SEEING THROUGH SYMBOLS: TRACING AND MITIGATING LABEL SYMBOL BIAS IN LARGE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision-Language Models (LVLMs) have demonstrated strong capabilities in multimodal reasoning, yet their predictions can be sensitive to the symbolic identifiers assigned to visual candidates. We investigate this phenomenon as *label symbol bias*, where changing the symbols labeling the candidate regions, points, or sub-images can alter model predictions despite leaving the underlying visual content and reasoning task unchanged. To uncover the mechanisms underlying this bias, we introduce Symbol Sensitivity Analysis (SSA), a path-patching-based framework that identifies and quantifies the sensitivity of attention heads to changes in label symbols. By tracing the attention patterns of these symbol-sensitive heads, we uncover how label information associated with visual candidates is propagated toward the final prediction, revealing a mechanism through which symbolic identifiers can influence multimodal reasoning beyond the visual content they denote. Building on these findings, we propose Sensitivity-Aware Activation Debiasing (SAAD), a training-free intervention that estimates the symbol-bias component of head activations and selectively amplifies their task-dependent component based on each head's measured label-symbol sensitivity. Extensive experiments across diverse domains show that SAAD reduces sensitivity to label-symbol assignments and improves task accuracy by up to 15% over the underlying model. These results demonstrate that mitigating label-symbol bias enables LVLMs to rely more faithfully on task-relevant visual evidence, leading to more robust multimodal reasoning.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) have made remarkable progress in understanding and reasoning over visual information. A common class of visual reasoning tasks requires a model to compare multiple candidate visual entities, such as points, bounding boxes, or sub-images, and identify the candidate that best satisfies the given task or question. To distinguish these candidates, symbolic identifiers are often embedded directly within the image. These identifiers are intended only to reference the corresponding visual entities. Therefore, the model should base its decision on the visual content and task-relevant relationships among the candidates, such that changing the identifiers assigned to the candidates while preserving the visual content and underlying reasoning problem should leave the selected visual candidate unchanged.

However, we find that this expected invariance does not always hold. Changing only the symbolic identifiers assigned to the visual candidates can substantially alter an LVLM's prediction. We refer to this phenomenon as *label symbol bias*. As illustrated in Figure 1, replacing the labels of two visual candidates (bounding boxes in this example) from (A/B) to (X/Y) can change the model's prediction, despite the candidates and their visual relationships remaining identical. Moreover, even within a fixed symbol set such as (A/B), models often exhibit a systematic preference for one symbol over the other across samples, resulting in a highly skewed prediction distribution despite a balanced distribution of ground-truth answers. These observations suggest that symbolic identifiers can interfere with visual reasoning, making LVLM predictions less reliable.

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

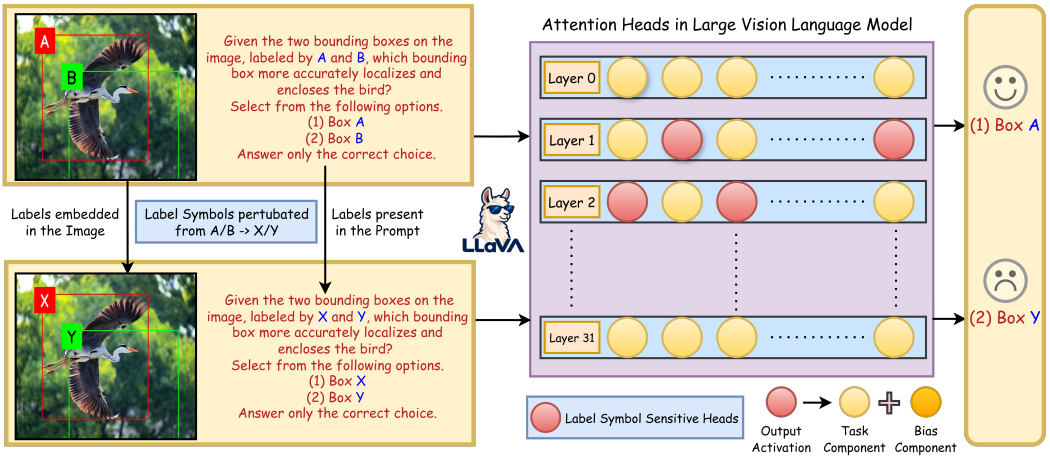


Figure 1: **Overview of Label Symbol Bias**. Changing the symbolic identifiers of visual candidates (e.g., A/B → X/Y) while preserving the underlying visual content can alter the LVLM's prediction, demonstrating label symbol bias. SSA identifies attention heads whose activations are sensitive to such symbol changes. SAAD then decomposes these head activations into bias and task-dependent components and selectively amplifies the latter according to the measured symbol sensitivity, reducing the influence of symbolic cues and promoting reliance on task-relevant visual evidence.

