ERASURE OR SILENCE? A PLACEBO-CONTROLLED EVALUATION OF TRAINING-FREE HALLUCINATION MITIGATION IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucinating objects has been a problem for LVLMs since the beginning. Recent advances in LVLMs have introduced notable hallucination-reduction methods: some introduce training-free hallucination mitigation by intervening in the weight space, and others introduce decoding-time interventions. However, most of these hallucination mitigation frameworks evaluate their success using evaluation metrics such as CHAIR and POPE scores, which measure the level of hallucination based on the number of claims made by the model, not on the ground truth. But the model can manipulate this hallucination score just by generating less. In this paper, we are introducing **SCAR** (Suppression-Corrected Assessment of Reduction), which proposes a placebo-controlled frontier for the unedited model to use as a baseline for their hallucination reduction methods, to check that their mitigation gain is not just coming from the verbosity cheat of the model output. Our intuition is that if the hallucination mitigation method does not improve hallucination risk at the same coverage compared to the results that can be achieved by mere dummy changes, then the method was not worth it in the first place. Our frontier shows how the risk (hallucination score, in this case CHAIR_i) can be achieved in different coverage (c) levels, just by changing the placebo controls (generation length, beam width, temperature, a logit bias on object words), interventions that do not add any visual info, just control how much the model asserts.

1 INTRODUCTION

In recent years, we have witnessed remarkable advances in Large Vision-Language Models (LVLMs) (Liang et al., 2024; Bai et al., 2023; Achiam et al., 2023; Liu et al., 2023; Alayrac et al., 2022). LVLMs achieve state-of-the-art performance in various tasks such as visual question answering, image captioning, and visual reasoning (Goyal et al., 2017; Agrawal et al., 2019; Hudson & Manning, 2019; Liu et al., 2023). Despite those successes, LVLMs are still facing the issue of hallucination. They often mention objects that do not exist in the image (Rohrbach et al., 2018; Li et al., 2023). Also, very often LVLMs prioritize the underlying textual embeddings over the visual information (Zhou et al., 2024; Leng et al., 2024; Liu et al., 2024b). For addressing those issues, different kinds of training-based (Zhou et al., 2024; Yu et al., 2024) and training-free (Leng et al., 2024; Liu et al., 2024b; Huang et al., 2024; Yang et al., 2025; Jiang et al., 2025) methods are available, and most of them are validated with CHAIR (Rohrbach et al., 2018) for captions and POPE (Li et al., 2023) for yes/no probes. The issue is that both metrics are normalized by what the model chooses to assert, not by what the image actually contains. CHAIR_i measures the fraction of hallucinated mentions among all object mentions; POPE evaluates accuracy under a fixed decision rule. As a result, a model can achieve a good result by asserting less, and comparison of models at different answering rates can be misleading, too. Having experimented with training-free hallucination-reduction methods on the MSCOCO dataset (Lin et al., 2014), we have seen that even without adding any visual information, the CHAIR score can be moved. Figure 1 shows this on one image: a recent weight-space projection-based mitigation (Yang et al., 2025) removes the hallucinated table, but only by naming fewer real ones, while a content-free sampling temperature on the unedited model is just as clean and names as many. Over 500 images, simply changing beam width achieves a CHAIR reduction comparable to the same mitigation (Fig. 2, Sec. 3.2).

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

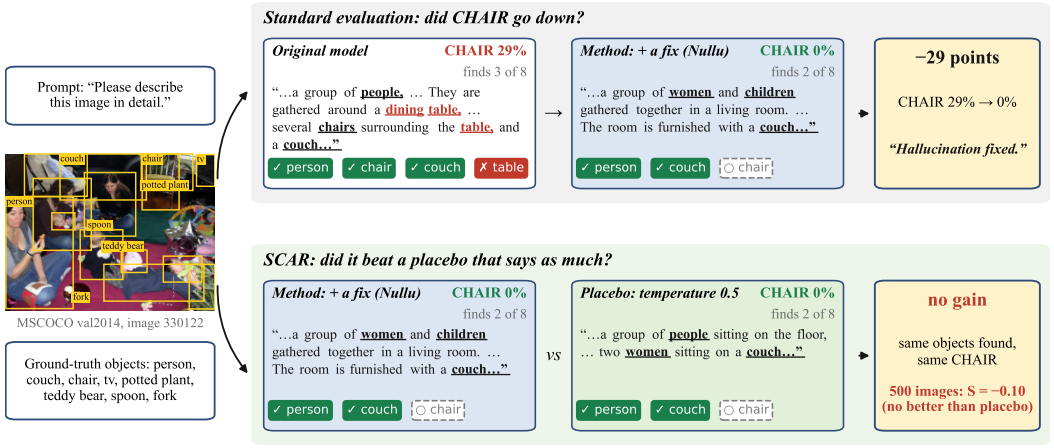


Figure 1: **Hallucinating less, or just generating less?** LVLMs often mention objects that are not in the picture. The standard metric, CHAIR, measures the hallucination rate by comparing the number of hallucinated objects against the total number of objects generated by the model, so a model can improve its score simply by generating fewer objects. **Top:** The model invents a dining table (CHAIR 29%). A published fix removes it (0%), but the model also stops mentioning the real chairs. **Bottom:** The unchanged model, with only a setting that adds no visual information adjusted (a *placebo*), names the same objects with the same 0%. SCAR credits a method only if it beats such placebos while mentioning as many real objects. This fix does not.

Existing works try to overcome this limitation in three ways. Open-vocabulary metrics (Petryk et al., 2024; Ben-Kish et al., 2024; Qiu et al., 2024) extend the object set beyond the 80 COCO categories, but they are still normalized by what the model asserts. Coverage-based analyses (Wang et al., 2023; Liu et al., 2026) report recall next to the hallucination rate and show that lower hallucination often comes with lower recall, but they leave the reader to weigh the two numbers. Null interventions (Yin et al., 2025; Bendre et al., 2026) show that the gains of contrastive decoding can come from a shifted decision bias or be matched by random noise, but they are built for decoding-time methods and have no counterpart for a method folded into the weights. So these methods do not measure how much of a method's reported gain can be explained by content-free effects. Specifically, what remains missing is a reference point indicating how much improvement can be achieved at the same coverage by the unedited model. Our work addresses this gap by proposing a reference curve that characterizes what the unedited model can attain at each coverage without adding any visual information.

We introduce SCAR (Suppression-Corrected Assessment of Reduction), a placebo-controlled evaluation protocol that places every configuration on a coverage/risk plane (fraction of ground-truth objects mentioned against CHAIR_i) and fits a placebo frontier as the lower envelope of content-free controls applied to the unedited model. The score is the method's vertical gap to the frontier at its own coverage, normalized into a bounded skill score S: $S = 0$ indicates placebo-equivalence, while $S < 0$ indicates that a content-free knob performs better. As it is built on the unedited model, this reference is applicable for both decoding-time methods and baked-in mitigation within the weights. Unlike a recall reported next to CHAIR, the comparison is made at the same coverage, and unlike a noise null, it needs nothing from the method except its output.

Contributions. Our contributions are summarized as follows.

- We propose SCAR, a placebo-controlled evaluation protocol. It compares a method against a placebo frontier built on the unedited model at the same coverage, and its score S shows how much of the reported improvement is real and how much is the placebo effect.
- We extend the evaluation metric to the discriminative setting with ΔAUROC on the restricted posterior, which separates improved ranking from a shifted decision threshold, and with $\varepsilon_{\mathrm{acc}}$, which measures what a method achieves beyond re-thresholding the unedited model to its own answering rate.

- We evaluate six training-free methods (VCD(Leng et al., 2024), PAI(Liu et al., 2024b), ProjectAway(Jiang et al., 2025), Nullu, OPERA(Huang et al., 2024), DeCo(Wang et al., 2025)), 20 configurations in total, on LLaVA-1.5-7B (Liu et al., 2024a). For most of them a content-free knob obtains a lower CHAIR score at the same coverage without applying the method, and none clears the band the knobs reach; on POPE, re-thresholding the unedited model recovers up to 97.6% of the reported accuracy gain. The same holds for a second model, MiniGPT-4 (Zhu et al., 2024), and on a second image set, AMBER.
- We show that SCAR can detect a real gain: an oracle bias that uses the image's annotation clears the band ($S = +0.64\ [+0.47, +0.75]$), while the same operator with permuted labels falls below the frontier ($S = -1.21$).

2 Related Work

How object hallucination is measured. Earlier attempts to evaluate object hallucination in image captioning relied mostly on lexical matching and under a fixed vocabulary. The CHAIR metric reported two rates: either the relative total number of objects mentioned or the generated caption using the fixed set of 80 categories from the MSCOCO dataset. Whereas POPE evaluates relying on multiple balanced yes/no probes within the fixed threshold boundary.

Then, newer benchmarks (such as ALOHa (Petryk et al., 2024) and OpenCHAIR (Ben-Kish et al., 2024)) expanded the scope from a limited set of objects to a broader one by using more advanced language models. In the same way, VALOR-EVAL (Qiu et al., 2024) widens the scope by adding attributes, relations, and reference-free factuality in longer captions.

Coverage is reported, and competed on. AMBER (Wang et al., 2023) addresses this issue to some extent by considering both how much the Model Hallucinates and how much it mentions real objects. They tried to trick the model by adding a content-free knob to force the model to talk less, and showed that it does not really make the model smarter. The author successfully portrayed the trade-off based on only the simple length trick. But they don't penalize the score if the model lowered the coverage, and also the model can be tricked in other ways too.

Training-free mitigation, with the guards already built in. The methods we audit are training-free, but the methods can be divided into two families: decoding-time changes that modify the model's thought process, and weight or representation changes that modify the model's internal knowledge or the subspace that leads to hallucinations (Yang et al., 2025; Jiang et al., 2025). Other methods add outside evidence, such as the detector-guided MARINE (Zhao et al., 2025), or build other contrastive branches, such as SID (Huo et al., 2025). We do not audit them here, but we test the case of added visual information with an oracle control in Sec. 5.

Even though for measuring hallucination, the recall and CHAIR value are being used. But for comparison, they don't provide any baseline to do it. ProjectAway comes closest: it sweeps edit strength, greys out every hyperparameter combination that reduces the count of correctly-detected objects, and reports that at large strength captions lose "both correctly detected and hallucinated objects" — the suppression regime, observed directly.

Null interventions. Mirage (Yin et al., 2025) has shown that the gain in contrastive decoding on POPE was due to a shift in decision bias. Also, a similar kind of performance can be achieved without even adding any visual support. Whereas Bendre et al. (2026) proposed a powerful null intervention. In this proxy method, Gaussian noise has been used in place of an amateur branch from contrastive decoding. As a result, On CHAIR, the performance of this proxy falls within the confidence intervals of both VCD and SID.

3 Suppression-Corrected Assessment of Reduction (SCAR)

SCAR is the evaluation framework that we propose, which incorporates the suppression of object mentions into the quantification of hallucination reduction, and consists of a plane, a family of placebo controls, a placebo frontier, and certain scalars that we obtain from this placebo frontier. The plane is the coverage/risk plane introduced in §3.1: coverage is a model's tendency to mention

Table 1: The placebo family: five knob families and the settings swept for each. The default cell (beam 3, 64 tokens, $\delta = 0$) is shared by π_{bias}, π_{beam} and π_{len}, so the 27 settings listed here give 25 distinct configurations.

knob	definition	settings
$\pi_{\text{bias}}(\delta)$	logit bias $\delta \leq 0$ on the first sub-token of every object surface form, at every step	$0, -0.25, -0.5, -0.75, -1, -1.25, -1.5, -2, -3, -4$
$\pi_{\text{beam}}(b)$	beamwidth	1, 2, 3, 5, 8
$\pi_{\text{temp}}(\tau)$	sampling temperature	0.5, 1.0, 1.5
$\pi_{\text{len}}(L)$	`max_new_tokens`	16, 32, 64, 128
$\pi_{\text{rand}}(k)$	random rank-k orthogonal fold into the edited layers	1, 4, 8, 16, 28

any given ground-truth object at all, whereas risk is computed using an instance-level variant of the hallucination-oriented CHAIR metric. Our placebo family in fact has 25 elements, as described in Tab. 1, and we run them all on identical computing hardware. The vertical distance between two curves at fixed coverage will be denoted by ε throughout this section; it is our plane of reference for evaluating hallucination reduction methods.

Selective prediction (El-Yaniv & Wiener, 2010) has already been used as a detection filter in LVLM hallucination detection; however, to our knowledge, its applicability in hallucination reduction, which does not abstain and where coverage shifts implicitly, has not been addressed before.

Two other tools are also adapted from concept erasure: guardedness, which tests whether edited-out concepts are still linearly decodable (Ravfogel et al., 2020; 2022), and the minimum-distortion oblique projection (Belrose et al., 2023), which only provides the denominator of the operator cost ρ (App. B).

3.1 The plane and the placebo family

CHAIR_i is the fraction of object mentions that are hallucinated over the total number of object mentions. Dropping ground-truth objects thus doesn't affect the hallucination metric. These two scores define a plane on which we can lay out hallucination-risk-object-coverage scores for image-captioning models, with each model represented by a point on the plane. For a model M,

$$c(M) = \frac{\#\{\text{correct mentions}\}}{\#\{\text{Actual objects present}\}}, \tag{1}$$

$$r(M) = \text{CHAIR}_i(M) = \frac{\#\{\text{hallucinated mentions}\}}{\#\{\text{all object mentions}\}}. \tag{2}$$

Placebo family: We apply controls to the reference model that alter its verbosity and tendency to assert, but do not incorporate any visual information. We call these placebo controls and denote by $\mathcal{K}$ the resulting *family of placebo models*. Given the reference model, each member of the placebo family is identified by a knob-setting $\mathbf{k}$. Four of the controls are decoding-stage adjustments: logit bias on the 403 counted object surface-forms, beam width, sampling temperature, and max_new_tokens. The remaining control is a new, weight-level placebo: folding down the model to a random orthogonal subspace, with details in App. A and Alg. 2.

3.2 The frontier

Placebo frontier: finally, given a family of placebo models $\mathcal{K}$ and a SCAR plane, the placebo frontier $\mathcal{F}$ is the lower envelope of the placebo family on the SCAR plane. While the frontier is not physically realizable by the reference model, its hull is: sampled from two placebo models at random. The frontier is drawn on the SCAR plane along with its member knobs, as well as the hallucination mitigations under study. Fig. 2 shows why it is needed: starting from the same unedited model, reducing beam width lowers CHAIR_i about as much as a rank-4 null-space edit (Yang et al., 2025), and the two reach it by moving coverage in opposite directions.

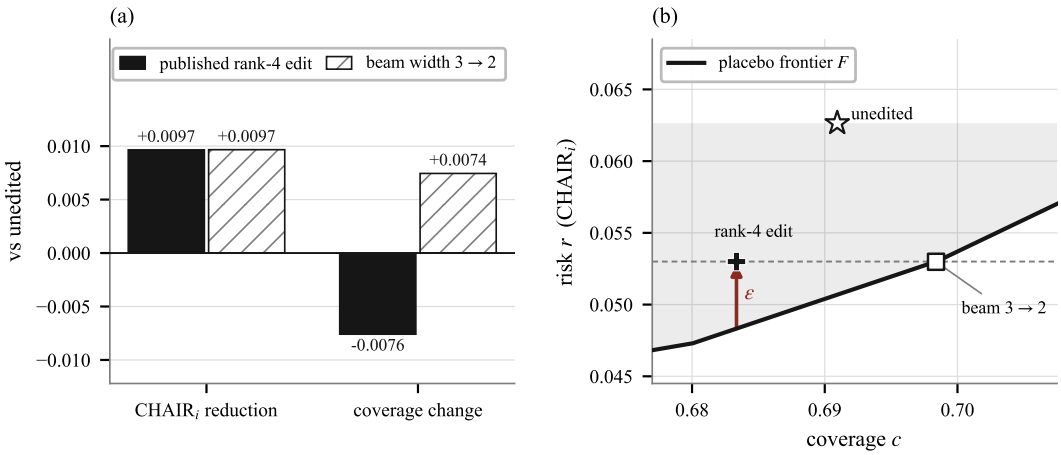


Figure 2: **We can lower the CHAIR score even by controlling the generation flag.** (a) From the same point, reducing beam width achieves comparable CHAIR reduction as the rank-4 null-space edit of Yang et al. (2025). (b) Reducing beam width and null-space editing change coverage in opposite directions. The solid curve is the placebo frontier F, and ε (Eq. 3) is the gap between it and the edited model at its own coverage. LLaVA-1.5-7B, beam 3 unless stated, 64 tokens, seed 0.

3.3 The scalars

Let r_{raw} be the CHAIR_i risk of the raw, unedited model and define the reported gain of a method as $\Delta(M) = r_{\text{raw}} - r_M$. We use three scalars. These are the excision gain (in CHAIR_i points), our proposed substitute for the widely reported ΔCHAIR,

$$\varepsilon(M) = r_F(c_M) - r_M, \tag{3}$$

$$\eta(M) = \varepsilon(M)/\Delta(M), \qquad \sigma(M) = \text{clip}_{[0,1]}(1 - \eta(M)) \,. \tag{4}$$

and the unclipped share of its reported gain that the placebo match attributes to suppression, $\eta(M)$, along with its clipped version, the suppression share $\sigma(M)$. A negative excision gain means the knob can achieve the method's coverage with a smaller risk. Note that clipping makes σ constant on the set $\{\varepsilon \leq 0\}$ and assigns $\sigma = 1.000$ to all methods with no gain. Both σ and η are undefined when $\Delta \leq 0$. Our headline scalar is the method's *score*

$$S(M) \;=\; \frac{\varepsilon(M)}{r_F(c_M)} \;=\; 1 - \frac{r_M}{r_F(c_M)} \tag{5}$$

following the nomenclature of skill scores such as R^2 or the Brier skill score. A method excising all risk relative to placebo gets $S = 1$; $S = 0$ marks placebo-equivalence, and $S < 0$ means that at its coverage, the content-free knob beats the method. As r_F is strictly positive, $\text{sign}(S) = \text{sign}(\varepsilon)$: S is a reparameterisation of ε and gives rise to no new hypotheses: replicate-wise, $\Pr[S \leq 0] = \Pr[\varepsilon \leq 0]$. Unlike the suppression share, the score is defined everywhere the method is inside the span of the frontier and doesn't saturate.

Scalars show 95% confidence intervals, calculated via a percentile bootstrap over evaluation samples, resampling question/image pairs within grid cells. σ additionally reports an *undefined fraction*: the fraction of replicates for which the resampled frontier does not cover this row. We also score each placebo cell against the frontier of the other cells; the largest of these leave-one-out scores is the edge of the *null band*, and we report $\Pr[S \leq \text{band}]$, with the band recomputed in each replicate.

3.4 The discriminative half

We use POPE questions to test discriminative ability. Its accuracy metric applies a fixed decision rule, which can lead to gains from changes to yes-bias alone. To avoid this, we compute accuracy on POPE questions using a restricted posterior: we extract scores $s = P(\text{yes})/(P(\text{yes}) + P(\text{no}))$ on the "yes" token, summed over its cased variants (Alg. 3). We read these scores at the first

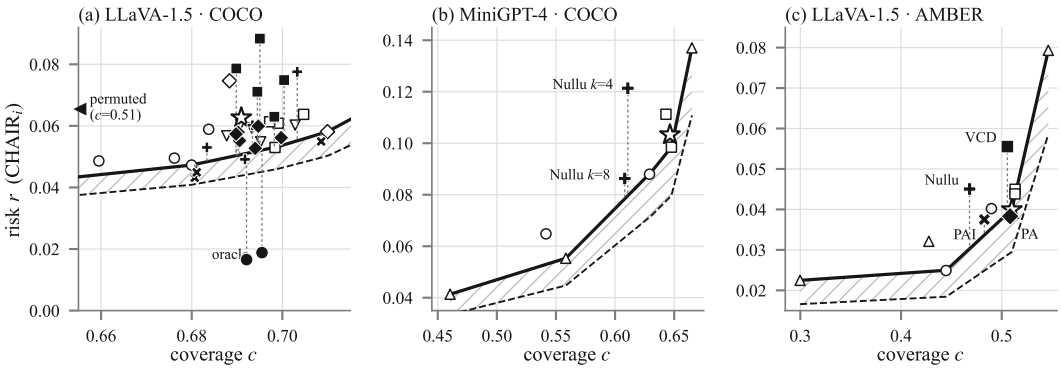


Figure 3: **The audit on three planes.** Hollow markers are placebo cells of the unedited model (star), filled markers are audited rows; the solid line is the frontier F, the hatched region is the null band, and a dotted drop marks each row's ϵ. (a) is enlarged around the unedited model; every scored row's S and interval is in Fig. 4. Bold labels are published settings. OPERA, DeCo and the MiniGPT-4 VCD and PAI rows of Tab. 2 are not drawn. Only the oracle clears its band.

answer position and use them to compute accuracy and AUROC. Since the latter is invariant to monotone rescalings of s, we find it useful to report ΔAUROC, which is thus only sensitive to object ranking changes. We compute a threshold $\hat{\tau}$ for the unedited model that matches the yes-rate of the intervened model at the published decision rule. Using this, we define a score

$$\varepsilon_{\text{acc}}(M) = \text{acc}(M) - \text{acc}_{\text{raw}}(\hat{\tau}), \tag{6}$$

is accuracy improvement beyond what's available by changing the decision threshold in the baseline.

