IS OBJECT HALLUCINATION IN LVLMS OVERESTIMATED? REVISITING EVALUATION, MODELS, AND MITIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (LVLMs) have become remarkably capable of describing images at an impressive level of detail. However, object hallucinations have been a long-standing problem and a wide variety of mitigation strategies have been proposed and evaluated. Interestingly, most methods including very recent work rely on one particular metric—the CHAIR metric on COCO—and often report results primarily for LLaVA1.5-7B model. In this work, we take a step back by carefully designing an evaluation to understand if the CHAIR metric on COCO is actually measuring object hallucinations properly and how significant they are for modern models. In particular, we evaluate 10 models (7B to 12B) from 5 different families, and find that measured hallucination rates significantly fall (16 to 38%) from the original CHAIR value for every model except LLaVA-1.5-7B. Since major mitigation methods, however, are still developed and tested on LLaVA-1.5 with this metric, we revisit them on current models and find that none of them reduce hallucination consistently across models on COCO: a large gain on LLaVA-1.5 does not carry over or results in modest gains on newer models. We also perform finer-grained analyses that go beyond CHAIR. Finally, to evaluate models on more challenging images, we repurpose the existing BEAF dataset and introduce the persistence metric to assess models and hallucination mitigation methods under counterfactual object removal. Taken together, these metrics help better understand the prevalence of hallucinations in modern LVLMs and the effectiveness of mitigation methods.

1 INTRODUCTION

Large vision-language models (LVLMs) are becoming increasingly capable of visual understanding, generating rich and detailed descriptions of images. However, they suffer from object hallucinations i.e., mentioning objects that are not there in the image. Many methods have been proposed to reduce object hallucinations (Leng et al., 2024; Huang et al., 2024; Liu et al., 2024b), and the progress is predominantly measured by the CHAIR metric (Rohrbach et al., 2018) on the COCO benchmark (Lin et al., 2015) for image description generation and by simple accuracy for visual question answering. Most mitigation methods are developed and tested in this setting, usually on LLaVA-1.5-7B (Liu et al., 2024a).

This setting has not kept pace with the models, which raises three concerns. First, CHAIR has known issues (Kaul et al., 2025; Ben-Kish et al., 2024; Petryk et al., 2024): it finds object mentions by matching words against a fixed synonym list for the 80 COCO classes, so it can count negated, hypothetical or ambiguous mentions as hallucinations, e.g., "no people" is read as *person*. Current LVLMs write long and detailed descriptions, which makes such errors more likely. Despite this, CHAIR continues to be widely used as a benchmark for evaluating hallucination mitigation methods. Alternatives such as THRONE (Kaul et al., 2025) and ALOHa (Petryk et al., 2024) have been proposed, but they are rarely used in the mitigation literature, and do not allow for direct comparison with CHAIR. Second, LLaVA-1.5-7B dates from 2023 and is several generations behind modern LVLMs. Despite this, it remains the most widely used benchmark for evaluating the effectiveness of proposed hallucination mitigation methods. Third, if the metric overestimates hallucination and the

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.

model used for development is outdated, it is unclear whether the reported gains of mitigation methods hold for current models. We therefore ask: (i) if CHAIR measure object hallucination correctly for current LVLMs, (ii) if existing mitigation methods still help on these models, and (iii) to what extent do current LVLMs still hallucinate on harder benchmarks, and if mitigation methods continue to help.

For the first question, we keep the CHAIR pipeline and propose a controlled setup, judge-CHAIR, that changes only how object mentions are found: an LLM reads each caption without seeing the image, and returns the objects that the caption describes as present (Section 3.2). This allows obtaining a more accurate estimate of the hallucination rate. Across modern LVLMs, we find that the hallucination rates are significantly lower than expected, while being similar as reported for LLaVA-1.5-7B. For the second question, we revisit three mitigation methods that use diverse mechanisms, and find that none of them reduce hallucination consistently across models. We also propose more fine-grained metrics to gain an understanding beyond what CHAIR provides. For the third question, we repurpose BEAF (Moon et al., 2024), a benchmark that consists of paired images that differ in exactly one object, as a controlled, harder setup for measuring hallucination. We find that mitigation methods help lower than on COCO, and they rank differently for descriptions and for yes/no questions.

Our main contributions are:

- An evaluation of ten LVLMs showing that rule-based CHAIR overestimates object hallucination of current models on COCO by more than half, while it changes little for LLaVA-1.5-7B (Section 3.3).
- A re-evaluation of three widely used mitigation methods on seven models, showing that their gains on LLaVA-1.5 does not always carry over to current models (Section 3.4).
- A repurposed BEAF setup for open-ended descriptions, with a metric, persistence, that measures whether a model keeps naming a removed object. With it, we show that current models hallucinate even more on these harder images and that mitigation methods help even less there (Section 4).

Together, these results suggest that conventional CHAIR metric is an unreliable measure of object hallucination and mitigation methods should be evaluated on current models and on harder settings than standard COCO images.

2 Related work

2.1 Object hallucination in LVLMs

Object hallucination, where a model mentions an object that is not in the image, was first studied in image captioning (Rohrbach et al., 2018) and remains one of the most studied failures of LVLMs. Co-occurrence statistics in the training data (Rohrbach et al., 2018; Biten et al., 2022; Zhou et al., 2024), strong language priors that override the image (Leng et al., 2024), and too little attention to image tokens (Liu et al., 2024b) are some of the widely believed hypothesis for object hallucination. However, HaloProbe (Zohrabi et al., 2026) finds that hallucinated objects tend to appear later in a description, and that once token position is taken into account, hallucinated tokens often get as much image attention as correct ones, or more. Most benchmarks test hallucination with yes/no questions, such as POPE (Li et al., 2023), AMBER (Wang et al., 2024) and HallusionBench (Guan et al., 2024), and then there are others which test it in free form caption generation (Rohrbach et al., 2018). Here, we are evaluating on both object hallucination in free form generation and visual question answering.

2.2 Mitigation techniques

Mitigation methods for object hallucination either use training based methods or test time interventions. Training-based methods use fine-grained human feedback (Yu et al., 2024), preference optimization (Yang et al., 2026), or control over how much the model imagines beyond the image (Zhai et al., 2024). LURE (Zhou et al., 2024) instead trains a separate model that revises the caption after it is written. On the other hand test time interventions are cheaper and are increasingly

common. VCD (Leng et al., 2024) assume hallucination happens due to co-occurrence pattern and language prior and contrasts the output with the output from a noisy image. OPERA (Huang et al., 2024) penalizes over-trust in a few summary tokens during beam search, which is computationally very expensive to deploy. PAI (Liu et al., 2024b) increases attention to image tokens. Several recent methods find specific attention heads linked to hallucination and change them (Yang et al., 2025; Sarkar et al., 2025). LocoRE (Zhang et al., 2026) strengthens attention to recently generated text to alleviate contextual forgetting. Almost all of these methods were developed on LLaVA-1.5 and are scored with CHAIR on COCO and with POPE. Only a few report results on current models. We re-evaluate VCD, PAI and LocoRE on seven models from four model families.

2.3 Evaluation metrics for object hallucination

Object hallucination in free form generation is predominantly evaluated using CHAIR (Rohrbach et al., 2018), which performs rigid word matching to find object mentions with a fixed COCO synonym list and checks them against COCO labels. Several works have flagged issues with it. For example, it is known to handle long, detailed descriptions poorly (Kaul et al., 2025), and flags negations as false positives (Petryk et al., 2024). Several works therefore use a language model to help decide whether a specific object is mentioned in the caption. Some work with the limited number of COCO classes (Yu et al., 2024; Kaul et al., 2025), and others move to an open vocabulary (Petryk et al., 2024; Ben-Kish et al., 2024). Few others check attributes, relations or all atomic facts (Wang et al., 2024; Qiu et al., 2024; Jing et al., 2024). These works change multiple things at once to the existing CHAIR pipeline, because of which it is difficult to compare them directly with published CHAIR numbers. On the other hand, we make a simple modification to the existing CHAIR metric to extract objects faithfully. Any change in the score then comes from the extraction alone, and the numbers stay comparable with prior work. We then ask whether conclusions about mitigation methods still hold for current LVLMs.

3 Revisiting object hallucination in LVLMs

Despite known issues (Ben-Kish et al., 2024; Kaul et al., 2025; Petryk et al., 2024), CHAIR (Rohrbach et al., 2018) on LLaVA-1.5 (Liu et al., 2024a) remains the predominant setup to evaluate hallucinations in open-ended generation and developing mitigation methods. In this section, we revisit object hallucinations in modern LVLMs. In Section 3.1, we first describe CHAIR (henceforth referred to as rule-CHAIR) and enumerate problems with it. Then, in Section 3.2, we propose judge-CHAIR to mitigate these issues. In Section 3.3, we compare these metrics and evaluate prevalence of object hallucination in modern LVLMs. Next, in Section 3.4, we revisit mitigation methods in this setup to evaluate if they remain effective. Finally, in Section 3.5, we perform additional analyses beyond CHAIR scores that provide insights in the workings of the models and mitigation methods.

3.1 Rule-CHAIR

The CHAIR metric Rohrbach et al. (2018) (henceforth referred to as rule-CHAIR) consists of the following two scores:

$$\text{CHAIR}_i = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|}, \qquad \text{CHAIR}_s = \frac{|\{\text{sentences with hallucinated object}\}|}{|\{\text{all sentences}\}|} \tag{1}$$

where, CHAIR_i is an object level and CHAIR_s is a sentence level metric, and lower is better for both. Typically applied on COCO (Lin et al., 2015), CHAIR extracts mentions of COCO object classes from the captions and matches them against the COCO ground truth, by first lemmatising the words in the caption and then performing a rule based matching with a fixed synonym list for the COCO categories. However, as shown in Figure 1, this results in a significant number of false positives, as the rule-based approach does not take into account context. For example, COCO annotations refer to ‘orange’ the the fruit, but rule-CHAIR also flags ‘orange’ the colour (Figure 1, top left).



Figure 1: **Failure modes of rule-CHAIR.** We report various failure modes caused by the keyword-based matching used by rule-CHAIR. For example, in the bottom left, a 'toothbrush holder' is flagged as a hallucination for 'toothbrush'. This causes hallucination rates to be significantly overestimated.

3.2 Proposed metric: Judge-CHAIR

Since the issue of the rule-CHAIR largely stems from the erroneous object extraction procedure, we propose to extract object mentions using an LLM as a judge that can take context and nuances into account. Crucially, for each sample, the judge is only shown the caption from the model and the list of COCO ground truth annotations, and is not shown the image, which prevents visual errors of the judge from affecting the evaluation. In our experiments, we use a `gpt-5.6-luna` model as the judge, and Section A.1 and Table 5 shows the prompt and an example respectively.

Recent models also often hedge when they are not sure what an object is, e.g., "a wooden surface, possibly a table or chair". The judge marks such mentions as *hedged*, and we leave them out of judge-CHAIR, so they count neither as real nor as hallucinated. Counting them can flag more hallucinations and lower precision; we discuss this in Section 3.5.

Prior works such as THRONE (Kaul et al., 2025) and ALOHA (Petryk et al., 2024) also use judge-based approaches to evaluate hallucinations in place of CHAIR. However, these approaches propose an alternative metric, while judge-CHAIR is intended to be a drop-in replacement for rule-CHAIR to allow direct comparison in a controlled manner.

Human Validation. To validate judge-CHAIR, we perform a human evaluation using a single annotator using 50 captions each from LLaVA-1.5-7B (Liu et al., 2024a) and Qwen2.5-VL-7B (Bai et al., 2025b). The annotator is provided a protocol (Section B.1) that asks for COCO classes to be labeled, including hedged mentions. We report results in Table 1, and find that judge-CHAIR agrees well with the human on both models. Interestingly, rule-CHAIR also shows high agreement on LLaVA-1.5, likely since its short captions do not flag many false positives, which highlights the risk of using the combination for understanding the true rate of hallucination. On Qwen2.5-VL, however, the captions are richer and rule-CHAIR has significantly lower agreement as it flags more false positives.

Table 1: **Agreement with a human annotator** on the CHAIR_s verdict, 50 COCO captions per model. In %.

Model	rule-CHAIR	judge-CHAIR
LLaVA-1.5-7B	82.0	**86.0**
Qwen2.5-VL-7B	54.0	**92.0**

3.3 Revisiting hallucinations in modern LVLMs

Next, we leverage rule-CHAIR to obtain a better estimate on the rate of object hallucinations in LVLMs. To understand how significant this remains in modern LVLMs, we sweep over several models and families.