Table 2 reports the performance of LLaVA-NeXT-Vicuna-7B (Liu et al. (2024)) on the Relative Depth (RD) task of the BLINK dataset (Fu et al. (2024)) under different symbol sets, using the same underlying visual reasoning instances. Changing the symbol set leads to substantial variations in overall accuracy. For example, with the (A/B) symbol set, the model achieves an accuracy of 44.2% while with (X/Y), the accuracy is 51.3%. Similar variations are observed across other symbol sets. Figure 3 shows the prediction distributions. Despite the approximately balanced ground-truth distribution, the model's predictions can be strongly skewed toward one symbol within a fixed symbol set. This skew also changes when the symbol set is changed, indicating that the model's prediction behavior is sensitive to the symbolic identifiers assigned to the visual candidates.

If changing only the label symbols alters the final prediction, the attention heads that transmit the resulting changes provide a natural locus for tracing the source of this bias. Motivated by this observation, we introduce *Symbol Sensitivity Analysis (SSA)*, a path-patching-based framework for identifying attention heads whose activations are sensitive to label-symbol perturbations. SSA compares a reference execution with a perturbed execution that differs only in the symbolic identifiers assigned to visual candidates. For each attention head, we replace its output activation in the reference execution with the corresponding activation from the perturbed execution, while keeping the outputs of all other heads fixed to their reference values. The resulting change in the model's logits measures the causal influence of that head on the prediction under the symbolic perturbation. Heads that induce substantial logit changes under the intervention are identified as *symbol-sensitive attention heads*.

Beyond identifying symbol-sensitive heads, we investigate how label symbols embedded in the image propagate through the model and ultimately influence its prediction. For highly symbol-sensitive heads, we observe that the final query token attends strongly to the option token corresponding to the model's prediction. Using SSA, we trace the upstream pathways contributing to the representation of this option token. We find that these upstream heads exhibit strong attention from the option token, serving as the query, to label symbols appearing early in the textual prompt. Tracing this pathway further upstream reveals attention heads that connect these textual label symbols to their corresponding symbols embedded in the image. Taken together, these observations reveal a three-stage propagation pathway: label symbols in the image are first associated with their textual counterparts, these textual symbols are then propagated to the corresponding candidate option, and this option-level representation ultimately contributes to the model's prediction.

In order to mitigate the bias, a natural strategy is to remove or suppress symbol-sensitive heads. However, these heads can encode both symbolic bias and task-relevant information, so indiscrimi-

nate suppression may discard useful reasoning signals. We therefore propose Sensitivity-Aware Activation Debiasing (SAAD), a training-free intervention that suppresses symbolic bias while preserving task-relevant information. SAAD estimates the bias component of each head's activation using its dataset-level mean, while the deviation from this mean captures instance-specific, task-dependent information. Instead of suppressing the entire head, SAAD selectively amplifies this task-dependent component according to the head's measured sensitivity to label-symbol perturbations. Thus, more symbol-sensitive heads receive a stronger intervention while retaining their instance-specific reasoning signals.

Our contributions are summarized as follows:

- We identify and systematically characterize *label symbol bias* in LVLMs, showing that changing only the symbolic identifiers of visual candidates can substantially alter model predictions despite an unchanged reasoning task.
- We introduce *Symbol Sensitivity Analysis (SSA)*, a method that shows how different attention heads causally contribute to this bias and reveals how visual labels are first linked to their text counterparts, then passed to the candidate options, and finally used when making the prediction.
- We propose *Sensitivity-Aware Activation Debiasing (SAAD)*, a training-free intervention that separates systematic symbol-related activation from task-dependent information and selectively amplifies the latter based on measured symbol sensitivity.
- We demonstrate across diverse visual reasoning domains that SAAD reduces sensitivity to label-symbol assignments while improving task accuracy.