Pre-registration. Hypothesis section pre-registered prior to taking measurements: We predict: **H1** at least one published mitigation method that quotes a headline CHAIR margin will have $\sigma \geq 0.5$; **H2** some methods' rankings will change under ε vs. ranking by Δ *CHAIR* as reported; **H3** the operator cost ρ (App. B) will be predictive of σ for the methods we consider. Failure of **H1** is a null result that would mean the hallucination literature passes this particular audit.

4 Experimental Setup

Model, data, protocol. Unless otherwise specified, all results are produced using SCAR on LLaVA-1.5-7B (Liu et al., 2024a) on a random sample of 500 images from the COCO Dataset. For hallucination mitigation methods we test VCD at $\alpha = 1.0$, PAI ($\alpha = 0.5$), ProjectAway ($\alpha = 3.5$), Nullu ($k = 4$), OPERA (Huang et al., 2024) ($\sigma = 50$, $\alpha = 1$, $r = 15$, $N_{\text{can}} = 5$) and DeCo (Wang et al., 2025) ($\alpha = 0.6$, top-p 0.9, layers 20–28), each at its published hyperparameters (Tab. 2). On MiniGPT-4, PAI uses its published $\alpha = 0.2$ for resampler models, while VCD and DeCo publish no MiniGPT-4 setting, so we reuse their LLaVA-1.5 values; these two rows are transfers, not published settings. For open-ended generation, we use the prompt query *"Please describe this image in detail."* with beam-width 3 (1 for *Greedy*) and a maximum of 64 new tokens. For closed-ended POPE we report accuracy, F1, yes-rate, AUROC and ε_{acc} over all questions, for a total of 9000 questions for each setting (Tab. 3). Experiments were all run with a random seed of 0. We repeat the audit on MiniGPT-4 with LLaMA-2-7B-chat (Zhu et al., 2024; Touvron et al., 2023) at the same protocol, where Nullu publishes rank $k = 8$, and on 500 images of AMBER, scored by AMBER's released scorer with its prompt *"Describe this image."*. There, the frontier uses 8 settings of the knob families that form the LLaVA frontier ($\delta \in \{0, -0.5, -2\}$, $b \in \{1, 2\}$, $L \in \{16, 32, 128\}$.

5 Results

The placebo frontier. There is currently no agreed-upon scale for the assertion knob. But we can get a sense of what's involved by implementing 25 unique settings that reach a range of coverages between 0.4323 and 0.8122, as shown in Tab. 6 in the appendix. The envelope is not monotone.

Table 2: **Main results.** Each row reports reproduced results, how far CHAIR moved, what coverage it cost, and whether a content-free placebo at the *same coverage* does as well (S, Eq. 5; higher is better. Each block begins with the unedited model (a placebo cell, thus $S \leq 0$). Subsequent methods use published hyperparameters (main setting in small type, full settings in Sec. 4). Subscripts give the change from the unedited model and, for S, the 95% interval over 2,000 paired image resamples. Null bands (largest leave-one-out placebo S), LLaVA / MiniGPT-4: COCO $+0.14$ / $+0.19$; AMBER $+0.26$ / $+0.32$. **Bold** S: the whole interval lies below 0 (a placebo does better) or above its band. Cov. is grounded recall (%); on AMBER, CHAIR_s and CHAIR_i are AMBER's Hal and CHAIR metrics. Oracle and permuted are controls: applying a logit bias ($\delta{=}-4$) to categories absent from the image's (oracle) or another image's (permuted) annotation. ‡ means point estimate. ProjectAway, Oracle, and Permuted were not evaluated on MiniGPT-4. Tab. 14 extends this table.

	Dataset	Method	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$	Cov.↑	$S \uparrow$
LLaVA-1.5-7B	**COCO**	*Unedited*	19.20	6.26	69.1	**−0.24** [−.56, −.12]
		VCD $\alpha{=}1$	22.00 $_{+2.80}$	7.11 $_{+0.84}$	69.4 $_{+0.4}$	**−0.37** [−.82, −.18]
		PAI $\alpha{=}0.5$	12.20 $_{-7.00}$	4.31 $_{-1.95}$	68.1 $_{-1.0}$	+0.09 [−.21, +.27]
		ProjectAway $\alpha{=}3.5$	17.00 $_{-2.20}$	5.73 $_{-0.53}$	69.0 $_{-0.1}$	**−0.14** [−.45, −.02]
		Nullu $k{=}4$	14.60 $_{-4.60}$	5.30 $_{-0.97}$	68.3 $_{-0.8}$	−0.10 [−.44, +.08]
		OPERA $\alpha{=}1, r{=}15$	16.80 $_{-2.40}$	5.89 $_{-0.38}$	69.1 $_{-0.0}$	**−0.16** [−.51, −.02]
		DeCo $\alpha{=}0.6$	14.60 $_{-4.60}$	5.68 $_{-0.58}$	65.5 $_{-3.6}$	**−0.31** [−.84, −.05]
		Oracle $\delta{=}-4$	6.60 $_{-12.60}$	1.88 $_{-4.38}$	69.6 $_{+0.5}$	**+0.64** [+.47, +.75]
		Permuted $\delta{=}-4$	12.60 $_{-6.60}$	6.55 $_{+0.28}$	51.0 $_{-18.1}$	**−1.21** [−2.81, −.56]
	AMBER	*Unedited*	17.20	4.00	51.0	0.00 [−.15, .00]
		VCD $\alpha{=}1$	22.20 $_{+5.00}$	5.55 $_{+1.56}$	50.6 $_{-0.4}$	**−0.43** [−.77, −.18]
		PAI $\alpha{=}0.5$	15.00 $_{-2.20}$	3.75 $_{-0.24}$	48.3 $_{-2.7}$	−0.11 [−.47, +.12]
		ProjectAway $\alpha{=}3.5$	16.60 $_{-0.60}$	3.83 $_{-0.16}$	50.8 $_{-0.2}$	+0.03 [−.12, +.11]
		Nullu $k{=}4$	17.20 $_{+0.00}$	4.50 $_{+0.51}$	46.8 $_{-4.2}$	**−0.49** [−.95, −.20]
		OPERA $\alpha{=}1, r{=}15$	21.20 $_{+4.00}$	5.23 $_{+1.23}$	50.5 $_{-0.5}$	**−0.35** [−.64, −.13]
		DeCo $\alpha{=}0.6$	18.00 $_{+0.80}$	4.64 $_{+0.64}$	49.1 $_{-1.9}$	−0.30 $_{\text{point}}$‡
MiniGPT-4	**COCO**	*Unedited*	25.80	10.33	64.6	−0.06 [−.21, .00]
		VCD $\alpha{=}1$	26.60 $_{+0.80}$	11.25 $_{+0.92}$	64.5 $_{-0.2}$	−0.16 [−.43, +.10]
		PAI $\alpha{=}0.2$	25.20 $_{-0.60}$	9.29 $_{-1.04}$	65.5 $_{+0.8}$	+0.18 [−.09, +.35]
		Nullu $k{=}8$	20.40 $_{-5.40}$	8.63 $_{-1.70}$	60.8 $_{-3.8}$	−0.10 [−.40, +.09]
		OPERA $\alpha{=}1, r{=}15$	26.60 $_{+0.80}$	10.58 $_{+0.24}$	64.6 $_{-0.0}$	−0.08 [−.25, +.10]
		DeCo $\alpha{=}0.6$	22.40 $_{-3.40}$	9.35 $_{-0.98}$	62.8 $_{-1.8}$	−0.07 $_{\text{point}}$‡
	AMBER	*Unedited*	53.80	12.93	61.4	0.00 [−.07, .00]
		VCD $\alpha{=}1$	61.20 $_{+7.40}$	16.44 $_{+3.51}$	59.4 $_{-2.0}$	**−0.38** [−.58, −.25]
		PAI $\alpha{=}0.2$	54.00 $_{+0.20}$	13.05 $_{+0.12}$	61.5 $_{+0.1}$	−0.00 [−.14, +.13]
		Nullu $k{=}8$	43.40 $_{-10.40}$	11.72 $_{-1.21}$	51.8 $_{-9.6}$	**−0.41** [−.69, −.23]
		OPERA $\alpha{=}1, r{=}15$	58.60 $_{+4.80}$	14.13 $_{+1.19}$	61.1 $_{-0.3}$	**−0.10** [−.21, −.03]
		DeCo $\alpha{=}0.6$	50.80 $_{-3.00}$	13.13 $_{+0.20}$	57.8 $_{-3.6}$	−0.17 $_{\text{point}}$‡

Lowering the assertion knob will lower assertion (equivalently coverage), but beyond a certain point may not improve CHAIR score, as seen in Tab. 6 (App. A). For instance, the logit-bias knob reaches its lowest CHAIR at $\delta = -3$ (CHAIR_i 0.0396, CHAIR_s 0.080) and gets worse at $\delta = -4$ (0.0617, 0.098), while coverage falls monotonically, to 0.4673 at $\delta = -4$. We have 18 rows that can be scored, shown with 95% confidence intervals computed using bootstrap (Fig. 4). We also have 2 rows that are refused. The full audit can be seen in Tab. 4 in App. A.

SCAR scores (H1). In the LLaVA-1.5-7B model, our threshold is to beat the SCAR score $+0.136$, that is, the score we can get with the knob. Among them PAI did best. It scored $+0.044$, $+0.092$ and $+0.057$ across three different configuration. In its published configuration it scores $+0.092$ that is well under the free knob threshold. Nullu scored -0.407, -0.170, -0.097, and $+0.036$ at $k = 1, 2, 4, 8$. OPERA (-0.16) and DeCo (-0.31) both fall below the frontier with their whole

Table 3: **POPE on COCO, pooled over the random, popular and adversarial splits**. Acc., F1, yes-rate and ε_{acc} are in percentage points. Subscripts show changes from the same model's unedited row, and for ε_{acc} (Eq. 6) its 95% paired image-bootstrap interval. In the unedited rows, the ε_{acc} cell is the gain from re-thresholding the unedited model alone. **Bold** indicates best in its column within each model; for ε_{acc}, interval above 0. * indicates interval below 0. † indicates published setting. ‡OPERA acts only while decoding, so its answer is read from the generated text with its own yes/no rule; on LLaVA its first decoding step is unchanged, so its AUROC equals the unedited model's. §Partial run: 2,466 random/popular-split questions, compared to the unedited model.

Model	Method		Acc. ↑	F1 ↑	Yes (%)	AUROC ↑	ε_{acc} [95% CI]
LLaVA-1.5-7B	*Unedited*		82.26	83.45	57.2	0.9250	$\leq +2.88_{\text{free}}$
	VCD	α=1†	80.59 $_{-1.67}$	82.14 $_{-1.30}$	58.7	0.9033 $_{-.0217}$	$-1.18^*_{[-1.67,\,-.72]}$
	PAI	α=0.5†	83.33 $_{+1.08}$	84.26 $_{+0.82}$	55.9	0.9140 $_{-.0110}$	$\mathbf{+0.76}_{[+.48,\,+1.09]}$
	Nullu	k=1	82.24 $_{-0.01}$	83.41 $_{-0.03}$	57.0	**0.9257** $_{+.0007}$	$+0.00_{[-.11,\,+.11]}$
		k=2	83.19 $_{+0.93}$	83.94 $_{+0.49}$	54.7	0.9255 $_{+.0005}$	$\mathbf{+0.16}_{[+.02,\,+.25]}$
		k=4†	83.23 $_{+0.98}$	83.92 $_{+0.48}$	54.3	0.9253 $_{+.0003}$	$-0.04_{[-.11,\,+.07]}$
		k=8	**85.00** $_{+2.74}$	**84.68** $_{+1.23}$	47.9	0.9249 $_{-.0001}$	$+0.07_{[-.07,\,+.18]}$
	OPERA‡		83.94 $_{+1.69}$	84.40 $_{+0.95}$	52.9	0.9250 $_{+.0000}$	$+0.31_{[-.02,\,+.58]}$
MiniGPT-4	*Unedited*		50.13	66.73	99.9	0.7663	$\leq +19.94_{\text{free}}$
	Nullu	k=8†	**52.38** $_{+2.24}$	**67.68** $_{+0.95}$	97.3	**0.7729** $_{+.0066}$	$+0.02_{[.00,\,+.07]}$
	OPERA‡§		50.85 $_{+0.57}$	67.01 $_{+0.22}$	99.0	0.8215 $_{-.0316}$	$-0.16_{[-.41,\,.00]}$

interval. The other three planes read the same way (Tab. 2): the best published row scores $+0.18$ on MiniGPT-4 COCO (PAI), $+0.03$ on LLaVA AMBER (ProjectAway) and -0.00 on MiniGPT-4 AMBER (PAI), against bands of $+0.19$, $+0.26$ and $+0.32$. No published row clears its band on any of the four planes; of the 5 OPERA and DeCo rows with an interval, 4 lie wholly below 0.

POPE: discrimination does not move, accuracy does. When we audit Nullu, AUROC shows no significant change. Relative to the baseline of 0.9250, the changes for $k = 1, 2, 4, 8$ are respectively $+0.0007$, $+0.0005$, $+0.0003$, and -0.0001 (Tab. 3). On the other hand, in the same runs, the measured accuracy increases by using the published decision rule. Specifically, accuracy rises by 0.98 points at k=4 and by 2.74 points at k=8, increasing from 82.26% to 85.00%.

In POPE, yes and no responses are equal. The unedited model initially gives 57.19% yes answers, decreasing to 47.91% after the edit. By simply raising the unedited model's confidence threshold to match the edited model's "yes" rate, we recover 97.6% of the largest reported accuracy gain (Nullu, k=8), and 100% at the published rank. Notably, at k=4, the edited model's accuracy (83.23%) almost perfectly matches the re-thresholded unedited model (83.28%).

On MiniGPT-4, the yes-bias is stronger (99.87% yes), yielding a chance-level 50.13% accuracy. Re-thresholding it to its optimal threshold yields 70.08% (a 19.94-point gain). In contrast, Nullu (k=8) gains only 2.24 points at the published rule and just 0.74 points at its best threshold (Tab. 18). (These posterior-based accuracies differ from generation-based numbers in Yang et al. (2025)). Across methods, only PAI ($\varepsilon_{\text{acc}} = +0.76$) and Nullu at k=2 ($+0.16$) maintain gains beyond re-thresholding. VCD falls below it (-1.18), and OPERA's interval ($+0.31$) includes zero on LLaVA.

A second model and a second image set. On MiniGPT-4, frontier vertices match LLaVA's knob settings (Fig. 3b). Nullu at its published rank reproduces its CHAIR (20.93 ± 1.67 / 9.11 ± 0.69 over three draws, vs. 21.40 ± 1.00 / 8.99 ± 0.36), though our unedited model hallucinates less than reported (App. C). It does not clear the frontier: $S = -0.102$ $[-0.401, +0.094]$, $\Pr[S \leq \text{band}] = 0.991$, and a placebo at its coverage reproduces $\geq 73\%$ of its CHAIR_i reduction (σ lower bound). A $\delta = -0.5$ logit bias reaches similar CHAIR (21.2 / 8.80 vs. 20.4 / 8.63), adding 2.1 coverage points. At k=4, Nullu loses coverage without lowering CHAIR_i ($S = -0.527$). On AMBER (337 object types, custom scorer), no method clears the band ($\Pr[S \leq \text{band}] \geq 0.992$), with VCD and Nullu below the frontier (Fig. 3c). PAI lowers AMBER's CHAIR ($4.00 \rightarrow 3.75$), but hallucinated mentions rise ($111 \rightarrow 115$): it names 10% more objects while covering fewer presents.

Can SCAR detect a real gain? Every row above is a null, so we check that SCAR can also return a positive. The oracle control is the same operator as π_{bias} ($\delta = -4$), applied only to the COCO categories absent from the image's annotation; the permuted control uses the annotation of another image, so it biases about the same number of tokens (448 of 532 on average) without knowing the image. Scored against the frozen LLaVA frontier (Tab. 2, Figs. 3a and 4), the oracle clears the band with its whole interval ($S = +0.638$ $[+0.474, +0.747]$, $\Pr[S \leq \text{band}] = 0.003$): it removes 69.5% of the hallucinated mentions and keeps coverage (0.6955 against 0.6909). The permuted control also suppresses objects that are present, loses 18 points of coverage and falls below the frontier ($S = -1.207$); the paired difference is $+1.845$ $[+1.201, +3.411]$. The oracle reads the answer key, so it is not a method, but a gain of this kind is visible to SCAR.

6 Discussion

S is the headline scalar; σ is not. Our pre-registered primary metric σ is informative when mitigation achieves an audit, but it saturates for null audits and is undefined if the method reports a null result (for example, 0% improvement). Thus we recommend the SCAR score S as a summary across all rows. We recommend presenting it with the CI and the probability of being non-positive; a point estimate such as $S = -0.10$ would overclaim the mitigation's benefit, since the CI includes 0. At the published rank, at least three-quarters of the reported mitigation is attainable by the placebo, and all of it if the placebo is permitted to make more claims. This is robust to a harder object set and an alternate coverage definition (App. A). Overall, none of the 20 configurations we test on LLaVA-1.5 (from six published methods) come close to the top of the null band; that is, we do not observe mitigation of hallucination beyond placebo effects in our empirical analysis.

How should researchers use SCAR? SCAR's placebo frontier is a resource-intensive computation, taking roughly 22 GPU-hours to construct, fortunately only once, but it reveals effects not apparent from common metrics. The frontier depends on the model, the prompt and the token budget; 8 settings were enough on a new model or image set (Sec. 4), and we release every frontier we computed. We hope that SCAR will serve as one interpretability tool among many to understand models' strengths and weaknesses, and help researchers avoid overstating mitigation methods' effects when comparing two models' performance at arbitrary, uncalibrated operating points.

6.1 Limitations

Our empirical analysis covers six published methods, examining the published metric at each method's published operating point, and, for four, at nearby strengths to observe dose-response curves. Larger or other models and detector-guided methods such as MARINE (Zhao et al., 2025) are future work. Our confidence intervals perturb the images, not seeds; cross-GPU noise provides the uncertainty interval, while our method is reproducible in producing bitwise-identical results across runs on the same GPU. Except for three temperature settings, listed in Tab. 6, results are bitwise-identical between runs, and these exceptions do not substantively change our results. Analysis of alternative definitions of coverage is limited by CHAIR's 80 categories, which are less semantically rich than OpenCHAIR (Ben-Kish et al., 2024); however, AMBER's 337 objects and an open-vocabulary re-score of our captions (Tab. 19) give the same verdict. SCAR thresholds the first answer-token probability at 0.5, unlike baselines that implicitly threshold by argmax; Tab. 18 varies that threshold. Finally, Π is not exhaustive: a sixth knob would push F down and every ε toward zero, so negative verdicts cannot be artifacts of too small a family, though a positive ε could be.