Models and evaluation setup. We test nine recent open source LVLMs apart from LLaVA-1.5-7B (Oct 2023; Liu et al., 2024a) in the 7-12 billion parameter scale across five families:

LLaVA-OneVision-2-8B (Apr 2026; An et al., 2026), Qwen2.5-VL-7B-Instruct (Jan 2025; Bai et al., 2025b), Qwen3-VL-8B (Oct 2025; Bai et al., 2025a), Qwen3.5-9B (Mar 2026; Qwen Team, 2026), InternVL3-8B (Apr 2025; Zhu et al., 2025), InternVL3.5-8B (Aug 2025; Wang et al., 2025), Molmo2-8B (Dec 2025; Clark et al., 2026), Molmo2-O-7B (Dec 2025; Clark et al., 2026) and Gemma-4-12B (Jun 2026; Gemma Team, 2026). We use greedy decoding to keep runs reproducible and use a maximum token length of 1024 for each model.

Observations. We report results in Table 2. We find that rule-CHAIR largely overestimates object hallucination in current LVLMs, and **the actual object hallucination rates estimated via judge-CHAIR are much lower**. For the nine recent models, CHAIR_s decreases significantly by 16.1–38.3 points and CHAIR_i by 3.4–7.1 points. Recall changes by at most 2.7 points, whereas precision, in contrast, increases by 7.7–18.3 points. As discussed in Section 3.1 and Figure 1, object extraction from rule-CHAIR flags a lot of false positives, resulting in significantly lower reported precision. These false positives concentrate on a few classes whose synonym lists contain common words with another meaning, e.g. 'orange' is read as the fruit even when it is a colour (Appendix A.3). We also observe that the largest gaps occur for Qwen3-VL and Qwen3.5, which generate the longest descriptions and hence amplify errors by rule-CHAIR. With judge-CHAIR, the nine recent models show significantly lower hallucination rates, in the range of 12.2–18.2 CHAIR_s and 3.1–5.1 CHAIR_i. Interestingly, LLaVA-1.5-7B shows similar values across both the fixed and the flawed metric, which further highlights the risk of evaluating primarily on it, as most mitigation methods do, which we discuss further next.

Table 2: **Rule-CHAIR and judge-CHAIR on vanilla on the 500-image COCO val2014 eval set.** Δ is judge-CHAIR minus rule-CHAIR. Length is the average caption length in words. All values in %, except Length. We see that judge-CHAIR removes a lot of false positives by rule-CHAIR and the actual hallucination rates drop significantly.

	$\text{CHAIR}_s \downarrow$			$\text{CHAIR}_i \downarrow$			Recall ↑			Precision ↑			F1 ↑			
Model	rule	judge	Δ	rule	judge	Δ	rule	judge	Δ	rule	judge	Δ	rule	judge	Δ	Length
Qwen3.5-9B	55.6	**17.3**	−38.3	10.2	**3.1**	−7.1	77.1	74.4	−2.7	74.0	92.3	+18.3	75.5	82.4	+6.9	303
Qwen3-VL-8B	55.2	**18.2**	−37.0	10.0	**3.5**	−6.5	76.2	74.2	−2.0	74.7	91.8	+17.1	75.4	82.1	+6.7	336
Qwen2.5-VL-7B	36.3	**12.7**	−23.6	9.8	**3.6**	−6.2	69.8	68.7	−1.1	82.5	93.9	+11.4	75.6	79.4	+3.8	164
Molmo2-O-7B	33.3	**13.5**	−19.8	7.9	**3.3**	−4.6	65.7	65.3	−0.4	83.3	93.2	+9.9	73.5	76.8	+3.3	171
Molmo2-8B	35.8	**16.3**	−19.5	8.7	**3.8**	−4.9	70.6	71.0	+0.4	82.7	92.5	+9.8	76.2	80.4	+4.2	151
InternVL3.5-8B	37.3	**18.1**	−19.2	10.3	**5.1**	−5.2	69.6	70.1	+0.5	81.3	91.1	+9.8	75.0	79.3	+4.3	133
InternVL3-8B	32.3	**13.3**	−19.0	8.5	**3.4**	−5.1	70.5	71.3	+0.8	84.8	93.8	+9.0	77.0	81.0	+4.0	115
LLaVA-OV-2-8B	29.7	**12.2**	−17.5	6.7	**3.3**	−3.4	65.9	66.9	+1.0	85.3	93.6	+8.3	74.4	78.0	+3.6	175
Gemma-4-12B	30.9	**14.8**	−16.1	7.8	**3.8**	−4.0	63.8	64.7	+0.9	84.6	92.3	+7.7	72.8	76.1	+3.3	143
LLaVA-1.5-7B	53.6	**49.0**	−4.6	13.5	**13.2**	−0.3	78.6	77.7	−0.9	74.0	75.4	+1.4	76.3	76.5	+0.2	91

3.4 Revisiting mitigation methods

As we discussed in Section 1, mitigation methods primarily use the rule-CHAIR metric and are built using LLaVA1.5, which can be misleading as discussed in Section 3.3. Therefore, in this section we focus on better understanding the effectiveness of mitigation methods and their impact on modern LVLMs. In particular we revisit the VCD (Leng et al., 2024), PAI (Liu et al., 2024b), and the LocoRE (Zhang et al., 2026) method. VCD and PAI are two popular mitigation methods, while LocoRE is a recent approahc with strong reported results. The three methods rest on different assumptions about where hallucinations come from. VCD and PAI both blame language priors, but in different ways. VCD assumes that a hallucinated token is one that the model would still predict if the image were degraded, so it contrasts the output against a noised copy of the image. PAI assumes that the model pays too little attention to the image, so it raises the attention to image tokens inside the model. LocoRE takes the opposite view to PAI: it links hallucinations to the model losing track of what it has just written, so it strengthens attention to recently generated text tokens rather than to the image. Appendix D provides more details on each method.

Models and evaluation setup. We use the same evaluation setup in Section 3.3. For all mitigation methods, we use the published hyperparameters for LLaVA-1.5 where they are specified. All unspecified settings, including LocoRE's window size for LLaVA-1.5 and the method settings for

the newer models—are chosen on 100 held-out COCO images, none of which overlap with the 500-image evaluation set. Appendix D.1 gives the candidate settings, the selection rule and the chosen values.

Table 3: **Mitigation methods on COCO.** We use the judge-CHAIR metric. *van.*: vanilla; *Loc.*: LocoRE. Best value per row and metric is in bold. Methods improving over vanilla are underlined. Length is the average caption length in words; all other values in %. Full results with recall, precision and the differences to the original model are in Section E. We find that VCD does not improve hallucination on any model, PAI works only on the LLaVA1.5 model it was built for, LocoRE works more consistently on modern models.

	$\text{CHAIR}_s \downarrow$				$\text{CHAIR}_i \downarrow$				F1 ↑				Length			
Model	van.	VCD	PAI	Loc.	van.	VCD	PAI	Loc.	van.	VCD	PAI	Loc.	van.	VCD	PAI	Loc.
Qwen3-VL-8B	18.2	20.5	17.7	**13.3**	3.5	3.8	3.4	**2.7**	82.1	83.0	83.3	81.6	336	335	304	332
Qwen2.5-VL-7B	12.7	14.9	14.3	**10.8**	3.6	3.7	4.0	**2.6**	79.4	80.5	77.5	79.1	164	175	136	159
Molmo2-O-7B	13.5	16.3	14.1	**11.4**	3.3	3.8	3.4	**3.0**	76.8	76.8	75.5	75.9	171	157	140	165
Molmo2-8B	16.3	18.5	16.1	**12.9**	3.8	4.5	3.8	**3.1**	80.4	82.2	81.4	80.0	151	148	152	164
InternVL3.5-8B	18.1	22.3	18.9	**15.9**	5.1	6.3	4.9	**4.5**	79.3	79.3	79.7	77.0	133	129	125	143
InternVL3-8B	13.3	16.7	13.7	**12.2**	3.4	4.6	3.5	**3.3**	81.0	80.4	82.6	79.4	115	116	113	113
LLaVA-1.5-7B	49.0	50.8	**30.8**	49.4	13.2	15.4	**8.0**	13.8	76.5	75.6	78.4	76.6	91	91	94	90

Observations. Table 3 shows that **the three mitigation methods behave differently, and that methods built on older models do not necessarily work on current LVLMs**. We call a change in CHAIR_s significant when the exact McNemar teston the captions the method changes gives $p < 0.05$. VCD does not reduce hallucination on any of the models, not even on LLaVA-1.5. It was originally shown to work on older LVLMs for VQA, but in our evaluation the benefit does not carry over to free-form generation: it increases CHAIR_s on all seven models by 1.8–4.2 points, and the increase is significant for InternVL3.5 ($+4.2$) and InternVL3 ($+3.4$). PAI has a large effect on LLaVA-1.5, the model it was originally built on, reducing CHAIR_s from 49.0 to 30.8 ($p < 0.001$) while increasing F1 by 1.9 points. On the six newer models, however, its change in CHAIR_s is only -0.5 to $+1.6$ points, and none is significant. LocoRE is directionally more consistent for descriptions. The reduction is significant only on Qwen3-VL (-4.9 points) and Molmo2-8B (-3.4); on the other four it is 1.1–2.2 points. (Zhang et al., 2026) originally report a large reduction on LLaVA1.5-7B, which we do not observe, which could be due to choosing the hyper-parameters on the final evaluation set (Table 3 and table 2 in (Zhang et al., 2026)). We get a similar reduction on CHAIR_s on the set we choose the hyperparameters, however the same gains do not translate to the final evaluations for LLaVA1.5.

3.5 Finer-grained evaluation beyond CHAIR

CHAIR scores only shows the net change in hallucination rate on aggregate. However, this can occur by a method fixing some hallucinated captions and breaking non-hallucinated ones (Appendix G), and mitigation methods can also change how the model phrases what it sees. Here we look at both.

Break share. First, we evaluate the change in hallucination behaviour at a per-caption level. Let f be the *fix rate*, the share of vanilla's hallucinating captions that the method fixes, and b the *break rate*, the share of vanilla's non-hallucinating captions that the method changes to hallucinating (Figure 6 in Appendix G shows both for every method and model). The *break share* is $\text{BS} = \frac{b}{f+b}$, the fraction of the method's changes that go the wrong way, counting a real and a hallucinating caption with equal weight. $\text{BS} = 0$ means the method only fixes and never breaks; $\text{BS} = 50\%$ means it breaks a real caption as often as it fixes a hallucinating one. A method lowers CHAIR_s exactly when its break share is below the vanilla CHAIR_s.[1] **So the lower a model hallucinates, the fewer breaks a method can afford.** We report results in Figure 2a. On COCO, PAI on LLaVA-1.5 has $\text{BS} = 23\%$: roughly one in four of its changes breaks a real caption. Despite this, CHAIR_s improves significantly, since there is far more to fix than to break. The six recent models hallucinate in only

[1]With h the vanilla CHAIR_s, the method fixes a share fh of all captions and breaks a share $b(1-h)$, and $fh > b(1-h)$ is the same as $b/(f+b) < h$.

13–18% of their captions, so even a few broken captions worsens the hallucination rate. LocoRE improves on all six, with BS $= 11\%$ on average (9–14% per model), but its gain is still small because there is little to fix. PAI's break share (14–20%) sits close to the vanilla rate, so its fixes and breaks nearly cancel, and VCD's (17–27%) is above it on all six, so it raises CHAIR_s on all of them (Table 3). Within a recent model, **what separates the three methods is mainly how many real captions they break**: reduction in CHAIR_s is correlated with breaks stronger than the fixes (Figure 2b,c).

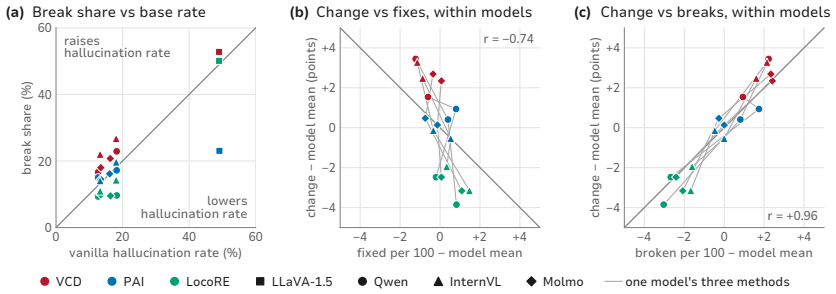


Figure 2: **Break share and the change in CHAIR_s.** Colour: method; marker: model family. (a) Break share of each method on each of the seven models, against the model's vanilla CHAIR_s. (b, c) The change in CHAIR_s (method minus vanilla) and the fixed (b) or broken (c) captions per 100 captions are shown relative to the model's mean over its three methods; grey lines join the three methods of one model. The fixed and broken captions behind each point are shown in Appendix G. Within a model, the change follows the breaks ($r = +0.96$) more closely than the fixes ($r = -0.74$).

