System-Mediated Attention Imbalances Make Vision-Language Models Say Yes

Tsan Tsai Chan[1], **Varsha Suresh**[1], **Anisha Saha**[1,2], **Michael Hahn**[1], **Vera Demberg**[1,2]

[1]Saarland Informatics Campus, Saarland University, Germany,
[2]Max Planck Institute for Informatics, Germany
tsch00001@stud.uni-saarland.de, vsuresh@lst.uni-saarland.de, ansaha@mpi-inf.mpg.de, mhahn@lst.uni-saarland.de, vera@lst.uni-saarland.de

Abstract

Vision-language model (VLM) hallucination is commonly linked to imbalanced allocation of attention across input modalities: system, image and text. However, existing mitigation strategies tend towards an image-centric interpretation of these imbalances, often prioritising increased image attention while giving less consideration to the roles of the other modalities. In this study, we evaluate a more holistic, system-mediated account, which attributes these imbalances to functionally redundant system weights that reduce attention to image and textual inputs. We show that this framework offers a useful empirical perspective on the yes-bias, a common form of hallucination in which VLMs indiscriminately respond 'yes'. Causally redistributing attention from the system modality to image and textual inputs substantially suppresses this bias, often outperforming existing approaches. We further present evidence suggesting that system-mediated attention imbalances contribute to the yes-bias by encouraging a default reliance on coarse input representations, which are effective for some tasks but ill-suited to others. Taken together, these findings firmly establish system attention as a key factor in VLM hallucination and highlight its potential as a lever for mitigation. The code for our main interventions can be found at: `https://github.com/anisha0325/vlm-hallucination-yes-bias/tree/main`.

1 Introduction

For mechanistic analysis, the inputs of vision-language models (VLMs) are generally taken to comprise three modalities: system, image and text (Chen et al., 2024a; Yang et al., 2025).[1] VLM hallucination has been interpreted as resulting from these modalities not being attended to in a balanced way, most often manifesting as insufficient image attention (cf. Liu et al., 2024c; Shu et al., 2025; Yang et al., 2025; Zheng and Zhang, 2025). This has motivated many existing mitigation strategies to focus primarily on increasing attention to the image modality (Liu et al., 2024c; Zhu et al., 2025; Liu et al., 2025b; Sarkar et al., 2025; Wang et al., 2025; Zhou et al., 2025; Jiang et al., 2025). As such, these approaches implicitly assume image attention to be the most important lever against hallucination. We refer to this position as the **image-centric hypothesis**.

However, in sidelining the other two modalities, image-centric approaches do not provide a holistic account of attention imbalances. The system modality, in particular, contains substantial amounts of functionally redundant attention in the form of attention sinks (Xiao et al., 2024; Gu et al., 2025): tokens which receive disproportionate attention despite being semantically uninformative. Prior work reports that this modality often accounts for over 70% of the total attention in VLM decoders (Chen et al., 2024a; Tang et al., 2025; Yang et al., 2025; Liang and Cai, 2026). This concentration suggests that system attention may play a central role in hallucination by contributing to attention imbalances, a possibility prior work has not rigorously examined.

To address this gap, we investigate an alternative, more holistic view of hallucination, which we term the **system-mediated hypothesis**. Instead of focussing only on image attention, this hypothesis posits that hallucination arises from the joint influence of two factors: one, *redundant* system attention and two, the resulting *insufficient* attention to both the image and textual modalities. We demon-

[1]We focus on adapter-based VLMs, where all three modalities are passed to the model in a fixed template. Loosely following Chen et al. (2024a), the system modality encompasses everything before the image tokens, including the <BOS> token and system message. The image modality comprises the image patches output by the image encoder and projected into the textual embedding space of the text decoder. The textual modality only consists of the user query for our short-answer benchmarks.

Findings of the Association for Computational Linguistics: ACL 2026, pages 38958–38979
July 2-7, 2026 ©2026 Association for Computational Linguistics

strate that a system-mediated account is empirically superior to an image-centric one for a widespread form of VLM hallucination, the yes-bias. This form of single-token hallucination involves VLMs responding 'yes' in binary yes/no settings irrespective of the prompt (Zhang et al., 2016; Ross et al., 2024; Li et al., 2023, 2024; Zhang et al., 2025). It occurs in a diverse range of binary tasks, most notably object detection (Li et al., 2023; Zheng and Zhang, 2025), reasoning (Zhang et al., 2025) and factual question-answering (Li et al., 2024).

Using causal interventions that redistribute attention weights between modalities in different parts of a VLM's text decoder, we find that redundant late-layer system attention plays a decisive role in promoting the yes-bias by causing localised attention imbalances. Consistent with the system-mediated hypothesis, correcting these imbalances by redistributing attention to the image and textual modalities substantially suppresses the bias and outperforms alternative image- and text-centric mitigation strategies. These gains do not exclusively depend on increasing image attention, undermining the universality of the image-centric hypothesis. We then present evidence suggesting that such late-layer attention imbalances contribute to the yes-bias by encouraging a default reliance on coarse input representations, which is detrimental for certain tasks.

Our contributions are twofold:

(1) We identify redundant late-layer system attention as an important contributor to the yes-bias, highlighting the system modality's relevance to hallucination.

(2) Using principled redistributions of attention weights, we formalise a system-mediated account of VLM hallucination as a more holistic alternative to prevailing image-centric interpretations of imbalanced attention. This motivates attention-based mitigation methods that do not focus on one modality in isolation.

2 Related Work

2.1 Attention-based hallucination mitigation in VLMs

Despite their wide acceptance (Liu et al., 2024c, 2025b; Wang et al., 2025; Sarkar et al., 2025; Zheng and Zhang, 2025), mitigation strategies that increase image attention typically posit that insufficient image attention is the foremost contributor to hallucination without objectively measuring the scale of the presupposed attention deficit. Any success in suppressing hallucination is then simplistically equated to a confirmation of the image-centric hypothesis. Reliance on arbitrary hyperparameters is common as a result, for example in scaling image weights by coefficients with questionable generalisability (cf. Liu et al., 2024c).

More fundamentally, this body of work has not shown that non-image modalities' attention patterns contribute *less* to VLM hallucination. This is a non-trivial omission, as increasing either the absolute magnitude or relative prominence of image attention weights implies reducing the influence of the other input modalities. Our targeted causal interventions directly address this gap by rigorously contrasting the trade-offs entailed by changes to the attention weights of each modality.

2.2 Role of system attention

Although prior work widely presumes that removing or redistributing weights from attention sinks suppresses VLM hallucination (Huang et al., 2024; Kang et al., 2025), the large attention sinks in the system modality discussed above have yet to be strategically exploited for hallucination mitigation. Instead, the modality as a whole has generally been dismissed as irrelevant to hallucination (Bi et al., 2025; Yang et al., 2025) or lumped together with the textual one (Sarkar et al., 2025). While Liang and Cai (2026) examine the functions of attention sinks within the system modality by zero-ablating their attention weights, the study largely neglects the potential causal role of such tokens in VLM hallucination. Our system-mediated interpretation enables us to directly investigate just this by testing a clear hypothesis: that large system weights induce hallucination through depriving both the image and textual modalities of attention.

2.3 The yes-bias in VLMs

Studies on the causes of VLM hallucination have chiefly evaluated models and interventions on a combination of yes/no — i.e. binary, mostly single-token generation — and autoregressive tasks, e.g. image captioning (Bi et al., 2025; Gong et al., 2024; Huang et al., 2024; Shu et al., 2025; Liu et al., 2025a; Zhou et al., 2025; Li et al., 2025). Although intended to test generalisability, this has in practice conflated hallucinatory behaviour in different settings and overlooks evidence that yes/no and autoregressive hallucination may have distinct underlying causes (Zheng and Zhang 2025, also cf.

Parcalabescu and Frank 2025).

As a result, the mechanistic causes of the yes-bias have not been adequately investigated beyond tentative references to imbalanced training data (Hu et al., 2023; Liu et al., 2024a; He et al., 2025; Hu et al., 2025). This neglect is difficult to justify as the yes-bias crops up in a wide variety of binary tasks and is specifically targeted by many commonly used benchmarks, including POPE (Li et al., 2023) and HallusionBench (Guan et al., 2024). It is likewise a missed opportunity to isolate clean mechanistic signals of hallucination that can be intervened on. This comes about because yes/no tasks permit objective evaluation in contrast to autoregressive settings, where hallucination-related signals are more ambiguous (Huang et al., 2024) and evaluation often relies on less objective methods such as GPT-as-judge (e.g. Yue et al., 2024; Ghosh et al., 2025; Liu et al., 2025a).

Collectively, the above makes the case for causal interventions on a VLM's text decoder to evaluate the feasibility of a system-mediated account of the yes-bias against an image-centric one. The next section details how we implement these.

3 Causal Intervention Framework

3.1 Disambiguating image-centric and system-mediated hypotheses

Our causal interventions are designed to differentiate the image-centric and system-mediated hypotheses through their effectiveness in suppressing the yes-bias. To operationalise these two hypotheses, we define a modality to have insufficient attention if increasing attention weights enhances performance. Conversely, a modality has functionally redundant attention if ablating the attention weights of all its tokens either does not degrade performance or leads to improvement.

To reiterate, the image-centric hypothesis assumes that hallucination arises mainly from insufficient image attention. This view accordingly predicts that the yes-bias is best mitigated by increasing attention to the image. In contrast, our system-mediated hypothesis attributes hallucination to functionally redundant system attention weights inducing deficits in both the image and textual modalities. As such, it predicts that redistributing attention from the system modality to boost attention to the image and/or textual modalities will be most effective.

3.2 Proportional redistribution of attention weights

As a minimal test of both hypotheses, we design an intervention to probe the role of each modality in producing the yes-bias. Specifically, we apply what we term *proportional redistribution* to the system and textual modalities in turn. Each of these serves as *source modalities*. From these, we reallocate *all* attention weights to the remaining two modalities, the *recipient modalities*,[2] in proportion to the latter's original attention weights.