7 Conclusion

Commonly used benchmarks for mitigating hallucination are sensitive to the strengths of mitigation, which leads to misleading claims of significant excision of hallucination at arbitrary strengths. SCAR introduces a new evaluation protocol that constructs a reference frontier of content-free placebos computed from the unedited model, enabling researchers to distinguish improvements over this frontier. Our experiments on two models and two image sets demonstrate that reported gains over this frontier are small or nonexistent, while a control with real image information clears it.

AI USE STATEMENT

Generative AI was used to aid and polish writing: to rephrase sentences for clarity and correct grammatical errors. Generative AI was also used to produce Python code to create plots and format tables based on results and measurements that we took. Generative AI was also used for retrieval and discovery: to suggest literature search terms, and to double-check that cited publications existed by querying the arXiv API. Generative AI was not used to propose research questions or ideas, design our pre-registration, perform experiments, generate data, or prove mathematical claims. We validated all content produced by the generative AI tools, including written text and Python code, using primary sources (publications) and results that we recorded. We take responsibility for the content of this paper.

ETHICS STATEMENT

This paper audits previously published methods for reducing hallucinations using publicly available datasets. We did not use personal or private data, involve human subjects, or include newly released model weights. Generated image captions were created using public model checkpoints on the public validation set of MS-COCO, and question-answering results were created using public model checkpoints on the publicly available POPE benchmark and public images from MS-COCO and AMBER. While this paper reports a null result for these six methods at fixed coverage, we do not claim that these methods are incorrect.

REPRODUCIBILITY STATEMENT

Every value across this paper comes from one protocol described in Sec. 4, performed using a random seed of 0, beam search with beam width 3, a max number of new tokens of 64, 500 validation set images from MS-COCO. Tab. 1 presents the placebo family and its configurations, Tabs. 4 and 5 display evaluated rows, and Tab. 6 contains all measured placebo entries. The MiniGPT-4 and AMBER planes use the same protocol with the changes listed in Sec. 4, and their cells are listed in App. C and App. D; the algorithms are given in App. B. Recomputing the frontier takes a one-off cost of $\sim$22 GPU-hours, or about 2 GPU-hours for an 8-cell frontier. All further calculations, such as producing numbers in the appendix and robustness checks in Tab. 7, can be computed on CPU using our saved captions and question-answering results. Code and saved results will be made publicly available.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8947–8956. IEEE, 2019.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. Mitigating open-vocabulary caption hallucinations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. arXiv:2312.03631.

Arnav Bendre, Guneesh Gupta, Kavish Grover, Chayan Aggarwal, and Shreyansh Modi. Rethinking CD: A reproducibility study and extension on the ineffectiveness of contrastive decoding at mitigating object hallucinations in MLLMs. *arXiv preprint arXiv:2607.25196*, 2026. Submitted to TMLR.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, May 2010. ISSN 1532-4435. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427, 2024.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702, 2019. doi: 10.1109/CVPR.2019.00686.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=rsZwwjYHuD.

Nicholas Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=94kQgWXojH.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering (CAICE 2024)*, New York, NY, USA, January 2024. Association for Computing Machinery. doi: 10.1145/3672758.3672824. URL https://dl.acm.org/doi/10.1145/3672758.3672824.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, 2024a.

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. arXiv:2407.21771.

Zhipeng Liu, Haochen Wang, and Zhaoxiang Zhang. CAPEval: A decoupled caption evaluation across understanding and generation. *arXiv preprint arXiv:2608.02589*, 2026.

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. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 342–357, 2024.

Haoyi Qiu, Wenbo Hu, Zi-Yi Dou, and Nanyun Peng. VALOR-EVAL: Holistic coverage and faithfulness evaluation of large vision-language models. In *Findings of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2404.13874.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7237–7256, 2020.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. Linear adversarial concept erasure. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18400–18421. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ravfogel22a.html.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4035–4045, 2018.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. MLLM can see? dynamic correction decoding for hallucination mitigation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4z3IguA4Zg.

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. *arXiv preprint arXiv:2311.07397*, 2023.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via HalluSpace projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14635–14645, 2025.

Hao Yin, Guangzong Si, and Zilei Wang. The mirage of performance gains: Why contrastive decoding fails to mitigate object hallucinations in MLLMs? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=wGeQr5cARm.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13807–13816, June 2024.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via image-grounded guidance. In *International Conference on Machine Learning (ICML)*, 2025.

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.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1tZbq88f27.

A SUPPLEMENTARY MEASUREMENTS

This appendix reports the measurements the main text summarises. Coverage is computed per image, over the distinct annotated objects a caption names, and then averaged over images; Eq. 1 pooled over all images is the alternate definition of Tab. 7. Tab. 6 lists the whole placebo family cell by cell, together with the leave-one-out score each cell receives when it is scored exactly the way an audited method is scored; the upper edge of that band, $+0.136$, is the threshold quoted in Sec. 5. The weight-level placebo $\pi_{\mathrm{rand}}(k)$ of Sec. 3 folds a random orthogonal rank-k subspace into the same layers the audited operator edits, and is run in two arms: $k \in \{1, 4, 8\}$ matches three of the four audited ranks, while $k = 16$ and $k = 28$ match the Frobenius distortion of the rank-1 and rank-4 edits. It contributes no vertex to the envelope. Tab. 7 re-scores the audit under a harder object vocabulary, under an alternate coverage definition, and under deletion of single knob families, all on CPU from the banked per-caption records: dropping π_{rand} leaves the envelope unchanged, dropping π_{temp} moves the rank-1 excision gain from -0.0224 to -0.0209, and no row changes the sign of its score. The primary frontier F is the lower convex hull of the cells ("hull" in Tab. 15); randomizing between adjacent hull vertices realizes each chord to within 0.00023 CHAIR_i over the coverage interval the audited rows occupy, two orders below the gaps read there. Tab. 8 reports the operator cost ρ of Sec. B: the published rank-4 fold pays 48.6 against 5.89 for a random projection of the same rank, an $8.3\times$ whitened surcharge, on an activation covariance whose stable rank is 10.9 to 29.8 out of 4096. Linear probes at the edited site read 0.9161 before the fold and 0.9159 after it, against a positive control of 0.0766 from removing one probe-aligned direction, so the concept stays linearly decodable. Tab. 9 reports what moves POPE accuracy. Within every family it is the shift in yes-bias, not the change in ranking; the pooled accuracy/AUROC correlation reverses sign against all three family clouds and is quoted only as the aggregation artefact it is. The same table gives the rank agreement between the SCAR score and the reported $\Delta\mathrm{CHAIR}_i$, which is the evidence on **H2**: inside one operator family the dose-response is monotone in both at once, so these rows cannot separate the two orderings. Tabs. 4 and 5 give the full audit of the 18 rows and the 2 refused rows, and Fig. 4 places them on the whole plane.

Table 4: **The full audit, part 1: operating points**, on one protocol: 500 COCO images, beam 3, 64 tokens, seed 0, one RTX 4090. $r_F(c)$ is the frontier's risk at the row's own coverage. †published setting; *greedy*, *sampling*, *beam 5* are non-primary decoding arms. [b]refused rows keep their operating point. The scores of the same rows are in Tab. 5.

		operating point				
family	setting	c	CHAIR_i	CHAIR_s	ment.	$r_F(c)$
unedited	—	0.6909	0.0626	0.192	2091	—
Nullu (Yang et al., 2025)	k=1	0.7033	0.0776	0.228	2230	0.0552
	k=2	0.6934	0.0602	0.168	2193	0.0514
	k=4†	0.6833	0.0530	0.146	2151	0.0483
	k=8	0.6918	0.0491	0.138	2158	0.0509
PAI (Liu et al., 2024b)	α=0.2	0.7086	0.0549	0.164	2148	0.0575
	α=0.5†	0.6807	0.0431	0.122	2295	0.0475
	α=0.5, greedy	0.6811	0.0449	0.120	2360	0.0476
ProjectAway (Jiang et al., 2025)	α=0.5	0.6905	0.0552	0.174	2065	0.0506
	α=1.5	0.6940	0.0527	0.166	2068	0.0516
	α=2.5	0.6902	0.0579	0.174	2072	0.0505
	α=3.5†	0.6898	0.0573	0.170	2075	0.0503
	α=1.5, beam 5	0.6997	0.0562	0.158	2119	0.0536
	α=3.5, beam 5	0.6947	0.0599	0.170	2120	0.0518
VCD (Leng et al., 2024)	α=0.5	0.6982	0.0630	0.188	2128	0.0529
	α=1†	0.6945	0.0711	0.220	2110	0.0518
	α=2	0.7004	0.0749	0.206	2070	0.0539
	α=3	0.6950	0.0883	0.242	2060	0.0519
	α=1, sampling	0.6898	0.0786	0.218	2009	0.0503
PAI (Liu et al., 2024b)	α=0.75	0.0298	0.0476	0.004	42	refused[b]
PAI (Liu et al., 2024b)	α=1	0.0000	1.0000	0.008	4	refused[b]

Table 5: **The full audit, part 2: scores** of the rows of Tab. 4, with 2,000 bootstrap resamples of the shared images. $\varepsilon = r_F(c) - \mathrm{CHAIR}_i$, $S = \varepsilon / r_F(c)$. †published setting. [a]undefined where a row reduces no CHAIR_i; undefined fraction 0.000 throughout. [b]refused: coverage outside the swept span $[0.4323, 0.8122]$, so not scored; refused rows keep their operating point (Tab. 4). **Bold**: best row per family. At its identity setting (α=0), PAI changes 18 of 100 captions (fp16 kernel drift); VCD and ProjectAway change none.

		excision gain		**SCAR score**	suppression
family	setting	ε [95% CI]	$P(\varepsilon \le 0)$	S [95% CI]	σ [95% CI]
Nullu	k=1	−0.0224 [−0.0375, −0.0113]	1.000	−0.407 [−0.809, −0.190]	undef.[a]
	k=2	−0.0088 [−0.0225, −0.0004]	0.978	−0.170 [−0.510, −0.007]	1.000 [0.906, 1.000]
	k=4†	−0.0047 [−0.0185, +0.0039]	0.878	−0.097 [−0.444, +0.084]	1.000 [0.757, 1.000]
	k=8	+0.0018 [−0.0137, +0.0122]	0.548	**+0.036** [−0.314, +0.228]	0.866 [0.395, 1.000]
PAI	α=0.2	+0.0026 [−0.0107, +0.0139]	0.462	+0.044 [−0.218, +0.225]	0.669 [0.000, 1.000]
	α=0.5†	+0.0044 [−0.0086, +0.0129]	0.323	**+0.092** [−0.210, +0.273]	0.777 [0.481, 1.000]
	α=0.5, greedy	+0.0027 [−0.0105, +0.0116]	0.446	+0.057 [−0.257, +0.239]	0.847 [0.496, 1.000]
ProjectAway	α=0.5	−0.0047 [−0.0166, +0.0005]	0.964	−0.092 [−0.376, +0.010]	1.000 [0.952, 1.000]
	α=1.5	−0.0011 [−0.0134, +0.0045]	0.793	**−0.021** [−0.290, +0.088]	1.000 [0.603, 1.000]
	α=2.5	−0.0075 [−0.0199, −0.0020]	0.994	−0.148 [−0.458, −0.040]	1.000 [1.000, 1.000]
	α=3.5†	−0.0070 [−0.0198, −0.0010]	0.990	−0.140 [−0.453, −0.022]	1.000 [1.000, 1.000]
	α=1.5, beam 5	−0.0026 [−0.0159, +0.0052]	0.820	−0.048 [−0.346, +0.095]	1.000 [0.331, 1.000]
	α=3.5, beam 5	−0.0081 [−0.0213, −0.0004]	0.983	−0.156 [−0.473, −0.008]	1.000 [1.000, 1.000]
VCD	α=0.5	−0.0100 [−0.0248, −0.0004]	0.980	**−0.190** [−0.554, −0.008]	undef.[a]
	α=1†	−0.0193 [−0.0346, −0.0096]	1.000	−0.373 [−0.823, −0.176]	undef.[a]
	α=2	−0.0210 [−0.0392, −0.0078]	0.997	−0.390 [−0.874, −0.130]	undef.[a]
	α=3	−0.0364 [−0.0541, −0.0236]	1.000	−0.701 [−1.276, −0.415]	undef.[a]
	α=1, sampling	−0.0283 [−0.0460, −0.0151]	1.000	−0.562 [−1.127, −0.275]	undef.[a]
PAI	α=0.75	**refused[b]:** not scored			
PAI	α=1	**refused[b]:** not scored			

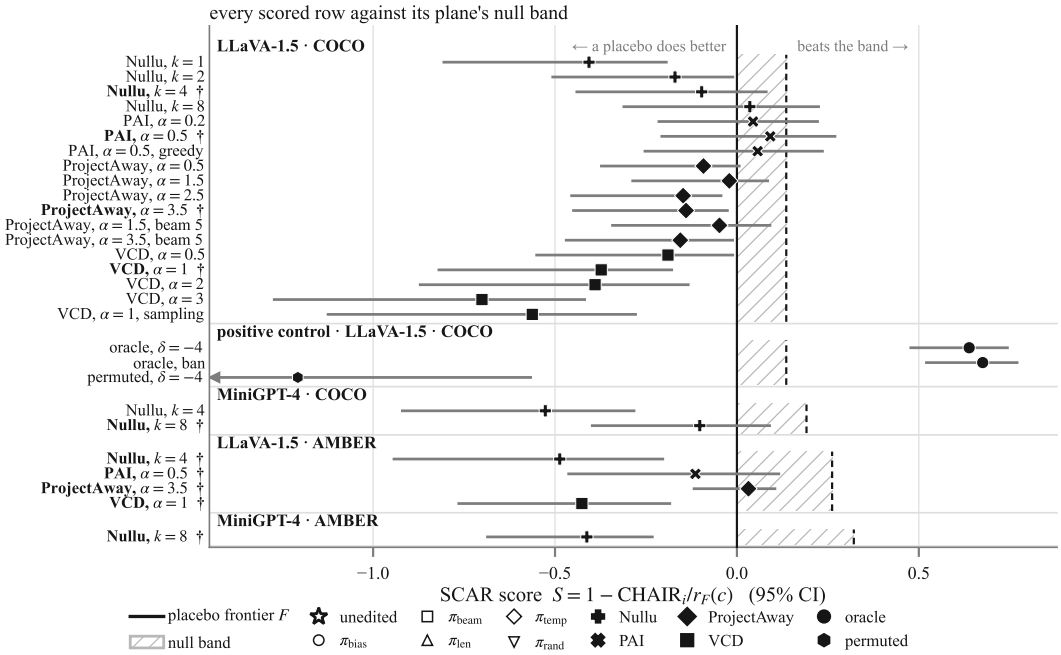


Figure 4: **The audit on four planes.** S with 95% intervals for the 28 scored rows of Tabs. 5 and 16, each against its own plane's band; the OPERA, DeCo and MiniGPT-4 VCD and PAI rows of Tab. 2 are not drawn. Bold labels are published settings. Only the oracle clears its band. Beam 3, 64 tokens, seed 0, one RTX 4090.

Table 6: The 25 distinct placebo cells, ordered by coverage. S (LOO) scores each cell against the frontier of the other 24. The two span endpoints are not scored, since the frontier is never extrapolated. Across the family CHAIR_s varies by $18.9\times$ and CHAIR_i by $5.6\times$ with no visual information added anywhere.

family	setting	c	CHAIR_i	CHAIR_s	ment.	S (LOO)
$\pi_{\mathrm{len}}(L)$	128	0.8122	0.1391	0.5280	3745	—
$\pi_{\mathrm{temp}}(\tau)$	0.5	0.7100	0.0581	0.1880	2151	$+0.059$
$\pi_{\mathrm{beam}}(b)$	1	0.7047	0.0637	0.1860	2165	-0.143
$\pi_{\mathrm{rand}}(k)$	28	0.7028	0.0603	0.1840	2091	-0.097
$\pi_{\mathrm{beam}}(b)$	5	0.6991	0.0609	0.1780	2136	-0.142
$\pi_{\mathrm{beam}}(b)$	2	0.6984	0.0530	0.1620	2095	$+0.017$
$\pi_{\mathrm{beam}}(b)$	8	0.6972	0.0613	0.1700	2103	-0.166
$\pi_{\mathrm{rand}}(k)$	16	0.6952	0.0548	0.1620	2115	-0.055
$\pi_{\mathrm{rand}}(k)$	4	0.6933	0.0592	0.1860	2110	-0.153
unedited	δ=0, beam 3	0.6909	0.0626	0.1920	2091	-0.236
$\pi_{\mathrm{rand}}(k)$	8	0.6908	0.0605	0.1940	2115	-0.195
$\pi_{\mathrm{temp}}(\tau)$	1.0	0.6883	0.0746	0.2340	2023	-0.497
$\pi_{\mathrm{rand}}(k)$	1	0.6877	0.0568	0.1780	2078	-0.143
$\pi_{\mathrm{bias}}(\delta)$	-0.25	0.6837	0.0589	0.1620	2038	-0.215
$\pi_{\mathrm{bias}}(\delta)$	-0.5	0.6800	0.0473	0.1400	1946	$+0.048$
$\pi_{\mathrm{bias}}(\delta)$	-0.75	0.6762	0.0495	0.1400	1898	-0.061
$\pi_{\mathrm{bias}}(\delta)$	-1	0.6595	0.0486	0.1340	1769	-0.102
$\pi_{\mathrm{bias}}(\delta)$	-1.25	0.6469	0.0488	0.1340	1723	-0.156
$\pi_{\mathrm{bias}}(\delta)$	-1.5	0.6274	0.0508	0.1280	1633	-0.297
$\pi_{\mathrm{bias}}(\delta)$	-2	0.6100	0.0401	0.0980	1470	-0.100
$\pi_{\mathrm{len}}(L)$	32	0.5889	0.0332	0.0620	1173	$\mathbf{+0.136}$
$\pi_{\mathrm{bias}}(\delta)$	-3	0.5500	0.0396	0.0800	1211	-0.260
$\pi_{\mathrm{temp}}(\tau)$	1.5	0.5087	0.1459	0.2880	1316	-3.932
$\pi_{\mathrm{bias}}(\delta)$	-4	0.4673	0.0617	0.0980	973	-1.227
$\pi_{\mathrm{len}}(L)$	16	0.4323	0.0261	0.0280	575	—

Table 7: *Left:* the audit re-scored under a harder object vocabulary and under an alternate coverage definition. *band* is the leave-one-out band edge under that variant, *best row* the highest S of the 18 audited rows, and *flips* how many rows change the sign of S. The one flip moves against the method. *Right:* the excision gain ε at each erased rank when a knob family is deleted from Π. [d]The piecewise interpolant through the placebo points is excluded from the primary reading: it is defined only where the swept family is monotone, and the scoring region reaches into the non-monotone stretch of Sec. 5.

variant	c span	band	best S	>band	flips
primary (Tab. 4)	0.432–0.812	$+0.136$	$+0.092$	**0**/18	—
80 names, no synonyms	0.288–0.669	$+0.235$	$+0.170$	**0**/18	1
pooled, not per-image, recall	0.331–0.761	$+0.097$	$+0.072$	**0**/18	0

Π variant	k=1	k=2	k=4	k=8
full (25 cells)	$-.0224$	$-.0088$	$-.0047$	$+.0018$
no π_{rand} (20)	$-.0224$	$-.0088$	$-.0047$	$+.0018$
no π_{temp} (22)	$-.0209$	$-.0088$	$-.0047$	$+.0018$
interpolant[d]	$-.0165$	$-.0012$	$+.0047$	$+.0123$

Table 8: Operator cost against audit outcome. [a]σ is undefined where a row reduces no CHAIR_i. [b]In the banked k=8 checkpoint, $W - W'$ has rank 4 and equals the k=4 fold, so its rank-8 subspace cannot be read from the weights. [c]VCD and PAI act at decoding time and have no linear operator; ProjectAway's scaled projection $I - \alpha\hat{u}\hat{u}^\top$ has no kernel at $\alpha \neq 1$ (App. B). Both give $\rho = \infty$ by definition. The three identifiable pairs are rank-concordant — the costlier the eraser, the worse it scores — which is a direction, not a correlation, and leaves **H3** untestable against this literature.

operator	ρ (L16–29)	ρ random	σ	S [95% CI]
Nullu $k = 1$	90.2	6.05	undef.[a]	-0.407 $[-0.809, -0.190]$
Nullu $k = 2$	56.7	6.02	1.000	-0.170 $[-0.510, -0.007]$
Nullu $k = 4$	48.6	5.89	1.000	-0.097 $[-0.444, +0.084]$
Nullu $k = 8$	—[b]	—	0.866	$+0.036$ $[-0.314, +0.228]$
PAI (3 rows), ProjectAway (6 rows), VCD (5 rows): $\rho = \infty$[c]				

Table 9: What moves POPE accuracy, and how the SCAR score ranks against the reported margins. Pearson r on the left, two-sided permutation p from 10,000 shuffles; Kendall τ on the right, over the audited rows of each family. ProjectAway has no POPE rows, since it targets the objects a first-pass caption mentions with low confidence. [e]The pooled accuracy/AUROC figure is an aggregation reversal: it is negative in all three families and positive only when they are pooled, so only the within-family reading may be quoted.