Hedging. Recent models often hedge, e.g. using words such as 'possibly'. Since judge-CHAIR also counts hedged mentions, we analyse them in Figure 4 (Appendix A.2). How often a model hedges varies a lot: Qwen3.5 and Qwen3-VL hedge about 5% of their mentions, Gemma-4 almost never (Figure 4a). Hedged mentions are wrong much more often than the rest. On the recent models except Gemma-4, 22–37% of hedged mentions are hallucinated, against 3–5% of the others (Figure 4b). Counting them, as rule-CHAIR does, would therefore flag more hallucinations and lower precision. For Qwen3.5 and Qwen3-VL, it raises CHAIR_s by 5.0 and 6.0 points and lowers precision by 2.3 and 2.9 points (Table 6 in Appendix A.2). On the recent models, judge-CHAIR stays far below rule-CHAIR either way.

Mitigation methods also change how often a model hedges (Figure 4c). VCD lowers the share of hedged mentions on all seven models, by 0.8 points on average, PAI changes it little (-0.2), and LocoRE raises it on all seven ($+0.4$). Since judge-CHAIR leaves hedged mentions out (Section 3.2), this can move the score without the model seeing anything new. When VCD turns "possibly a table or chair" into "a wooden table" (Figure 4d), a mention that was not scored becomes one that is. When LocoRE turns "a white game controller" into "a white object, possibly a game controller" (Figure 4e), it does the opposite. On COCO this effect is small: over the seven models, only 3–6% of each method's fixed or broken captions involve a hedge being added to or removed from a hallucinated object. Still, a method can change CHAIR by changing how sure a caption sounds, not only which objects it names, so we suggest reporting the hedging rate next to CHAIR.

4 Revisiting Mitigation Methods on Harder Benchmark

The CHAIR metric on the COCO benchmark only evaluates hallucinations on ordinary COCO images. To test whether mitigation methods show same benefits on hard images i.e., where co-occurring objects are removed, we repurpose an existing benchmark BEAF (Moon et al., 2024), which was primarily proposed for visual question answering. The results in (Moon et al., 2024) show that it is a harder setup for models as this dataset captures the tendency of LVLMs to keep giving the same answer instead of responding to the image. We empirically evaluate LVLMs on these images for open ended generation and as shown in Figure 3, this provides a harder setup to evaluate object hallucinations. Naturally, this suggests that object hallucination is much more of a prevalent issue in harder settings in the above mentioned 7-12B LVLMs.

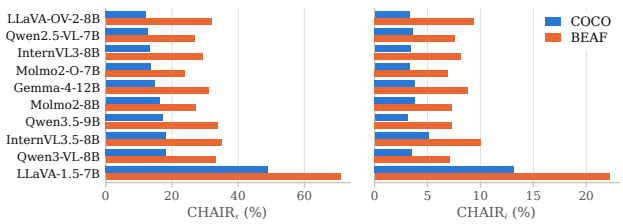


Figure 3: **Judge-CHAIR on COCO and on BEAF.** CHAIR$_s$ (left) and CHAIR$_i$ (right) of the ten models on the 500 COCO images and on all 2,223 BEAF images. Models are sorted by judge-CHAIR$_s$ on COCO. We find that, counterfactual object removal is a harder set up for current LVLMs and judge-CHAIR is consistently high on the BEAF dataset for all the ten models.

4.1 Evaluation setup

BEAF starts from 500 COCO images and makes 1,723 edited copies. In each copy one object (in a few cases two or three) has been painted out, and the rest of the image is left unchanged. We call one (original image, edited image, removed object) triple a *unit*. There are 1,778 units. For every original and edited image, we ask the model to describe the image in detail (*"Please help me describe the image in detail."*), and we prompt an LLM judge to read the description and decide whether it says the removed object is present. In the VQA setting, we instead ask the question *"Is there a ⟨object⟩ in the image?"* and use the yes/no answer. While Figure 3 uses the full set, the mitigation experiments (Tables 4 and 17) use a hard subset of 450 units, i.e., 250 edited images and the 200 originals they come from. We gather units for this subset if, with vanilla decoding, the model still claims the removed object on the edited image in both settings (the description still mentions it and the VQA answer is still "yes") for at least 2 of the 7 models. This gives 302 candidate units. From these, we draw 250 units spread evenly over five size bins of the removed object (quintiles of its area over all 1,778 units), so that a method cannot look better only because it helps on small or on large objects. Since the subset is selected on original model failures, it serves as a set of hard cases for comparing mitigation methods, not as an estimate of how often these models hallucinate in general.

Therefore, we want to evaluate whether the model keeps claiming an object, even after it has been removed from the image. To measure this, we propose a metric called *persistence*.

Eligibility and persistence. A unit is *eligible* if the model says the object is there in the original image. Persistence is measured only on eligible units:

$$\text{eligibility} = \frac{\#\text{eligible units}}{\#\text{units}}, \qquad \text{persistence} = \frac{\#\text{eligible units still claimed after the edit}}{\#\text{eligible units}}.$$

As the object is no longer present in the edited image, it is a hallucination if the model still claims it to be present on the edited image. In the VQA setup, a model is always asked about all the objects and it is forced to answer. However, when a model describes an image in detail, it might miss mentioning specific objects in the description. And it is unclear whether the model did not describe it because it did not perceive it or if it perceived but decided not to mention it. Therefore, to measure persistence in the open ended generation, we condition on the original image because otherwise a model that never mentions an object would score perfectly on it without looking at the image. With this condition, persistence asks a direct question: the model saw the object, we removed it, does the model notice? We report eligibility next to persistence, because it shows how many units the persistence score is based on. Persistence results on open ended generation and VQA are in Table 4 and Table 17 (Appendix H.2).

4.2 Persistence of LVLMs

In this part, we show the benefits of mitigation methods on persistence in current LVLMs on the BEAF dataset. Table 4 is on open ended generation and Table 17 (Appendix H.2) is on VQA.

Table 4: **Caption persistence of the mitigation methods on the BEAF hard subset** (%, lower is better), over matched units (n, out of the 250 units). Δ: method minus vanilla (van.); Fix./Brk.: units the method fixes/breaks relative to vanilla. Only PAI on LLaVA-1.5 and LocoRE on InternVL3.5 lower persistence significantly (exact McNemar test, $p < 0.05$).

	VCD						PAI						LocoRE					
Model	n	van.	VCD	Δ	Fix.	Brk.	n	van.	PAI	Δ	Fix.	Brk.	n	van.	Loc.	Δ	Fix.	Brk.
Qwen3-VL-8B	235	71.5	73.6	+2.1	18	23	237	70.5	70.5	+0.0	24	24	232	71.6	68.1	−3.5	32	24
Qwen2.5-VL-7B	217	60.8	62.7	+1.9	23	27	196	63.8	61.7	−2.1	25	21	214	61.7	57.9	−3.8	25	17
Molmo2-O-7B	180	48.9	53.3	+4.4	29	37	176	47.7	52.3	+4.6	24	32	185	49.7	46.5	−3.2	26	20
Molmo2-8B	215	52.1	54.0	+1.9	26	30	221	52.0	57.9	+5.9	17	30	214	52.8	55.6	+2.8	18	24
InternVL3.5-8B	222	68.9	67.1	−1.8	25	21	224	68.8	75.0	+6.2	26	40	216	69.4	54.6	−14.8	42	10
InternVL3-8B	226	68.6	67.7	−0.9	19	17	226	68.1	66.8	−1.3	21	18	222	68.5	61.7	−6.8	34	19
LLaVA-1.5-7B	229	89.5	84.7	−4.8	20	9	221	88.7	81.0	−7.7	26	9	239	89.5	90.0	+0.5	7	8

Observations. Tables 4 and 17 show that under object removal, none of the three methods consistently lowers persistence, and that the gains on COCO do not carry over. We use the same significance test as in Section 3.4. In captions, only two of the 21 method–model pairs show a significant drop; on the other 19, the McNemar test finds no significant difference, so the change could be due to noise. VCD does not significantly change caption persistence on any model, but in VQA it is the only method that significantly helps recent models: it lowers persistence on Molmo2-8B (-11.6 points), InternVL3.5 (-4.8) and InternVL3 (-2.8). PAI again helps only on LLaVA-1.5, where it significantly lowers persistence in both captions (-7.7) and VQA (-4.4). Even there, the gain is much smaller than on COCO: caption persistence falls by 9% relative (88.7 to 81.0), against 37% for CHAIR_s. For VQA, we also boost the attention to the image on the answer token, since the yes/no answer is the first generated token and the official implementation leaves the boost off there. On the recent models, PAI then significantly raises VQA persistence on Qwen2.5-VL ($+12.4$), Molmo2-O ($+5.7$), InternVL3.5 ($+2.8$) and InternVL3 ($+2.4$), fixing no unit and by only breaking. LocoRE's only significant caption gain is on InternVL3.5 (-14.8), a model on which its COCO reduction is not significant, while its two significant COCO gains, on Qwen3-VL and Molmo2-8B, do not carry over (-3.5 and $+2.8$). LocoRE is not reported for VQA because it strengthens attention to previously generated output tokens, whereas the yes/no answer is the first generated token. Overall, none of the three methods is uniformly effective across models and output formats.

Findings. PAI is the only method that helps on both benchmarks and in both querying formats, but only on LLaVA-1.5, the model it was built on, and its benefit shrinks as the setting gets harder. VCD helps only when the model answers a yes/no question, and only on some models. LocoRE is the most consistent method on COCO, but its gains do not carry over to the harder BEAF setting. No method therefore reduces hallucination consistently across models, benchmarks and querying formats. Mitigation methods need to be tested on recent models, on harder benchmarks and across querying formats before a general benefit is claimed. PAI is a clear example: it helps an older model in every setting we test, but not the recent ones, and in VQA it even makes several of them worse.

5 Conclusion

We asked how reliably object hallucination is measured on current LVLMs and whether mitigation methods built on LLaVA-1.5 still help. First, we saw that rule-CHAIR overestimates object hallucination on current models. We kept the CHAIR pipeline and only replaced its word matching with an LLM judge that extracts the object mention from caption. With this change, we saw that the hallucination rate significantly dropped for all the recent models. Second, the gains reported on LLaVA-1.5 do not carry over to current models. Third, current models hallucinate more on harder images, where the realised benefits of mitigation methods shrinks. A lower rule-CHAIR score on LLaVA-1.5 is therefore not enough to show that a mitigation method works on current LVLMs. Since LVLMs are becoming more powerful, we suggest testing new methods on current models, and on harder settings. We discuss the limitations of our study in Appendix I.

AI USE STATEMENT

AI tools supported several stages of the research process. They were used as a sounding board for refining hypotheses and evaluating methodological and experimental choices. In addition, AI assistance was used to examine experimental outputs and support their interpretation during preparation of the manuscript. AI coding agents were used for helping in implementing the code at various levels of autonomy based on the type of task and was checked for plausibility at a high level. AI tools were used creating and refining figures, editing code, identifying relevant literature and improving the presentation of the manuscript. AI-assisted text and figures were reviewed by the authors. Regardless of the use of AI assistance, the authors assume full responsibility for the methodology, results, claims, and overall content of the paper.

ETHICS STATEMENT

This research was conducted in accordance with the ICLR Code of Ethics and established standards for responsible research. Human validation of the LLM-based evaluation was carried out by a single annotator, who was also an author of the paper. To maintain independence from the automated evaluation, the annotator was not shown the LLM judge's predictions during annotation. Our experiments use datasets that are either publicly accessible or collected following appropriate ethical practices. We have aimed to maintain transparency and integrity throughout the collection, evaluation, and reporting of our results, and have identified no conflicts of interest related to this work.