This approach operationalises the predictions outlined in Section 3.1. If the system-mediated hypothesis is correct, proportionally redistributing attention from the system modality towards both the image and textual modalities should produce the largest performance gains. Conversely, under the image-centric hypothesis, proportional redistributions from either the system or textual modalities should yield comparable improvements as both would increase image attention.

Formally, we let $\mathcal{M} = $ system, image, text represent the set of input modalities. For each modality $m \in \mathcal{M}$, let A_i denote the post-softmax attention weight assigned to token $i \in m$, aggregated across heads and decoder query positions within the decoder layers we intervene on. The total attention mass assigned to modality m is then defined as:

$$\alpha_m = \sum_{i \in m} A_i, \tag{1}$$

Further, let $s \in \{\text{system}, \text{text}\}$ denote the source modality. The set of recipient modalities is then defined as $R = \mathcal{M} \setminus \{s\}$, comprising the two remaining modalities. On this basis, proportional redistribution is implemented as a two-step process: first, zero-ablation of the source modality's attention weights (Equation 2) and second, reallocation of the ablated weights to the recipient modalities in proportion to their original attention weights (Equation 3).

$$\alpha'_s = 0, \tag{2}$$

$$\alpha'_r = \alpha_r + \alpha_s \cdot \frac{\alpha_r}{\sum_{r' \in R} \alpha_{r'}}, \quad r \in R, \tag{3}$$

[2]Unlike Yang et al. (2025), we perform redistributions instead of just rescaling modality-specific weights without renormalisation for numerical stability.

4 Main Experiments

4.1 Evaluation set-up

Architectures. Our interventions focus on the text decoder of LLaVA-1.5 7B (Liu et al., 2023) as it is widely used in interpretability studies and to test hallucination-mitigating techniques. Appendix A.5 reports supplementary results from interventions on additional models from the same family, LLaVA-NeXT-Vicuna 7B (Liu et al., 2024b) and LLaVA-1.5 13B. The applicability of our conclusions to Qwen2-VL-Instruct (Team et al., 2024), which does not belong to the same family, is also critically evaluated in Appendix A.6.

Scope of interventions. Prior work situates model components responsible for hallucination in the later layers of the text decoder (Bai et al., 2024; Chen et al., 2024b; Wang et al., 2025). Accordingly, we partition the 32-layer text decoder of LLaVA-1.5 7B into four quarters, each consisting of eight attention layers. We refer to these as Q1-4 and focus on Q4 (layers 25 to 32), taking it to encompass the decoder's late layers.[3] Appendix A.2 shows that our focus on Q4 — rather than other model components — is empirically well founded.

Benchmarks. To investigate the degree of yes-bias with and without intervention, we perform evaluations on six benchmarks specifically designed for this purpose (Table 1). Each benchmark consists of paired prompts that differ minimally from each other either in their image or in the user query such that the ground-truth answer is 'yes' for one prompt and 'no' for the other. Where this is not already enforced, we explicitly request yes/no answers in the user prompts.

Evaluation metrics. We use three complementary metrics to quantify the yes-bias:

(1) **Simple accuracy**: Proportion of prompts answered correctly

(2) **Paired accuracy**: Proportion of prompt pairs with both answers correct

(3) **Yes-rate**: Proportion of 'yes' out of the total number of responses generated, ignoring correctness (cf. Guan et al., 2024)

The difference between the yes-rate and the actual proportion of ground-truth-yes prompts is our most direct measure of the yes-bias following Guan et al. (2024). Simple and paired accuracy quantify its impact on performance: default 'yes' responses harm simple accuracy and are especially penalised by paired accuracy given the paired-prompt design of our benchmarks. Unless otherwise stated, we report only simple accuracy and exclude benchmark-specific metrics (e.g. Accuracy+ for MME).

4.2 The system-mediated hypothesis is valid

We first apply proportional redistribution of Q4 system and text attention as described in Section 3.2. As a shorthand, we refer to these as *Q4 system* and *Q4 text redistribution*, respectively. To verify the need for positing two factors under our system-mediated account of hallucination (see Section 1), we also perform Q4 system ablation. This removes system attention without redistributing or renormalising weights across modalities, enabling us to distinguish the effects of removing system attention from those of redistributing it to the other modalities.

We compare these interventions against a no-intervention baseline, interventions that boost image attention (PAI, Liu et al., 2024c, and Image × 2.0) and a method that suppresses text attention like Q4 text redistribution (AD-HH, Yang et al., 2025). Additional implementation details for these alternative methods are provided in Appendix A.1.

Results. Compared to image-boosting and text-suppressing alternatives, Q4 system redistribution mitigates the yes-bias much more effectively on our three metrics and in five of our six benchmarks (Table 2). The severity of this bias is particularly evident in HallusionBench, SugarCrepe and Winoground, where the no-intervention baselines exhibit yes-rates exceeding 90% despite ground-truth yes proportions being closer to 50%. It is for these benchmarks that Q4 system redistribution produces the largest improvements in paired accuracy, which more than doubles for Winoground.

Within Q4, these findings confirm the system-mediated hypothesis' prediction, i.e. removing system weights to boost both image and text attention effectively reduces the yes-bias. MME exhibits a different pattern, which we analyse in Section 5. As a sanity check, Appendix A.4 further presents evidence that such interventions affect yes/no responses much more strongly than other types of generation, justifying our view of the yes-bias as a distinct form of VLM hallucination.

[3] Quartering the model is a compromise as per-head and layer interventions are computationally infeasible. Given our still developing understanding of how VLM functions are spread across decoder layers, segmenting the model this way is arguably more principled than the customary practice of dividing it into early, middle and late segments, each comprising different numbers of layers (Bi et al., 2025; Neo et al., 2025; Wang et al., 2025).

Benchmark	Purpose	No. of prompts
BEAF (Moon et al., 2024)	Adversarial object detection (objects digitally edited out)	26,000
HallusionBench (Guan et al., 2024)	Various (e.g. optical illusions, OCR, reasoning)	951
MME (Fu et al., 2025)	Various, with emphasis on knowledge extraction	2,374
NaturalBench (Li et al., 2024)	Verification of the same captions against minimally different images	5,800 (yes/no subset)
SugarCrepe (Hsieh et al., 2023)	Verification of minimally different captions against the same image	1,820 (subset)
Winoground (Thrush et al., 2022)	Verification of captions only differing in word order against images	5,664

Table 1: Overview of the six yes/no paired-prompt vision-language benchmarks used in this study.

Likewise, Q4 system ablation clearly indicates the necessity of such a two-factor account of hallucination. Without also increasing both image and text attention, removing system attention fails to bring about substantially better performance, yielding only negligible improvements (again Table 2). In sum, these results affirm the empirical soundness of our system-mediated hypothesis as operationalised in our set-up.

4.3 The image-centric hypothesis is insufficient

To more cleanly evaluate the predictions of both hypotheses as set out in Section 3.1, we next examine the relative impact of selectively boosting image versus text attention. For this purpose, we perform *pairwise* redistributions that transfer removed weights in Q4 to only one recipient modality rather than splitting them between two. This yields transfers in both directions between the system and image, system and text, and image and text modalities. Based on the account above, the image-centric hypothesis predicts that image-boosting transfers (system-to-image and text-to-image) will suppress the yes-bias more effectively than text-boosting ones (here, system-to-text). The system-mediated hypothesis, by contrast, predicts system-suppressing transfers (system-to-image and system-to-text) will be superior to image-boosting ones (text-to-image).

Results. Overall, the greatest improvements in all three metrics are obtained with system-to-text transfers, followed by the system-to-image and text-to-image settings. Table 3 illustrates that for HallusionBench. The pattern generalises to the remaining benchmarks (Table 4), again with MME as the only exception (Table 5). With system-to-text transfers outperforming both image-boosting interventions, the prediction of our system-mediated hypothesis is validated. At the same time, the gains brought about by boosting text attention show a strong formulation of the image-centric hypothesis — that the image modality is the only one with an attention deficit — to be untenable in our set-up.

Collectively, these findings imply a substantial attention imbalance in Q4, characterised by redundant attention allocated to system tokens and insufficient attention directed towards both image and text tokens, rather than image tokens alone.

5 Error Analysis

5.1 Distinguishing coarse- from fine-grained tasks

Having shown that Q4 system-mediated attention imbalances are major contributors to the yes-bias, we now examine why redistributing this attention yields gains in all our paired-prompt benchmarks except MME. We start with two critical clues: one, Q4 system redistribution suppresses the yes-rate across all benchmarks — it is only in MME that this fails to translate into performance gains; two, Q4 system ablation does not suppress the yes-rate despite removing system attention (Table 2).

A salient commonality of our non-MME benchmarks is that they explicitly require compositional analysis of multimodal inputs to distinguish between minimally different prompt pairs (see Table 1). In other words, these tasks prioritise fine-grained, local detail, and systematically penalise reliance on high-level representations of inputs.

By contrast, MME instantiates a class of tasks where such representations are often appropriate. Many MME prompts involve verifying if images depict specific types of artwork, landmarks, films, or scenes. This arguably requires matching high-level representations of known referents across modalities more than local visual detail.