		POPE, per-split rows		Kendall τ vs. S	
family	n	Δacc vs. yes-bias↓	Δacc vs. ΔAUROC	$\Delta\mathrm{CHAIR}_i$	$\Delta\mathrm{CHAIR}_s$
Nullu	12	$+0.897$ (p=.0002)	-0.695 (p=.0101)	$+1.000$	$+1.000$
PAI	6	$+0.968$ (p=.0045)	-0.406 (p=.3908)	$+1.000$	$+0.333$
ProjectAway	—	—	—	$+0.867$	$+0.358$
VCD	6	$+0.955$ (p=.0081)	-0.083 (p=.8737)	$+1.000$	$+0.600$
pooled	24	$+0.837$ (p=.0001)	$+0.622$[e] (p=.0009)	$+0.922$	$+0.757$

B ALGORITHMS AND THE OPERATOR COST ρ

ρ: what an operator costs, before generating anything. Finally, when comparing interventions, orthogonal projectors may unfairly receive penalties for obliqueness. Removing a concept requires cutting across the space, including high-variance directions of activation covariance Σ_{XX} and destroying more than necessary. Given a kernel k, the minimally-distorting linear map with that kernel is an oblique projector, $P^* = I - ua^\intercal$. We can compute this from the covariance of activations. Given this, we define a score

$$\rho(P) = \mathcal{D}(P)/\mathcal{D}(P^*), \tag{7}$$

where $\mathcal{D}(P)$ is the distortion from intervention by P, which removes k dimensions. The score, which will be ≥ 1, can be computed as $\mathcal{D}(P)/k$ and is entirely determined by token moments banked before generation. An operator with no kernel erases nothing while still paying $\mathcal{D} > 0$, so we set $\rho = \infty$; a decoding-time method has no linear operator at all, and a scaled projection $I - \alpha\hat{u}\hat{u}^\top$ has a kernel only at $\alpha = 1$.

Alg. 1 gives the computation. Σ_{XX} is the covariance of the activations at the edited site (the output of `down_proj` in layers 16–31), accumulated in float64 over the token activations of the 1,000 LURE captions (500 pairs) that Nullu uses to build its edit; no shrinkage is applied. For Nullu the erased subspace is read from the checkpoints, since the edit folds $W' = (I - P)W$ and so $W - W'$ spans it. We average ρ over layers 16–29, where the condition number of Σ_{XX} is between 1.1×10^4 and 5.2×10^4, and report layers 30 and 31 apart: their stable rank is 1.05 and 2.58, so whitening there amplifies a near-null direction.

Algorithm 1 Operator cost ρ of a linear edit.

Require: raw and edited `down_proj` weights W_ℓ, W'_ℓ; token activations X_ℓ at that site; rank k

for each edited layer ℓ **do**
  $\Sigma_{XX} \leftarrow \mathrm{cov}(X_\ell)$ ▷ float64, Cholesky factor kept
  $U \leftarrow$ top-k left singular vectors of $W_\ell - W'_\ell$ ▷ the erased subspace
  write the edit as $P = I + QGQ^\top$ with Q orthonormal ($P = I - UU^\top$ for Nullu)
  $\mathcal{D}(P) \leftarrow \mathrm{tr}\left((P-I)\,\Sigma_{XX}\,(P-I)^\top \Sigma_{XX}^{-1}\right) = \mathrm{tr}\left(G\,Q^\top \Sigma_{XX} Q\, G\, Q^\top \Sigma_{XX}^{-1} Q\right)$
  $P^* \leftarrow I - U\left(\Sigma_{XX}^{-1} U (U^\top \Sigma_{XX}^{-1} U)^{-1}\right)^\top$, for which $\mathcal{D}(P^*) = k$
  $\rho_\ell \leftarrow \mathcal{D}(P)/k$ ▷ $\rho_\ell = \infty$ if $k = 0$
end for
return the mean of ρ_ℓ over layers 16–29; layers 30–31 are reported apart

Algorithm 2 Weight-level placebo $\pi_{\mathrm{rand}}(k)$.

Require: language-model `down_proj` weights W_ℓ for $\ell \in \{16, \ldots, 31\}$ (the layers Nullu edits); rank k; seed s

for each ℓ **do**
  draw $Z \in \mathbb{R}^{4096\times k}$ with i.i.d. $\mathcal{N}(0,1)$ entries, seed $1000\,s + \ell$
  $Q \leftarrow \mathrm{QR}(Z)$ ▷ a uniformly random orthonormal basis
  $W_\ell \leftarrow (I - QQ^\top)\,W_\ell$ ▷ in memory; no image or caption is used
end for

The rank-matched arm uses $k \in \{1, 4, 8\}$. The distortion-matched arm uses the rank at which a random projection has the same Frobenius distortion $\mathbb{E}\|(P-I)x\|^2$ as Nullu, averaged over layers 16–29: $k = 16$ for Nullu k=1 and $k = 28$ for Nullu k=4.

Algorithm 3 Restricted posterior, ΔAUROC and $\varepsilon_{\mathrm{acc}}$ on POPE.

Require: model M, unedited model R, POPE questions $q_1, \ldots, q_n$ with labels y_i

for each question q_i and each model **do**
  one forward pass; $p \leftarrow$ softmax of the logits at the first answer position
  $s_i \leftarrow p(\mathcal{Y})/\big(p(\mathcal{Y}) + p(\mathcal{N})\big)$ ▷ $\mathcal{Y}, \mathcal{N}$: first sub-tokens of `Yes`, `yes` and `No`, `no`
end for
$\Delta\mathrm{AUROC} \leftarrow \mathrm{AUROC}(s^M, y) - \mathrm{AUROC}(s^R, y)$
$\mathrm{acc}(M) \leftarrow$ accuracy of the rule $s^M > 0.5$; $u \leftarrow$ its yes-rate
$\hat{\tau} \leftarrow$ the threshold at which R answers yes on $\mathrm{round}(u\,n)$ questions
$\varepsilon_{\mathrm{acc}} \leftarrow \mathrm{acc}(M) - \mathrm{acc}(s^R > \hat{\tau})$
bootstrap: resample images with all their questions, and recompute $\hat{\tau}$ in each replicate

Tab. 18 also reports the reverse direction (the edited model re-thresholded to the unedited yes-rate), the best-threshold accuracy of each score vector, and accuracy at 0.5 after a 5-fold cross-fitted isotonic calibration.

C MiniGPT-4

MiniGPT-4 uses LLaMA-2-7B-chat and Nullu's published harness. The k=8 edit follows the published setting (layers 16–31, 5,000 LURE pairs); rebuilding the activations and the edit on a second machine reproduced all 18,000 POPE scores byte for byte. Tab. 10 lists the frontier cells, and Tab. 11 repeats the unedited model and both ranks on two more image draws (seeds 42 and 123), since the published CHAIR is a mean over three draws. Our edited row and BLEU match the published ones, but our unedited CHAIR is lower. On POPE, the first answer token of MiniGPT-4 puts 0.796 of its mass on yes or no (LLaVA: 0.985), so the posterior is a weaker proxy for its generated answer; our accuracies at the published rule (50.37/50.03/50.00 by split) are not the generation-based ones of Yang et al. (2025) (64.33/56.63/55.17), while the yes-bias itself reproduces (recall 100.0 against 97.13).

Table 10: MiniGPT-4 on COCO: the 8 placebo cells and the two Nullu rows, draw 0. • marks a vertex of the frontier; the vertices are the same knob settings as on LLaVA (π_{len} 16, 32, 128; π_{beam} 2; π_{bias} -0.5). S scores each cell against the frontier of the other 7, as in Tab. 6; the largest is the band edge (**bold**).

family	setting	c	CHAIR_i	CHAIR_s	ment.	vertex	S
$\pi_{\text{len}}(L)$	128	0.6650	0.1370	0.348	2204	•	—
$\pi_{\text{beam}}(b)$	2	0.6477	0.0985	0.266	1767	•	$+0.067$
unedited	δ=0, beam 3	0.6464	0.1033	0.258	1771		-0.057
$\pi_{\text{beam}}(b)$	1	0.6431	0.1112	0.278	1663		-0.161
$\pi_{\text{bias}}(\delta)$	-0.5	0.6292	0.0880	0.212	1637	•	$+0.017$
$\pi_{\text{len}}(L)$	32	0.5583	0.0553	0.108	1121	•	$+\mathbf{0.191}$
$\pi_{\text{bias}}(\delta)$	-2	0.5415	0.0648	0.130	1219		-0.225
$\pi_{\text{len}}(L)$	16	0.4604	0.0413	0.052	653	•	—
Nullu	k=8	0.6082	0.0863	0.204	1750		-0.102
Nullu	k=4	0.6108	0.1214	0.308	1755		-0.527

Table 11: MiniGPT-4 over three image draws of 500 COCO images (mean $\pm$ sd), against the published rows of Yang et al. (2025). Bottom: pooled change against the unedited model.

	CHAIR_s		CHAIR_i		BLEU	coverage
arm	ours	Yang et al. (2025)	ours	Yang et al. (2025)	ours	ours
unedited, beam 3	26.20 ± 0.69	30.10 ± 0.30	10.73 ± 0.38	11.87 ± 0.37	15.16 ± 0.30	63.63 ± 0.94
Nullu k=8†	20.93 ± 1.67	21.40 ± 1.00	9.11 ± 0.69	8.99 ± 0.36	14.89 ± 0.12	60.28 ± 1.00
pooled change against the unedited model, 1,500 images, resampled within each draw						
	ΔCHAIR_s		ΔCHAIR_i		Δcoverage	
Nullu k=8†	-5.27 [−7.67, −2.93]		-1.62 [−2.69, −0.57]		-3.35 [−4.32, −2.35]	
Nullu k=4	$+1.93$ [−0.47, +4.20]		$+0.54$ [−0.47, +1.50]		-2.25 [−3.22, −1.30]	

D AMBER

We use 500 of AMBER's 1,004 generative images, its prompt, and its released scorer; our adapter reproduces that scorer's Cover, Hal and Cog on the unedited captions to 0.1 point. AMBER's CHAIR is hallucinated over all mentions, so it is our risk axis; for coverage we use the per-image mean grounded recall, as on COCO. At 64 tokens the unedited model scores CHAIR 4.00, below AMBER's published LLaVA-1.5 row (7.8, with untruncated responses (Wang et al., 2023)); our 128-token cell gives 7.93 / 49.61 / 37.4 / 4.58 against the published 7.8 / 51.0 / 36.4 / 4.2, so the gap is the token budget. MiniGPT-4 uses the same 500 images and the same 8 settings. Its unedited cell at 64 tokens gives CHAIR 12.93, close to AMBER's published MiniGPT-4 row (13.6). Its vertices are π_{len} 16, 32, 128 and the unedited cell, and the band edge is $+0.321$, set by π_{len} 32.

Table 12: 500 AMBER images: AMBER's own metrics and the SCAR coverage for the 8 placebo cells and the audited methods, LLaVA-1.5 (top) and MiniGPT-4 (bottom). • marks a frontier vertex; unlike COCO, the unedited cell is a vertex on both, and π_{bias} -2 on LLaVA-1.5. S of a cell is its leave-one-out score; the largest is the band edge (**bold**).

		AMBER's own metrics				SCAR plane			
family	setting	CHAIR	Cover	Hal	Cog	c	ment.	vertex	S
$\pi_{\text{len}}(L)$	128	7.93	49.61	37.4	4.58	0.5463	4060	•	—
$\pi_{\text{beam}}(b)$	1	4.50	45.97	18.4	1.72	0.5133	2912		-0.034
$\pi_{\text{beam}}(b)$	2	4.38	45.90	18.2	2.00	0.5131	2806		-0.012
unedited	δ=0, beam 3	4.00	45.58	17.2	1.59	0.5100	2778	•	$+0.070$
$\pi_{\text{bias}}(\delta)$	-0.5	4.02	43.79	16.6	1.68	0.4898	2562		-0.139
$\pi_{\text{bias}}(\delta)$	-2	2.49	39.62	8.8	0.86	0.4447	1967	•	$+\mathbf{0.262}$
$\pi_{\text{len}}(L)$	32	3.21	37.62	8.8	0.94	0.4276	1619		-0.304
$\pi_{\text{len}}(L)$	16	2.25	25.89	3.4	0.41	0.2997	801	•	—
VCD	α=1	5.55	45.19	22.2	1.84	0.5056	2773		-0.426
PAI	α=0.5	3.75	42.97	15.0	1.02	0.4827	3066		-0.114
ProjectAway	α=3.5	3.83	45.36	16.6	1.47	0.5083	2767		$+0.032$
Nullu	k=4	4.50	41.69	17.2	1.39	0.4681	2664		-0.487

		AMBER's own metrics				SCAR plane			
family	setting	CHAIR	Cover	Hal	Cog	c	ment.	vertex	S
$\pi_{\text{len}}(L)$	128	17.03	61.59	67.4	10.67	0.6555	4092	•	—
unedited	δ=0, beam 3	12.93	57.10	53.8	6.33	0.6141	3317	•	$+0.038$
$\pi_{\text{beam}}(b)$	2	13.31	57.28	56.2	6.91	0.6126	3358		-0.035
$\pi_{\text{beam}}(b)$	1	16.14	56.95	61.8	8.66	0.6113	3228		-0.261
$\pi_{\text{bias}}(\delta)$	-0.5	12.33	55.46	51.2	6.17	0.5987	3073		-0.012
$\pi_{\text{bias}}(\delta)$	-2	11.03	46.93	37.0	4.66	0.5059	2411		-0.432
$\pi_{\text{len}}(L)$	32	5.79	42.12	18.2	1.14	0.4663	1848	•	$+\mathbf{0.321}$
$\pi_{\text{len}}(L)$	16	5.16	30.42	10.0	0.94	0.3490	1007	•	—
Nullu	k=8	11.72	47.18	43.4	4.62	0.5182	2807		-0.412

E POSITIVE CONTROL

The oracle reads the absent categories from the same annotation CHAIR scores against, so it is an upper bound on the information a method could add, not a method. The permuted control takes the annotation of another image by a fixed derangement. With $\delta = 0$, the patched decoder reproduced the stored unedited captions byte for byte on 100 images. The pre-registered pass rule was that the oracle removes at least 60% of the unedited model's hallucinated mentions, the point from which its whole interval was forecast to clear the band.

Table 13: Positive control on LLaVA-1.5 and 500 COCO images, scored against the frozen 25-cell frontier. *suppressed*: share of the unedited model's 131 hallucinated mentions removed. S and $\Pr[S \leq \text{band}]$ are in Tab. 14, ε and σ in Tab. 16.

arm	δ	c	CHAIR_s	mentions	halluc.	suppressed [95% CI]
unedited	0	0.6909	0.192	2091	131	reference
oracle	δ=-4	0.6955	0.066	2124	40	69.5% [57.3, 79.5]
oracle	ban	0.6921	0.056	2114	35	73.3% [62.0, 82.7]
permuted	δ=-4	0.5104	0.126	1268	83	36.6% [15.1, 53.5]

F SCAR ROWS, SENSITIVITY, POWER AND OPEN VOCABULARY

Tab. 14 is the full version of Tab. 2: it adds the frontier's risk $r_F(c)$, $\Pr[S \leq \text{band}]$, and the non-published rows (MiniGPT-4 at $k{=}4$ and the oracle that bans absent categories outright). The OPERA and DeCo rows and the MiniGPT-4 VCD and PAI rows are reported in Tab. 2 only.

Table 14: **Main results, full version.** Each method at its published setting on two models and two image sets, and the positive control. The CHAIR ratio ignores the change in coverage, which is captured by grounded recall of the ground-truth objects. S compares each row with the best content-free placebo *at the row's own recall* c (Eq. 5); higher is better, and $S \leq 0$ means a placebo does at least as well. The shaded rows give the unedited model and the plane's null band. Underlined: best S among the rows listed for the plane; **bold**: the whole 95% interval clears the band; *: the whole interval is below 0; †published setting. On AMBER the first two columns are AMBER's Hal and CHAIR. The OPERA, DeCo and MiniGPT-4 VCD and PAI rows are in Tab. 2 only. The 28 rows scored here and in Tab. 5 are in Fig. 4 and Tab. 16.

	operating point			reported	SCAR (ours)		
Method	$\text{CHAIR}_s\downarrow$	$\text{CHAIR}_i\downarrow$	Recall↑	ΔCHAIR_i	$r_F(c)$	$S\uparrow$ [95% CI]	$\Pr[S\leq\text{band}]$
LLaVA-1.5-7B, COCO, *unedited*	19.20	6.26	69.09			*null band* $S \leq +0.136$	
VCD, $\alpha{=}1^\dagger$	22.00	7.11	69.45	$+0.84$	5.18	-0.373^* [$-0.823, -0.176$]	1.000
PAI, $\alpha{=}0.5^\dagger$	12.20	4.31	68.07	-1.95	4.75	$\underline{+0.092}$ [$-0.210, +0.273$]	0.778
ProjectAway, $\alpha{=}3.5^\dagger$	17.00	5.73	68.98	-0.53	5.03	-0.140^* [$-0.453, -0.022$]	1.000
Nullu, $k{=}4^\dagger$	14.60	5.30	68.33	-0.97	4.83	-0.097 [$-0.444, +0.084$]	0.989
MiniGPT-4, COCO, *unedited*	25.80	10.33	64.64			*null band* $S \leq +0.191$	
Nullu, $k{=}8^\dagger$	20.40	8.63	60.82	-1.70	7.83	$\underline{-0.102}$ [$-0.401, +0.094$]	0.991
Nullu, $k{=}4$	30.80	12.14	61.08	$+1.80$	7.95	-0.527^* [$-0.923, -0.279$]	1.000
LLaVA-1.5-7B, AMBER, *unedited*	17.20	4.00	51.00			*null band* $S \leq +0.262$	
VCD, $\alpha{=}1^\dagger$	22.20	5.55	50.56	$+1.56$	3.89	-0.426^* [$-0.768, -0.181$]	1.000
PAI, $\alpha{=}0.5^\dagger$	15.00	3.75	48.27	-0.24	3.37	-0.114 [$-0.466, +0.119$]	0.993
ProjectAway, $\alpha{=}3.5^\dagger$	16.60	3.83	50.83	-0.16	3.96	$\underline{+0.032}$ [$-0.121, +0.108$]	0.992
Nullu, $k{=}4^\dagger$	17.20	4.50	46.81	$+0.51$	3.03	-0.487^* [$-0.946, -0.200$]	1.000
MiniGPT-4, AMBER, *unedited*	53.80	12.93	61.41			*null band* $S \leq +0.321$	
Nullu, $k{=}8^\dagger$	43.40	11.72	51.82	-1.21	8.30	$\underline{-0.412^*}$ [$-0.689, -0.229$]	1.000
Control, *unedited*	19.20	6.26	69.09	LLaVA-1.5-7B, COCO; *null band* $S \leq +0.136$			
oracle, $\delta{=}-4$	6.60	1.88	69.55	-4.38	5.21	$+\mathbf{0.638}$ [$+0.474, +0.747$]	**0.003**
oracle, ban	5.60	1.66	69.21	-4.61	5.10	$+\mathbf{0.676}$ [$+0.517, +0.774$]	**0.001**
permuted, $\delta{=}-4$	12.60	6.55	51.04	$+0.28$	2.97	-1.207^* [$-2.810, -0.563$]	1.000

Table 15: S on the three new planes when one knob family is removed from Π, and with the piecewise-linear (stairs) frontier. *The 95% interval excludes zero. No row clears its band under any variant, and only one changes its verdict: MiniGPT-4 Nullu on AMBER moves from below 0 to placebo-equivalent without π_{len} or with stairs. On LLaVA-1.5 AMBER the stairs frontier is not monotone over the scoring region, so, as on LLaVA COCO (Tab. 7), it is excluded from the primary reading there.