REPRODUCIBILITY STATEMENT

We have taken extensive care to ensure others can reproduce our results. Keeping that in mind, our experimental pipeline is documented with the goal of making the reported results straightforward to replicate. The paper specifies the models, evaluation methodology, and construction of the evaluation data, with further technical details, including implementation choices and hyperparameters provided in the appendix. The experiments rely exclusively on datasets that are openly accessible. Following publication, we will additionally make the complete experimental code base available with the necessary details needed to rerun the experiments.

REFERENCES

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. LLaVA-OneVision-2: Towards Next-Generation Perceptual Intelligence, 2026. URL https://arxiv.org/abs/2605.25979. arXiv preprint arXiv:2605.25979.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report, 2025a. URL https://arxiv.org/abs/2511.21631. arXiv preprint arXiv:2511.21631.

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

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. Mitigating open-vocabulary caption hallucinations, 2024.

Ali Furkan Biten, Lluis Gomez, and Dimosthenis Karatzas. Let there be a clock on the beach: Reducing Object Hallucination in Image Captioning. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. URL https://arxiv.org/abs/2110.01705.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open Weights and Data for Vision-Language Models with Video Understanding and Grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://arxiv.org/abs/2601.10611.

Gemma Team. Gemma 4 Technical Report, 2026. URL https://arxiv.org/abs/2607.02770. arXiv preprint arXiv:2607.02770.

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. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models, March 2024. URL http://arxiv.org/abs/2310.14566. arXiv:2310.14566 [cs.CV].

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2311.17911.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. FaithScore: Fine-grained Evaluations of Hallucinations in Large Vision-Language Models, September 2024. URL http://arxiv.org/abs/2311.01477. arXiv:2311.01477 [cs.CV].

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, C. J. Taylor, and Stefano Soatto. THRONE: An Object-based Hallucination Benchmark for the Free-form Generations of Large Vision-Language Models, April 2025. URL http://arxiv.org/abs/2405.05256. arXiv:2405.05256 [cs.CV].

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2311.16922.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models, October 2023. URL http://arxiv.org/abs/2305.10355. arXiv:2305.10355 [cs.CV].

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a. URL https://arxiv.org/abs/2310.03744.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying More Attention to Image: A Training-Free Method for Alleviating Hallucination in LVLMs. In *European Conference on Computer Vision (ECCV)*, 2024b. URL https://arxiv.org/abs/2407.21771.

Ye-Bin Moon, Hyeon-Woo Nam, Wonseok Choi, and Tae-Hyun Oh. Beaf: Observing before-after changes to evaluate hallucination in vision-language models. In *Computer Vision – ECCV 2024*. Springer, 2024.

Suzanne Petryk, David M. Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph E. Gonzalez, and Trevor Darrell. ALOHa: A New Measure for Hallucination in Captioning Models, April 2024. URL http://arxiv.org/abs/2404.02904. arXiv:2404.02904 [cs.CV].

Haoyi Qiu, Wenbo Hu, Zi-Yi Dou, and Nanyun Peng. VALOR-EVAL: Holistic Coverage and Faithfulness Evaluation of Large Vision-Language Models, October 2024. URL http://arxiv.org/abs/2404.13874. arXiv:2404.13874 [cs.CL].

Qwen Team. Qwen3.5: Towards Native Multimodal Agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5. Blog post.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object Hallucination in Image Captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. URL https://arxiv.org/abs/1809.02156.

Sreetama Sarkar, Yue Che, Alex Gavin, Peter A. Beerel, and Souvik Kundu. Mitigating Hallucinations in Vision-Language Models through Image-Guided Head Suppression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. URL https://arxiv.org/abs/2505.16411.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation, February 2024. URL http://arxiv.org/abs/2311.07397. arXiv:2311.07397 [cs.CL].

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency, 2025. URL https://arxiv.org/abs/2508.18265. arXiv preprint arXiv:2508.18265.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and Mitigating Hallucination in Large Vision-Language Models via Modular Attribution and Intervention. In *International Conference on Learning Representations (ICLR)*, 2025.

Yilin Yang, Zhenghui Guo, Yuke Wang, Omprakash Gnawali, Sheng Di, and Chengming Zhang. HII-DPO: Eliminate Hallucination via Accurate Hallucination-Inducing Counterfactual Images, 2026. URL https://arxiv.org/abs/2602.10425. arXiv preprint arXiv:2602.10425.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2312.00849.

Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, Chunyuan Li, and Manling Li. HallE-Control: Controlling Object Hallucination in Large Multimodal Models, March 2024. URL http://arxiv.org/abs/2310.01779. arXiv:2310.01779 [cs.CV].

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Feilong Tang, Sinan Fan, Yaomin Shen, Chen Shen, and Hao Tang. Hallucination begins where saliency drops. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=sjnErRHXf3.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.00754.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models, 2025. URL https://arxiv.org/abs/2504.10479. arXiv preprint arXiv:2504.10479.

Reihaneh Zohrabi, Hosein Hasani, Akshita Gupta, Mahdieh Soleymani Baghshah, Anna Rohrbach, and Marcus Rohrbach. HaloProbe: Bayesian Detection and Mitigation of Object Hallucinations in Vision-Language Models, 2026. URL https://arxiv.org/abs/2604.06165. arXiv preprint arXiv:2604.06165.

APPENDIX CONTENTS

A Judge-CHAIR details 14
A.1 The judge prompt 14
A.2 Hedged mentions 16
A.3 Which classes rule-CHAIR flags 17
B Human validation 19
B.1 Annotation guideline 19
C Rule-CHAIR results on COCO 21
D Details of the mitigation methods 23
D.1 Hyperparameters 23
E Full judge-CHAIR results for the mitigation methods 24
F Statistical tests 26
G How the methods change captions 27
H BEAF details 27
H.1 Persistence judge prompt 27
H.2 VQA persistence of the mitigation methods 28
H.3 Official PAI and PAI with the boost on the answer token 29
I Limitations 29

A JUDGE-CHAIR DETAILS

A.1 THE JUDGE PROMPT

Judge-CHAIR uses `gpt-5.6-luna` with the reasoning effort set to none. Each caption is judged in a separate call. The judge sees the caption and the COCO ground-truth classes of the image, but not the image itself. The ground-truth list is there only to resolve vague references such as "the animal"; the prompt tells the judge not to use it to decide whether an object is mentioned. The system prompt below is the same for every caption; the list of the 80 COCO categories is shown as a placeholder. For every mention of a COCO category, the judge returns the span, its head word, the COCO class, whether the caption asserts that the object is in the image (*asserted present*), whether the mention is *hedged*, and a short reason, in a fixed JSON format. Only mentions asserted present enter CHAIR. As in rule-CHAIR, such a mention is real if its class is in the ground truth of the image and hallucinated otherwise, so the two metrics differ only in which words count as object mentions. Captions caught in a repetition loop are left out of both metrics. Table 5 shows the judge's output on one caption.

System prompt.

```
You are a careful linguistic annotator supporting CHAIR-style object-hallucination
evaluation on MSCOCO images.

Your only job is to find which objects are mentioned in the *text*. You cannot see the
image and you must NEVER GUESS what is really in it. For every mention in the caption you
decide two things: does the caption assert that a physical instance of an MSCOCO category
is present in the depicted scene (asserted_present), and did the caption commit to what
that object is (hedged)?

# The 80 MSCOCO categories
<the 80 COCO category names, one per line>


# The `hedged` flag
This is recorded SEPARATELY from asserted_present and never changes it -- a hedged mention
is still asserted. It only records whether the caption committed to what the object is, so
the metric can be reported with and without these.
Set hedged = true when the caption does not commit to WHAT the object is and is not sure
about the object it is mentioning: "possibly a bird", "what looks like a bench or a piece
of equipment", "a white appliance, possibly a refrigerator or freezer", "a plate with food
(possibly pizza)", "perhaps a bench".
Set hedged = false when the caption is really sure about the object. E.g., A small black
pouch with a strap, possibly for carrying a camera or other gear. Pouch is not hedged here.
Wallet, possibly for carrying cards and money - wallet is not hedged here.
Judge what the hedge scopes over, not whether a hedge word appears somewhere in the
sentence. The doubt has to be about the object's identity: "a man likely holding a phone"
(the phone is certain), "possibly a farm, as there are horses visible" (the horses are
```

```
certain). A mention with no hedge at all is hedged = false, and so is a shortened or
coreferent repeat of something already named ("two younger ones" after "giraffes", "the
toy" after "a plush toy") -- the caption committed when it first named the thing.


# Mark asserted_present = true when
- The caption states or clearly implies that a physical instance is in the scene, including
via a synonym, hyponym, or plural/collective noun ("guy" -> person, "a crowd" -> person,
"sofa" -> couch, "pup" -> dog).
- The mention is hedged but sincere: "what appears to be a cat", "possibly a bird".
- The instance is partial, occluded, small, or in the background: "a hand holding a fork",
"cars parked in the distance".
- The mention is a specific subcategory of the COCO categories: "wine glass", "school bus",
"police officer", "office chair", "dining table".
- The caption names a bigger object that always carries the category on it, so the category
is visible too: "clock tower" -> clock, "traffic light pole" -> traffic light. This works
one way only -- naming a part never asserts the whole it belongs to, see the false list
below.
- The caption mentions the instance separately, even if the compound on its own would not
count: "bike rack with bikes" -> bicycle, "a shelf full of books" -> book.
- If the mention is uncertain "a bench or table", "cameras or phones", "a stove or oven"
name both the objects, with the hedged flag True.


# Mark asserted_present = false when
- MODIFIER-ONLY COMPOUND. The word appears only as a modifier of a different head noun, and
the compound denotes a place, container, accessory, product, or activity *for* the category
rather than the category itself, and no instance is separately asserted. Examples:
"toothbrush holder" (toothbrush), "bus stop" (bus), "dog food" / "dog bowl" (dog), "mouse
pad" (mouse), "baseball field" (baseball bat), "basketball hoop" / "basketball court"
(sports ball), "horse trailer" (horse), "cat tree" (cat), "book shelf" (book), "train
station" / "train tracks" (train), "boat ramp" (boat), "chair lift" (chair), "TV stand"
(tv), "cake pan" (cake), "bird cage" with no bird (bird).
- NAMING A PART DOES NOT PUT THE WHOLE IN THE SCENE: "windshield", "a car dashboard", "the
seat" do not add a car, because the caption never said car and you must not guess. "shows a
close-up of a motorcycle side mirror" -> does not say motorcycle is visible in image.
- NEGATED OR ABSENT: "there is no clock on the wall", "a man without a tie", "the table has
no bottles on it".
- NON-LITERAL: simile, metaphor, idiom, shape, pattern, logo, or brand name. "a rock shaped
like a bear", "a bear hug", "a teddy-bear-print blanket", "Apple laptop" still counts as a
laptop but "Apple" is not the fruit.
- DEPICTION OR EFFIGY, NOT INSTANCE: pictures, photos, paintings, drawings, murals, screens
showing the object, reflections, and equally statues, sculptures, figurines, ornaments,
toys and plush animals. "a painting of a horse", "a cartoon cat on the mug", "a toy cow",
"a stone figure holding a bottle", "a red bird ornament". A later bare mention keeps the
category it was introduced with: "one bear" pointing back at a teddy bear is teddy bear,
not bear. EXCEPTION: teddy bear is itself a category of plush object, so a teddy, plush
panda or other stuffed animal IS an asserted teddy bear -- it is only the *animal* it
depicts that is not.
- THE DEPICTION FRAMING CARRIES OVER. Once the caption frames something as a picture,
mural, carving, sculpture group or toy arrangement, everything described inside it is part
of that depiction too, however it is later named. "a stone carving of two figures ... the
right figure holds a bottle" asserts no person and no bottle.
- HYPOTHETICAL, GENERIC, OR NON-PRESENT: "you could park a car here", "the kind of place
where boats dock", "he used to own a dog".
- ROLE, PURPOSE, OR RULE -- not a sighting. The noun says what something is for, or what
someone is about to do, rather than what is visible: "the red light tells vehicles to
stop", "a cab, likely for the operator". In particular a bare "the <noun>" governed by a
verb of intent -- about to, ready to, preparing to, winding up to, waiting to -- is a role
filler, not a sighting: "preparing to hit the ball", "ready to catch the ball" assert no
ball. Assert only if the caption separately says the thing is visible.
- NEAREST-MATCH ROUNDING. Do not forcefully fit an object to a nearest COCO category. The
object is real but no category truly fits, so emit nothing rather than the closest one. A
bathtub is not a bed and neither is a pet bed; a frying pan, a washing basin or a
skate-park bowl is not a bowl; a railway carriage or tanker car is not a car; a workbench
is not a bench; a plate is not a bowl; bread is not a sandwich, meat is not a hot dog,
fruit is not an orange; a helmet is not a person, a magazine is not a book. Do not force
mushroom to be broccoli. Do not force the stool to be a chair. Magazines and booklets are
not books. A sign is a stop sign only if it is the octagonal STOP sign.
- WRONG WORD SENSE. Several MSCOCO category names collide with common non-object words.
Check the sense before asserting:
  * orange = the fruit only. "an orange cat", "an orange shirt" -> not asserted.
  * mouse = the computer mouse only. An animal mouse is not this category.
  * keyboard = computer keyboard
  * remote = a remote control only, never the adjective ("a remote village").
  * tie = a necktie only, never the verb ("tie the laces") and never "the game was a tie".
  * bowl = the dish only, not "Super Bowl" and not a bowling action.
  * sink = the basin only, never the verb ("the boat began to sink").
  * bear = the animal; "teddy bear" is a separate category, do not report both.
  * glass = a drinking vessel, never the material ("frosted glass", "a glass countertop",
"glass doors"). A plain tumbler is cup; wine glass needs a stem.
```

```
  * kite, bench, spoon, fork, book: only when the physical object is present.


# Choosing the right category
Look at context. "there is another smaller red stuffed animal, possibly a bear or a similar
plush toy" -> do not flag bear rather it is a teddy bear. In context of a car if it says
"color with black leather seats" -> not chair. A smaller, pastel-colored plush toy
resembling a bear or a similar creature -> teddy bear not real bear. There is a strap from
a bag or backpack - does not say backpack is there, rather says the strap is there from
backpack.


# Output
EMIT ONE ENTRY PER DISTINCT TEXTUAL MENTION, in order of appearance in the caption. If the
same category is mentioned multiple times in separate places, emit the mention those many
times, as it is critical to calculate the metrics downstream. Include entries with
asserted_present = false as well -- they are the audit trail for the false positives being
removed. `span` must be copied verbatim from the caption. `head_word` must be the single
noun (or the minimal noun compound, e.g. "teddy bear", "hot dog", "fire hydrant") inside `
span` that actually names the object, copied verbatim from the caption including its
surface form (plural, casing) -- e.g. span "passengers visible leaning out of the windows"
-> head_word "passengers". `reason` must be at most 15 words.
```