To show the relevance of this distinction, we introduce a seventh benchmark, POPE (Li et al.,

Benchmark (% prompts with ground-truth yes)	**Metric**	**PAI**	**Image× 2.0**	**AD-HH**	**Q4 text redistr (prop)**	**Q4 system abl**	**Q4 system redistr (prop)**	**No-intervention baseline**
BEAF (34.11% yes)	Simple acc$^{\Uparrow}$	83.01 (–0.87)	85.73 (+2.38)	83.67 (-0.08)	85.55 (+2.16)	84.44 (–3.76)	**88.40** (+5.56)	83.74
	Paired acc$^{\Uparrow}$	77.72 (–1.14)	81.01 (+3.04)	78.51 (–0.14)	80.75 (+2.71)	79.45 (+1.05)	**84.74** (+7.78)	78.62
	Yes-rate$^{\Delta\downarrow}$	45.73 (+34.07)	41.55 (+21.81)	44.74 (+31.16)	41.68 (+22.19)	43.58 (+27.77)	**32.62** (–4.37)	44.69 (+31.02)
HallusionBench (42.17% yes)	Simple acc$^{\Uparrow}$	45.64 (0.00)	45.95 (+0.68)	45.22 (–0.92)	46.58 (+2.06)	45.74 (+0.22)	**53.31** (+16.81)	45.64
	Paired acc$^{\Uparrow}$	9.53 (–2.46)	10.00 (+2.35)	9.30 (–4.81)	9.77 (0.00)	10.00 (+2.35)	**19.07** (+95.19)	9.77
	Yes-rate$^{\Delta\downarrow}$	90.22 (+113.94)	88.22 (+109.20)	90.43 (+114.44)	84.23 (+99.74)	89.27 (+111.70)	**59.62** (+41.38)	89.80 (+112.95)
MME (50.00% yes)	Simple acc$^{\Uparrow}$	78.73 (–0.32)	78.35 (–0.80)	78.73 (–0.32)	78.26 (–0.92)	78.90 (–0.11)	72.83 (–8.44)	**78.98**
	Paired acc$^{\Uparrow}$	**30.77** (0.00)	**30.77** (0.00)	29.23 (–5.00)	29.23 (–5.00)	30.00 (–2.50)	21.54 (–30.00)	**30.77**
	Yes-rate$^{\Delta\downarrow}$	42.62 (-14.76)	40.14 (–19.72)	**42.88** (-14.24)	41.07 (–17.86)	41.20 (–17.61)	28.98 (–42.04)	42.29 (-15.42)
NaturalBench (50.00% yes)	Simple acc$^{\Uparrow}$	59.97 (–0.42)	61.28 (+1.76)	59.79 (–0.71)	62.17 (+3.24)	60.97 (+1.24)	**66.02** (+9.63)	60.22
	Paired acc$^{\Uparrow}$	21.10 (–2.09)	23.72 (+10.07)	20.66 (–4.13)	25.52 (+18.42)	23.00 (+6.73)	**34.59** (+60.51)	21.55
	Yes-rate$^{\Delta\downarrow}$	86.03 (+72.06)	83.83 (+67.66)	86.34 (+72.68)	82.07 (+64.14)	84.55 (+69.10)	**66.67** (+33.34)	85.67 (+71.34)
SugarCrepe (50.00% yes)	Simple acc$^{\Uparrow}$	57.47 (–0.10)	58.18 (+1.13)	57.14 (–0.68)	58.79 (+2.19)	57.97 (+0.76)	**62.75** (+9.07)	57.53
	Paired acc$^{\Uparrow}$	15.05 (–0.73)	16.37 (+7.98)	14.40 (–5.01)	17.58 (+15.96)	16.04 (+5.83)	**25.93** (+71.04)	15.16
	Yes-rate$^{\Delta\downarrow}$	91.98 (+83.96)	91.37 (+82.74)	92.53 (+85.06)	90.66 (+81.32)	91.59 (+83.19)	**85.71** (+71.42)	92.03 (+84.06)
Winoground (50.00% yes)	Simple acc$^{\Uparrow}$	53.50 (+0.22)	53.69 (+0.58)	53.06 (–0.60)	55.00 (+3.03)	53.56 (+0.34)	**57.88** (+8.43)	53.38
	Paired acc$^{\Uparrow}$	7.63 (+3.39)	8.85 (+19.92)	6.63 (–10.16)	12.00 (+62.60)	8.13 (+10.09)	**19.50** (+164.23)	7.38
	Yes-rate$^{\Delta\downarrow}$	95.25 (+90.5)	94.56 (+89.12)	95.94 (+91.88)	92.00 (+84.00)	94.81 (+89.63)	**75.38** (+50.76)	95.38 (+90.76)

Table 2: A comparison of different interventions on simple accuracy, paired accuracy and yes-rate across six paired-prompt benchmarks. Best results for each metric are in bold. Our simple redistribution of system attention weights (Q4 system redistr (prop)) consistently outperforms the image and text-based alternatives by a sizeable margin on all three metrics in five of the six benchmarks. Moreover, simply removing Q4 system attention without proportional redistribution (Q4 system ablation) does not bring about the sharp improvements in performance that Q4 system redistribution does. For simple and paired accuracy, percentage (not percentage-point) differences relative to the no-intervention baseline are shown in brackets. The differences for yes-rate are with reference to the actual proportion of prompts with ground truth 'yes, shown bracketed in the Benchmark column — the best yes-rate is the one that deviates least from that ground-truth value.

Source	Recipient								
	Simple acc$^{\Uparrow}$			Paired acc$^{\Uparrow}$			Yes-rate$^{\Delta\downarrow}$ (ground truth 42.17%)		
	System	*Image*	*Text*	*System*	*Image*	*Text*	*System*	*Image*	*Text*
System	45.64	51.52 (+12.88)	**58.04** (+27.17)	9.77	15.12 (+54.76)	**23.26** (+138.08)	89.80 (+112.95)	65.82 (+56.08)	**39.12** (–7.23)
Image	45.43 (–0.46)	45.64	45.53 (–0.24)	9.30 (–4.81)	9.77	9.53 (–2.46)	88.75 (+110.46)	89.80 (+112.95)	89.7 (+112.71)
Text	46.69 (+2.30)	47.21 (+3.44)	45.64	10.00 (+2.35)	10.23 (+4.71)	9.77	84.12 (+99.48)	83.39 (+97.75)	89.80 (+112.95)

Table 3: Changes in performance in HallusionBench under pairwise attention-weight transfers between all combinations of source and recipient modalities in Q4 of LLaVA-1.5 7B. System-to-text transfers consistently yield the biggest improvements in performance, followed by system-to-image transfers. Performance is measured in simple and paired accuracy alongside yes-rate, with percentage differences from the no-intervention baseline (simple/ paired accuracy) or the ground truth (yes-rate) indicated in brackets. The best score for each metric is bolded, the second best underlined and baseline configurations are marked by shading.

Source	Recipient											
	BEAF			**NaturalBench (yes/no)**			**SugarCrepe**			**Winoground**		
	System	*Image*	*Text*	*System*	*Image*	*Text*	*System*	*Image*	*Text*	*System*	*Image*	*Text*
System	82.38	88.06	**88.08**	60.22	65.59	**67.50**	57.53	62.47	**64.95**	53.38	57.19	**58.44**
Image	84.17	82.38	83.71	60.60	60.22	60.10	57.74	57.53	57.53	53.63	53.38	53.31
Text	85.56	85.76	82.38	62.10	62.47	60.22	58.79	58.96	57.53	55.06	54.88	53.38

Table 4: Changes in simple accuracy in BEAF, NaturalBench (yes/no), SugarCrepe, and Winoground under pairwise attention-weight transfers between all combinations of source and recipient modalities in Q4 of LLaVA-1.5 7B. System-to-text transfers consistently yield the biggest improvements in performance, followed by system-to-image transfers. The best score is bolded, the second best underlined and baseline configurations are marked by shading.

Source	Recipient		
	MME		
	System	*Image*	*Text*
System	78.98	66.39	61.25
Image	78.85	78.98	78.48
Text	78.22	78.18	78.98

Table 5: Changes in simple accuracy in MME under pairwise attention-weight transfers between all combinations of source and recipient modalities in Q4 of LLaVA-1.5 7B. Exactly contrary to the other benchmarks, the system-to-text transfer consistently yields the *worst* performance, followed by the system-to-image transfer.

2023). POPE is a balanced yes/no task that requires models to verify if an object is in an image. It comprises three splits, differing in how absent objects are selected. Q4 system redistribution improves performance only in the adversarial split, while degrading it in the other two despite identical ground-truth yes-rates throughout (Table 6).

Crucially, POPE's adversarial split resembles the non-MME benchmarks in that correct responses hinge on identifying subtle differences in objects present, hence necessitating fine-grained scrutiny of visual inputs. Here, absent objects are selected to be related to the scene in the image, making reliance on high-level visual representations actively misleading. The other two splits, on the other hand, do not penalise such reliance as harshly, as absent objects are chosen either randomly or by their dataset frequency. Example prompts from representative benchmarks are included in Appendix A.3.

5.2 Late-layer system attention and cross-modal aggregation

This pattern motivates a unified interpretation of the seven benchmarks: system-mediated attention imbalances in the baseline model encourage the reliance on coarse, aggregated representations of its inputs as a learnt default. This reliance reflects attention allocation patterns learnt during training that are helpful for tasks where fine-grained analysis is not crucial, but detrimental where it is.

Mirroring the two factors under our system-mediated hypothesis, we submit that this reliance arises through a dual mechanism. First, attention sinks such as the <BOS> token in the system modal-

POPE split	Metric	PAI	Image×2.0	AD-HH	Q4 text (prop)	Q4 system (prop)	No-intervention baseline
Overall (50.00% yes)	Simple acc$^{\Uparrow}$	86.56 (–0.35)	**87.67** (+0.92)	86.77 (–0.12)	87.56 (+0.79)	87.20 (+0.38)	86.87
	Yes-rate$^{\Delta\downarrow}$	54.00 (+8.00)	**50.53** (+1.06)	53.19 (+6.38)	50.71 (+1.42)	43.49 (–13.02)	53.22 (+6.44)
Adversarial (50.00% yes)	Simple acc$^{\Uparrow}$	81.73 (–0.61)	83.67 (+1.75)	82.03 (–0.24)	83.43 (+1.45)	**85.10** (+3.49)	82.23
	Yes-rate$^{\Delta\downarrow}$	58.87 (+17.74)	54.53 (+9.06)	57.97 (+15.94)	54.83 (+9.66)	**45.50** (–9.00)	57.90 (+15.80)
Popular (50.00% yes)	Simple acc$^{\Uparrow}$	87.50 (–0.53)	**88.60** (+0.72)	87.83 (–0.16)	88.47 (+0.57)	87.80 (–0.19)	87.97
	Yes-rate$^{\Delta\downarrow}$	53.03 (+6.06)	49.60 (–0.80)	52.10 (+4.20)	**49.80** (–0.40)	42.93 (–14.14)	52.10 (+4.20)
Random (50.00% yes)	Simple acc$^{\Uparrow}$	90.43 (+0.03)	90.73 (+0.37)	90.43 (+0.03)	**90.77** (+0.41)	88.70 (–1.88)	90.40
	Yes-rate$^{\Delta\downarrow}$	**50.10** (+0.20)	47.47 (–5.06)	49.50 (–1.00)	47.50 (–5.00)	42.03 (–15.94)	49.67 (-0.66)

Table 6: Shifts in simple accuracy and yes-rate for each split of POPE (MSCOCO) under different interventions. Q4 system redistribution only benefits performance in the adversarial split, while harming it in the two others. Paired accuracy does not apply to this benchmark because while balanced, prompts are not explicitly grouped into positive/negative pairs.

ity encourage the coarse, cross-modal aggregation of information (Barbero et al., 2025). Second, image and text attention deficits hinder the extraction of fine-grained information from the inputs (Bi et al., 2025; Wang et al., 2025). When coarse, aggregated representations dominate, responses rely chiefly on high-level information; where this does not suffice for an adequate response, the model falls back on a second default behaviour, the yes-bias (Hu et al., 2023; Liu et al., 2024a; Hu et al., 2025).