2 Related Work

2.1 Multimodal Visual Reasoning Benchmarks

Many vision-language benchmarks do not sufficiently test whether models can make precise visual judgments about fine-grained image properties. BLINK (Fu et al., 2024) addresses this by recasting 14 classic computer vision tasks, such as visual correspondence, relative depth and reflectance, and object localization, into 3,807 multiple-choice questions, revealing a substantial gap between LVLM and human performance. Several BLINK tasks present multiple visual candidates within one image and distinguish them with symbolic identifiers, such as labels on points or bounding boxes. Similar designs appear elsewhere: representing multiple images as labeled subimages lets single-image models perform multi-image reasoning (Zhao et al., 2024), training datasets such as M3CoT (Chen et al., 2024) present candidate options as labeled subimages, and visual prompting approaches draw markers directly on the image (Shtedritski et al., 2023; Cai et al., 2024; Zheng et al., 2024a). In all these settings, models must associate identifiers embedded in the visual input with the entities they denote, raising the possibility that they rely on the identifiers as shortcuts rather than on visual content. This motivates our study of *label symbol bias*: whether changing the identifiers alone can alter LVLM predictions while the visual content and reasoning task stay unchanged.

2.2 Selection Bias in Reasoning

Multiple-choice question answering (MCQA) is widely used to evaluate language and vision-language models, yet models exhibit *selection bias*, preferring particular option positions or symbols independently of the underlying answer (Zhao et al., 2021). Zheng et al. (2024b) traced this preference for option identifiers partly to token-level biases and proposed PriDe, which estimates the identifier prior by permuting option contents and removes it from predictions. Xue et al. (2024) attributed selection bias to limitations in *Multiple Choice Symbol Binding (MCSB)* (Robinson & Wingate, 2023), showing that it persists after supervised fine-tuning, and Yang et al. (2025) showed that changing only the option symbol set can substantially alter MCQA performance.

In multimodal settings, LVLM predictions depend on the positions of images in multi-image inputs, which Tian et al. (2025) mitigate with SoFt Attention (SoFA), and on the spatial location of identical visual content within an image, which Zhu et al. (2026) address with Adaptive Global Context Injection (AGCI). Atabuzzaman et al. (2025) showed that LVLM selection bias in MCQA grows as

the semantic distinction between candidate answers changes and corrected it with an ensemble bias estimate, while SCICON (Roh et al., 2026) subtracts text-only option logits from image-conditioned logits. These methods operate on the model's output rather than identifying the internal source of the bias. Our work instead studies a distinct source of symbolic bias: *label symbols* embedded in the visual input, which identify visual entities such as points, bounding boxes, or sub-images rather than answer choices, and we locate the attention heads through which they influence the prediction.

2.3 Mechanistic Interpretability for Activation Intervention

Behavioral analyses reveal that a model is sensitive to a factor but not why; mechanistic interpretability addresses this by identifying the components and pathways that causally drive predictions, including how language models bind entities and answer symbols (Feng & Steinhardt, 2024; Wiegreffe et al., 2025). Path Patching (Wang et al., 2023) isolates a component's causal effect by replacing its activation with its counterpart from a corrupted input while keeping all other activations fixed, with the resulting logit change measuring its influence (Zhang & Nanda, 2024).

Such causal analysis has recently been extended to vision-language models (Neo et al., 2025); for example, NOTICE (Golovanevsky et al., 2025) uses controlled corruption to identify attention heads that consistently contribute across tasks and modalities. We build on this principle but use *symbolic perturbation* as the intervention, changing only the identifiers embedded in the visual input while preserving the visual content and task. We use head-level patching to identify components causally sensitive to these changes and, following prior inference-time interventions on model activations (Li et al., 2023; Wang et al., 2025), leverage their measured sensitivity to develop SAAD, which intervenes on their activations while preserving task-relevant information.

3 Investigation on Label Symbol Bias

In this section, we aim to identify the LVLM attention heads that are responsible for label symbol bias. We then analyze the internal attention patterns of these heads. The analysis is presented on Relative Reflectance (RR) domain of the BLINK dataset (Fu et al. (2024)) with LLaVA-NeXT-Vicuna-7B (Liu et al. (2024)).