User prompt.

```
Caption:
"""
<caption>
"""

Ground-truth objects annotated for this image (scene context only):
- <ground-truth COCO class>
- ...

Use this list ONLY to resolve ambiguous references (e.g. "the animal" when the ground truth
says zebra). Do NOT use it to decide asserted_present. A genuinely asserted mention stays
asserted even when its category is missing from this list, and a modifier-only mention
stays non-asserted even when its category is in this list.

List every mention of an MSCOCO category in this caption, in order of appearance, with your
asserted_present decision for each.
```

Table 5: **Worked example** (Qwen2.5-VL-7B, COCO image 36349; ground truth: car, carrot, truck). Rule-CHAIR reads both "orange" (the colour of the carrots and of the stripes) as the fruit and flags the caption as hallucinating. The judge does not list "orange" at all, since the caption never names the fruit, and rules on the three mentions below. Carrot and car are both in the ground truth, so under judge-CHAIR the caption does not hallucinate; the hedged "vehicle" is left out under the paper setting. Reasons are the judge's own words.

Caption: The image shows a street scene with a variety of fresh vegetables displayed on the ground. In the foreground, there are two distinct piles of vegetables. On the left is a pile of white root vegetables, possibly salsify or burdock roots, with their green leafy tops still attached. On the right is a pile of orange carrots, also with some green tops remaining.
In the background, part of a blue vehicle is visible, including its front wheel and part of the bumper. The vehicle has yellow and orange diagonal stripes painted on its side, which may indicate it is used for public services or municipal work. The ground where the vegetables are placed appears to be paved with gray bricks, and there are yellow lines painted on the pavement, likely indicating a no-parking zone or a designated area. The setting suggests this could be a market or a roadside sale of fresh produce.

Mention	COCO class	asserted present	hedged	reason
"carrots"	carrot	yes	no	A pile of orange carrots is explicitly described in the scene.
"vehicle"	car	yes	yes	A blue vehicle is visibly present, though its exact type is unspecified.
"front wheel"	car	no	no	A wheel is mentioned, but wheels are not an MSCOCO category and do not assert a car.

A.2 HEDGED MENTIONS

The judge marks a mention as hedged when the caption is not sure *what* the object is, as in "possibly a bird". In the paper, hedged mentions are left out: they count neither as real nor as hallucinated.

Table 6 shows how the numbers change when they are counted like any other mention. Qwen3.5 and Qwen3-VL hedge the most (about 5% of their mentions), and counting hedged mentions raises their $CHAIR_s$ by 5.0 and 6.0 points and lowers their precision by 2.3 and 2.9 points; LLaVA-1.5, LLaVA-OV-2 and Gemma-4 change by at most 0.8 points. Precision drops because hedged mentions are wrong much more often than the rest: on the recent models except Gemma-4 (two hedged mentions), 22–37% of hedged mentions are hallucinated, against 3–5% of the others (Figure 4b). Either way, judge-CHAIR on the recent models stays far below rule-CHAIR (Table 2), so the conclusion of Section 3.3 does not depend on this choice.

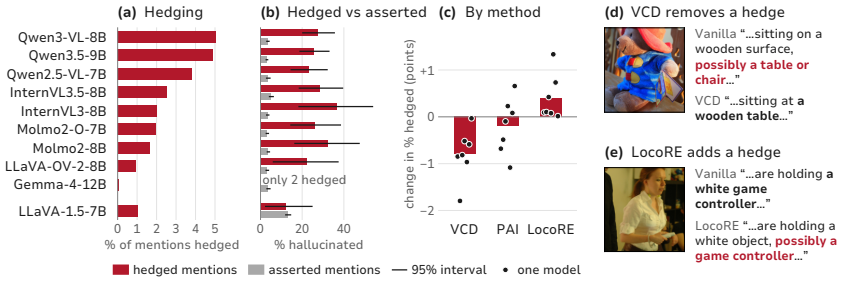


Figure 4: **Hedged mentions.** (a) Share of each model's mentions that the judge marks as hedged (vanilla). (b) Share of hedged and of the other (asserted) mentions that are hallucinated, with 95% intervals; Gemma-4 hedges only twice. (c) Change in the share of hedged mentions, method minus vanilla; bars are the mean over the seven models of Table 3, dots are single models. (d, e) Two InternVL3.5-8B captions where VCD drops a hedge and LocoRE adds one. Hedged mentions are hallucinated far more often than the rest, so a method that changes how often a model hedges can change judge-CHAIR without the model seeing more.

Table 6: **Judge-CHAIR with and without hedged mentions** (vanilla, 500 COCO val2014 images). *Paper*: hedged mentions are left out, as in Table 2; *all*: they are counted like any other mention. Δ is all minus paper. *Hedged*: share of the judge's asserted mentions that it marks as hedged. †vLLM runs, as in Table 2. All values in %.

		$CHAIR_s \downarrow$			$CHAIR_i \downarrow$			Precision ↑			F1 ↑		
Model	Hedged	paper	all	Δ	paper	all	Δ	paper	all	Δ	paper	all	Δ
Qwen3.5-9B†	4.9	17.3	22.3	+5.0	3.1	4.2	+1.1	92.3	90.0	−2.3	82.4	83.1	+0.7
Qwen3-VL-8B	5.0	18.2	24.2	+6.0	3.5	4.8	+1.3	91.8	88.9	−2.9	82.1	82.1	+0.0
Qwen2.5-VL-7B	3.8	12.7	15.9	+3.2	3.6	4.3	+0.7	93.9	92.6	−1.3	79.4	80.3	+0.9
Molmo2-O-7B	2.0	13.5	15.1	+1.6	3.3	3.7	+0.4	93.2	92.4	−0.8	76.8	77.0	+0.2
Molmo2-8B	1.6	16.3	18.1	+1.8	3.8	4.2	+0.4	92.5	91.6	−0.9	80.4	80.3	−0.1
InternVL3.5-8B	2.5	18.1	20.3	+2.2	5.1	5.7	+0.6	91.1	90.2	−0.9	79.3	79.5	+0.2
InternVL3-8B	2.0	13.3	15.1	+1.8	3.4	4.1	+0.7	93.8	93.0	−0.8	81.0	81.2	+0.2
LLaVA-OV-2-8B†	0.9	12.2	13.0	+0.8	3.3	3.4	+0.1	93.6	93.3	−0.3	78.0	78.2	+0.2
Gemma-4-12B†	0.1	14.8	14.8	+0.0	3.8	3.8	+0.0	92.3	92.3	+0.0	76.1	76.1	+0.0
LLaVA-1.5-7B	1.1	49.0	49.8	+0.8	13.2	13.2	+0.0	75.4	75.2	−0.2	76.5	76.5	+0.0

A.3 WHICH CLASSES RULE-CHAIR FLAGS

Figure 5 breaks down the hallucinations that rule-CHAIR reports on original captions by COCO class. For every (image, class) pair that rule-CHAIR marks as hallucinated, we check whether judge-CHAIR marks the same class as hallucinated in the same caption. On the nine recent models, the judge does not confirm 59–82% of rule-CHAIR's flags; on LLaVA-1.5 it does not confirm 10%. The unconfirmed flags concentrate on a few classes whose synonym lists contain common words with another meaning. "Orange" is read as the fruit even when it is a colour, and makes up 23% of all unconfirmed flags over the ten models. Other examples are "seat" and "stool" (chair), "bow" (tie), "container" (bowl) and "table" or "desk" (dining table).



Figure 5: **Rule-CHAIR's hallucination flags per COCO class** (vanilla, 500 COCO val2014 images). One panel per model, in the order of Table 2. Each bar is the share of the model's rule-CHAIR flags, i.e. (image, class) pairs marked hallucinated, that fall in the class, so the bars of a panel add up to 100%. *Judge also flags*: judge-CHAIR marks the same class as hallucinated in the same caption; *rule-CHAIR only*: it does not. Classes are sorted by their number of flags over all ten models; seven classes are never flagged.

B HUMAN VALIDATION

Table 1 compares rule-CHAIR and judge-CHAIR with a human annotator, one of the authors. The annotator labelled the object mentions in the vanilla captions of LLaVA-1.5-7B and Qwen2.5-VL-7B on 50 COCO images (100 captions) in a web tool that shows each image with its caption, following the guideline below. For every mention of one of the 80 COCO categories, the annotator selected the words, picked the category and labelled the mention real, hallucinated or ambiguous from the image.

B.1 ANNOTATION GUIDELINE

We give the guideline word for word.

Read this once before you start, and keep it open while you annotate.

Your task Each item shows an image and a caption. Find every place where the caption mentions an object from the 80 categories listed in the tool. For each one:

1. Select the word(s) that name the object.
2. Pick its category.
3. Choose a label by looking at the image:
   - **Real:** the object is in the image, even if it's small, partly hidden or in the background.
   - **Hallucinated:** the caption says it's there, but it isn't in the image.
   - **Ambiguous:** you can't tell from the image.
4. Add **Hedged** if the caption isn't sure *what* the object is (see Hedged).

When you've read the whole caption, mark it done, even if there was nothing to annotate.

Main rule: text first, then image Use only the caption's words to decide whether a mention counts and which category it gets. Look at the image only to choose real, hallucinated or ambiguous.

> The caption says "a stool" and the image shows a chair. A stool isn't one of the categories, so there's nothing to annotate.

What to select

- Select only the word(s) that name the object: "dog" in "a small brown dog", "teddy bear" in "a fluffy teddy bear".
- Annotate every mention, including repeats. "a man … the man" is two mentions.
- Words that refer back to something already named count too: "the animal" after "a dog" is a dog.
- Skip pronouns (he, she, it, they).

These count

- Synonyms and more specific names:
  - "guy", "crowd", "police officer" → person
  - "sofa" → couch
  - "pup" → dog
  - "school bus" → bus
  - "office chair" → chair
- Objects that are partly visible, small or far away: "cars parked in the distance".
- Something bigger that always includes the object: "clock tower" → clock, "traffic light pole" → traffic light.
- An object named on its own inside a longer phrase. In "a bike rack with bikes", "bikes" → bicycle. In "a shelf full of books", "books" → book.
- Uncertain mentions:
  - "possibly a bird" counts; add Hedged.
  - "a bench or table": annotate both, and add Hedged to each.
- Teddy bears, plush toys and stuffed animals → teddy bear.