Such an interpretation provides a principled explanation for the performance differences between Q4 system ablation and redistribution. As it only addresses aggregation by removing system attention, Q4 system ablation falls short of increasing the use of fine-grained multimodal information, thereby failing to disable the default yes-bias. Q4 system redistribution, on the other hand, both reduces cross-modal aggregation *and* facilitates access to finer-grained image and text information. Provided with information helpful in compositional tasks, the model relies less on default-yes responses, thus suppressing the yes-bias.

5.3 Behavioural evidence

LLaVA-1.5 7B's yes-rates support linking coarse representations to default model behaviour. Tasks focussing on such representations exhibit markedly smaller disparities between no-intervention yes-rates and the ground truth than compositional tasks. For MME, this disparity is only 15.42%, compared to gaps ranging from 31.02% to 112.95% for the remaining benchmarks (Table 2). A similar pattern holds among the three POPE splits, where the adversarial split exhibits a 15.80% disparity, in stark contrast to 4.20% and 0.66% respectively for the other two splits (Table 6). Consequently, it is precisely the tasks least aligned with default behaviour that benefit from intervention.

5.4 Representational evidence

There is likewise mechanistic evidence more directly in favour of our account. If indeed late-layer system attention encourages cross-modal aggregation of representations — i.e. representations that are coarser-grained and more global, we predict that the presence of this attention would make image and text representations more similar. By implication, *removing* this attention should make them *less* similar.

Confirming our predictions, redistributing system attention significantly lowers image-text cosine similarity in late decoder layers across our benchmarks. Using a paired t-test and Wilcoxon sign-rank test, we consistently obtain Holm-corrected p-values of less than 0.001 as shown in Table 7.

Dataset	No. (layers × prompts)	Mean cos sim (no intervention)	Mean cos sim (Q4 redistribution)	Paired t-stat	Holm p (t-test)	Wilcoxon stat	Holm p (Wilcoxon)
BEAF	182,448	0.139	0.100	762.154	< 0.001	271,170,836.5	< 0.001
HallusionBench	6,657	0.127	0.087	129.159	< 0.001	593,527.5	< 0.001
MME	16,618	0.136	0.090	187.214	< 0.001	4,711,829.5	< 0.001
POPE	63,000	0.139	0.095	454.681	< 0.001	30,506,750	< 0.001
NaturalBench	46,400	0.135	0.092	391.512	< 0.001	27,924,061	< 0.001
SugarCrepe	790,622	0.136	0.092	1634.860	< 0.001	4,407,350,131	< 0.001
Winoground	11,200	0.137	0.090	198.562	< 0.001	740,974	< 0.001

Table 7: Redistributing late-layer (Q4) system attention brings about a clearly lower image-text cosine similarity in the layers affected by intervention. This is shown by the similarity values under intervention (Mean cos sim (Q4 redistribution)) being significantly lower than those without intervention (Mean cos sim (no intervention)) for all seven datasets covered here. All p-values obtained are consistently under 0.001. For each dataset and condition (with or without intervention), cosine similarity values are averaged across Q4 layers and prompts within that dataset.

Layers without intervention show no change.

Both these lines of evidence underscore the close association between system-mediated attention imbalances and the yes-bias. An image-centric account, by contrast, fails to predict the specific failure modes of system-based interventions.

6 Conclusion

Above, we investigated the yes-bias, a widespread form of VLM hallucination, and showed that attributing it solely to insufficient image attention is empirically inadequate. Our causal interventions on LLaVA-1.5 7B revealed that redundant system attention in late decoder layers drives imbalances that harm both image and text processing, possibly by promoting reliance on coarse input representations. Redistributing this system attention substantially reduces the yes-bias, outperforming existing image- and text-oriented mitigation strategies for compositional tasks.

Building on this, future work should explore the implications of our findings for autoregressive hallucination, complementing recent work relating VLM answers to their explanations (Parcalabescu and Frank, 2025). It would be especially instructive to apply holistic causal interventions, similar to what we used, to other manifestations of VLM hallucination in order to make out whether hallucination-inducing attention imbalances share any commonalities. Further, given the robust improvements system-to-text redistributions yielded in our set-up, the prospect of redundant system attention driving the yes-bias in text-only LLMs should also be investigated.

Limitations

Our work focusses on LLaVA-based architectures, chosen for their widespread use in interpretability studies. We also restrict our scope to single-token generation in yes/no contexts. The generality of our system-mediated account should therefore be tested on a broader variety of VLM families and in a wider range of settings. The latter ought to cover other types of default single-token responses in multiple-choice and counting tasks for instance (cf. Li et al., 2024; Sim et al., 2025), as well as autoregressive generation.

Next, our interventions mostly employ zero-ablation with complete redistribution to avoid arbitrary assumptions about how much attention to transfer between modalities. This has prevented us from methodically exploring the effects of partial redistributions, to investigate for example how monotonic their impacts on performance are. We likewise concentrate on the quarter level owing to computational constraints, precluding a per-layer study analogous to Shi et al. (2025).

Finally, our claims regarding representation-level mechanisms should be further validated through directly probing a model's internal states, e.g. to quantify whether tasks we claim to be more closely aligned with its default state induce less uncertainty than those less well aligned. This would supplement the behavioural and representational evidence we adduced above.

Acknowledgements

Varsha Suresh is funded by Deutsche Forschungsgemeinschaft, Funder ID: http://dx.doi.org/10.13039/501100001659, SFB 1102: 'Informa-

tion Density and Linguistic Encoding'.

References

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of Multimodal Large Language Models: A Survey. *CoRR*, abs/2404.18930.

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Petar Veličković, Razvan Pascanu, and Michael M. Bronstein. 2025. Why do LLMs attend to the first token? In *Second Conference on Language Modeling*.

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. 2025. Unveiling visual perception in language models: An attention head analysis approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4135–4144.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking Common Sense: WHOOPS! A Vision-and-Language Benchmark of Synthetic and Compositional Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2616–2627.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024a. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. 2024b. Multi-Object Hallucination in Vision Language Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2025. MME: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. 2025. Visual Description Grounding Reduces Hallucinations and Boosts Reasoning in LVLMs. In *The Thirteenth International Conference on Learning Representations*.

Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. 2024. DAMRO: Dive into the Attention Mechanism of LVLM to Reduce Object Hallucination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7696–7712, Miami, Florida, USA. Association for Computational Linguistics.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2025. When Attention Sink Emerges in Language Models: An Empirical View. In *The Thirteenth International Conference on Learning Representations*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.

Yixiao He, Haifeng Sun, Pengfei Ren, Jingyu Wang, Huazheng Wang, Qi Qi, Zirui Zhuang, and Jing Wang. 2025. Evaluating and Mitigating Object Hallucination in Large Vision-Language Models: Can They Still See Removed Objects? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6841–6858.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. 2023. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36:31096–31116.

Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. 2023. CIEM: Contrastive Instruction Evaluation Method for Better Instruction Tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Rui Hu, Yahan Tu, Shuyu Wei, Dongyuan Lu, and Jitao Sang. 2025. Prescribing the Right Remedy: Mitigating Hallucinations in Large Vision-Language Models via Targeted Instruction Tuning. *Information Sciences*, page 122361.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25004–25014.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See What You Are Told: Visual Attention Sink in Large Multimodal Models. In *The Thirteenth International Conference on Learning Representations*.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. 2024. NaturalBench: Evaluating vision-language models on natural adversarial samples. *Advances in Neural Information Processing Systems*, 37:17044–17068.

Shawn Li, Jiashu Qu, Yuxiao Zhou, Yuehan Qin, Tiankai Yang, and Yue Zhao. 2025. Treble Counterfactual VLMs: A Causal Approach to Hallucination. *CoRR*, abs/2503.06169.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Haiji Liang and Yujun Cai. 2026. Not Errors but Guardians: Understanding Sink Tokens in Multimodal LLMs. OpenReview.

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

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A Survey on Hallucination in Large Vision-Language Models. *CoRR*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sheng Liu, Haotian Ye, and James Zou. 2025a. Reducing Hallucinations in Large Vision-Language Models via Latent Space Steering. In *The Thirteenth International Conference on Learning Representations*.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024c. Paying more attention to image: A training-free method for alleviating hallucination in LVLMs. In *European Conference on Computer Vision*, pages 125–140. Springer.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Joy Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Benoit Dumoulin, and Hanghang Tong. 2025b. Seeing but Not Believing: Probing the Disconnect Between Visual Attention and Answer Correctness in VLMs. *Preprint*, arXiv:2510.17771.

Ruilin Luo, Chufan Shi, Yizhen Zhang, Cheng Yang, Songtao Jiang, Tongkun Guan, Ruizhe Chen, Ruihang Chu, Peng Wang, Mingkun Yang, Lei Wang, Yujiu Yang, Junyang Lin, and Zhibo Yang. 2026. From narrow to panoramic vision: Attention-guided cold-start reshapes multimodal reasoning. In *The Fourteenth International Conference on Learning Representations*.