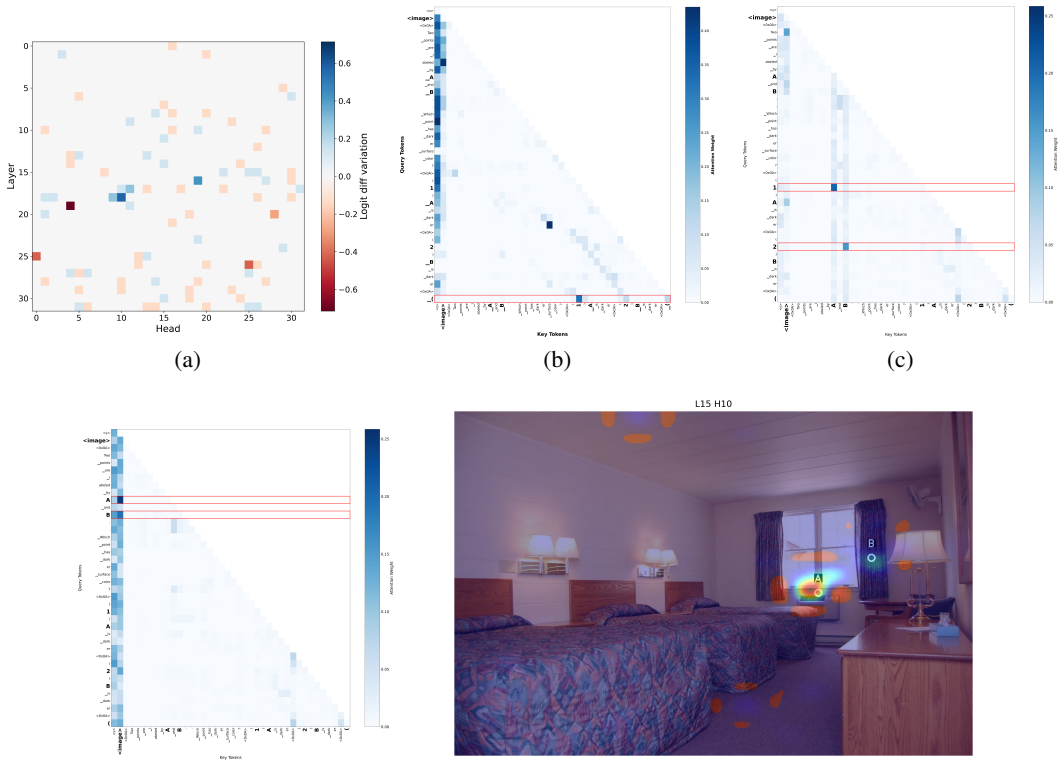
3.1 Label Symbol Sensitivity of Attention Heads

Based on the intervention method of Path Patching (Wang et al. (2023)), we introduce *Symbol Sensitivity Analysis (SSA)* to identify the internal attention heads that make model predictions more susceptible to label symbol bias.

Consider a sample x whose visual candidates are associated with label symbols. We independently sample two label-symbol sets from $\mathrm{A/B}, \mathrm{X/Y}, \mathrm{I/II}, \mathrm{a/b}, \mathrm{x/y}, \mathrm{i/ii}$: one is designated as the *reference* symbol set and the other as the *perturbed* symbol set. We then construct x_{ref} and x_{pert} by assigning the reference and perturbed symbol sets, respectively, to the same underlying sample. The replacement is applied consistently to both the label symbols embedded in the image and those appearing in the textual prompt, while the underlying visual content and reasoning problem remain unchanged.

We perform a forward pass on both x_{ref} and x_{pert} and cache the output activation for every attention head. Let $h_{i,j}$ denote the j-th attention head in layer i. To measure the influence of $h_{i,j}$ on the model's prediction, we perform a patched forward pass using the reference input x_{ref}. Specifically, we patch the output activation of the targeted head $h_{i,j}$ by replacing its original output activation with the corresponding cached activation obtained from the perturbed input x_{pert}. The output activations of all other attention heads are simultaneously fixed to their cached values from the reference forward pass. Consequently, the residual stream between the original and patched executions differs only through the contribution of the targeted attention head.

To quantify the extent to which an attention head is sensitive to the label symbol change, we define the *label symbol sensitivity* of head $h_{i,j}$ as



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

Figure 2: **Causal tracing of label-symbol propagation.** (a) Symbol-sensitive attention heads identified by SSA. (b) The final query token attends to the predicted option. (c) The option representation attends to its associated label symbol. (d) The label-symbol representation attends to image patches. (e) The attended patches correspond to regions containing the embedded label symbols, revealing a causal pathway from visual label symbols to the model's output. (Zoom in for better visibility.)

$$\sigma_{i,j} = \frac{d^{\text{orig}} - d^{\text{patch}}_{i,j}}{d^{\text{orig}} - d^{\text{pert}}}, \tag{1}$$

where d denotes the logit difference between the first option and the second option. Specifically, d^{orig}, d^{pert}, and $d^{\text{patch}}_{i,j}$ correspond to the logit differences obtained from the original, perturbed, and head-patched executions, respectively.