These don't count

- **Words that only describe something else:** "bus stop", "dog food", "dog bowl", "toothbrush holder", "mouse pad", "train station", "train tracks", "chair lift", "TV stand", "book shelf", "basketball court", or "a bird cage" when no bird is mentioned. These are places or things *for* the object, not the object itself.
- **Parts:** "windshield", "dashboard" or "side mirror" don't mean a car or motorcycle is there. A hand or a helmet isn't a person.
- **Negations:** "there is no clock", "a man without a tie".
- **Non-literal uses:** "a rock shaped like a bear", "a bear hug", "a teddy-bear-print blanket", logos and brand names. "Apple laptop" is still a laptop, but "Apple" isn't the fruit.
- **Pictures and figures:** paintings, photos, drawings, murals, screens, reflections, statues, figurines, ornaments and toys, such as "a painting of a horse", "a cartoon cat on the mug" or "a toy cow".
  - Everything described inside the picture belongs to the picture. In "a carving of two figures ... one holds a bottle", there is no person and no bottle.
  - Teddy bears are the exception (see above).
- **Hypothetical or absent objects:** "you could park a car here", "he used to own a dog".
- **Purpose or intent, not an actual sighting:** "ready to catch the ball", "preparing to hit the ball", "the red light tells vehicles to stop". These count only if the caption also says the object is visible.
- **No close-enough matches:** if no category truly fits, don't annotate. Examples:
  - A plate or a frying pan is not a bowl.
  - Bread is not a sandwich, meat is not a hot dog, and fruit is not an orange.
  - A magazine is not a book, and a stool is not a chair.
  - A workbench is not a bench, and a bathtub is not a bed.
  - A railway carriage is not a car, and a mushroom is not broccoli.
  - A sign is a stop sign only if the caption makes clear it's the octagonal STOP sign.

Words with two meanings Only these meanings count:

- **orange:** the fruit, not the colour
- **mouse:** the computer mouse
- **keyboard:** the computer keyboard
- **remote:** a remote control, not "a remote village"
- **tie:** a necktie, not "tie the laces"
- **bowl:** the dish, not "Super Bowl" or a skate-park bowl
- **sink:** the basin, not "began to sink"
- **glass:** a drinking glass, not the material ("glass doors"). A plain glass or tumbler is a cup. A wine glass needs a stem.
- **bear:** the animal. Teddy bear is a separate category; never annotate both for one mention.
- **kite, bench, spoon, fork, book:** only the physical object.

Pick the category from context

- "a small stuffed animal, possibly a bear" $\rightarrow$ teddy bear, not bear.
- A word that refers back keeps its first category: "one bear" referring to a teddy bear is a teddy bear.
- "black leather seats" in a car $\rightarrow$ not a chair.
- "a strap from a bag" mentions only the strap, not the bag.

Hedged Add Hedged when the caption isn't sure *what* the object is. Examples: "possibly a bird", "perhaps a bench", "what looks like a bench", "a white appliance, possibly a refrigerator or freezer", "a plate with food (possibly pizza)".

Don't add it when:

- The doubt is about something else. In "a man likely holding a phone", the phone is certain. In "possibly a farm, as there are horses", the horses are certain.
- The mention repeats something already named: "the toy" after "a plush toy".

Table 7: **Rule-CHAIR on vanilla and VCD** on the 500-image COCO val2014 eval set. Δ is VCD minus vanilla; Δ is negative if VCD reduces hallucination. Length is the average caption length in words. All values in %, except Length. Rows are sorted by how much VCD lowers the judge's CHAIR_s.

	$\text{CHAIR}_s \downarrow$			$\text{CHAIR}_i \downarrow$			Recall ↑			Precision ↑			F1 ↑			Length	
Model	van.	VCD	Δ	van.	VCD	Δ	van.	VCD	Δ	van.	VCD	Δ	van.	VCD	Δ	van.	VCD
LLaVA-1.5-7B	53.6	54.4	+0.8	13.5	15.8	+2.3	78.6	78.2	−0.4	74.0	72.4	−1.6	76.3	75.2	−1.1	91	91
Qwen2.5-VL-7B	36.3	43.0	+6.7	9.8	10.3	+0.5	69.8	70.5	+0.7	82.5	79.7	−2.8	75.6	74.8	−0.8	164	175
Molmo2-8B	35.8	45.4	+9.6	8.7	11.8	+3.1	70.6	72.8	+2.2	82.7	78.3	−4.4	76.2	75.4	−0.8	151	148
Qwen3-VL-8B	55.2	59.6	+4.4	10.0	10.6	+0.6	76.2	76.9	+0.7	74.7	72.6	−2.1	75.4	74.7	−0.7	336	335
Molmo2-O-7B	33.3	38.6	+5.3	7.9	10.7	+2.8	65.7	64.2	−1.5	83.3	79.4	−3.9	73.5	71.0	−2.5	171	157
InternVL3-8B	32.3	36.1	+3.8	8.5	10.1	+1.6	70.5	70.7	+0.2	84.8	82.5	−2.3	77.0	76.1	−0.9	115	116
InternVL3.5-8B	37.3	41.6	+4.3	10.3	11.0	+0.7	69.6	70.1	+0.5	81.3	80.1	−1.2	75.0	74.8	−0.2	133	129

Table 8: **Rule-CHAIR on vanilla and PAI** on the 500-image COCO val2014 eval set. Δ is PAI minus vanilla; Δ is negative if PAI reduces hallucination. Length is the average caption length in words. All values in %, except Length. Rows are sorted by how much PAI lowers the judge's CHAIR_s.

	$\text{CHAIR}_s \downarrow$			$\text{CHAIR}_i \downarrow$			Recall ↑			Precision ↑			F1 ↑			Length	
Model	van.	PAI	Δ	van.	PAI	Δ	van.	PAI	Δ	van.	PAI	Δ	van.	PAI	Δ	van.	PAI
LLaVA-1.5-7B	53.6	34.4	−19.2	13.5	8.5	−5.0	78.6	73.7	−4.9	74.0	84.1	+10.1	76.3	78.5	+2.2	91	94
Qwen3-VL-8B	55.2	56.9	+1.7	10.0	10.5	+0.5	76.2	76.8	+0.6	74.7	73.2	−1.5	75.4	75.0	−0.4	336	304
Molmo2-8B	35.8	43.3	+7.5	8.7	10.0	+1.3	70.6	71.7	+1.1	82.7	79.9	−2.8	76.2	75.6	−0.6	151	152
InternVL3-8B	32.3	34.3	+2.0	8.5	9.2	+0.7	70.5	72.8	+2.3	84.8	83.5	−1.3	77.0	77.8	+0.8	115	113
Molmo2-O-7B	33.3	33.3	+0.0	7.9	8.9	+1.0	65.7	62.6	−3.1	83.3	82.2	−1.1	73.5	71.1	−2.4	171	140
InternVL3.5-8B	37.3	41.6	+4.3	10.3	10.4	+0.1	69.6	70.6	+1.0	81.3	80.1	−1.2	75.0	75.0	+0.0	133	125
Qwen2.5-VL-7B	36.3	36.3	+0.0	9.8	11.7	+1.9	69.8	64.2	−5.6	82.5	80.6	−1.9	75.6	71.5	−4.1	164	136

Choosing the label

- Only presence matters: "two dogs" when the image has one dog is real.
- If the caption names a real object but the image has only a picture, statue or toy of it, the mention is hallucinated. Example: "a horse" when the image has only a painting of a horse.
- Use Ambiguous only when the image genuinely doesn't let you tell.

C RULE-CHAIR RESULTS ON COCO

Section E reports the three mitigation methods under judge-CHAIR, where the object mentions are pulled out by an LLM. Here we give the same three comparisons scored with rule-CHAIR instead, with the fixed synonym list for the 80 COCO categories. Everything else – the models, the 500 COCO val2014 images, the decoding settings and the chosen hyperparameters – is identical, so each table below lines up row for row with its judge-CHAIR counterpart in Section E.

Table 9: **Rule-CHAIR on vanilla and LOCORE** on the 500-image COCO val2014 eval set. Δ is LOCORE minus vanilla; Δ is negative if LOCORE reduces hallucination. Length is the average caption length in words. All values in %, except Length. Rows are sorted by how much LOCORE lowers the judge's CHAIR_s.

	$\text{CHAIR}_s \downarrow$			$\text{CHAIR}_i \downarrow$			Recall ↑			Precision ↑			F1 ↑			Length	
Model	van.	LOC.	Δ	van.	LOC.	Δ	van.	LOC.	Δ	van.	LOC.	Δ	van.	LOC.	Δ	van.	LOC.
Qwen3-VL-8B	55.2	49.3	−5.9	10.0	8.8	−1.2	76.2	75.1	−1.1	74.7	77.1	+2.4	75.4	76.1	+0.7	336	332
Molmo2-8B	35.8	33.5	−2.3	8.7	8.0	−0.7	70.6	69.4	−1.2	82.7	83.2	+0.5	76.2	75.7	−0.5	151	164
InternVL3.5-8B	37.3	36.3	−1.0	10.3	8.6	−1.7	69.6	67.5	−2.1	81.3	82.7	+1.4	75.0	74.3	−0.7	133	143
Molmo2-O-7B	33.3	31.1	−2.2	7.9	7.4	−0.5	65.7	65.1	−0.6	83.3	83.9	+0.6	73.5	73.3	−0.2	171	165
Qwen2.5-VL-7B	36.3	34.7	−1.6	9.8	8.6	−1.2	69.8	69.9	+0.1	82.5	83.4	+0.9	75.6	76.1	+0.5	164	159
InternVL3-8B	32.3	32.9	+0.6	8.5	8.9	+0.4	70.5	67.6	−2.9	84.8	84.0	−0.8	77.0	74.9	−2.1	115	113
LLaVA-1.5-7B	53.6	53.8	+0.2	13.5	14.5	+1.0	78.6	78.6	+0.0	74.0	74.1	+0.1	76.3	76.3	+0.0	91	90

D DETAILS OF THE MITIGATION METHODS

This section gives more details on the three mitigation methods revisited in Section 3.4.

Visual Contrastive Decoding (VCD) VCD (Leng et al., 2024) is a training-free decoding strategy designed to reduce hallucinations arising from language priors and statistical biases. The method first constructs a distorted version of the input image by adding Gaussian noise, which weakens the available visual evidence and consequently amplifies the model's reliance on such priors. During generation, VCD performs two forward passes, conditioned on the original and distorted images, respectively, and contrasts their next-token logits as $\ell_{\mathrm{VCD}} = (1+\alpha)\ell(y \mid v, x) - \alpha\ell(y \mid v', x)$, where v and v' denote the original and distorted images and α controls the contrastive strength. Intuitively, tokens that remain likely when the visual evidence is degraded are downweighted, while tokens supported specifically by the original visual input are promoted.

Paying more attention to image (PAI) PAI (Liu et al., 2024b) is a training-free inference-time intervention motivated by the observation that LVLMs may under-attend to visual tokens and consequently rely excessively on language priors. PAI addresses this through two complementary interventions. First, during autoregressive generation, it amplifies the pre-softmax self-attention scores from the current token to image tokens, $\tilde{A}n, j \leftarrow \tilde{A}n, j + \alpha|\tilde{A}n, j|, \quad j \in \mathcal{I}$, where $\mathcal{I}$ denotes the set of image-token positions and α controls the attention amplification, thereby increasing the contribution of visual information to the resulting hidden representations. Second, it performs an image-centric logit refinement by contrasting the output distribution obtained from the multimodal input with that obtained from a text-only input: $\ell_{\mathrm{PAI}} = \gamma\,\ell(y \mid v, x) - (\gamma - 1)\,\ell(y \mid x)$, where $\gamma \geq 1$ controls the contrastive strength ($\gamma = 1$ turns it off). This suppresses predictions that can be explained primarily by the language context while promoting predictions that are more dependent on the visual input, with the goal of reducing hallucinations caused by excessive reliance on language priors.