Ye-Bin Moon, Hyeon-Woo Nam, Wonseok Choi, and Tae-Hyun Oh. 2024. BEAF: Observing BEfore-AFter Changes to Evaluate Hallucination in Vision-language Models. In *European Conference on Computer Vision (ECCV)*.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. 2025. Towards Interpreting Visual Information Processing in Vision-Language Models. In *The Thirteenth International Conference on Learning Representations*.

Letitia Parcalabescu and Anette Frank. 2025. Do Vision & Language Decoders use Images and Text equally? How Self-consistent are their Explanations? In *The Thirteenth International Conference on Learning Representations*.

Candace Ross, Melissa Hall, Adriana Romero-Soriano, and Adina Williams. 2024. What makes a good metric? Evaluating automatic metrics for text-to-image consistency. In *First Conference on Language Modeling*.

Sreetama Sarkar, Yue Che, Alex Gavin, Peter A. Beerel, and Souvik Kundu. 2025. Mitigating Hallucinations in Vision-Language Models through Image-Guided Head Suppression. *Preprint*, arXiv:2505.16411.

Cheng Shi, Yizhou Yu, and Sibei Yang. 2025. Vision Function Layer in Multimodal LLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Dong Shu, Haiyan Zhao, Jingyu Hu, Weiru Liu, Ali Payani, Lu Cheng, and Mengnan Du. 2025. Large Vision-Language Model Alignment and Misalignment: A Survey Through the Lens of Explainability. *Preprint*, arXiv:2501.01346.

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. 2025. Can VLMs actually see and read? A survey on modality collapse in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24452–24470.

Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. 2025. Intervening Anchor Token: Decoding Strategy in Alleviating Hallucinations for MLLMs. In *The Thirteenth International Conference on Learning Representations*.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. 2025. MLLM can see? Dynamic Correction Decoding for Hallucination Mitigation. In *The Thirteenth International Conference on Learning Representations*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient Streaming Language Models with Attention Sinks. In *The Twelfth International Conference on Learning Representations*.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. 2025. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *The Thirteenth International Conference on Learning Representations*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5014–5022.

Yiming Zhang, Zicheng Zhang, Xinyi Wei, Xiaohong Liu, Guangtao Zhai, and Xiongkuo Min. 2025. IllusionBench: A Large-scale and Comprehensive Benchmark for Visual Illusion Understanding in Vision-Language Models. *arXiv preprint arXiv:2501.00848*.

Haohan Zheng and Zhenguo Zhang. 2025. Modality Bias in LVLMs: Analyzing and Mitigating Object Hallucination via Attention Lens. *arXiv preprint arXiv:2508.02419*.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. 2025. Mitigating Modality Prior-Induced Hallucinations in Multimodal Large Language Models via Deciphering Attention Causality. In *The Thirteenth International Conference on Learning Representations*.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2025. IBD: Alleviating hallucinations in large vision-language models via image-biased decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1624–1633.

A Appendix

A.1 Baseline methods used in Q4 proportional redistribution

As stated in Section 4, we include two methods that boost image attention and one that suppresses text attention:

Image-boosting:

- **PAI** (Liu et al., 2024c) scales image weights by 1.5 *before* softmax, then ensures the output logits are sufficiently distinct from those of a unimodal model without access to image inputs. It targets all decoder heads by default. We use a reimplementation of the original algorithm, with the recommended settings of $\alpha = 0.5$ and $\gamma = 1.1$.
- **Image** $\times$ 2.0 is an ad-hoc intervention on the decoder attention heads identified by Yang et al. (2025) to contribute most to object hallucination in the MSCOCO (Lin et al., 2014) image captioning task: their 'MSCOCO hallucination heads'. It simply scales image attention on each by 2.0.

Text-suppressing:

- **AD-HH** (Yang et al., 2025) similarly targets Yang et al.'s (2025) MSCOCO hallucination heads. Following the recommended settings, text attention is zeroed out on a head if it exceeds 40% of the total weights post softmax.

A.2 The yes-bias is a late-layer phenomenon

This section provides empirical justification for our focus on Q4. We first rule out the presence of any substantial attention imbalances at the global (i.e. whole-model) level and subsequently in Q1-3 (i.e. layers 1 to 24) of LLaVA-1.5 7B's text decoder.

A.2.1 Global redistributions

To assess whether the yes-bias arises from global attention imbalances as a cause of the yes-bias, we perform both pairwise and proportional redistributions of attention weights at the global level, intervening simultaneously across all decoder heads.

Unlike the 100% redistributions employed in earlier experiments, we adopt graduated transfers that reallocate 10%, 20% and 30% of the source modality's attention weights respectively. This is because fine-tuning showed that transferring more than 30% of weights from the source modality across all decoder heads tended to lead to a collapse of model functionality.

Overall, these global interventions predominantly degrade performance. In the few cases where performance does not deteriorate, the observed improvements are marginal and substantially smaller than those achieved through the Q4 interventions introduced in Sections 3 and 4. This suggests that global attention imbalances are neither robust nor the primary driver of the yes-bias, thereby supporting the localised intervention strategy adopted in this work. Where proportional redistributions of attention weights are concerned, Figure 1 shows these to be mostly harmful regardless of the source and recipient modalities involved and not lead to consistent gains. Figure 2 tells a similar story for proportional redistributions in the case of the SugarCrepe benchmark. The remaining plots and metrics beyond simple accuracy are not shown for space considerations, >100 interventions having been performed here. However, the observed trends are consistent across our benchmarks and evaluation metrics, also extending to the pairwise setting.

These results justify our local approach to interventions, emphasising redistributions performed at the quarter level rather than across all decoder heads.

A.2.2 Per-quarter redistributions

To verify that Q4 is indeed where the largest attention imbalances are found, we carried out pairwise and proportional redistribution of attention weights between all possible combinations of modalities in Q1-3 in turn. As in prior quarter-level experiments, these interventions involve redistributing 100% of the attention weights from the source modality. In terms of proportional redistribution, Figure 3 shows that across all six benchmarks, interventions in Q1 and Q2 are consistently detrimental, regardless of the source or target modality. These trends hold even for MME, the one benchmark that Q4 system redistribution failed to yield enhanced performance for. Interventions in Q3 similarly fail to yield consistent improvements.

Accuracy gains are almost exclusively concentrated in Q4 (layers 25-32), implicating the system modality most of all but to a much smaller extent also the textual one. This indicates that the earlier portions of the decoder do not contain stable reservoirs of modality-specific redundant attention that can be reallocated for performance gains comparable to those achieved in Q4.

Unexpectedly by our assumption that there is insufficient image attention in Q4, redistributing image weights in that quarter does not make a substantial difference to performance. At first glance, this appears to confirm observations by Shu et al. (2025) and Bi et al. (2025) that image attention in the later decoder layers may be largely redundant. What is problematic about such an account is that boosting image attention by redistributing system and text attention in Q4 benefits performance, suggesting that there may be more subtle deficits in image attention there after all.

Now turning to pairwise redistributions of weights, Figure 4 shows that gains through these interventions are similarly almost exclusively concentrated in Q4. While only results for SugarCrepe are reported, the same trends obtain across our other paired-prompt benchmarks.

A.3 Example prompts

In this appendix, we present example prompts from MME (Table 8), the three splits of POPE (Table 9) and the remaining paired-prompt benchmarks (Table 12) to illustrate the contrast between MME and the adversarial POPE split on the one hand and the rest of the benchmarks on the other, as described in Section 5.

A.4 Yes/no vs other responses

Here, we put forth additional evidence for a causal relationship between the yes-bias and redundant late-layer system attention. This is important in practical terms because it suggests that the yes-bias can be mitigated by targeting one clearly defined underlying factor.

To this end, we adduce two further sets of prompts to demonstrate that Q4 system redistribution most strongly impacts on yes/no responses. Other types of response, on the other hand, do not exhibit the same changes under intervention.

NaturalBench (A/B prompts). The first set of prompts is the A/B subset of the NaturalBench dataset. NaturalBench comprises two types of binary-choice prompt, the first yes/no — covered