We apply SSA to 10% of the samples in the dataset and compute the average absolute sensitivity of each attention head across the samples. Figure 2 (a) visualizes the resulting average sensitivity of heads across layers.

3.2 Attention Patterns of Symbol-Sensitive Heads

Having identified attention heads that are sensitive to label-symbol perturbations, we first examine their attention patterns to understand how these heads contribute to the model's prediction. We then trace the upstream causal pathways through which label symbols embedded in the visual input are propagated through intermediate representations and ultimately influence the model's output.

We first examine the attention patterns of symbol-sensitive heads at the final query position (Figure 2 (b)). We observe that the final query token attends strongly to the option token corresponding to the model's prediction. This behavior is reminiscent of the *Name Mover Heads* identified in Wang et al. (2023). This observation raises a natural question: *what information is encoded in the key representation of the attended option token that causes the final query to attend to it?*

To answer this question, we trace the causal pathways contributing to the key representation of the attended option token. Given a target symbol-sensitive head h_t, we identify upstream attention heads whose activations may causally influence its key vector. For each upstream head h_u, we perform a path-patching intervention from h_u to the key vector of h_t. Specifically, we replace the activation of h_u with its counterpart from the perturbed-symbol input, while keeping the activations of all intermediate heads between h_u and the target key vector fixed to their values from the reference forward pass. This isolates the causal path from h_u to the target key representation. We then recompute the target key vector under this intervention and perform a regular forward pass downstream of the recomputed key vector, allowing all subsequent computations to proceed normally. The resulting change in logit difference is used to quantify the causal effect of h_u on the key representation of the symbol-sensitive head.

As shown in Figure 2 (c), the attention visualization reveals that the causally influential heads exhibit strong attention from the option token, serving as the query, toward the corresponding label symbols appearing earlier in the prompt. This pattern suggests a second level of symbolic propagation, where information about the label identifier present in the prompt is incorporated into the option token representation.

We then extend the same path-patching procedure one level further upstream by tracing the causal inputs to the key representation of the textual label-symbol token. As shown in Figure 2 (d), we identify attention heads whose query at the textual label-symbol position attends strongly to a small number of image patches. By projecting these influential patches back onto the input image, we find that they correspond to regions containing the label symbols embedded in the visual input (Figure 2 (e)). Thus, these heads provide a pathway through which the visual occurrence of a label symbol influences the representation of its corresponding textual symbol.

Taken together, the traced causal pathways reveal a three-level mechanism for label-symbol propagation: (1) visual label symbols influence the representations of their corresponding textual label symbols; (2) these textual symbols influence the representations of the corresponding candidate options; and (3) the final query attends to these candidate representations, which in turn contributes to the model's output logits. Importantly, these pathways are identified through controlled activation interventions rather than attention patterns alone, providing mechanistic evidence that label-symbol information is causally propagated from the visual input to intermediate candidate representations and ultimately to the model's prediction, independently of the underlying visual evidence.

4 Mitigating the Bias

The SSA introduced in Section 3 identifies a subset of attention heads that are sensitive to label symbol perturbations. A straightforward mitigation strategy is to deactivate these heads entirely. Specifically, for a selected head (i, j), its output activation at the prediction position can be replaced by

$$a_{i,j} = \mathbf{0}. \tag{2}$$

Although this strategy can suppress the contribution of a symbol-sensitive head, it can also be destructive. A symbol-sensitive head may encode both undesirable symbolic information and useful task-dependent information. Completely removing its activation may therefore eliminate useful reasoning signals along with the bias. This motivates a more targeted intervention that suppresses the bias component of a head activation while preserving its task-relevant information.

4.1 Decomposition of Head Activation

Let $a_{i,j}^{(n)}$ denote the output activation of attention head (i, j) at the prediction position for the n-th input sample. We conceptualize the head activation as a combination of a systematic bias component and a task-dependent component:

$$a_{i,j}^{(n)} = a_{i,j}^{\text{bias}} + a_{i,j}^{\text{task},(n)}, \tag{3}$$

Here, $a_{i,j}^{\text{bias}}$ represents the systematic information associated with label symbols encoded by the head, while $a_{i,j}^{\text{task},(n)}$ represents instance-specific information required to solve the particular reasoning problem.

