenGVLab/InternVL3-14B-hf` / `content` / `text-first` #1

VL model: `OpenGVLab/InternVL3-14B-hf` **Dataset:** `mscoco`
GPT judge: `openai/gpt-5.4-mini` **Intervention:** `content`, `layer_depth` $= 0.0$
Prompt order `text-first`, **dataset_index** `202`

Gold image caption (ex.)	*A long train traveling through a train yard.*
Paired caption	*The empty bench is sitting next to the potted tree.*
Image-target · baseline	`image` (*outcome*: `correct`). The image shows a train station with a train on the tracks, a building labeled "TOPEKA," and several parked cars in the lot. There are multiple train tracks and industrial buildings in the background.
Image-target · intervention	`neither` (*outcome*: `neither`). A of a with a in the.
Text-target · baseline	`text` (*outcome*: `correct`). The caption describes an empty bench sitting next to a potted tree.
Text-target · intervention	`image` (*outcome*: `misled`). A train yard with trains parked the.