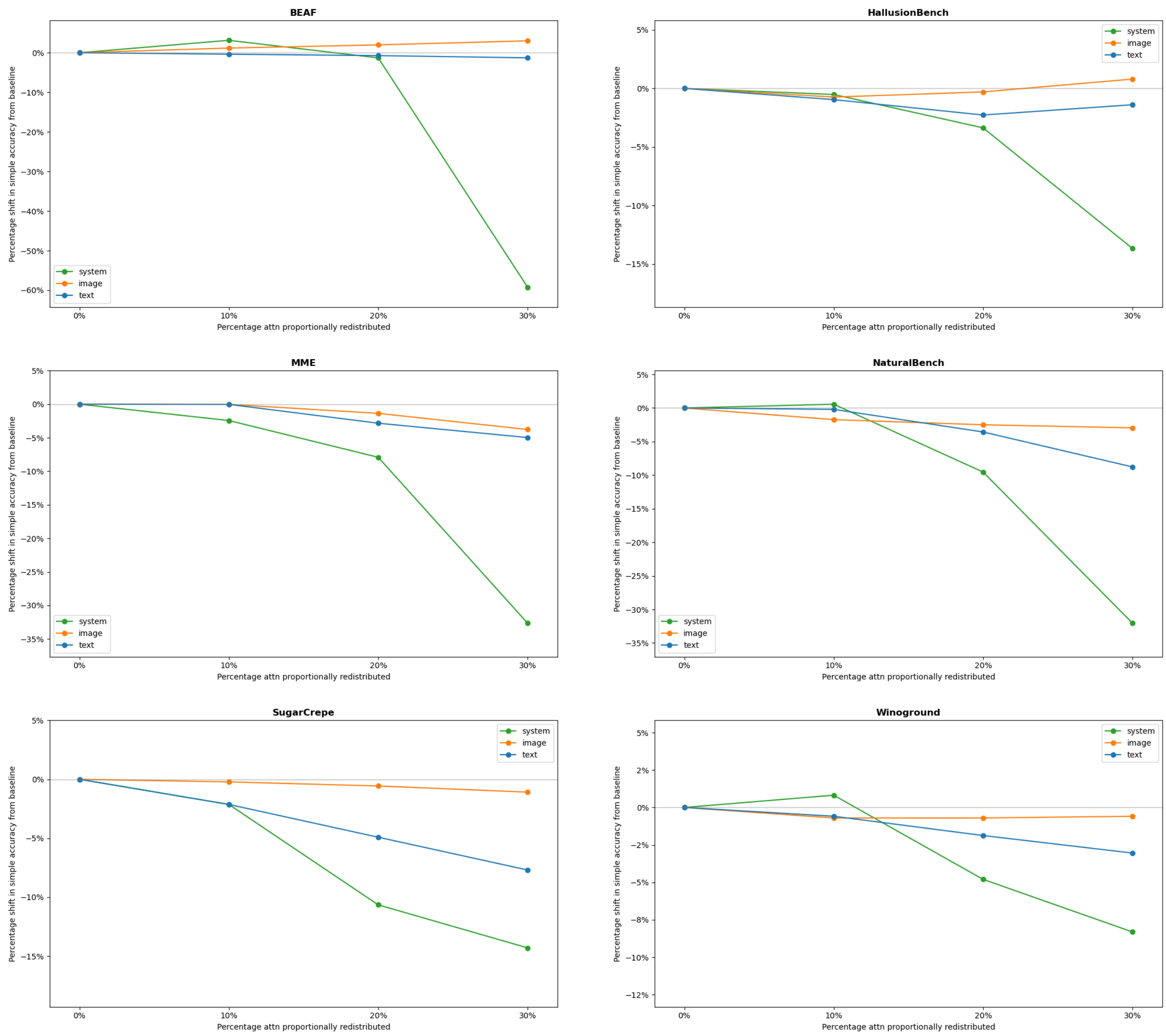


Figure 1: Percentage shift in simple accuracy against percentage of attention weights from each source modality proportionally redistributed to the remaining modalities *across all decoder heads* (i.e. globally), for all six paired-prompt benchmarks in our evaluation suite. We observe that the majority of these interventions are harmful and gains are not consistent across benchmarks. Despite not exhibiting the same types of gains elsewhere, MME patterns with the rest of the benchmarks as far as the harmfulness of global interventions is concerned.

in the main body (see Table 1) — and the second A/B.

Despite both involving binary choice, Q4 system redistribution only exerts a markedly beneficial effect on the former (+60.51% paired accuracy, Table 10). For A/B prompts, on the other hand, the intervention only marginally improves performance (+1.01%). This strongly indicates that the mechanism described in this study, redundant late-layer system attention, is highly specific to yes-bias hallucination.

Whoops!. Our second set of prompts is based on the Whoops! benchmark (Bitton-Guetta et al., 2023) and designed to simultaneously elicit both yes/no and open-ended responses.

The Whoops! benchmark comprises AI-generated images that violate common sense. Accordingly, one of the benchmark's chief purposes is to test whether LLMs are able to detect that. To evaluate to what extent Q4 system redistribution exerts the same effect on yes/no responses on the one hand and autoregressively generated ones on the other, for each image in this benchmark we elicited: i) a yes/ no response ('*Does this image depict a normal scene?*') and ii) an open-ended explanation for the response, mirroring the methodology used by Parcalabescu and Frank (2025) (see Section 6).

This reveals cases where Q4 system redistribution changed the response from 'yes' to 'no' but

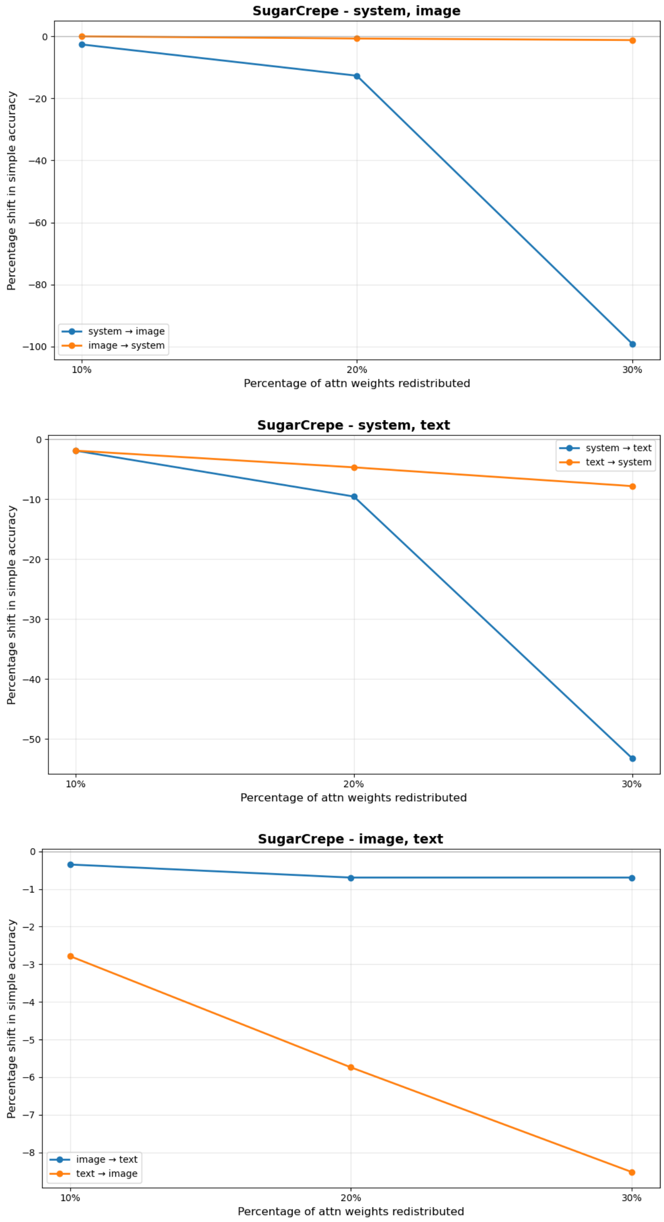


Figure 2: Global (i.e. applied across all decoder heads) graduated pairwise redistributions of attention weights between all possible combinations of source and recipient modalities in SugarCrepe show whole-model interventions to mostly harm performance. These soft-touch interventions redistribute 10%, 20% and 30% of weights from source to recipient at a time. The same trend is reflected across all six benchmarks studied here, but results are not shown owing to space constraints.

not the semantics of the explanation. An example is shown in Table 11, involving an image of strawberries on a pizza. In this case, although Q4 system redistribution changed the response from ‘yes’ to ‘no’, the explanation given for the response in both cases described the image using terms such as ’not typical’ and ’unconventional’.

This buttresses our argument that removing redundant late-layer system attention could potentially serve as a targeted means of mitigating the yes-bias.

A.5 Results for other LLaVA architectures

To examine how system-mediated attention imbalances are affected by differences in training regime and model scale, we perform interventions on two architectures closely related to LLaVA-1.5 7B:

- **LLaVA-NeXT-(Vicuna-)7B**. Results are presented in Tables 13 to 15. All benchmarks used for the main experiments on LLaVA-1.5 7B are included, but a subset of BEAF containing approx. 50% of the original number of

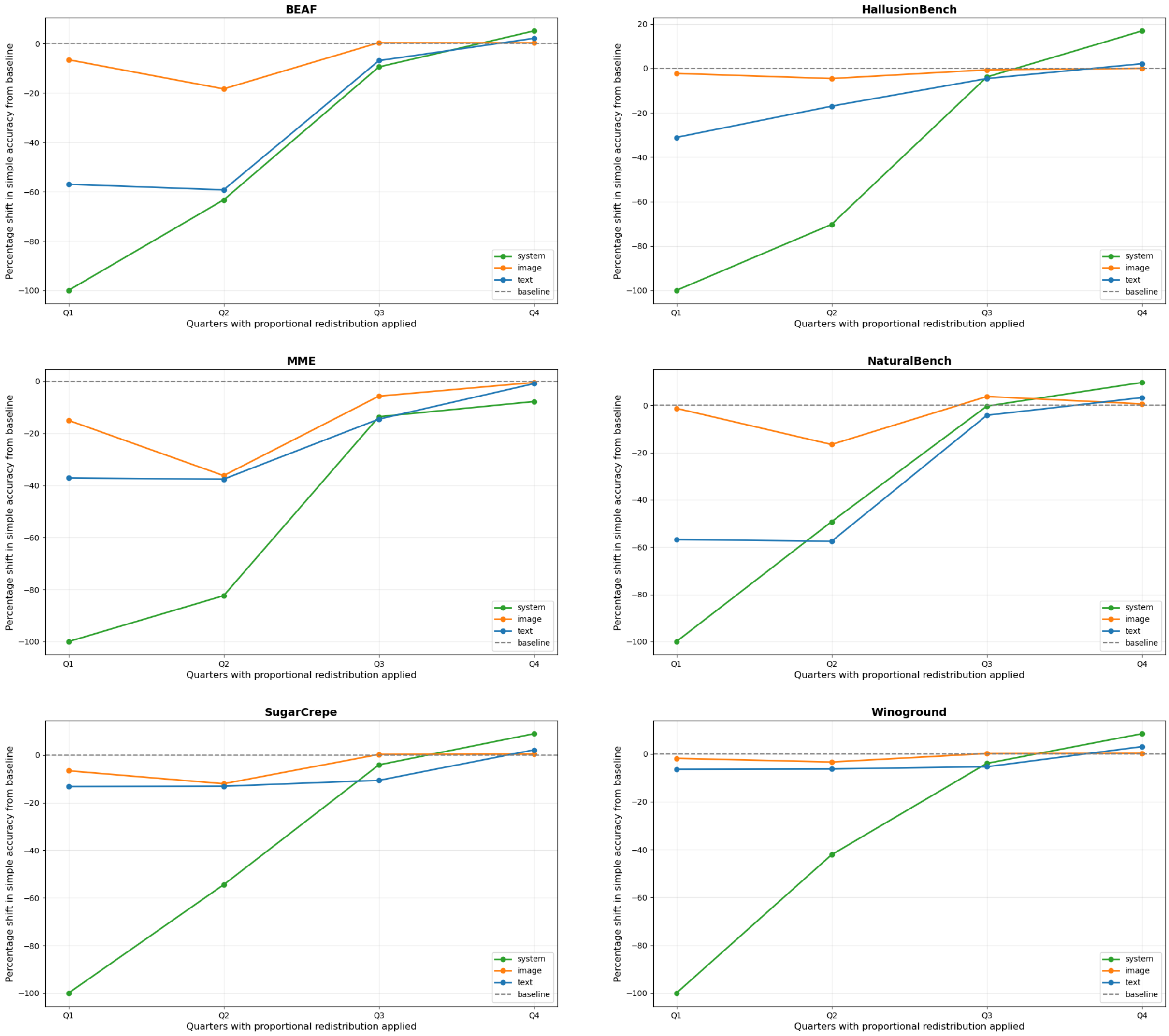


Figure 3: Percentage shift in simple accuracy against percentage of attention weights from each source modality proportionally redistributed to the remaining modalities *per quarter*, for all six paired-prompt benchmarks in our evaluation suite. These plots demonstrate that for five of our six benchmarks, proportional redistribution of system attention in Q4 yields the largest gains. MME is the sole exception to this pattern.

samples (14,024) needed to be used instead of the full benchmark due to memory issues. Follow-up experiments show our baseline architecture LLaVA-1.5 7B to exhibit the analogous shifts in performance with and without intervention on this subset.