The key intuition is that the task-dependent component varies across different reasoning instances, whereas the bias component is comparatively consistent. Consequently, when activations are averaged over a sufficiently diverse set of samples, the instance-specific components are expected to partially cancel:

$$\frac{1}{N}\sum_{n=1}^{N} a_{i,j}^{\text{task},(n)} \approx \mathbf{0}. \tag{4}$$

Under this assumption, the mean activation of a head provides an estimate of its systematic bias component:

$$\bar{a}_{i,j} = \frac{1}{N}\sum_{n=1}^{N} a_{i,j}^{(n)} \approx a_{i,j}^{\text{bias}}, \tag{5}$$

Thus, $\bar{a}_{i,j}$ serves as a representative estimate of the bias component encoded by head (i, j).

4.2 Sensitivity-Aware Activation Debiasing

Building on the above decomposition, we introduce *Sensitivity-Aware Activation Debiasing (SAAD)*, which modifies the activation of each attention head according to its sensitivity to label symbol perturbations. Unlike directly suppressing or removing the activation of symbol-sensitive heads, SAAD exploits the deviation of an individual activation from its dataset-level mean to enhance its task-dependent component.

For a given head (i, j) and input sample n, we first compute the deviation of its activation from the estimated bias component:

$$\delta_{i,j}^{(n)} = a_{i,j}^{(n)} - \bar{a}_{i,j}, \tag{6}$$

where $a_{i,j}^{(n)}$ denotes the activation of head (i, j) for the n-th sample and $\bar{a}_{i,j}$ denotes its average activation across the dataset. Based on the decomposition in Eq. 3, this deviation primarily captures the instance-dependent component of the head activation. We therefore use $\delta_{i,j}^{(n)}$ as a proxy for the task-relevant information contained in the head.

Rather than suppressing the original activation, SAAD adds a sensitivity-weighted portion of this task-dependent deviation to the original activation:

$$\tilde{a}_{i,j}^{(n)} = a_{i,j}^{(n)} + \alpha w_{i,j} \delta_{i,j}^{(n)}, \tag{7}$$

where α controls the overall intervention strength and $w_{i,j}$ determines the contribution of each head based on its measured sensitivity to label symbol perturbations. The sensitivity weight is computed over the complete set of attention heads:

$$w_{i,j} = \text{softmax}_{(i,j)}\left(\sigma_{i,j}\right), \tag{8}$$

Thus, heads with higher sensitivity to label symbol perturbations receive larger intervention weights, while less sensitive heads receive smaller weights. This enables the intervention to adapt its strength continuously across all attention heads rather than selecting a fixed subset of heads for modification.

Substituting Eq. 6 into Eq. 7, the intervention can be expressed as

$$\tilde{a}_{i,j}^{(n)} = a_{i,j}^{(n)} + \alpha w_{i,j}\left(a_{i,j}^{(n)} - \bar{a}_{i,j}\right), \tag{9}$$

Table 1: Main results on the BLINK benchmark (%). All values denote accuracy. AVG denotes the average accuracy across the six domains. Best results are shown in **bold**.

Method	VS	AS	JS	RR	RD	OL	AVG
Zero-shot Baseline							
LLaVA-NeXT	46.42	54.71	62.82	49.33	52.42	55.28	53.50
Logit-level Debiasing							
Text-only Bias Estimation	45.87	55.16	61.94	50.08	51.96	56.03	53.51
Non-contextual Bias Estimation	47.31	53.92	63.17	48.76	53.08	54.61	53.48
Activation-level Debiasing							
Zero Ablation	44.83	53.47	56.76	47.92	51.68	54.12	51.46
SAAD	**53.63**	**56.47**	**65.86**	**56.67**	**55.34**	**62.93**	**58.48**
SAAD (Shared)	53.63	56.02	65.43	56.67	54.91	62.93	58.27

Equivalently, under the decomposition in Eq. 3, the intervention becomes

$$\tilde{a}_{i,j}^{(n)} \approx (1 + \alpha w_{i,j})\, a_{i,j}^{\mathrm{task},(n)} + a_{i,j}^{\mathrm{bias}}, \tag{10}$$

This formulation highlights the key intuition behind SAAD. Instead of eliminating the entire activation of a symbol-sensitive head, the method amplifies the component that deviates from the head's systematic activation pattern. Since this deviation is used as a proxy for task-dependent information, the intervention preferentially enhances instance-specific reasoning signals while preserving the underlying model representation.