Local Coherence Reinforcement (LocoRE) LocoRE (Zhang et al., 2026) is a training-free inference-time intervention motivated by the observation that hallucinations are associated with reduced dependence on previously generated tokens, indicating a loss of contextual information during generation. This is different from the commonly adopted view that excessive reliance on textual tokens relative to visual information contributes to hallucination. However, rather than uniformly increasing or suppressing attention to textual tokens, LocoRE selectively reinforces recent output-token dependencies to maintain local contextual coherence, in particular, LocoRE strengthens the attention from the current token to recently generated tokens within a local window by scaling their attention weights as $\gamma_j^{(P)} = 1 + \beta\,\mathbb{I}(P - j \leq w_s)$, where P denotes the current generation position, w_s determines the local context window, and β controls the intervention strength. The original work additionally proposes Saliency-Guided Rejection Sampling (SGRS), which uses gradient-based saliency to filter candidate tokens during decoding. We do not employ SGRS, as computing saliency during autoregressive inference introduces substantial computational overhead and is less practical for deployment; instead, we use only the lightweight LocoRE intervention.

D.1 HYPERPARAMETERS

For LLaVA-1.5 we keep the published settings where they are given. Every other setting is chosen on 100 held-out COCO val2014 images that do not overlap with the 500-image evaluation set. We select the setting with the lowest judge-CHAIR$_s$, provided that its F1 is no more than 1.5 points below the original model and that it does not introduce repetition loops. VCD mainly uses the sampling based decoding, however they show that VCD works with greedy decoding as well (Table 7, (Leng et al., 2024)), although the benefit is smaller in case of greedy as it provides a stronger baseline than sampling. The candidates are:

- **VCD:** the 12 settings of the one-at-a-time ablation in the VCD paper. Starting from $T = 500$, $\alpha = 1$, $\beta = 0.1$, one value changes at a time: $T \in \{200, 500, 700, 999\}$, $\alpha \in \{0.25, 0.5, 0.75, 1\}$ and $\beta \in \{0, 0.001, 0.01, 0.1, 0.2, 0.5\}$.
- **PAI:** first $\alpha \in \{0.1, 0.2, 0.4, 0.5, 0.6, 0.7\}$ at $\gamma = 1.1$, then $\gamma \in \{1.1, 1.2, 1.3, 1.5, 2.0\}$ at the chosen α.

Table 11: **Judge-CHAIR on vanilla and VCD** on the 500-image COCO val2014 eval set. Δ is VCD minus vanilla; Δ is negative if VCD reduces hallucination. Length is the average caption length in words. All values in %, except Length. Rows are sorted by how much VCD lowers the judge's CHAIR_s.

	$\text{CHAIR}_s \downarrow$			$\text{CHAIR}_i \downarrow$			Recall ↑			Precision ↑			F1 ↑			Length	
Model	van.	VCD	Δ	van.	VCD	Δ	van.	VCD	Δ	van.	VCD	Δ	van.	VCD	Δ	van.	VCD
LLaVA-1.5-7B	49.0	50.8	+1.8	13.2	15.4	+2.2	77.7	77.6	−0.1	75.4	73.7	−1.7	76.5	75.6	−0.9	91	91
Qwen2.5-VL-7B	12.7	14.9	+2.2	3.6	3.7	+0.1	68.7	70.8	+2.1	93.9	93.3	−0.6	79.4	80.5	+1.1	164	175
Molmo2-8B	16.3	18.5	+2.2	3.8	4.5	+0.7	71.0	74.6	+3.6	92.5	91.5	−1.0	80.4	82.2	+1.8	151	148
Qwen3-VL-8B	18.2	20.5	+2.3	3.5	3.8	+0.3	74.2	76.5	+2.3	91.8	90.8	−1.0	82.1	83.0	+0.9	336	335
Molmo2-O-7B	13.5	16.3	+2.8	3.3	3.8	+0.5	65.3	65.9	+0.6	93.2	92.0	−1.2	76.8	76.8	+0.0	171	157
InternVL3-8B	13.3	16.7	+3.4	3.4	4.6	+1.2	71.3	71.4	+0.1	93.8	92.0	−1.8	81.0	80.4	−0.6	115	116
InternVL3.5-8B	18.1	22.3	+4.2	5.1	6.3	+1.2	70.1	71.0	+0.9	91.1	89.7	−1.4	79.3	79.3	+0.0	133	129

Table 12: **Judge-CHAIR on vanilla and PAI** on the 500-image COCO val2014 eval set. Δ is PAI minus vanilla; Δ is negative if PAI reduces hallucination. Length is the average caption length in words. All values in %, except Length. Rows are sorted by how much PAI lowers the judge's CHAIR_s.

	$\text{CHAIR}_s \downarrow$			$\text{CHAIR}_i \downarrow$			Recall ↑			Precision ↑			F1 ↑			Length	
Model	van.	PAI	Δ	van.	PAI	Δ	van.	PAI	Δ	van.	PAI	Δ	van.	PAI	Δ	van.	PAI
LLaVA-1.5-7B	49.0	30.8	−18.2	13.2	8.0	−5.2	77.7	72.4	−5.3	75.4	85.6	+10.2	76.5	78.4	+1.9	91	94
Qwen3-VL-8B	18.2	17.7	−0.5	3.5	3.4	−0.1	74.2	76.1	+1.9	91.8	91.9	+0.1	82.1	83.3	+1.2	336	304
Molmo2-8B	16.3	16.1	−0.2	3.8	3.8	+0.0	71.0	72.8	+1.8	92.5	92.4	−0.1	80.4	81.4	+1.0	151	152
InternVL3-8B	13.3	13.7	+0.4	3.4	3.5	+0.1	71.3	73.9	+2.6	93.8	93.7	−0.1	81.0	82.6	+1.6	115	113
Molmo2-O-7B	13.5	14.1	+0.6	3.3	3.4	+0.1	65.3	63.7	−1.6	93.2	92.6	−0.6	76.8	75.5	−1.3	171	140
InternVL3.5-8B	18.1	18.9	+0.8	5.1	4.9	−0.2	70.1	71.0	+0.9	91.1	90.8	−0.3	79.3	79.7	+0.4	133	125
Qwen2.5-VL-7B	12.7	14.3	+1.6	3.6	4.0	+0.4	68.7	66.4	−2.3	93.9	92.9	−1.0	79.4	77.5	−1.9	164	136

- **LocoRE:** $\beta \in \{0.1, 0.15, 0.2\}$ and $w_s \in \{5, 10, 20\}$, all nine combinations. No code has been released for LocoRE, so we implemented it from the paper. The paper gives no window size, so for LLaVA-1.5 we keep its $\beta = 0.15$ and choose only w_s.

Table 10 lists the chosen settings.

Table 10: **Settings used for each model.** PAI and LocoRE act on LLM layers 2 to $L-1$ (L layers in all, the first two are left out). VCD: noise step T, contrast strength α, plausibility cut-off β. PAI: attention boost α, text-only contrast γ. LocoRE: boost β, window w_s. *LLaVA-1.5 keeps the published settings where they are given (all of VCD and PAI, LocoRE's β); every other value is chosen on the 100 held-out images.

		VCD			PAI		LocoRE	
Model	L	T	α	β	α	γ	β	w_s
Qwen3-VL-8B	36	500	1	0.1	0.1	1.5	0.1	20
Qwen2.5-VL-7B	28	700	1	0.1	0.5	1.1	0.15	5
Molmo2-O-7B	32	500	1	0.01	0.4	1.2	0.1	5
Molmo2-8B	36	500	1	0.001	0.1	1.1	0.1	10
InternVL3.5-8B	36	500	1	0.1	0.5	1.1	0.2	10
InternVL3-8B	28	500	0.5	0.1	0.1	1.1	0.15	10
LLaVA-1.5-7B*	32	500	1	0.1	0.5	1.1	0.15	10

E FULL JUDGE-CHAIR RESULTS FOR THE MITIGATION METHODS

Table 3 in Section 3.4 shows only CHAIR_s, CHAIR_i, F1 and length. Here we give the full results for each method on COCO, including recall, precision and the difference to vanilla decoding.

Table 13: **Judge-CHAIR on vanilla and LOCORE** on the 500-image COCO val2014 eval set. Δ is LOCORE minus vanilla; Δ is negative if LOCORE reduces hallucination. Length is the average caption length in words. All values in %, except Length. Rows are sorted by how much LOCORE lowers the judge's CHAIR_s.

	$\text{CHAIR}_s \downarrow$			$\text{CHAIR}_i \downarrow$			Recall ↑			Precision ↑			F1 ↑			Length	
Model	van.	LOC.	Δ	van.	LOC.	Δ	van.	LOC.	Δ	van.	LOC.	Δ	van.	LOC.	Δ	van.	LOC.
Qwen3-VL-8B	18.2	13.3	−4.9	3.5	2.7	−0.8	74.2	72.2	−2.0	91.8	93.8	+2.0	82.1	81.6	−0.5	336	332
Molmo2-8B	16.3	12.9	−3.4	3.8	3.1	−0.7	71.0	69.8	−1.2	92.5	93.7	+1.2	80.4	80.0	−0.4	151	164
InternVL3.5-8B	18.1	15.9	−2.2	5.1	4.5	−0.6	70.1	66.2	−3.9	91.1	92.1	+1.0	79.3	77.0	−2.3	133	143
Molmo2-O-7B	13.5	11.4	−2.1	3.3	3.0	−0.3	65.3	63.8	−1.5	93.2	93.8	+0.6	76.8	75.9	−0.9	171	165
Qwen2.5-VL-7B	12.7	10.8	−1.9	3.6	2.6	−1.0	68.7	67.8	−0.9	93.9	94.9	+1.0	79.4	79.1	−0.3	164	159
InternVL3-8B	13.3	12.2	−1.1	3.4	3.3	−0.1	71.3	68.8	−2.5	93.8	94.0	+0.2	81.0	79.4	−1.6	115	113
LLaVA-1.5-7B	49.0	49.4	+0.4	13.2	13.8	+0.6	77.7	77.7	+0.0	75.4	75.4	+0.0	76.5	76.6	+0.1	91	90

F STATISTICAL TESTS

We compare each method with vanilla decoding on the same items: the same COCO images for CHAIR_s, and the same matched units for persistence. Each item is either unchanged, *fixed* (vanilla hallucinates and the method does not) or *broken* (the reverse). The change in CHAIR_s or persistence is $(b - f)/n$, where f and b are the numbers of fixed and broken items and n the number of items, so only the changed items carry information about the difference. If the method had no effect, each changed item would be fixed or broken with probability $1/2$. The exact McNemar test checks this with a two-sided binomial test:

$$p = \min\Big(1,\ 2 \sum_{i=0}^{\min(f,b)} \binom{f+b}{i} 2^{-(f+b)}\Big). \tag{2}$$

For example, LocoRE on Qwen3-VL fixes 46 COCO captions and breaks 22, which gives $p = 0.005$; PAI on Qwen3-VL fixes 4 BEAF VQA units and breaks none, which gives $p = 0.125$. We call a change significant when $p < 0.05$. Tables 14 to 16 give f, b and p for every method–model pair of Tables 3, 4 and 17. A table has 14 to 21 pairs, so even if no method had any effect, about one pair per table would reach $p < 0.05$ by chance; a single borderline result should be read with this in mind.

Table 14: **Significance of the CHAIR_s changes in Table 3** (COCO, 500 images). n: captions judged for both vanilla and the method. *Fix.*: vanilla's caption hallucinates and the method's does not; *Brk.*: the reverse. p: exact McNemar test; bold if $p < 0.05$.

	VCD				PAI				LocoRE			
Model	n	Fix.	Brk.	p	n	Fix.	Brk.	p	n	Fix.	Brk.	p
Qwen3-VL-8B	494	36	48	0.230	494	44	41	0.828	493	46	22	**0.005**
Qwen2.5-VL-7B	498	29	40	0.228	498	36	44	0.434	498	31	22	0.272
Molmo2-O-7B	498	34	48	0.151	498	33	36	0.810	498	34	24	0.237
Molmo2-8B	497	32	43	0.248	496	30	30	1.000	495	39	21	**0.027**
InternVL3.5-8B	498	32	53	**0.029**	498	36	40	0.731	498	45	34	0.260
InternVL3-8B	498	20	37	**0.033**	498	27	29	0.894	498	26	21	0.560
LLaVA-1.5-7B	498	55	64	0.463	497	132	41	**<0.001**	498	45	47	0.917

Table 15: **Significance of the caption persistence changes in Table 4** (BEAF hard subset). n: matched units. *Fix.*: vanilla still names the removed object on the edited image and the method does not; *Brk.*: the reverse. p: exact McNemar test; bold if $p < 0.05$.