- **LLaVA-1.5 13B**. Results are contained in Tables 16 and 17. BEAF is not included because of memory issues analogous to those for LLaVA-NeXT.

These two architectures share LLaVA-1.5 7B's Vicuna decoder but each differ from it in one of two aspects — LLaVA-NeXT(-Vicuna)-7B with substantially improved fine-tuning, and LLaVA-1.5 13B, which differs in scale. This reveals system-mediated attention imbalances to behave consistently across training regimes but vary more with scale.

LLaVA-NeXT 7B's improvements include more effective curation of datasets and also higher input image resolution to enhance its multimodal question-answering and reasoning capabilities (Liu et al., 2024b). Despite these, Tables 13, 14 and 15 show that redistributing Q4 system attention both proportionally and in a pairwise fashion yields qualitatively similar suppression of yes-bias across most benchmarks. These findings indicate on the whole that late-layer system-mediated attention imbalances are not meaningfully reduced by improved image processing or training alone. Subtle differences from LLaVA-1.5 7B, most notably in

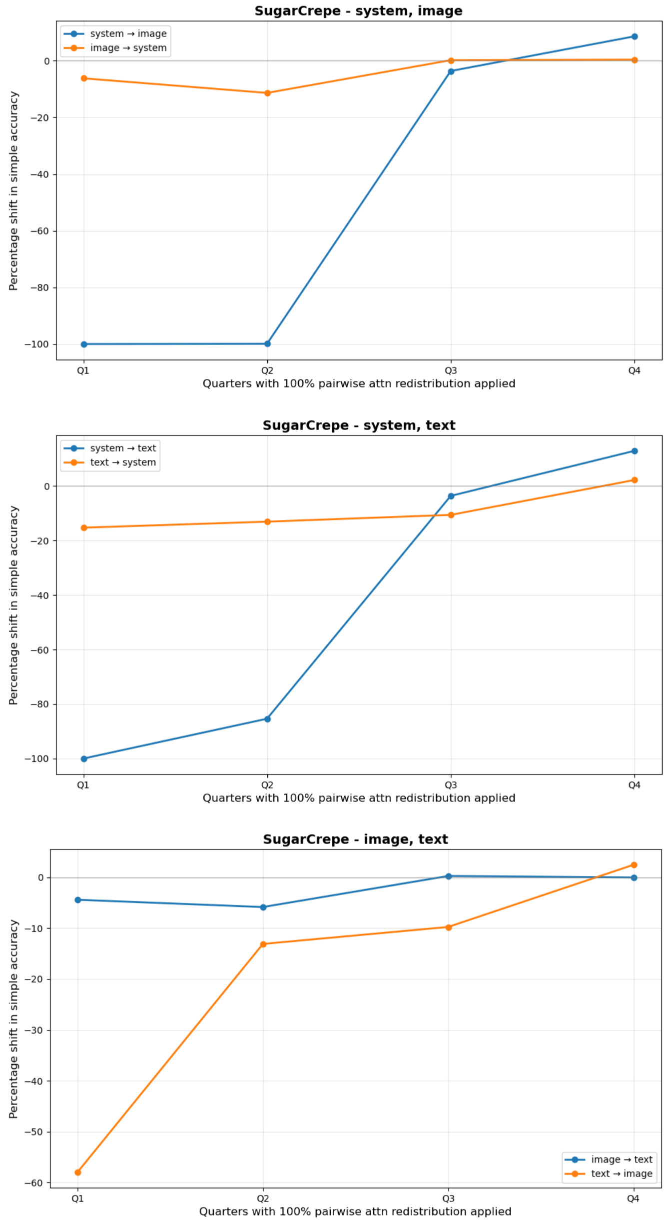


Figure 4: Per-quarter 100% pairwise redistributions of attention weights between all possible combinations of source and recipient modalities in SugarCrepe show whole-model interventions to mostly harm performance. The same trend is reflected across all six benchmarks studied here, but results are not shown owing to space constraints.

system-to-text and system-to-image transfers being better matched in their performance, suggest that these improvements partly address image attention deficits, but without resolving the underlying redundant system attention.

Scale appears to have a larger impact on attention imbalances. Despite lacking LLaVA-NeXT's enhancements, LLaVA-1.5 13B achieves stronger baseline performance in most benchmarks, surpassing even the best interventions on both 7B models (Table 16). In this regime, Q4 system redistribution no longer produces consistent gains and can in many cases worsen the yes-bias. Nonetheless, targeted system-to-image and system-to-text redistributions still yield approx. 9-20% improvements in simple accuracy in HallusionBench and SugarCrepe respectively (Table 17). This indicates that system-mediated attention imbalances remain potentially useful levers against hallucination at larger scales, but assume different characteristics which warrant further investigation.

A.6 Results for Qwen2-VL-7B-Instruct

For the purposes of cross-architectural comparison, the Qwen family of models is especially interesting because Luo et al. (2026) note Qwen2.5-VL-

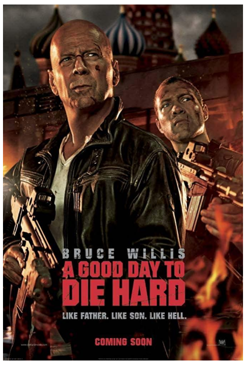
Category	Image	Positive prompt	Negative prompt
Artwork		Does this artwork exist in the form of painting? Please answer yes or no.	Does this artwork exist in the form of sculpture? Please answer yes or no.
Landmark		Is this a photo of De Bataaf, Winterswijk? Please answer yes or no.	Is this a photo of Porte Guillaume-Lion? Please answer yes or no.
Posters		Is this movie originated from the country or region of usa? Please answer yes or no.	Is this movie originated from the country or region of uk? Please answer yes or no.
Scene		Does this image describe a place of stable? Please answer yes or no.	Does this image describe a place of junkyard? Please answer yes or no.

Table 8: Example prompts from MME, which may permit reliance on high-level, global cross-modal representations. Positive prompts refer to those with ground truth ‘yes’ and negative to those with ground truth ‘no’.

Instruct to not be very receptive to training-free interventions of the kind used in this study. We speculate that that may be due to, among various other factors, Qwen’s joint training of the image encoder and text decoder leading to more efficient allocation of attention weights in the text decoder as compared to LLaVA-family architectures without joint training (Team et al., 2024).

Indeed, as shown in Table 18, Q4 system redistribution only benefits performance negligibly in Qwen2-VL-7B-Instruct, never by more than 0.5%. Only simple accuracy is reported.

Nonetheless, as the same table shows, the results broadly agree with our findings for LLaVA-1.5 7B, providing this study’s core arguments with compelling support across model families:

1. Proportional Q4 system redistribution, an operationalisation of the system-mediated hypothesis, is consistently more beneficial than Q4 text-to-image redistribution, which represents the image-centric hypothesis (see Sections 1 and 4.3)

2. Q4 system redistribution brings about a dip in performance for MME but not for the other datasets

The first point crucially demonstrates that our system-mediated account is superior to the prevail-

Category	Image	Positive prompts	Negative prompt
Adversarial		Is there a {chair/ person/ dining table} in the image? Answer only 'yes' or 'no'.	Is there a couch in the image? Answer only 'yes' or 'no'.
Popular			Is there a car in the image? Answer only 'yes' or 'no'.
Random			Is there an airplane in the image? Answer only 'yes' or 'no'.

Table 9: Example prompts from POPE demonstrating different sampling strategies.

NaturalBench	Yes/no prompts		A/B prompts	
	Q4 system redistr (prop)	*No-intervention baseline*	*Q4 system redistr (prop)*	*No-intervention baseline*
Simple acc $^{\Uparrow}$	66.02 (**+9.63**)	60.22	69.33 (+0.57)	68.94
Paired acc $^{\Uparrow}$	34.59 (**+60.51**)	21.55	41.46 (+1.05)	39.49

Table 10: Compared against the no-intervention baseline, Q4 system redistribution improves performance by a much larger margin in yes/no prompts as compared to A/B prompts. This is despite both involving binary choice, and especially visible in the paired accuracy metric.

Prompt	Image	Q4 system redistr (prop)	No-intervention baseline
Does this image show a normal scene? Explain.		**No**, this image does not show a normal scene. The image shows a pizza with a unique and *unusual* topping: a large, fresh, and freshly-picked strawberry. This is *not a typical topping* for a pizza, as it is more commonly topped with various types of cheese, meats, and other typical toppings. The image is a close-up of the pizza, which is placed on a table, and the focus is on the unique and *unusual* topping.	**Yes**, this image shows a normal scene. It features a pizza with strawberries as toppings, placed on a wooden table. The pizza is cut into slices, and the strawberries are scattered on top of the cheese. This combination of ingredients is *not typical* for a traditional pizza, but it is an example of a creative and unique pizza topping. The image captures a moment where someone has prepared and served this *unconventional* pizza, making it a part of a regular meal or gathering.