	MiniGPT-4, COCO		LLaVA-1.5, AMBER				MiniGPT-4, AMBER
Π variant	Nullu k=8	Nullu k=4	VCD α=1	PAI α=0.5	PA α=3.5	Nullu k=4	Nullu k=8
primary (hull)	-0.102	-0.527^{*}	-0.426^{*}	-0.114	$+0.032$	-0.487^{*}	-0.412^{*}
no π_{bias}	-0.087	-0.505^{*}	-0.405^{*}	-0.004	$+0.037$	-0.252^{*}	-0.412^{*}
no π_{len}	-0.047	-0.460^{*}	-0.426^{*}	-0.114	$+0.032$	-0.487^{*}	-0.046
no π_{beam}	-0.102	-0.527^{*}	-0.426^{*}	-0.114	$+0.032$	-0.487^{*}	-0.412^{*}
stairs	-0.102	-0.527^{*}	-0.388^{*}	$+0.008$	$+0.042$	-0.372^{*}	-0.046

Table 16: **The rest of the SCAR row** for the MiniGPT-4, AMBER and control rows of Tab. 14 (for the 18 LLaVA rows, see Tab. 5): excision gain ε, suppression share σ (undefined where CHAIR_i does not fall), and the smallest ε that 500 images resolve against the band. On AMBER, CHAIR_i is AMBER's CHAIR. **Bold**, * and † as in Tab. 14; 2,000 paired image resamples.

row	c	$r_F(c)$	ε [95% CI]	$P(\varepsilon \le 0)$	S [95% CI]	σ [95% CI]	resolv. ε
MiniGPT-4, COCO *null band* $S \le +0.191$							
Nullu, k=8†	0.6082	0.0783	-0.0080 [$-0.0268, +0.0078$]	0.863	-0.102 [$-0.401, +0.094$]	1.000 [0.732, 1.000]	0.0344
Nullu, k=4	0.6108	0.0795	-0.0419 [$-0.0626, -0.0244$]	1.000	-0.527^{*} [$-0.923, -0.279$]	undef.	0.0401
LLaVA-1.5-7B, AMBER *null band* $S \le +0.262$							
VCD, α=1†	0.5056	0.0389	-0.0166 [$-0.0265, -0.0078$]	0.999	-0.426^{*} [$-0.768, -0.181$]	undef.	0.0214
PAI, α=0.5†	0.4827	0.0337	-0.0038 [$-0.0138, +0.0043$]	0.833	-0.114 [$-0.466, +0.119$]	1.000 [0.530, 1.000]	0.0193
ProjectAway, α=3.5†	0.5083	0.0396	$+0.0013$ [$-0.0045, +0.0044$]	0.357	$+0.032$ [$-0.121, +0.108$]	0.238 [0.000, 1.000]	0.0179
Nullu, k=4†	0.4681	0.0303	-0.0148 [$-0.0244, -0.0065$]	1.000	-0.487^{*} [$-0.946, -0.200$]	undef.	0.0202
MiniGPT-4, AMBER *null band* $S \le +0.321$							
Nullu, k=8†	0.5182	0.0830	-0.0342 [$-0.0515, -0.0201$]	1.000	-0.412^{*} [$-0.689, -0.229$]	1.000 [1.000, 1.000]	0.0470
Control, LLaVA-1.5-7B, COCO *null band* $S \le +0.136$							
oracle, δ=-4	0.6955	0.0521	$+0.0333$ [$+0.0196, +0.0424$]	0.000	$\mathbf{+0.638}$ [$+0.474, +0.747$]	0.241 [0.089, 0.488]	0.0169
oracle, ban	0.6921	0.0510	$+0.0345$ [$+0.0210, +0.0431$]	0.000	$\mathbf{+0.676}$ [$+0.517, +0.774$]	0.252 [0.119, 0.482]	0.0163
permuted, δ=-4	0.5104	0.0297	-0.0358 [$-0.0571, -0.0194$]	1.000	-1.207^{*} [$-2.810, -0.563$]	undef.	0.0340

Table 17: S of all 18 LLaVA-1.5 rows under four scorings of the same captions: the primary CHAIR scorer, its 80 category names without synonyms, recall pooled over images instead of averaged per image, and the open-vocabulary scorer of Tab. 19. Each column is read against its own band (shaded row). Underlined: best row in the column.

row	primary	80 names, no synonyms	pooled recall	open vocabulary
null band edge	$+0.136$	$+0.235$	$+0.097$	$+0.072$
Nullu, k=1	-0.407	-0.676	-0.392	-0.035
Nullu, k=2	-0.170	-0.299	-0.113	-0.081
Nullu, k=4†	-0.097	-0.353	-0.101	-0.150
Nullu, k=8	$+0.036$	-0.170	$+0.033$	-0.041
PAI, α=0.2	$+0.044$	$+0.065$	$+0.004$	$\underline{+0.034}$
PAI, α=0.5†	$\underline{+0.092}$	$\underline{+0.170}$	$\underline{+0.072}$	-0.009
PAI, α=0.5, greedy	$+0.057$	$+0.118$	$+0.063$	-0.014
ProjectAway, α=0.5	-0.092	-0.106	-0.090	-0.017
ProjectAway, α=1.5	-0.021	-0.064	-0.014	-0.022
ProjectAway, α=2.5	-0.148	-0.179	-0.136	-0.033
ProjectAway, α=3.5†	-0.140	-0.178	-0.118	-0.032
ProjectAway, α=1.5, beam 5	-0.048	-0.025	-0.074	$+0.002$
ProjectAway, α=3.5, beam 5	-0.156	-0.110	-0.199	-0.029
VCD, α=0.5	-0.190	-0.353	-0.204	-0.047
VCD, α=1†	-0.373	-0.475	-0.359	-0.079
VCD, α=2	-0.390	-0.639	-0.418	-0.264
VCD, α=3	-0.701	-1.028	-0.728	-0.332
VCD, α=1, sampling	-0.562	-0.778	-0.548	-0.280
rows above the band	0/18	0/18	0/18	0/18

Table 18: POPE accuracy gain after four ways of moving the decision rule. *forward*: Eq. 6; *reverse*: the edited model moved to the unedited yes-rate; *ceiling*: best threshold of each score vector; *isotonic*: 5-fold cross-fitted calibration read at 0.5. 2,000 image resamples. †published setting.

row	reported Δacc	ε forward	ε reverse	ε ceiling	ε isotonic
LLaVA-1.5-7B, pooled over the three splits					
Nullu k=1	-0.0001	$+0.0000$ [-0.0011, $+0.0011$]	$+0.0002$ [-0.0011, $+0.0014$]	$+0.0006$ [-0.0002, $+0.0016$]	-0.0004 [-0.0030, $+0.0036$]
Nullu k=2	$+0.0093$	$+0.0016$ [$+0.0002$, $+0.0025$]	$+0.0000$ [-0.0013, $+0.0011$]	$+0.0008$ [-0.0002, $+0.0016$]	$+0.0014$ [-0.0027, $+0.0040$]
Nullu k=4†	$+0.0098$	-0.0004 [-0.0011, $+0.0007$]	$+0.0000$ [-0.0011, $+0.0009$]	$+0.0006$ [-0.0006, $+0.0013$]	$+0.0013$ [-0.0029, $+0.0039$]
Nullu k=8	$+0.0274$	$+0.0007$ [-0.0007, $+0.0018$]	-0.0013 [-0.0027, -0.0002]	$+0.0006$ [-0.0009, $+0.0011$]	$+0.0009$ [-0.0028, $+0.0042$]
PAI α=0.2	$+0.0037$	$+0.0029$ [$+0.0013$, $+0.0045$]	$+0.0011$ [-0.0007, $+0.0029$]	$+0.0004$ [-0.0007, $+0.0018$]	-0.0004 [-0.0039, $+0.0045$]
PAI α=0.5†	$+0.0108$	$+0.0076$ [$+0.0048$, $+0.0109$]	$+0.0049$ [$+0.0022$, $+0.0076$]	$+0.0026$ [-0.0003, $+0.0046$]	$+0.0042$ [-0.0038, $+0.0080$]
VCD α=1†	-0.0167	-0.0118 [-0.0167, -0.0072]	-0.0122 [-0.0171, -0.0071]	-0.0107 [-0.0155, -0.0056]	-0.0093 [-0.0182, -0.0034]
VCD α=2	-0.0357	-0.0273 [-0.0335, -0.0215]	-0.0251 [-0.0315, -0.0197]	-0.0108 [-0.0163, -0.0057]	-0.0091 [-0.0160, -0.0031]
MiniGPT-4, Nullu k=8†					
pooled	$+0.0224$	$+0.0002$ [$+0.0000$, $+0.0007$]	$+0.0000$ [$+0.0000$, $+0.0000$]	$+0.0074$ [$+0.0036$, $+0.0097$]	$+0.0091$ [$+0.0004$, $+0.0148$]
random	$+0.0470$	$+0.0000$ [$+0.0000$, $+0.0000$]	$+0.0000$ [$+0.0000$, $+0.0000$]	$+0.0073$ [-0.0003, $+0.0113$]	$+0.0153$ [-0.0077, $+0.0250$]
popular	$+0.0097$	$+0.0007$ [$+0.0000$, $+0.0020$]	$+0.0000$ [$+0.0000$, $+0.0000$]	$+0.0060$ [$+0.0013$, $+0.0107$]	$+0.0020$ [-0.0070, $+0.0197$]
adversarial	$+0.0107$	$+0.0000$ [$+0.0000$, $+0.0000$]	$+0.0000$ [$+0.0000$, $+0.0000$]	$+0.0077$ [$+0.0007$, $+0.0110$]	$+0.0163$ [-0.0067, $+0.0243$]

Table 19: The null band and what each plane can resolve. *resolvable* ε: band edge plus 1.645 bootstrap standard deviations of $S-$ band, in CHAIR_i units. The open vocabulary plane re-scores the same LLaVA captions against objects extracted from the five human COCO captions, matched by sentence embeddings (all-MiniLM-L6-v2) at a cosine threshold of 0.6 calibrated against CHAIR; its risk is higher than CHAIR_i because the reference set is incomplete.

plane	cells	band edge [95% CI]	rows	best S	$\min \Pr[S \le \text{band}]$	resolvable ε	r_{raw}
LLaVA-1.5, COCO	25	$+0.136$ [$+0.055$, $+0.341$]	18	$+0.092$	0.778	0.0173–0.0285	0.0626
LLaVA-1.5, COCO, open vocab.	25	$+0.072$ [$+0.016$, $+0.116$]	18	$+0.034$	0.876	0.0337–0.0491	0.2800
MiniGPT-4, COCO	8	$+0.191$ [$+0.069$, $+0.316$]	2	-0.102	0.991	0.0344–0.0401	0.1033
LLaVA-1.5, AMBER	8	$+0.262$ [$+0.080$, $+0.447$]	4	$+0.032$	0.992	0.0179–0.0214	0.0400
MiniGPT-4, AMBER	8	$+0.321$ [$+0.205$, $+0.413$]	1	-0.412	1.000	0.0470–0.0470	0.1293

G CAPTIONS UNDER EVERY CONFIGURATION

Figs. 6–19 show what the audit measures in aggregate, one image at a time: 16 COCO and 16 AMBER images for LLaVA-1.5-7B, and 17 COCO and 18 AMBER images for MiniGPT-4. Every block is one image and one prompt ("Please describe this image in detail." on COCO, "Describe this image." on AMBER; beam 3, 64 tokens unless the row says otherwise). The rows are the unedited model, each published method at the setting audited in the main text (LLaVA: Nullu k=4, VCD α=1, PAI α=0.5, ProjectAway α=3.5; MiniGPT-4: Nullu at its published rank k=8), and one placebo row on the *unedited* weights. Captions are verbatim; "…" marks the token cap. Fig. 5 first works through one COCO image in full: the unedited model, Nullu at its published rank and one fixed placebo side by side, next to the same comparison over all 500 images.

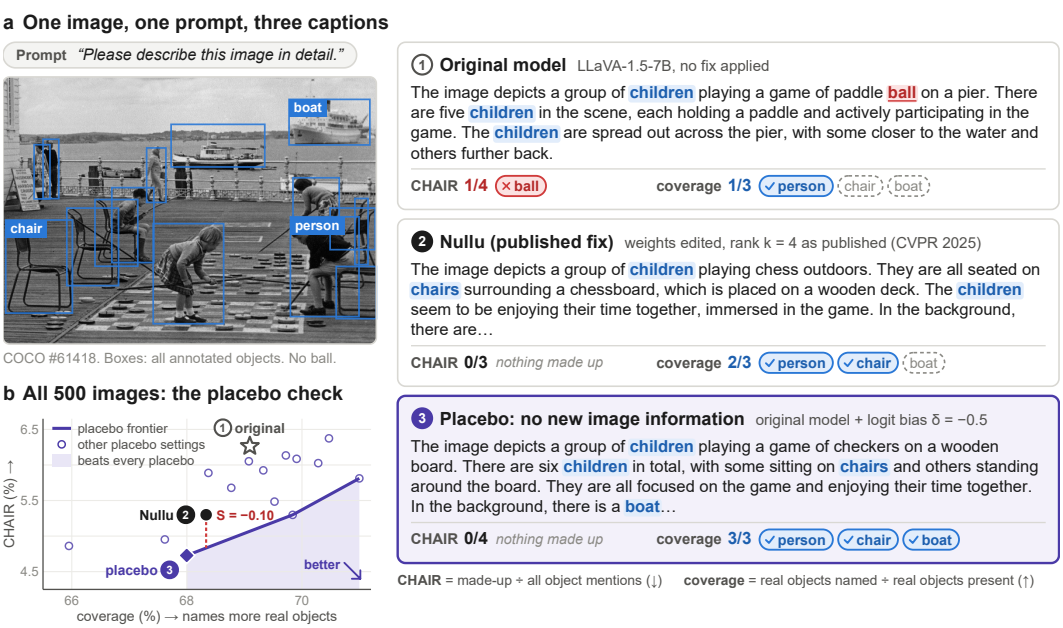


Figure 5: **A published fix against a placebo.** (a) One COCO image, one prompt shared by all three captions ("Please describe this image in detail."), and three LLaVA-1.5-7B captions (verbatim, 64-token cap). Highlights are CHAIR's own matches: blue for an annotated object, red for a hallucinated one. The original model invents a ball (CHAIR $1/4$) and names one of the three annotated object types (coverage $1/3$). Nullu (Yang et al., 2025) at its published rank (k=4) removes the ball and names two ($0/3$, $2/3$). A logit bias of δ=−0.5 on object words of the *unedited* model, a placebo that adds no information about the image, also removes it and names all three ($0/4$, $3/3$). (b) Over 500 images, the placebo frontier is the lower envelope of 25 such settings (Sec. 3.2); the shaded region below it beats every placebo, and the dashed red line is Nullu's gap above it. **Takeaway:** Nullu and the placebo both reach CHAIR 0 on this image, yet the placebo names more real objects with no new information about the image. Over 500 images, Nullu sits above the placebo frontier at its own coverage: SCAR score $S = -0.097$ $[-0.444, +0.084]$ (Eq. 5); only $S > 0$ would beat every placebo.

Reading a block. Highlights and counts come from the scorer, not from us: CHAIR on COCO and AMBER's own matcher on AMBER (App. D). A blue word is a mention the scorer matched to an annotated object; a red, underlined word is one it counted as hallucinated. Under each row name, × counts the distinct hallucinated objects and ✓ the annotated object types named, out of those in the chips beside the photo. A star marks a row that no other row beats on either count. COCO photos carry the instance boxes of every annotated object; AMBER ships no boxes, so its photos are shown plain.

The placebo row is chosen per image. It is the best setting, for that image, of the plane's placebo family: 20 settings on LLaVA/COCO (the cells of Tab. 6 other than the reference itself, the 16- and

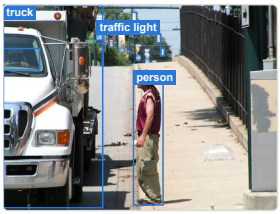
128-token caps and sampling at $T \geq 1$) and 5 elsewhere (beam 1, beam 2, $\delta{=}-0.5$, $\delta{=}-2$, 32 tokens). A method gets one setting; the placebo gets the best of several. The blocks therefore favour the placebo, and they are illustrations, not evidence. The evidence is the SCAR score at matched coverage over all 500 images (Tab. 5). Tab. 20 gives, for the same rows over all 500 images of each plane, how often each side gets the star.

Table 20: Who gets the star, over all 500 images of each plane, with the rows the gallery shows. "Both" is mostly images where several captions name the same objects and none hallucinates.

Model	Data	placebo only	a method only	both	neither	unedited hallucinates
LLaVA-1.5-7B	COCO	103	18	367	12	96
LLaVA-1.5-7B	AMBER	98	77	304	21	86
MiniGPT-4	COCO	183	24	278	15	129
MiniGPT-4	AMBER	273	28	112	87	269

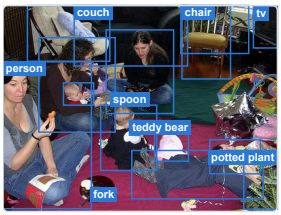
Selection. We screened every image automatically and then looked at each candidate photo. Most shown images are ones where the placebo gets the star, and in five images per plane a published method gets it instead (Nullu, VCD, PAI and ProjectAway all appear). We dropped images where the scorer's verdict is wrong, since it shows matcher noise and not suppression. Examples: a mention inside a negation ("no cars or people visible") counts as named; "bike" for a motorcycle is matched to bicycle; CHAIR's singulariser never matches "bus"; "skateboard's trucks" is credited as the annotated truck; a colour ("the carrot is orange") is counted as the fruit; and AMBER annotations miss visible objects such as a saddle or a dog's collar. Two such images remain, and their captions say so (Figs. effig:gal-llava-coco-2 and effig:gal-minigpt4-coco-1). We also dropped images where the placebo caption is the unedited caption. The figures are generated from the banked scorer outputs by `make_app_gallery.py`, which also writes Tab. 20.

dog annotated, named cat made up ✕made-up objects ✓objects named / annotated ★best row: fewest made up, most named placebo



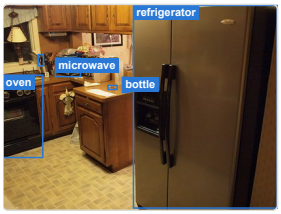
COCO #69827 · COCO boxes
annotated (3): truck traffic light person

Unedited	✕1 ✓2/3 beam 3	In the image, a man wearing a red shirt is standing on the sidewalk next to a white truck. He appears to be talking on a cell phone, with the phone held to his ear. The truck is parked on the side of the road, occupying a significant portion of the scene…
Nullu	✕0 ✓2/3 rank k = 4	The image features a man wearing a maroon shirt and a baseball cap, standing on the sidewalk next to a white truck. He appears to be listening to music through his headphones. The truck is parked on the street, and the man seems to be admiring it.
VCD	✕0 ✓2/3 α = 1	The image features a man wearing headphones, standing next to a large white truck parked on the side of the road. The man appears to be working on the truck, possibly as a mechanic or technician. He is wearing a maroon shirt and is focused on his task.
PAI	✕0 ✓2/3 α = 0.5	In the image, there is a man standing on the sidewalk next to a white truck. He is wearing headphones and appears to be engaged in a conversation with the truck driver. The truck is parked on the side of the road, and the man seems to be interested in the situation…
ProjectAway	✕0 ✓2/3 α = 3.5	In the image, a man is standing on the sidewalk next to a white truck. He appears to be wearing headphones, possibly listening to music or an audio guide. The truck is parked on the side of the road, occupying a significant portion of the scene. There are…
Placebo	✕0 ✓3/3 ★ beam 1	The image features a man standing on the sidewalk next to a white truck, which is parked on the side of the road. The man appears to be wearing headphones and is looking at the truck. There are several traffic lights in the scene, with one close to the man and others further…