The sensitivity-dependent weighting further ensures that the amplification is strongest for heads whose behavior is most affected by label symbol perturbations. In contrast, heads with low symbol sensitivity receive only a small modification. Importantly, SAAD is applied directly during the forward pass and does not modify the model parameters. Consequently, it requires neither additional training nor fine-tuning, making it a lightweight intervention for mitigating label symbol bias.

5 Experiments

5.1 Experimental Setup

Benchmark. We evaluate our approach on six diverse domains of BLINK (Fu et al. (2024)), a multimodal visual reasoning benchmark for fine-grained visual perception and reasoning: Relative Reflectance (RR), Relative Depth (RD), Object Localization (OL), Visual Similarity (VS), Art Style (AS), and Jigsaw (JS). For VS, AS, and JS, the original tasks provide a reference image and two candidate images as separate inputs. We combine the reference and candidate images into a single image and explicitly label the candidate subimages. For RR and RD, the visual candidates are points, while for OL they are bounding boxes.

Across all six domains, the candidate visual entities are explicitly identified by symbolic labels embedded in the input image, providing a common setting for analyzing label-symbol bias.

Implementation Details. We conduct our experiments using the recent LLaVA-family architecture, LLaVA-NeXT-Vicuna-7B (Liu et al. (2024)). We set the inference temperature to 0 and the hyperparameter α to 15. All experiments are conducted on a single NVIDIA A6000 GPU with 48 GB of VRAM. Additional experimental details are provided in the SM.

5.2 Main Results

Our proposed SAAD mitigates label symbol bias at its source by intervening on the internal activations of LVLM attention heads. We compare SAAD with two classes of debiasing approaches: logit-level methods that estimate bias from ambiguous inputs and subtract it from the output logits, including image-removal (Roh et al. (2026)) and non-contextual prompt (Atabuzzaman et al. (2025)) baselines, and an activation-level baseline that zeros the outputs of label-symbol-sensitive heads.

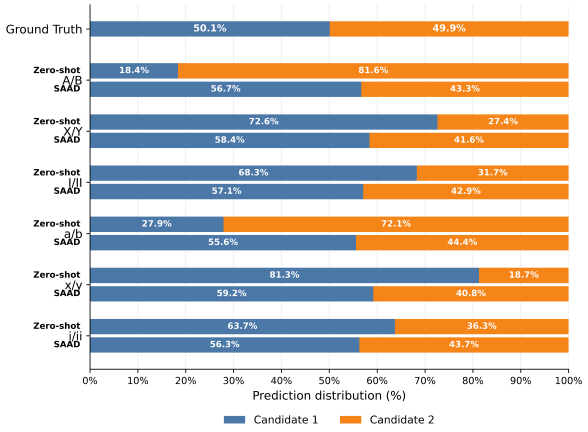


Figure 3: Prediction distribution of LLaVA-NeXT-Vicuna under different label-symbol assignments on the same instances. Although the ground truth is roughly balanced, zero-shot predictions shift substantially with the symbol set. SAAD reduces this sensitivity, yielding more balanced predictions.

Table 2: Effect of label-symbol assignments on LLaVA-NeXT-Vicuna's prediction behavior. We compare zero-shot inference with SAAD using overall accuracy under different symbol sets.

Symbol Set	Zero-shot	SAAD
A/B	49.63	56.67
X/Y	46.92	56.21
a/b	47.84	56.43
x/y	45.87	55.98
I/II	43.91	55.61
i/ii	44.63	55.82

Table 1 shows that logit-level debiasing provides limited improvements, as it operates only on the final logits. Zeroing sensitive heads also degrades performance by discarding task-relevant information. In contrast, SAAD selectively suppresses the bias component while preserving task-dependent information, achieving the strongest overall performance and a more balanced prediction distribution.

We further evaluate cross-domain generalization using *SAAD (Shared)*. We group domains by label structure: *Point Labels* (RR, RD) and *Subimage Labels* (VS, AS, JS). Due to substantial overlap in head sensitivity within each group, we use the RR sensitivity matrix for RD and the VS sensitivity matrix for AS and JS. As shown in Table 1, SAAD (Shared) improves performance across these related domains, demonstrating that label-symbol sensitivity transfers across domains with similar label structures.