	VCD				PAI				LocoRE			
Model	n	Fix.	Brk.	p	n	Fix.	Brk.	p	n	Fix.	Brk.	p
Qwen3-VL-8B	235	18	23	0.533	237	24	24	1.000	232	32	24	0.350
Qwen2.5-VL-7B	217	23	27	0.672	196	25	21	0.659	214	25	17	0.280
Molmo2-O-7B	180	29	37	0.389	176	24	32	0.350	185	26	20	0.461
Molmo2-8B	215	26	30	0.689	221	17	30	0.079	214	18	24	0.441
InternVL3.5-8B	222	25	21	0.659	224	26	40	0.109	216	42	10	**<0.001**
InternVL3-8B	226	19	17	0.868	226	21	18	0.749	222	34	19	0.053
LLaVA-1.5-7B	229	20	9	0.061	221	26	9	**0.006**	239	7	8	1.000

Table 16: **Significance of the VQA persistence changes in Table 17** (BEAF hard subset). Columns as in Table 15; *Fix.* and *Brk.* are answers that change from "yes" to "no" and back. PAI is the version with the image boost on the answer token, as in the main text.

	VCD				PAI			
Model	n	Fix.	Brk.	p	n	Fix.	Brk.	p
Qwen3-VL-8B	247	9	3	0.146	250	4	0	0.125
Qwen2.5-VL-7B	240	9	15	0.307	242	0	30	**<0.001**
Molmo2-O-7B	242	17	9	0.169	244	0	14	**<0.001**
Molmo2-8B	242	33	5	**<0.001**	246	1	3	0.625
InternVL3.5-8B	249	13	1	**0.002**	250	0	7	**0.016**
InternVL3-8B	249	7	0	**0.016**	249	0	6	**0.031**
LLaVA-1.5-7B	231	18	16	0.864	228	13	3	**0.021**

G How the methods change captions

CHAIR_s gives only the net change. Figure 6 splits it up: for every model and method, each vanilla caption is paired with the method's caption of the same image, and the pair falls into one of four cells. The fixed and broken cells hold the f and b of Table 14. Every method both fixes and breaks captions on every model. On the six recent models, the three methods fix a similar share of vanilla's hallucinating captions (30–57%) and differ mainly in how many real captions they break: VCD 9–13%, PAI 7–10% and LocoRE 5–8%. On LLaVA-1.5, PAI fixes 54% of the hallucinating captions and breaks 16% of the real ones.

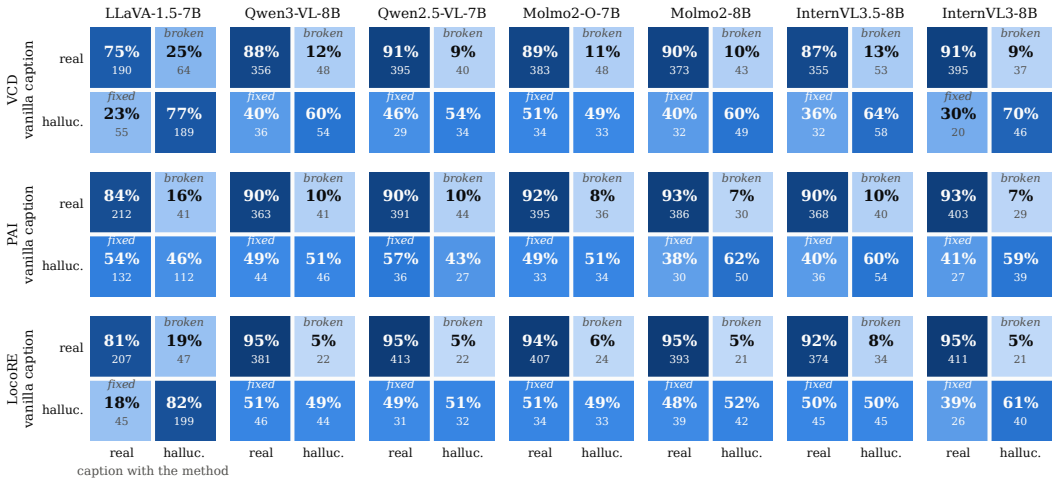


Figure 6: **How each mitigation method changes vanilla's captions** (COCO, 500 images). One confusion matrix per method (rows) and model (columns). The rows of a matrix say whether vanilla's caption is real or hallucinated, the columns say the same for the method's caption. Each cell shows the share of its row and, below it, the number of captions. Top right, *broken*: vanilla is real and the method hallucinates. Bottom left, *fixed*: vanilla hallucinates and the method is real. Judge-CHAIR, hedged mentions left out.

H BEAF details

H.1 Persistence judge prompt

For caption persistence (Section 4), a separate judge decides whether a description says that an object is present. It uses `gpt-5.6-luna` with the default reasoning effort (medium) and the system prompt below. Each call gets one description and the objects that BEAF asks about for that image (about 12), in a fixed random order that is the same for an original image and its edit. The judge sees neither the image, nor the other description of the pair, nor which object was removed. For each object it rules present, absent or hedged, and a hedged ruling counts as not mentioned. In the VQA setting no judge is needed: an answer counts as "yes" if it contains "yes", following BEAF.

System prompt.

```
You decide what a written image description claims is in a scene. You are given the
description only -- never the image -- so every ruling is about what the text asserts, not
about what is really there. You are not judging whether the description is correct.

For each object in the QUERIED OBJECTS list, rule exactly one of:

  present   The description asserts that a physical instance of this object is in the
            scene. Synonyms, hypernyms and plurals count: "locomotive" asserts a train,
            "van" asserts a car, "kids" asserts a person, "luggage" asserts a suitcase.
            A part or a clearly implied whole counts when the description commits to it
            ("the cat's paw" asserts a cat).
```

```
            * orange = the fruit only. "an orange cat", "an orange shirt" -> not asserted.
            * mouse = the computer mouse only. An animal mouse is not this category.
            * keyboard = computer keyboard
            * remote = a remote control only, never the adjective ("a remote village").
            * tie = a necktie only, never the verb ("tie the laces") and never "the game was
a tie".
            * bowl = the dish only, not "Super Bowl" and not a bowling action.
            * sink = the basin only, never the verb ("the boat began to sink").
            * bear = the animal; "teddy bear" is a separate category, do not report both.
            * glass = a drinking vessel, never the material ("frosted glass", "a glass
countertop",             "glass doors"). A plain tumbler is cup; wine glass needs a stem.

  absent    The description does not assert it, or actively denies it. This covers the
            object simply never coming up, and also:
              - explicit negation: "there is no clock on the wall"
              - a depiction rather than the thing: a painting of a horse, a photo of a
                dog, a statue of a bird, a cartoon bear on a lunchbox, a logo, a toy
              - a modifier that is not the object: "a toothbrush holder" is not a
                toothbrush, "bus stop" is not a bus, "horsepower" is not a horse
              - the word used for something else entirely: COCO's "orange" is the fruit,
                so "an orange cat" is not an orange; "mouse" the animal is not the
                computer mouse
              - a counterfactual, question, or instruction rather than an assertion
                ("if there were a bench here", "you would expect a table")

  hedged    The description raises the object but will not commit: "what appears to be a
            dog", "possibly a bench", "either a cat or a small dog" (hedged for BOTH), "it
            is hard to tell if there is a bottle". Use this only for genuine uncertainty
            about whether the object is there. A confident statement with a vague adjective
            ("some kind of large truck") is present, not hedged.

Rule on every queried object exactly once, in the order given. On present and hedged, keep
reason to one short clause quoting the description's own words. On absent, leave reason an
empty string -- there is nothing to quote and "never mentioned" is not information.
```

User prompt.

```
QUERIED OBJECTS:
- <queried object>
- ...

DESCRIPTION:
<description>
```

H.2 VQA PERSISTENCE OF THE MITIGATION METHODS

Table 17 is the VQA counterpart of Table 4: on the BEAF hard subset, the model is asked "Is there a ⟨object⟩ in the image?" about the removed object, on the original and on the edited image, and persistence is the share of units still answered "yes" on the edit among those answered "yes" on the original. The results are discussed in Section 4.

Table 17: **VQA persistence of the mitigation methods on the BEAF hard subset** (%, lower is better). Columns as in Table 4. PAI boosts image attention on the answer token as well. VCD significantly lowers persistence on Molmo2-8B and InternVL3/3.5; PAI significantly lowers it only on LLaVA-1.5 and raises it on Qwen2.5-VL, Molmo2-O and InternVL3/3.5 (exact McNemar test, $p < 0.05$).

	VCD						PAI					
Model	n	van.	VCD	Δ	Fix.	Brk.	n	van.	PAI	Δ	Fix.	Brk.
Qwen3-VL-8B	247	85.4	83.0	−2.4	9	3	250	84.8	83.2	−1.6	4	0
Qwen2.5-VL-7B	240	63.3	65.8	+2.5	9	15	242	62.8	75.2	+12.4	0	30
Molmo2-O-7B	242	57.0	53.7	−3.3	17	9	244	57.0	62.7	+5.7	0	14
Molmo2-8B	242	70.7	59.1	−11.6	33	5	246	70.3	71.1	+0.8	1	3
InternVL3.5-8B	249	96.4	91.6	−4.8	13	1	250	96.4	99.2	+2.8	0	7
InternVL3-8B	249	93.6	90.8	−2.8	7	0	249	93.6	96.0	+2.4	0	6
LLaVA-1.5-7B	231	72.7	71.9	−0.8	18	16	228	72.8	68.4	−4.4	13	3

H.3 Official PAI and PAI with the boost on the answer token

The yes/no answer is the first token the model generates. The official PAI code keeps the attention boost off while the prompt is read, and that pass produces the first token, so in official PAI only the contrast with the text-only prompt acts on the answer. In Table 17 we therefore also apply the boost to the answer token. Table 18 compares the two versions, with the same α, γ and layers. Official PAI never significantly lowers persistence: over the seven models it changes 45 answers on the edited image from "no" to "yes" and only 2 from "yes" to "no", and it raises persistence significantly on Qwen2.5-VL, Molmo2-O and LLaVA-1.5. With the boost on the answer token, PAI significantly lowers persistence on LLaVA-1.5 (13 fixed, 3 broken), but raises it significantly on Qwen2.5-VL (30 broken, none fixed), Molmo2-O and both InternVL models.

Table 18: **Official PAI and PAI with the image boost on the answer token** (BEAF hard subset, VQA persistence in %, lower is better). *Official*: the boost is off while the prompt is read, so only PAI's contrast with the text-only prompt acts on the yes/no answer. *Boost on*: the boost also acts on the answer token; this is the version in Table 17. All other settings are the same. n: matched units; Δ: PAI minus vanilla (van.); Fix./Brk.: answers PAI changes from "yes" to "no" and back; p: exact McNemar test, bold if $p < 0.05$.

	Official PAI							PAI, boost on the answer token						
Model	n	van.	PAI	Δ	Fix.	Brk.	p	n	van.	PAI	Δ	Fix.	Brk.	p
Qwen3-VL-8B	250	84.8	84.8	+0.0	1	1	1.000	250	84.8	83.2	−1.6	4	0	0.125
Qwen2.5-VL-7B	242	62.8	66.5	+3.7	0	9	**0.004**	242	62.8	75.2	+12.4	0	30	**<0.001**
Molmo2-O-7B	244	57.0	63.5	+6.5	0	16	**<0.001**	244	57.0	62.7	+5.7	0	14	**<0.001**
Molmo2-8B	246	70.3	71.1	+0.8	1	3	0.625	246	70.3	71.1	+0.8	1	3	0.625
InternVL3.5-8B	250	96.4	97.6	+1.2	0	3	0.250	250	96.4	99.2	+2.8	0	7	**0.016**
InternVL3-8B	249	93.6	94.8	+1.2	0	3	0.250	249	93.6	96.0	+2.4	0	6	**0.031**
LLaVA-1.5-7B	234	73.1	77.4	+4.3	0	10	**0.002**	228	72.8	68.4	−4.4	13	3	**0.021**

I Limitations

We designed judge-CHAIR for a direct comparison with CHAIR, since CHAIR is the most widely used metric for object hallucination. It therefore keeps the 80 COCO classes and the COCO labels, and inherits their missing annotations; objects outside these classes are not scored. The LLM judge is not perfect either, and our human validation covers only 100 captions from two models, labelled by one annotator. We study 7–12B models with greedy decoding and three mitigation methods, so other model sizes, decoding strategies and methods may behave differently.