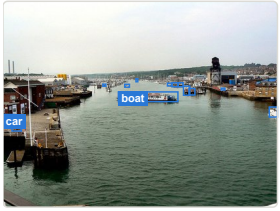
COCO #330122 · COCO boxes
annotated (8): teddy bear person couch chair spoon tv potted plant fork

Unedited	✕1 ✓3/8 beam 3	The image depicts a group of people, including both adults and children, sitting on the floor and enjoying a meal together. They are gathered around a dining table, which is placed in the center of the room. There are several chairs surrounding the table, and a couch can be seen…
Nullu	✕0 ✓2/8 rank k = 4	The image depicts a group of women and children gathered together in a living room. They are all seated on the floor, eating food and enjoying each other's company. There are at least five women and five children present in the scene. The room is furnished with a couch…
VCD	✕1 ✓2/8 α = 1	The image depicts a group of women and children gathered in a living room, enjoying a meal together. There are three women sitting on the floor, with one of them eating a sandwich, while the other two are sitting on a couch. A baby is also present in the scene, sitting…
PAI	✕0 ✓1/8 α = 0.5	The image depicts a group of women sitting on the floor, enjoying a meal together. They are surrounded by their children, who are also sitting on the floor, enjoying a meal. The women are sitting on a red carpet, while the children are sitting on a green blanket.
ProjectAway	✕1 ✓3/8 α = 3.5	The image depicts a group of people, including both adults and children, sitting on the floor and enjoying a meal together. They are gathered around a dining table, which is placed in the center of the room. There are several chairs surrounding the table, and a couch can be seen…
Placebo	✕0 ✓3/8 ★ beam 2	The image depicts a group of people, including both adults and children, gathered in a living room. They are sitting on the floor and enjoying a meal together. There are several people sitting on couches and chairs, while others are sitting on the floor. Various food items…



COCO #503772 · COCO boxes
annotated (6): bottle spoon oven refrigerator microwave knife

Unedited	✕2 ✓2/6 beam 3	The image depicts a kitchen with a large stainless steel refrigerator positioned in the middle of the room. The refrigerator is open, revealing its contents. The kitchen counter is cluttered with various items, including several bottles, cups, and a bowl.
Nullu	✕0 ✓1/6 rank k = 4	The image features a kitchen with a large stainless steel refrigerator freezer prominently displayed. The refrigerator takes up a significant portion of the scene, extending from the left-hand side all the way to the right-hand side of the image. In addition to the…
VCD	✕1 ✓3/6 α = 1	The image features a kitchen with a large stainless steel refrigerator standing prominently in the center of the room. The refrigerator is open, revealing its contents. The kitchen counter is cluttered with various items, including several cups, a knife, a spoon,…
PAI	✕0 ✓1/6 α = 0.5	In the image, there is a large stainless steel refrigerator standing in a kitchen. The refrigerator is open, revealing its contents. The refrigerator is located in the middle of the kitchen, surrounded by wooden cabinets. In the kitchen, there is…
ProjectAway	✕0 ✓4/6 ★ α = 3.5	The image depicts a kitchen with a stainless steel refrigerator standing in the middle of the room. The refrigerator is open, revealing its contents. The kitchen counter is cluttered with various items, including several bottles, a knife, a spoon, and a…
Placebo	✕0 ✓3/6 δ = −0.75	The image depicts a kitchen scene with a large metallic refrigerator standing prominently in the center of the room. The refrigerator is open, revealing its contents. The kitchen is well-equipped with various appliances, including a microwave and an oven…

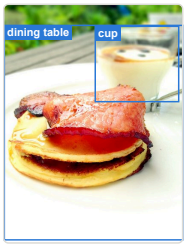


COCO #353231 · COCO boxes
annotated (2): car boat

Unedited	✕1 ✓2/2 beam 3	The image depicts a large body of water with several boats of various sizes floating on it. The boats are scattered throughout the scene, with some closer to the shore and others further out in the water. In addition to the boats, there are a few cars and trucks parked near the…
Nullu	✕0 ✓1/2 rank k = 4	The image depicts a harbor filled with numerous boats of various sizes docked in the water. The boats are lined up in rows, creating an organized appearance. Some of the boats are moored at docks, while others are floating freely in the harbor. In addition to the boats,…
VCD	✕0 ✓1/2 α = 1	The image depicts a harbor filled with numerous boats of various sizes docked in the water. The boats are scattered throughout the scene, with some closer to the shore and others further out in the water. The harbor is surrounded by a lush green landscape, creating a serene atmosphere. In…
PAI	✕0 ✓1/2 α = 0.5	In the image, there is a large boat traveling down the river, surrounded by several smaller boats. The large boat is situated in the middle of the scene, while the smaller boats are scattered around it. Some of the smaller boats are closer to the large boat, while others are further away. The…
ProjectAway	✕0 ✓1/2 α = 3.5	The image depicts a large body of water with several boats floating on it. There are at least 12 boats visible in the scene, varying in size and position. Some of the boats are closer to the foreground, while others are further away, creating a sense of depth in the water.
Placebo	✕0 ✓2/2 ★ beam 1	The image features a large body of water with several boats floating on it. There is a large boat in the middle of the water, surrounded by smaller boats. Some of the smaller boats are located closer to the shore, while others are further away. In addition to the boats, there are a few cars visible…

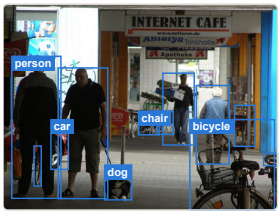
Figure 6: **LLaVA-1.5-7B, COCO (1/4).** Rows: unedited model; Nullu, VCD, PAI and ProjectAway at the settings audited in the main text; best placebo for the image. Starred: the placebo on 3 of 4 images, a published method on 1. Legend and selection: App. G.

dog annotated, named cat made up ×made-up objects ✓objects named / annotated ★best row: fewest made up, most named placebo



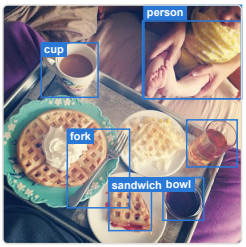
COCO #55772 · COCO boxes
annotated (3): cup cake dining table

Unedited	×1 ✓2/3 beam 3	The image features a delicious breakfast meal consisting of a bacon and egg **sandwich** on a white plate. The **sandwich** is cut in half, revealing the bacon and egg filling inside. The plate is placed on a dining **table**. In addition to the **sandwich**, there is a **cup**…
Nullu	×0 ✓2/3 rank k = 4	The image features a delicious breakfast meal served on a white plate. On the plate, there's a bacon-covered **pancake** with syrup drizzled over it, creating a mouth-watering combination. Beside the **pancake**, there's a **cup** of coffee, which…
VCD	×0 ✓1/3 α = 1	The image features a white plate filled with a delicious breakfast meal. On the plate, there are two **pancakes**, one of which is topped with bacon and syrup. The bacon is placed on top of the **pancakes**, adding a savory touch to the dish.
PAI	×0 ✓1/3 α = 0.5	In the image, there is a delicious-looking breakfast meal on a white plate. The meal consists of two **pancakes**, bacon, and butter. The **pancakes** are placed on top of the bacon and butter, creating a delightful combination of flavors and textures.
ProjectAway	×1 ✓2/3 α = 3.5	The image features a delicious breakfast meal consisting of a bacon and egg **sandwich** on a white plate. The **sandwich** is cut in half, revealing the bacon and egg filling inside. The plate is placed on a dining **table**. In addition to the **sandwich**, there is a **cup**…
Placebo	×0 ✓3/3 ★ δ = −1.5	The image features a delicious breakfast meal consisting of a bacon and maple syrup-covered **pancake** on a white plate. The plate is placed on a dining **table**, and the meal is accompanied by a **cup** of coffee, which is positioned near the top right corner of the plate…



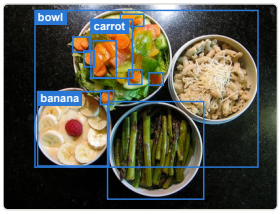
COCO #27519 · COCO boxes
annotated (5): dog person car bicycle chair

Unedited	×1 ✓2/5 beam 3	In the image, a **man** is walking his black and white **dog** on a leash down a sidewalk. The **dog** appears to be sitting on the sidewalk as the **man** walks by. There are several other **people** in the scene, some of whom are carrying **handbags**. In the background,…
Nullu	×0 ✓2/5 rank k = 4	The image depicts a **man** walking his black and white **dog** down a city street. The **dog** is on a leash, and the **man** is holding its collar. They are passing by an Internet Cafe, which is advertised on a sign above them. There are several other **people** in the scene…
VCD	×0 ✓2/5 α = 1	In the image, a black and white **dog** is sitting on a sidewalk next to its owner, who is wearing a black shirt. The **dog** appears to be patiently waiting for its owner to finish a conversation with another **person**. There are several other **people** in the scene, with one **person** standing…
PAI	×0 ✓2/5 α = 0.5	In the image, a **man** is walking his black and white **dog** on a leash down a city street. The **dog** appears to be well-behaved, sitting calmly on the sidewalk. The **man** and his **dog** seem to be the main focus of the scene. In the background, there…
ProjectAway	×0 ✓2/5 α = 3.5	In the image, a **man** is walking his black and white **dog** on a leash down a sidewalk. The **dog** is positioned close to the **man**'s feet, and they appear to be enjoying their walk together. There are several other **people** in the scene, some of whom are walking or…
Placebo	×0 ✓3/5 ★ T = 0.5	The image depicts a busy city street with several **people** and a **dog**. The **dog** is sitting on the sidewalk, waiting patiently for its owner. The **people** in the scene are engaged in various activities, such as walking, standing, and talking. There are multiple **bicycles** in the scene…



COCO #392048 · COCO boxes
annotated (6): cup person sandwich bed bowl fork

Unedited	×2 ✓2/6 beam 3	The image features a dining **table** with a tray filled with a variety of breakfast foods. On the tray, there is a waffle with whipped cream, a slice of pie, and a **cup** of coffee. A **fork** and a **knife** are also placed on the tray, ready for…
Nullu	×1 ✓1/6 rank k = 4	The image features a tray filled with delicious breakfast foods, including waffles with whipped cream and syrup, scrambled eggs, and corn on the cob. The tray is placed on a dining **table**, and a **baby** is seated in a highchair, eagerly…
VCD	×0 ✓3/6 ★ α = 1	The image depicts a **baby** sitting on a **bed** next to an **adult** and a tray of food. The tray is filled with various dishes, including a waffle with whipped cream, corn on the cob, and a slice of pie. There is also a **cup** of coffee on the…
PAI	×0 ✓3/6 ★ α = 0.5	In the image, a **person** is sitting on a **bed** with a tray of food in front of them. On the tray, there is a waffle, a piece of pie, and a **cup** of coffee. The **person** is sitting next to a **baby**, who is watching the food being prepared.
ProjectAway	×2 ✓2/6 α = 3.5	The image features a dining **table** with a tray filled with a variety of breakfast foods. On the tray, there is a waffle with whipped cream, a slice of pie, and a **cup** of coffee. A **fork** and a **knife** are also placed on the tray, ready for…
Placebo	×0 ✓1/6 δ = −2	The image depicts a family enjoying a meal together. There are three plates of food on a tray, including a piece of pie, a plate of waffles, and a plate of corn. Two **cups** are also present on the tray, one containing coffee and the other filled with…



COCO #364399 · COCO boxes
annotated (4): banana carrot bowl dining table

Unedited	×1 ✓3/4 beam 3	The image features a dining **table** with four white **bowls** filled with various types of food. The **bowls** are arranged in a semi-circle formation, with each **bowl** containing a different dish. In the first **bowl**, there is a salad consisting of **carrots** and **broccoli**…
Nullu	×0 ✓3/4 rank k = 4	The image features a dining **table** with four different **bowls** filled with various types of food. Each **bowl** contains different ingredients, creating a colorful and appetizing display. The first **bowl** contains **banana** slices and raspberries, while the second **bowl** is filled with…
VCD	×1 ✓4/4 α = 1	The image showcases a dining **table** with four white **bowls** filled with a variety of healthy foods. The **bowls** contain **bananas**, asparagus, **carrots**, and macaroni. The **bananas** are placed in two of the **bowls**, while the asparagus, **car**…
PAI	×0 ✓3/4 α = 0.5	The image features a dining **table** with four white **bowls** filled with a variety of healthy foods. The **bowls** contain a mix of fruits, vegetables, and pasta. There are several **carrots** in the **bowls**, with some of them being sliced into smaller pieces.
ProjectAway	×0 ✓3/4 α = 3.5	The image features a **table** with four **bowls** filled with various types of food. The first **bowl** on the left contains a delicious-looking salad, while the second one on the left is filled with **bananas**. The third **bowl** on the left is filled with **bananas**, and the fourth one on the…
Placebo	×0 ✓4/4 ★ beam 5	The image features a dining **table** with four **bowls** filled with a variety of healthy foods. Each **bowl** contains different types of vegetables and fruits, creating a colorful and appetizing display. In the first **bowl**, there are **bananas** and **carrots**, while the…

Figure 7: **LLaVA-1.5-7B, COCO (2/4).** Rows: unedited model; Nullu, VCD, PAI and ProjectAway at the settings audited in the main text; best placebo for the image. Starred: the placebo on 3 of 4 images, a published method on 1. On COCO #364399 the token cap cuts VCD's "carrots" to "car", which CHAIR counts as a hallucinated car; without it VCD would share the star. Legend and selection: App. G.

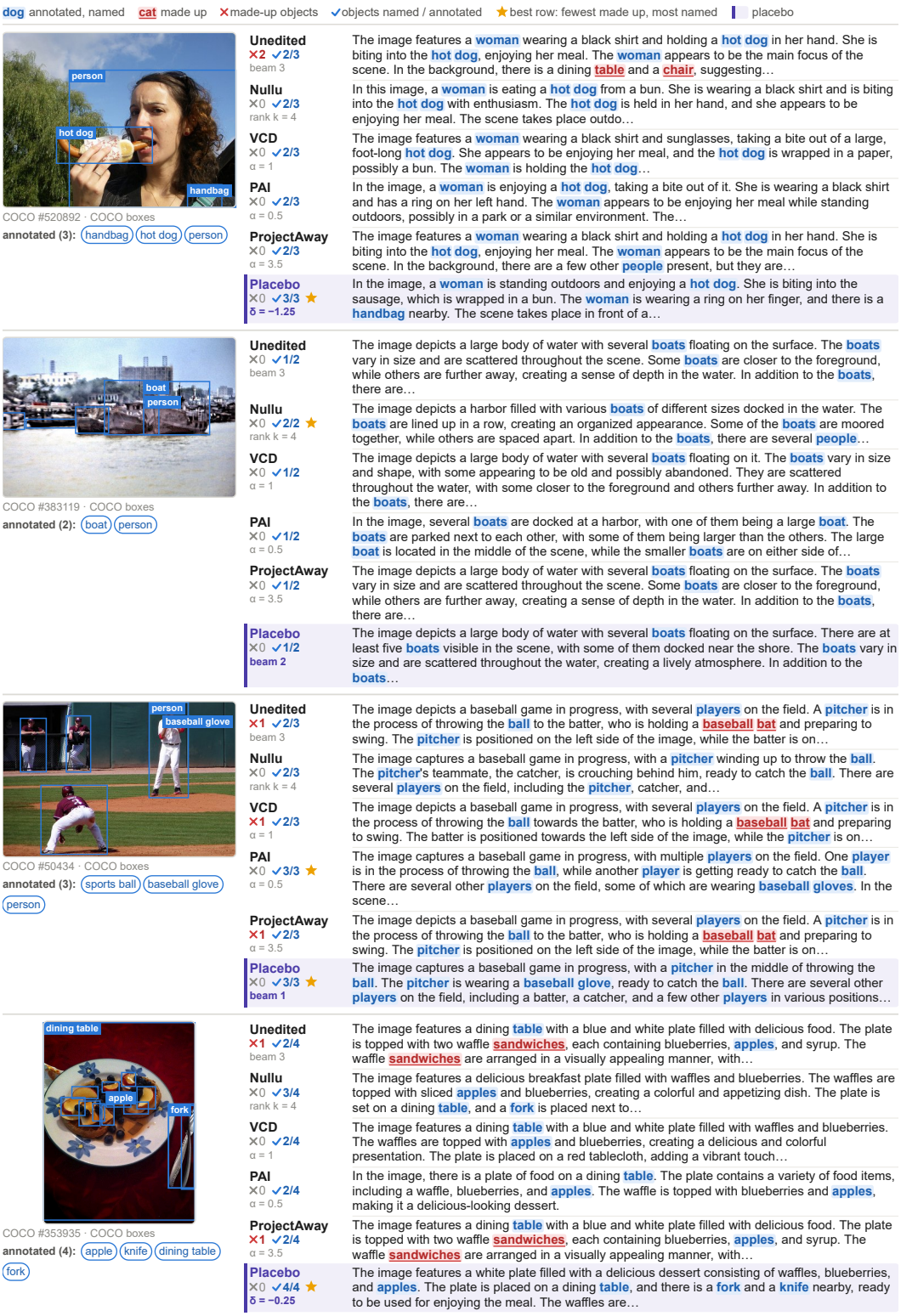
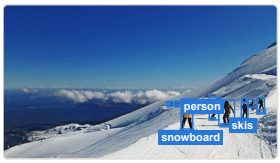


Figure 8: **LLaVA-1.5-7B, COCO (3/4).** Rows: unedited model; Nullu, VCD, PAI and ProjectAway at the settings audited in the main text; best placebo for the image. Starred: the placebo on 3 of 4 images, a published method on 2. Legend and selection: App. G.

dog annotated, named cat made up ✕made-up objects ✓objects named / annotated ★best row: fewest made up, most named placebo



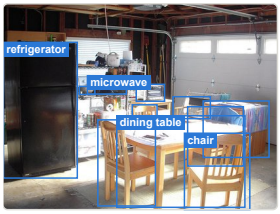
COCO #516601 · COCO boxes
annotated (3): snowboard skis person

Unedited ✕0 ✓1/3 beam 3	The image captures a group of **people** skiing and snowboarding on a snow-covered mountain slope. There are at least nine **people** visible in the scene, with some of them skiing and others snowboarding. The **skiers** and **snowboarders** are scattered across the slope, with some closer to the…
Nullu ✕0 ✓1/3 rank k = 4	The image depicts a group of **people** skiing and snowboarding down a snowy mountain. There are six **people** in total, with five of them skiing and one snowboarding. The **skiers** and **snowboarder** are spread out across the slope, with some of them skiing in pairs.
VCD ✕0 ✓1/3 α = 1	The image captures a group of **people** skiing and snowboarding on a snow-covered mountain. There are several **skiers** and **snowboarders** scattered across the slope, enjoying their time on the mountain. Some of them are closer to the top of the slope, while others are further down.
PAI ✕0 ✓3/3 ★ α = 0.5	The image captures a group of **people** skiing and snowboarding on a snow-covered mountain. There are six **people** in the scene, with some of them wearing **skis** and others wearing **snowboards**. They are standing at the top of the mountain, enjoying their time on the slopes…
ProjectAway ✕0 ✓1/3 α = 3.5	The image captures a group of **people** skiing and snowboarding on a snow-covered mountain slope. There are at least nine **people** visible in the scene, with some of them skiing and others snowboarding. The **skiers** and **snowboarders** are scattered across the slope, enjoying their time on…
Placebo ✕0 ✓2/3 beam 2	The image captures a group of **people** skiing and snowboarding on a snow-covered mountain. There are at least nine **people** visible in the scene, with some of them skiing and others snowboarding. They are scattered across the slope, enjoying their time on the mountain. The **ski**…



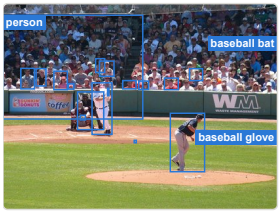
COCO #146126 · COCO boxes
annotated (4): cup remote tv dining table