Table 11: Example prompt from the Whoops! benchmark illustrating how Q4 system redistribution causes LLaVA-1.5 7B to respond 'no' instead of 'yes' to the same prompt, while the explanation for the response remains semantically identical under both settings.

ing image-centric alternative for both the LLaVA-family models tested and Qwen2-VL-7B-Instruct.

The second point lends credence to our hypothesis in Section 5 that the presence of late-layer system attention encourages the use of coarse-grained aggregated input representations, benefiting tasks such as MME. Performance in this benchmark is therefore harmed when this attention is redistributed, albeit by a much smaller margin where Qwen2-VL-Instruct is concerned, potentially for the reasons mentioned at the start of this appendix.

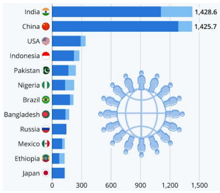
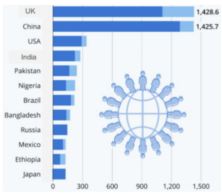
Benchmark	Category	Image	Positive prompt	Negative prompt
HallusionBench	Chart		According to the chart, does India have the second largest population in the world?	According to the chart, does UK have the largest population in the world, followed by China and USA?
	Chart		According to the chart, does UK have the largest population in the world, followed by China and USA?	According to the chart, does India have the second largest population in the world?
SugarCrepe	Add attribute		Two zebras are battling each other on hind legs.	Two striped-and-spotted zebras are battling each other on hind legs.
	Add object		A nicely decorated living room and dining area.	A nicely decorated living room with a bookshelf and dining area.

Table 12: Examples from HallusionBench and SugarCrepe requiring fine-grained multimodal discrimination. Here, the correct answer hinges on minimal differences in user prompts or the image input.

Benchmark (% prompts with ground-truth yes)	Metric	Q4 system redistr (prop)	No-intervention baseline
BEAF (sample) (33.71% yes)	Simple acc$^{\Uparrow}$	**90.67** (+0.82)	89.93
	Paired acc$^{\Uparrow}$	**87.72** (+1.07)	86.79
	Yes-rate$^{\Delta\downarrow}$	31.50 (–6.55)	**35.34** (+4.83)
HallusionBench (42.17% yes)	Simple acc$^{\Uparrow}$	**52.89** (+11.04)	47.63
	Paired acc$^{\Uparrow}$	**16.74** (+35.85)	12.33
	Yes-rate$^{\Delta\downarrow}$	**61.51** (+45.87)	87.17 (+106.71)
MME (50.00% yes)	Simple acc$^{\Uparrow}$	40.61 (–46.36)	**75.70**
	Paired acc$^{\Uparrow}$	0.77 (–95.00)	**15.38**
	Yes-rate$^{\Delta\downarrow}$	20.56 (–58.89)	**56.44** (+12.89)
NaturalBench (50.00% yes)	Simple acc$^{\Uparrow}$	**67.81** (+4.05)	65.17
	Paired acc$^{\Uparrow}$	**37.86** (+19.22)	31.76
	Yes-rate$^{\Delta\downarrow}$	**65.74** (+31.48)	77.17 (+54.34)
SugarCrepe (50.00% yes)	Simple acc$^{\Uparrow}$	**64.56** (+4.35)	61.87
	Paired acc$^{\Uparrow}$	**29.23** (+22.58)	23.85
	Yes-rate$^{\Delta\downarrow}$	**83.24** (+66.48)	86.81 (+73.63)
Winoground (50.00% yes)	Simple acc$^{\Uparrow}$	**58.81** (+6.21)	55.38
	Paired acc$^{\Uparrow}$	**21.50** (+70.30)	12.63
	Yes-rate$^{\Delta\downarrow}$	**76.81** (+53.63)	91.25 (+82.50)

Table 13: Changes in simple accuracy, paired accuracy and yes-rate across six paired-prompt benchmarks for LLaVA-NeXT-Vicuna-7B, with and without proportional Q4 system redistribution. Like for LLaVA-1.5 7B, the intervention yields gains in all benchmarks but MME. Best results for each metric are in bold. For simple and paired accuracy, percentage differences relative to the no-intervention baseline are shown in brackets. The differences for yes-rate are with reference to the actual proportion of prompts with ground-truth 'yes', shown bracketed in the Benchmark column — the best yes-rate is the one that deviates least from that ground-truth value.

Source	Recipient											
	HallusionBench			NaturalBench (yes/no)			SugarCrepe			Winoground		
	System	*Image*	*Text*	*System*	*Image*	*Text*	*System*	*Image*	*Text*	*System*	*Image*	*Text*
System	47.63	58.15	**58.36**	65.17	**69.93**	69.84	61.87	67.58	**67.69**	55.36	**61.12**	61.00
Image	47.53	47.63	47.42	65.17	65.17	65.21	61.98	61.87	61.87	55.19	55.36	55.31
Text	50.37	50.58	47.63	67.78	66.67	65.17	64.23	64.29	61.87	57.88	57.50	55.36

Table 14: Changes in simple accuracy in HallusionBench, NaturalBench (yes/no), SugarCrepe, and Winoground under pairwise attention-weight transfers between all combinations of source and recipient modalities in Q4 (layers 25-32) of LLaVA-NeXT-Vicuna 7B. Although transfers with system as the source modality consistently yield the biggest improvements in performance, system-to-text and system-to-image transfers perform more similarly to each other than in LLaVA-1.5 7B.

Source	Recipient					
	BEAF			MME		
	System	*Image*	*Text*	*System*	*Image*	*Text*
System	89.93	89.28	89.33	75.70	61.58	64.20
Image	89.92	89.93	89.97	75.74	75.70	75.82
Text	90.68	**90.84**	89.93	**76.16**	75.06	75.70

Table 15: Changes in simple accuracy in BEAF and MME under pairwise attention-weight transfers between all combinations of source and recipient modalities in Q4 (layers 25-32) of LLaVA-NeXT-Vicuna 7B. Transfers with system as the source modality consistently underperform the no-intervention baseline (shaded).

Benchmark (% prompts with ground-truth yes)	Metric	Q4 system redistr (prop)	No-intervention baseline
HallusionBench (42.17% yes)	Simple acc$^{\Uparrow}$	**51.31** (+0.83)	50.89
	Paired acc$^{\Uparrow}$	15.58 (–1.47)	**15.81**
	Yes-rate$^{\Delta\downarrow}$	**65.62** (+55.60)	67.93 (+61.08)
MME (50.00% yes)	Simple acc$^{\Uparrow}$	79.32 (–0.05)	**79.36**
	Paired acc$^{\Uparrow}$	27.69 (–14.29)	**32.31**
	Yes-rate$^{\Delta\downarrow}$	54.93 (+9.86)	**47.89** (–4.21)
NaturalBench (50.00% yes)	Simple acc$^{\Uparrow}$	68.36 (–1.59)	**69.46**
	Paired acc$^{\Uparrow}$	39.05 (–4.57)	**40.92**
	Yes-rate$^{\Delta\downarrow}$	65.83 (+31.66)	**61.25** (+22.50)
SugarCrepe (50.00% yes)	Simple acc$^{\Uparrow}$	**68.96** (+0.88)	68.35
	Paired acc$^{\Uparrow}$	**38.46** (+3.24)	37.25
	Yes-rate$^{\Delta\downarrow}$	**76.54** (+53.08)	77.36 (+54.73)
Winoground (50.00% yes)	Simple acc$^{\Uparrow}$	58.56 (–2.19)	**59.88**
	Paired acc$^{\Uparrow}$	21.50 (–12.69)	**24.63**
	Yes-rate$^{\Delta\downarrow}$	79.31 (+58.63)	**74.38** (+48.75)

Table 16: Changes in simple accuracy, paired accuracy and yes-rate across five paired-prompt benchmarks for LLaVA-1.5 13B, with and without proportional Q4 system redistribution. Compared to LLaVA-1.5 7B, the intervention produces very limited gains and increases the model's yes-rate for three benchmarks (MME, NaturalBench, Winoground). Best results for each metric are in bold. For simple and paired accuracy, percentage differences relative to the no-intervention baseline are shown in brackets. The differences for yes-rate are with reference to the actual proportion of prompts with ground-truth 'yes', shown bracketed in the Benchmark column — the best yes-rate is the one that deviates least from that ground-truth value.

Source	Recipient					
	HallusionBench			SugarCrepe		
	System	*Image*	*Text*	*System*	*Image*	*Text*
System	50.89	**61.30**	60.99	68.35	74.40	**74.62**
Image	51.63	50.89	51.21	68.46	68.35	68.35
Text	49.32	50.68	50.89	68.35	68.79	68.35

Table 17: Changes in simple accuracy in HallusionBench and SugarCrepe under pairwise attention-weight transfers between all combinations of source and recipient modalities in Q4 (layers 25-32) of LLaVA-1.5 13B. Transfers with system as the source modality consistently underperform the no-intervention baseline (shaded).

Benchmark	Q4 system redistr (prop)	Q4 text-to-image redistr	No-intervention baseline
BEAF	**89.16** (+0.01)	86.08 (-3.44)	89.15
HallusionBench	**69.36** (+0.37)	68.59 (-0.74)	69.10
MME	86.41 (-0.17)	86.22 (-0.39)	**86.56**
NaturalBench	**74.90** (+0.20)	74.67 (-0.11)	74.75
SugarCrepe	**77.31** (+0.29)	72.69 (-5.71)	77.09
Winoground	**75.00** (0.00)	73.83 (-1.56)	**75.00**

Table 18: Simple accuracy of Qwen2-VL-7B-Instruct in six benchmarks under two different intervention settings, alongside the no-intervention baseline. Interventions here impact much less considerably on performance as compared to the LLaVA-family models. Nevertheless, in accordance with this study's system-mediated hypothesis, proportionally redistributing Q4 system attention (Q4 system redistr (prop)) is consistently less harmful than Q4 text-to-image transfer, demonstrating the former's superiority as an empirical account. This is despite the performance differences arising under intervention being much smaller than in the LLaVA-family models shown previously. Best results per benchmark bolded.