6 Conclusion

In this work, we systematically investigate *label symbol bias* in Large Vision-Language Models (LVLMs) during multimodal multiple-choice reasoning. Through extensive experiments, we demonstrate the prevalence of this bias across different models and label symbol configurations. Using causal intervention via path patching, we identify a subset of attention heads that are highly sensitive to label symbol perturbations and reveal that these heads exhibit excessive attention to label symbols in both the textual prompt and visual regions. Based on these findings, we propose *Sensitivity-Aware Activation Debiasing (SAAD)*, an inference-time intervention that adjusts head activations according to their causal sensitivity. Experiments demonstrate that SAAD effectively mitigates label symbol bias while preserving task-relevant information. Our findings highlight the importance of understanding and mitigating superficial symbolic dependencies in LVLM reasoning.

AI Use Statement

We used generative AI tools for limited assistance with language editing, manuscript refinement, and code debugging. Generative AI was not used to generate experimental results, formulate the research methodology, interpret results, or make scientific claims. All AI-assisted content and code were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

Ethics Statement

This work does not involve human subjects, personally identifiable information, or the collection of new personal or sensitive data. We use publicly available benchmarks and pretrained models for research purposes. Our work focuses on identifying and mitigating symbolic biases in vision-language models and does not introduce applications involving high-stakes or autonomous decision-making.

Reproducibility Statement

In Section 5.1, we provide the methodological and experimental details necessary to reproduce our results, including the formulation of SSA and SAAD (Section 3, Section 4), experimental settings, datasets, model configuration, and evaluation protocols. The complete codebase, including implementations and scripts required to reproduce the experiments, will be released upon acceptance.

References

Md. Atabuzzaman, Ali Asgarov, and Chris Thomas. Benchmarking and mitigating MCQA selection bias of large vision-language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33548–33562, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1703. URL https://aclanthology.org/2025.emnlp-main.1703/.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12914–12923, 2024.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M^3CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. 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. 8199–8221, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.446. URL https://aclanthology.org/2024.acl-long.446/.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? In *The Twelfth International Conference on Learning Representations*, 2024.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXIII*, pp. 148–166, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73336-9. doi: 10.1007/978-3-031-73337-6_9. URL https://doi.org/10.1007/978-3-031-73337-6_9.

Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. What do VLMs NOTICE? a mechanistic interpretability pipeline for Gaussian-noise-free text-image corruption and evaluation. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11462–11482, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.571. URL https://aclanthology.org/2025.naacl-long.571/.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Joshua Robinson and David Wingate. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=yKbprarjc5B.

Taeyun Roh, Eun-yeong Jo, Wonjune Jang, and Jaewoo Kang. When choices become priors: Contrastive decoding for scientific figure multiple-choice qa. *arXiv preprint arXiv:2603.28026*, 2026.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does CLIP know about a red circle? visual prompt engineering for VLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11987–11997, 2023.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10599–10609, 2025. URL https://api.semanticscholar.org/CorpusID:277104177.

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.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach. In *The Thirteenth International Conference on Learning Representations*, 2025.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. Answer, assemble, ace: Understanding how LMs answer multiple choice questions. In *The Thirteenth International Conference on Learning Representations*, 2025.

Mengge Xue, Zhenyu Hu, Liqun Liu, Kuo Liao, Shuang Li, Honglin Han, Meng Zhao, and Chengguo Yin. Strengthened symbol binding makes large language models reliable multiple-choice selectors. 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. 4331–4344, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.237. URL https://aclanthology.org/2024.acl-long.237/.

Zhen Yang, Ping Jian, and Chengzhi Li. Option symbol matters: Investigating and mitigating multiple-choice option symbol bias of large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1902–1917, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.95. URL https://aclanthology.org/2025.naacl-long.95/.

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.

Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning, 2024. URL https://arxiv.org/abs/2406.12742.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706, 2021.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. GPT-4V(ision) is a generalist web agent, if grounded. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024a.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=shr9PXz7T0.

Yingjie Zhu, Xuefeng Bai, Kehai Chen, Yang Xiang, Youcheng Pan, Yongshuai Hou, Weili Guan, Jun Yu, and Min Zhang. Beyond the vision encoder: Identifying and mitigating spatial bias in large vision-language models, 2026. URL https://arxiv.org/abs/2509.21984.