Unedited ✕1 ✓1/4 beam 3	The image features a dining **table** with a variety of food and drink items on it. There is a **bowl** filled with noodles, which appears to be the main focus of the meal. Next to the **bowl** of noodles, there is a box of cigarettes and a pack of…
Nullu ✕1 ✓3/4 rank k = 4	The image features a dining **table** with a variety of foods and drinks. On the **table**, there's a box of instant noodles, a **bowl** of noodles, and a **cup** of soda. A **remote** control is also present on the **table**, likely used for entertainment purposes.
VCD ✕0 ✓1/4 α = 1	The image features a white dining **table** with a variety of food and drink items on it. There is a box of noodles, specifically ramen noodles, prominently placed in the center of the **table**. Beside the noodles, there is a glass of beer and a carton…
PAI ✕2 ✓2/4 α = 0.5	The image features a dining **table** with a variety of food and beverages. On the **table**, there is a box of noodles, a **bottle** of Saka soda, and a **remote** control. The noodles are placed in a **bowl**, and the soda is in a **bottle**…
ProjectAway ✕1 ✓1/4 α = 3.5	The image features a dining **table** with a variety of food and drink items on it. There is a **bowl** filled with noodles, which appears to be the main focus of the meal. Next to the **bowl** of noodles, there is a box of cigarettes and a pack of…
Placebo ✕0 ✓3/4 ★ δ = −0.75	The image features a dining **table** with a variety of food and drink items on it. There is a box of instant noodles, which appears to be the main focus of the meal, accompanied by a **cup** of soda. Additionally, there is a **remote** control and a pack of cigarettes placed…



COCO #110027 · COCO boxes
annotated (5): refrigerator microwave bowl dining table chair

Unedited ✕1 ✓3/5 beam 3	The image depicts a cluttered kitchen with a wooden dining **table** in the center, surrounded by several **chairs**. A **refrigerator** can be seen on the left side of the room, while an **oven** is located on the right side. Various items are scattered throughout the kitchen…
Nullu ✕2 ✓2/5 rank k = 4	The image depicts a cozy dining room with a wooden dining **table** surrounded by six **chairs**. The **chairs** are arranged around the **table**, creating an inviting atmosphere for meals and gatherings. The dining **table** is adorned with various items such as **books**, a **laptop**, and a…
VCD ✕0 ✓3/5 α = 1	The image depicts a cozy and cluttered garage turned into a makeshift classroom. The room features a dining **table** surrounded by several **chairs**, with some **chairs** placed around the **table** and others scattered around the room. A **refrigerator** can be seen on the left side of…
PAI ✕0 ✓2/5 α = 0.5	The image depicts a cluttered kitchen area with a wooden dining **table** surrounded by several **chairs**. The **table** is surrounded by **chairs**, with one on the left side, another on the right side, and two more **chairs** on the right side of the **table**. There is a re…
ProjectAway ✕0 ✓2/5 α = 3.5	The image depicts a small, cluttered room with a wooden **table** in the center, surrounded by several **chairs**. The **table** appears to be a dining **table**, and it is surrounded by **chairs** on all sides. There are three **chairs** on the left side, three on the right side, and…
Placebo ✕0 ✓4/5 ★ random fold k = 28	The image depicts a cluttered kitchen with a dining **table** surrounded by several **chairs**. The **table** is positioned in the center of the room, and the **chairs** are placed around it. There is a **refrigerator** on the left side of the room, and a **microwave** is located…



COCO #300753 · COCO boxes
annotated (4): sports ball baseball glove baseball bat person

Unedited ✕0 ✓3/4 beam 3	The image captures an exciting moment during a baseball game, with a **pitcher** in the process of throwing the **ball** towards the batter. The batter is holding a **baseball bat**, ready to swing at the incoming pitch. There are several other **people** in the scene, including teammates, opponents,…
Nullu ✕0 ✓3/4 rank k = 4	The image captures a thrilling moment during a professional baseball game. The **pitcher** is in the process of throwing the **ball** to the batter, who is poised and ready to swing. The catcher is crouched behind the batter, wearing a **baseball glove**, ready to catch the **ball** if necessary.
VCD ✕0 ✓4/4 ★ α = 1	The image captures a thrilling moment during a baseball game. A **pitcher** is in the process of throwing the **ball** towards the batter, who is holding a **baseball bat** and preparing to swing. The catcher is positioned behind the batter, wearing a **baseball glove**, ready to catch the **ball** if the…
PAI ✕0 ✓3/4 α = 0.5	The image captures an exciting moment during a baseball game, with a batter getting ready to hit the **ball**. The batter is holding a **baseball bat**, preparing to swing at the incoming **ball**. There are several other **players** on the field, including a catcher, a **pitcher**, and a batter.
ProjectAway ✕0 ✓3/4 α = 3.5	The image captures an exciting moment during a baseball game, with a **pitcher** in the process of throwing the **ball** to the batter. The batter is holding a **baseball bat**, ready to swing at the incoming pitch. There are several other **people** in the scene, including teammates, opponents,…
Placebo ✕0 ✓3/4 beam 2	The image depicts a baseball game in progress, with a **pitcher** in the middle of throwing the **ball** to the batter. The batter is holding a **baseball bat**, ready to swing at the incoming pitch. The **pitcher** is positioned on the left side of the image, while the batter is on the right side…

Figure 9: **LLaVA-1.5-7B, COCO (4/4).** Rows: unedited model; Nullu, VCD, PAI and ProjectAway at the settings audited in the main text; best placebo for the image. Starred: the placebo on 2 of 4 images, a published method on 2. Legend and selection: App. G.

dog annotated, named cat made up ×made-up objects ✓objects named / annotated ★best row: fewest made up, most named placebo



AMBER #175 · AMBER has no boxes
annotated (7): keyboard table mouse mousepad computer apple cable

Unedited ×1 ✓3/7 beam 3
The image features a white Apple computer mouse sitting on a desk next to a white keyboard. The mouse is positioned towards the right side of the keyboard, while the keyboard occupies the majority of the left side of the desk. In addition to the mouse and keyboard, there is a cell phone…

Nullu ×0 ✓3/7 rank k = 4
The image features a white Apple computer mouse with a clear plastic casing, sitting on a desk. The mouse is positioned perfectly in the center of the desk. Next to the mouse, there's a white keyboard, which is an essential part of the computer setup. In addition to the…

VCD ×2 ✓3/7 α = 1
The image features a white Apple computer mouse sitting on a desk next to a white computer keyboard. The mouse is positioned towards the right side of the keyboard, while the keyboard occupies the majority of the left side of the desk. In the background, there is a TV mounted on the wall…

PAI ×0 ✓3/7 α = 0.5
In the image, there is a white Apple computer mouse sitting next to a white computer keyboard. The computer mouse has a clear plastic case, which adds a touch of elegance to the device. The computer mouse and keyboard are placed on a desk, with the mouse sitting on the left side of the keyboard.

ProjectAway ×1 ✓3/7 α = 3.5
The image features a white Apple computer mouse sitting on a desk next to a white keyboard. The mouse is positioned towards the right side of the keyboard, while the keyboard occupies the majority of the left side of the desk. In addition to the mouse and keyboard, there is a cell phone…

Placebo ×0 ✓4/7 ★ beam 2
The image features a white Apple computer mouse sitting on a desk next to a white keyboard. The mouse is positioned slightly to the right of the keyboard, and both devices are placed on a mouse pad. The desk appears to be made of wood, providing a sturdy surface for the computer peripherals…



AMBER #818 · AMBER has no boxes
annotated (4): sky sea bird beach

Unedited ×1 ✓3/4 beam 3
The image captures a beautiful beach scene with a seagull standing prominently on the wet sand near the water's edge. The bird appears to be walking along the shoreline, enjoying its time by the ocean. In the background, there are a few people scattered across the beach,…

Nullu ×0 ✓1/4 rank k = 4
The image captures a beautiful beach scene with a seagull standing prominently on the wet sandy beach. The seagull's reflection can be seen in the wet sand, creating an interesting visual effect. In the background, waves are crashing onto the shore, adding to the serene atmosphere of…

VCD ×1 ✓2/4 α = 1
The image captures a beautiful beach scene with a seagull standing on the wet sand near the water's edge. The seagull is facing the ocean, possibly observing the waves or searching for food. In the background, a person can be seen swimming in the ocean, enjoying…

PAI ×0 ✓2/4 α = 0.5
In the image, there is a seagull standing on the beach near the ocean. The seagull is walking along the shoreline, with the ocean in the background. The seagull appears to be the only bird in the scene, as it stands alone on the beach. The se…

ProjectAway ×0 ✓2/4 α = 3.5
The image captures a beautiful beach scene with a seagull standing on the wet sand near the water's edge. The seagull appears to be walking along the shoreline, enjoying its time by the ocean. The beach is bathed in sunlight, creating a warm and inviting atmosphere.

Placebo ×0 ✓3/4 ★ beam 1
The image features a seagull standing on the wet sand of a beach, close to the water's edge. The bird is facing the ocean, possibly observing the waves or searching for food. The beach is lined with waves, creating a serene and picturesque scene.



AMBER #460 · AMBER has no boxes
annotated (4): grass sky sheep forest

Unedited ×0 ✓2/4 beam 3
The image depicts a peaceful scene of a herd of sheep grazing on a lush green field. There are at least 14 sheep scattered throughout the field, with some closer to the foreground and others further away. The sheep are enjoying the abundance of grass in the pasture,…

Nullu ×0 ✓3/4 ★ rank k = 4
The image depicts a peaceful scene of a herd of sheep grazing in a lush green field. There are thirteen sheep in total, spread out across the field. Some sheep are grazing on the grass, while others are resting. The field is dotted with trees, providing sh…

VCD ×1 ✓1/4 α = 1
The image depicts a peaceful scene of a large herd of sheep grazing on a lush green field. There are at least 14 sheep scattered throughout the field, with some closer to the foreground and others further in the background. A person can be seen standing near the right side of the…

PAI ×0 ✓1/4 α = 0.5
The image depicts a peaceful scene of a large herd of sheep grazing in a lush green field. There are at least 14 sheep in the field, with some of them standing close to each other, while others are scattered throughout the field. The sheep are grazing on the lush…

ProjectAway ×0 ✓2/4 α = 3.5
The image depicts a peaceful scene of a herd of sheep grazing on a lush green field. There are at least 14 sheep scattered throughout the field, with some closer to the foreground and others further away. The sheep are enjoying the abundance of grass in the pasture,…

Placebo ×0 ✓2/4 beam 2
The image depicts a peaceful scene of a herd of sheep grazing on a lush green field. There are at least 14 sheep scattered throughout the field, with some closer to the foreground and others further away. The sheep are enjoying the abundance of grass in the pasture,…



AMBER #910 · AMBER has no boxes
annotated (8): sky cloud grass house mountain tree window road

Unedited ×0 ✓3/8 beam 3
The image depicts a lush green hillside with a small white house nestled in the middle of a grassy field. The house is surrounded by a serene landscape, with a cloudy blue sky serving as the backdrop. The hillside is dotted with several trees, adding to the picturesque…

Nullu ×1 ✓0/8 rank k = 4
The image depicts a lush green hillside with a charming farmhouse nestled in the middle of it. The farmhouse is surrounded by a lush green field, creating a serene and picturesque scene. In addition to the farmhouse, there are several cows grazing peacefully throughout the…

VCD ×0 ✓3/8 α = 1
The image depicts a lush green hillside with a picturesque village nestled in the valley below. The village is surrounded by a beautiful, grassy field, and the sky above is filled with clouds, creating a serene atmosphere. There are several houses scattered throughout the valley, with some…

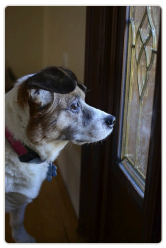
PAI ×0 ✓1/8 α = 0.5
The image depicts a picturesque scene of a small house situated in the middle of a lush green field. The house is surrounded by a lush green field, and it appears to be situated in the middle of a valley. The house is surrounded by a beautiful landscape, with a hillside in the background…

ProjectAway ×0 ✓3/8 α = 3.5
The image depicts a lush green hillside with a small white house nestled in the middle of a grassy field. The house is surrounded by a serene landscape, with a cloudy blue sky serving as the backdrop. The hillside is dotted with several trees, adding to the picturesque…

Placebo ×0 ✓5/8 ★ beam 2
The image depicts a lush green hillside with a small white house nestled in the middle of a grassy field. The house is surrounded by a beautiful landscape, featuring a mix of green grass and clouds in the sky. The hillside is dotted with a few trees, adding to the serene…

Figure 10: **LLaVA-1.5-7B, AMBER (1/4).** Rows: unedited model; Nullu, VCD, PAI and ProjectAway at the settings audited in the main text; best placebo for the image. Starred: the placebo on 3 of 4 images, a published method on 1. Legend and selection: App. G.

dog annotated, named cat made up ×made-up objects ✓objects named / annotated ★best row: fewest made up, most named placebo



AMBER #401 · AMBER has no boxes
annotated (6): dog wall door floor window necklace

Unedited	×0 ✓3/6 beam 3	The image features a brown and white dog standing in a room, looking out of a window. The dog appears to be focused on something outside, possibly observing its surroundings or waiting for its owner. The dog is wearing a pink collar, adding a touch of color to the scene.
Nullu	×2 ✓3/6 rank k = 4	The image features a brown and white dog standing in a living room, looking out of a sliding glass door. The dog's gaze is directed outdoors, capturing the viewer's attention. The living room is furnished with a couch and a chair, creating a cozy atmosphere.
VCD	×1 ✓3/6 α = 1	The image features a brown and white dog standing in a living room, attentively looking through a glass door. The dog appears to be curious about what is happening on the other side of the door. In the room, there is a couch located in the background, and a chair can be seen…
PAI	×0 ✓3/6 α = 0.5	In the image, a brown and white dog is standing in front of a doorway, looking out of the window with a curious expression. The dog is wearing a pink collar, which adds a touch of color to its white and brown fur. The dog appears to be waiting for someone to let it in…
ProjectAway	×0 ✓3/6 α = 3.5	The image features a brown and white dog standing in a room, looking out of a window. The dog appears to be focused on something outside, possibly observing its surroundings or waiting for its owner. The dog is wearing a pink collar, adding a touch of color to the scene.
Placebo	×0 ✓4/6 ★ δ = −0.5	The image features a brown and white dog standing in front of a door, looking out through the window with anticipation. The dog is wearing a pink collar and appears to be eagerly waiting for someone or something to arrive. The scene captures the dog's attention as it gazes out the window…



AMBER #37 · AMBER has no boxes
annotated (5): person forest snow ground road

Unedited	×0 ✓3/5 beam 3	The image depicts a man walking down a snow-covered path between two rows of trees. The path is surrounded by trees on both sides, creating a serene and peaceful atmosphere. The man appears to be the only person in the scene, as he makes his way through the snowy landscape.
Nullu	×0 ✓2/5 rank k = 4	The image depicts a man walking down a snow-covered pathway lined with trees. He is the only person visible in the scene, making his way through the snowy landscape. The snow-covered pathway stretches out into the distance, creating a serene and peaceful atmosphere.
VCD	×1 ✓3/5 α = 1	The image depicts a man walking down a snow-covered path in a park, surrounded by trees. The path is lined with snow-covered trees, creating a picturesque winter scene. The man appears to be walking alone, enjoying the serene atmosphere. There are two benches…
PAI	×0 ✓4/5 ★ α = 0.5	In the image, a person is walking down a snow-covered path in a park, surrounded by a row of trees. The person appears to be the only person in the scene, as they are walking down the path in the middle of the scene. The path is covered in a light dusting of snow,…
ProjectAway	×0 ✓3/5 α = 3.5	The image depicts a man walking down a snow-covered path between two rows of trees. The path is lined with trees on both sides, creating a serene and picturesque scene. The man appears to be enjoying the peaceful atmosphere as he strolls through the snow-covered landscape…
Placebo	×0 ✓3/5 beam 2	The image depicts a man walking down a snow-covered path in a park, surrounded by trees. The path is lined with trees on both sides, creating a serene and peaceful atmosphere. The man appears to be enjoying his walk in the snowy landscape. In addition to the…



AMBER #299 · AMBER has no boxes
annotated (4): sky cloud ship sea

Unedited	×0 ✓3/4 beam 3	The image features a small boat floating on a body of water, possibly the ocean. The boat is positioned in the middle of the scene, surrounded by a vast expanse of water. The sky above the boat appears to be cloudy, adding to the overall atmosphere of the scene.
Nullu	×0 ✓3/4 rank k = 4	The image features a small boat floating gently on the water. The boat appears to be old and weathered, and it's drifting alone in the ocean. The sky above the boat is cloudy, casting a moody atmosphere over the scene. The boat's presence in the vast expanse of water…
VCD	×1 ✓2/4 α = 1	The image features a small wooden boat floating on a body of water, possibly the ocean. The boat appears to be old and weathered, suggesting it has been used for quite some time. It is anchored in the middle of the water, with a rope tied to it, ensuring it stays in place.
PAI	×0 ✓3/4 α = 0.5	In the image, there is a small wooden boat floating on a body of water, possibly the ocean. The boat appears to be a canoe, and it is situated in the middle of the body of water. The boat is surrounded by a cloudy sky, which adds a dramatic backdrop to the scene.
ProjectAway	×0 ✓3/4 α = 3.5	The image features a small boat floating on a body of water, possibly the ocean. The boat is positioned in the middle of the scene, surrounded by a vast expanse of water. The sky above the boat appears to be cloudy, adding to the overall atmosphere of the scene.
Placebo	×0 ✓4/4 ★ δ = −2	The image depicts a small boat floating on a large body of water, possibly the ocean. The boat is positioned in the middle of the scene, surrounded by a vast expanse of water. The sky above the scene is filled with clouds, creating a dramatic and moody atmosphere. There are…



AMBER #428 · AMBER has no boxes
annotated (3): tree grass sheep

Unedited	×1 ✓1/3 beam 3	The image depicts a lush green field where a herd of horned animals, possibly goats or antelopes, are grazing on the grass. The animals are scattered throughout the field, with some closer to the foreground and others further in the background. In total, there are nine animals…
Nullu	×0 ✓1/3 rank k = 4	The image depicts a group of five horned animals grazing peacefully in a lush green field. The animals are evenly spaced throughout the field, with some of them eating grass and others looking around. The animals' horns are clearly visible as they graze. The field is…
VCD	×0 ✓0/3 α = 1	The image depicts a group of horned animals, likely rams or goats, grazing in a lush green field. There are five rams in total, spread out across the field, with some closer to the foreground and others further in the background. They are all focused on eating the…
PAI	×0 ✓1/3 α = 0.5	The image features a herd of horned animals, likely goats, grazing in a lush green field. There are five goats in the scene, with some of them eating grass and others standing around. The goats are scattered throughout the field, with some standing closer to the foreground and others…
ProjectAway	×1 ✓1/3 α = 3.5	The image depicts a lush green field where a herd of horned animals, possibly goats or antelopes, are grazing on the grass. The animals are scattered throughout the field, with some closer to the foreground and others further in the background. In total, there are nine animals…
Placebo	×0 ✓2/3 ★ beam 1	The image features a group of five horned animals, likely goats or sheep, grazing on a lush green field. They are spread out across the field, with some closer to the foreground and others further in the background. The animals are peacefully eating the grass, enjoying the serene environment…

Figure 11: **LLaVA-1.5-7B, AMBER (2/4).** Rows: unedited model; Nullu, VCD, PAI and ProjectAway at the settings audited in the main text; best placebo for the image. Starred: the placebo on 3 of 4 images, a published method on 1. Legend and selection: App. G.

dog annotated, named cat made up ✕made-up objects ✓objects named / annotated ★best row: fewest made up, most named placebo



AMBER #40 · AMBER has no boxes

annotated (5): person, forest, grass, flower, dog

Unedited	✕0 ✓1/5 beam 3	The image features a brown dog running through a lush green field, with its mouth open and tongue hanging out. The dog appears to be enjoying its time outdoors, possibly playing or exploring the area. In the background, there are several potted plants scattered throughout the scene, adding a…
Nullu	✕0 ✓2/5 ★ rank k = 4	The image features a brown dog running through a lush green field. The dog's tongue is hanging out, and it appears to be enjoying its time outdoors. The field is dotted with trees, creating a serene environment for the dog's playtime. Aside from the…
VCD	✕0 ✓1/5 α = 1	The image features a brown dog running through a lush green field, with its mouth open and tongue hanging out. The dog appears to be enjoying its time in the grassy area, possibly playing or exploring. In the background, there are several potted plants scattered throughout the scene, adding…
PAI	✕1 ✓1/5 α = 0.5	In the image, there is a brown dog running through a grassy field. The dog appears to be enjoying a game of chase with a frisbee, as it runs through the grassy field. The dog is in the middle of the scene, with its mouth open and tongue hanging out,…
ProjectAway	✕0 ✓1/5 α = 3.5	The image features a brown dog running through a lush green field, with its mouth open and tongue hanging out. The dog appears to be enjoying its time outdoors, possibly playing or exploring the area. In the background, there are several potted plants scattered throughout the scene, adding a…
Placebo	✕0 ✓1/5 beam 2	The image captures a brown dog running through a lush green field, with its mouth open and tongue hanging out. The dog appears to be enjoying its time outdoors, possibly playing or exploring the area. In the background, there are several potted plants scattered throughout the scene, adding…



AMBER #584 · AMBER has no boxes

annotated (6): person, racket, grass, court, necklace, tennis

Unedited	✕1 ✓5/6 beam 3	The image captures a woman playing tennis on a grass court. She is in the middle of a swing, holding a tennis racket and preparing to hit the ball. The woman is wearing a white outfit, which is typical attire for tennis players. The tennis ball can be seen in the…
Nullu	✕1 ✓3/6 rank k = 4	The image captures a female tennis player in action on a tennis court. She is wearing a white dress and is swinging her tennis racket skillfully. The player's body is twisted as she hits the ball, showcasing her agility and athleticism. The tennis ball can be seen…
VCD	✕0 ✓5/6 ★ α = 1	The image captures a woman playing tennis on a grass court. She is wearing a white dress and is in the middle of a swing, holding a tennis racket in her hand. The woman appears to be focused and determined as she plays the game. The tennis court is surrounded by a lush green…
PAI	✕0 ✓2/6 α = 0.5	In the image, a female tennis player is in the middle of her swing, with her tennis racket raised in the air. She is wearing a white dress and a white headband. The tennis player appears to be in the middle of her swing, displaying her athleticism and skill in the sport. The…
ProjectAway	✕1 ✓5/6 α = 3.5	The image captures a woman playing tennis on a grass court. She is in the middle of a swing, holding a tennis racket and preparing to hit the ball. The woman is wearing a white outfit, which is typical attire for tennis players. The tennis ball can be seen in the…
Placebo	✕0 ✓5/6 ★ 32 tokens	The image captures a woman playing tennis on a grass court. She is in the middle of a swing, holding a tennis racket with both hands, and…



AMBER #76 · AMBER has no boxes

annotated (8): sky, cloud, mountain, river, person, ship, paddle, hat

Unedited	✕0 ✓3/8 beam 3	In the image, a person is paddling a small yellow boat on a large body of water, possibly a lake or a river. The person appears to be enjoying their time on the water, surrounded by a serene environment. The boat is positioned in the middle of the scene, and the person is…
Nullu	✕0 ✓3/8 rank k = 4	The image captures a man paddling a yellow boat through calm waters. He is skillfully navigating his boat through the lake, enjoying his time outdoors. The man's boat is the main focus of the scene, and he is the only person visible in the image. In the…
VCD	✕1 ✓3/8 α = 1	In the image, a person is paddling a small yellow boat on a large body of water, possibly a lake or a river. The person is wearing a backpack and appears to be enjoying their time on the water. The boat is positioned in the middle of the scene, and the person is…
PAI	✕0 ✓4/8 ★ α = 0.5	In the image, a person is paddling a small yellow boat on a large body of water, likely a lake, surrounded by mountains. The person appears to be enjoying a leisurely day on the water, surrounded by a beautiful mountain range. The person is wearing a life jacket, which…
ProjectAway	✕0 ✓3/8 α = 3.5	In the image, a person is paddling a small yellow boat on a large body of water, possibly a lake or a river. The person appears to be enjoying their time on the water, surrounded by a serene environment. The boat is positioned in the middle of the scene, and the person is…
Placebo	✕0 ✓4/8 ★ beam 2	In the image, a person is paddling a small yellow boat on a calm lake, surrounded by mountains. The boat is positioned in the middle of the lake, and the person appears to be enjoying their time on the water. The mountains in the background provide a picturesque and serene setting for this…



AMBER #542 · AMBER has no boxes

annotated (4): grass, chicken, fence, wire

Unedited	✕0 ✓2/4 beam 3	The image features two brown and white chickens standing next to each other in a fenced-in area. The chickens are positioned close to a wooden fence, with one chicken on the left side and the other on the right side of the fence. Both chickens appear to be…
Nullu	✕0 ✓2/4 rank k = 4	The image features two chickens standing in a grassy enclosure. One chicken is on the left-hand side of the enclosure, and the other is on the right-hand side. Both chickens are pecking at the grassy ground, exploring their surroundings. The en…
VCD	✕0 ✓3/4 ★ α = 1	The image features two brown and white chickens standing next to each other in a grassy field. They are positioned close to a wire fence, with one chicken on the left side and the other on the right side of the fence. The chickens seem to be enjoying their time in…
PAI	✕1 ✓2/4 α = 0.5	In the image, there are two chickens standing next to each other in a fenced-in area. The chickens are standing near a wooden fence, possibly in a pen. One chicken is on the left side of the image, while the other chicken is on the right side. Both…
ProjectAway	✕0 ✓2/4 α = 3.5	The image features two brown and white chickens standing next to each other in a fenced-in area. The chickens are positioned close to a wooden fence, with one chicken on the left side and the other on the right side of the fence. Both chickens appear to be…
Placebo	✕0 ✓2/4 beam 2	The image features two brown and white chickens standing next to each other in a grassy field. They are positioned close to a wooden fence, which serves as a barrier between them and the viewer. The chickens appear to be enjoying their time outdoors, possibly grazing or…

Figure 12: **LLaVA-1.5-7B, AMBER (3/4).** Rows: unedited model; Nullu, VCD, PAI and ProjectAway at the settings audited in the main text; best placebo for the image. Starred: the placebo on 2 of 4 images, a published method on 4. Legend and selection: App. G.

dog annotated, named cat made up ×made-up objects ✓objects named / annotated ★best row: fewest made up, most named placebo



AMBER #491 · AMBER has no boxes
annotated (6): wall sofa woman shoe laptop floor

Unedited	×1 ✓3/6 beam 3	In the image, a woman is sitting on a red couch with her legs crossed. She is wearing a suit and tie, giving her a professional appearance. A laptop is placed on the couch in front of her, indicating that she might be working or browsing the internet. Additionally,…
Nullu	×2 ✓3/6 rank k = 4	In this image, a woman is sitting on a red couch with her feet propped up. She is wearing a suit and tie, giving her a professional appearance. A laptop computer is placed on her lap, and she seems to be working or browsing the internet. There are several books scattered…
VCD	×1 ✓3/6 α = 1	The image features a woman sitting on a red couch, wearing a suit and tie. She appears to be working on her laptop, which is placed on the couch in front of her. The woman seems to be posing for the camera, possibly for a professional photo. In addition to the…
PAI	×1 ✓3/6 α = 0.5	In the image, a young woman is sitting on a red couch, wearing a suit and tie. She appears to be looking at the camera, possibly posing for a photo. The couch she is sitting on is red in color. A laptop computer is placed on the couch next to…
ProjectAway	×1 ✓3/6 α = 3.5	In the image, a woman is sitting on a red couch with her legs crossed. She is wearing a suit and tie, giving her a professional appearance. A laptop is placed on the couch in front of her, indicating that she might be working or browsing the internet. There are two…
Placebo	×0 ✓3/6 ★ δ = −0.5	The image features a young woman sitting on a red couch in a living room. She is wearing a business suit and has her feet propped up on the couch. There is a laptop placed on the couch in front of her, indicating that she might be working or browsing the internet.



AMBER #39 · AMBER has no boxes
annotated (6): sky cloud person bicycle sea beach

Unedited	×1 ✓3/6 beam 3	The image features a man riding a bicycle on a sandy beach near the ocean. He is wearing a backpack and appears to be enjoying his time outdoors. The bicycle is positioned in the center of the scene, with the man riding it towards the right side of…
Nullu	×0 ✓3/6 rank k = 4	The image features a man riding a bicycle on a beach. He is wearing a white shirt and appears to be enjoying his time outdoors. The bicycle is prominently visible in the scene, with its wheels touching the sandy beach. The man is riding…
VCD	×0 ✓3/6 α = 1	The image features a man riding a bicycle on a sandy beach near the ocean. He is enjoying a leisurely ride along the shoreline, taking in the beautiful surroundings. The beach appears to be relatively empty, with no other people visible in the scene. The…
PAI	×0 ✓3/6 α = 0.5	In the image, a man is riding his bicycle on a beach near the ocean. He appears to be enjoying a leisurely bike ride along the shoreline. The man is wearing a white shirt and riding a bicycle on the beach. The…
ProjectAway	×1 ✓3/6 α = 3.5	The image features a man riding a bicycle on a sandy beach near the ocean. He is wearing a backpack and appears to be enjoying his time outdoors. The bicycle is positioned in the center of the scene, with the man riding it towards the right side of…
Placebo	×0 ✓5/6 ★ δ = −2	The image depicts a man riding a bicycle on a beach near the ocean. He is wearing a white shirt and appears to be enjoying his ride along the coastline. The scene is serene, with the ocean in the background and a few clouds in the sky. The…



AMBER #3 · AMBER has no boxes
annotated (3): child grass flower

Unedited	×0 ✓1/3 beam 3	The image depicts a young child standing in a grassy field, wearing a striped shirt. The child appears to be crying or upset, possibly due to a fall or an uncomfortable situation. The child is surrounded by a lush green field, which provides a serene and…
Nullu	×0 ✓1/3 rank k = 4	The image features a young child standing in a grassy field, wearing a striped shirt. The child appears to be crying or upset as they stand in the grassy area. The child's emotional state is evident from their facial expression and body language.
VCD	×0 ✓2/3 ★ α = 1	The image features a young child standing in a grassy field, wearing a striped shirt and grey pants. The child appears to be crying or upset, possibly due to a fall or an uncomfortable situation. The field is filled with grass, providing a natural and open environment for the…
PAI	×1 ✓1/3 α = 0.5	In the image, there is a small child standing in a lush green field, wearing a striped shirt. The child appears to be crying while standing in the grassy field. The baby appears to be upset, possibly due to the presence of a dog nearby. The dog is located on the…
ProjectAway	×0 ✓1/3 α = 3.5	The image depicts a young child standing in a grassy field, wearing a striped shirt. The child appears to be crying or upset, possibly due to a fall or an uncomfortable situation. The child is surrounded by a lush green field, which provides a serene and…
Placebo	×0 ✓1/3 beam 2	The image features a young child standing in a grassy field, wearing a striped shirt and grey pants. The child appears to be crying or upset, possibly due to a fall or an uncomfortable situation. The grassy field provides a natural and open environment for the child to explore…

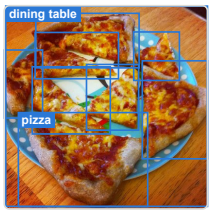


AMBER #581 · AMBER has no boxes
annotated (8): sky cloud train ground grass tree wire flag

Unedited	×0 ✓1/8 beam 3	The image features a yellow and black train car with the words "Chesie System" written on its side. The train car is sitting on a set of train tracks, and it appears to be an old-fashioned passenger train. The train car is positioned in the middle of the scene, with a…
Nullu	×0 ✓1/8 rank k = 4	The image features an old-fashioned yellow train engine with "Chesie System" written on its side, traveling down train tracks. The train engine is the main focus of the scene, taking up a significant portion of the image. In addition to the train engine, there are a few…
VCD	×1 ✓1/8 α = 1	The image features a yellow and red train car with the words "Chesie System" written on its side. The train car is sitting on a set of train tracks, and it appears to be an old-fashioned locomotive. In addition to the train car, there are two people visible…
PAI	×0 ✓1/8 α = 0.5	The image features a yellow train car with the words "Chesie System" written on the side of the train car. The train car is sitting on the tracks, and it appears to be an old-fashioned train car. The train car has a red and yellow color scheme, giving it a vint…
ProjectAway	×0 ✓1/8 α = 3.5	The image features a yellow and black train car with the words "Chesie System" written on its side. The train car is sitting on a set of train tracks, and it appears to be an old-fashioned passenger train. The train car is positioned in the middle of the scene, with a…
Placebo	×0 ✓2/8 ★ δ = −2	The image depicts a yellow and black train traveling down the tracks, with a red, white, and blue flag flying above it. The train is positioned in the middle of the scene, and it appears to be an old-fashioned locomotive. In the background, there is a…

Figure 13: **LLaVA-1.5-7B, AMBER (4/4).** Rows: unedited model; Nullu, VCD, PAI and ProjectAway at the settings audited in the main text; best placebo for the image. Starred: the placebo on 3 of 4 images, a published method on 1. Legend and selection: App. G.

dog annotated, named cat made up ×made-up objects ✓objects named / annotated ★best row: fewest made up, most named placebo



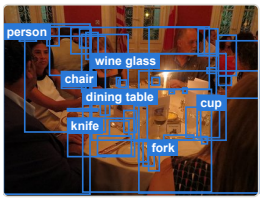
COCO #297574 · COCO boxes

annotated objects (2): dining table, pizza

Unedited ×2 ✓1/2 beam 3
The image is a plate with three slices of pizza on it. The pizza slices are topped with cheese, pepperoni, and mushrooms. The plate is on a wooden surface, and there is a fork and knife next to it.

Nullu ×0 ✓1/2 rank k = 8
This is an image of a pizza with various toppings on a plate. The pizza has a crispy crust and melted cheese on top. The toppings include pepperoni, sausage, mushrooms, and onions. The pizza is cut into slices and served…

Placebo ×0 ✓2/2 ★ beam 1
The image is a plate with three slices of pizza on it. Each slice has cheese and toppings on it, such as pepperoni and mushrooms. The plate is sitting on top of a wooden table.



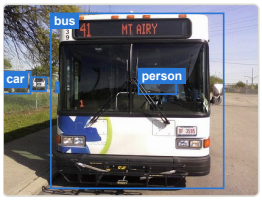
COCO #63939 · COCO boxes

annotated objects (8): spoon, dining table, cup, knife, chair, fork, person, wine glass

Unedited ×1 ✓2/8 beam 3
The image shows a group of people sitting around a table in a dimly lit room. They are dressed in formal attire and appear to be enjoying a meal together. There are candles on the table and a vase of flowers in the center. The walls are painted a warm red color and there are…

Nullu ×0 ✓3/8 ★ rank k = 8
The image shows a group of people seated around a table at a dinner party. The table is set with white linens and the chairs are upholstered in red and white striped fabric. The room is dimly lit by candles and the walls are adorned with portraits of…

Placebo ×0 ✓2/8 beam 1
The image shows a group of people sitting around a table at a dinner party. They are dressed in formal attire and appear to be enjoying their meal. There is a candle on the table, casting a warm glow over the scene. The walls of the room are painted a deep red color and there…



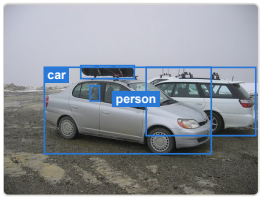
COCO #20273 · COCO boxes

annotated objects (3): bus, person, car

Unedited ×2 ✓0/3 beam 3
This is an image of a bus stop with a bus parked in front of it. The bus has the words "City Transit" written on the side in white letters. The bus stop sign has the words "Bus Stop" written on it in black letters. There is a bench in front of the bus…

Nullu ×1 ✓1/3 rank k = 8
This is an image of a bus stop. The bus stop is located on a city street with a lot of greenery around it. The bus stop has a sign with the number 41 on it, indicating that it is a bus stop for route 41. The bus stop has a bench for passengers…

Placebo ×0 ✓1/3 ★ δ = −0.5
This is an image of a city bus parked on the side of the road. The bus is white with green and blue stripes on the sides and has the number 41 on the front. There are people standing on the sidewalk in front of the bus, looking at it. The sky is cloudy…



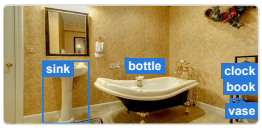
COCO #97679 · COCO boxes

annotated objects (4): skis, person, snowboard, car

Unedited ×2 ✓1/4 beam 3
The image shows two cars parked on a dirt road in front of a mountain range. The cars are a silver sedan and a white SUV. The sedan has a roof rack with two bicycles attached to it, while the SUV has a roof rack with two kayaks attached to it…

Nullu ×1 ✓1/4 rank k = 8
This is an image of two cars parked on a dirt road at the base of a mountain. One car is a silver Toyota Corolla and the other is a white Honda Civic. Both cars have roof racks with bicycles attached to them. The Toyota Corolla has…

Placebo ×0 ✓1/4 ★ beam 2
The image shows two cars parked on the side of a dirt road. One car is a silver sedan with a roof rack on top, while the other is a white SUV with a roof rack on top as well. Both cars are parked next to each other on the side of the road,…



COCO #472621 · COCO boxes

annotated objects (6): vase, book, sink, bowl, bottle, clock

Unedited ×1 ✓1/6 beam 3
The image appears to be a bathroom with a white bathtub in the center of the room. The walls are painted a light beige color and there is a white sink on one side of the room. There is also a toilet on the other side of the room. The floor is made of hard…

Nullu ×1 ✓1/6 rank k = 8
This is a bathroom with a bathtub, sink, and toilet. The walls and floor are made of beige-colored tiles, and the bathtub and sink are made of white porcelain. The toilet is made of white porcelain and has a seat and…

Placebo ×0 ✓1/6 ★ δ = −2
This is a bathroom with a white bathtub in the center of the room. The walls are painted a light yellow color and there is a large mirror on one wall. The floor is made of hardwood and there is a rug in front of the bathtub. There is also a sink on one side…



COCO #398884 · COCO boxes

annotated objects (3): truck, car, bed

Unedited ×1 ✓2/3 beam 3
This is an image of a green and black pickup truck parked on a grassy field. The truck has a flatbed and is parked next to several other vehicles. There are people standing around the truck, looking at it. The image appears to be taken at a classic car show or event…

Nullu ×0 ✓2/3 ★ rank k = 8
This is an image of a green, vintage truck parked on a grassy field at a classic car show. The truck has a flatbed and is surrounded by other vintage vehicles.

Placebo ×0 ✓1/3 beam 1
This is an image of a green truck parked on grass. The truck has a flatbed with a red and white striped tarp covering it. There are two other trucks in the background, one red and one blue.

Figure 14: **MiniGPT-4, COCO (1/3).** Rows: unedited model; Nullu at its published rank (k=8); best placebo for the image. Starred: the placebo on 4 of 6 images, a published method on 2. On COCO #20273 CHAIR never matches "bus", so the unedited and placebo rows each miss one named object; the star does not change. Legend and selection: App. G.

dog annotated, named cat made up ×made-up objects ✓objects named / annotated ★best row: fewest made up, most named placebo



COCO #548843 · COCO boxes

annotated objects (2): donut person

Unedited ×3 ✓1/2 beam 3
The image shows a young boy wearing a blue jacket and holding a slice of pizza in his hand. He is standing in front of a group of people who are also eating pizza. The background appears to be a restaurant or food court, with tables and chairs set up for customers to eat.

Nullu ×1 ✓1/2 rank k = 8
The image shows a young boy wearing a blue jacket and holding a slice of pizza in his hand. He is standing in front of a group of people, all of whom are smiling and looking at him. The boy's hair is messy and he is wearing glasses, giving him a…

Placebo ×0 ✓1/2 ★ δ = −2
The image shows a young boy standing in front of a group of people, holding a plate of food in his hands. The boy is wearing a blue jacket and glasses, and has a big smile on his face. The people in the background are dressed in various colors and are standing in a line, looking…



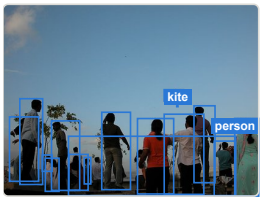
COCO #560111 · COCO boxes

annotated objects (2): train car

Unedited ×1 ✓1/2 beam 3
This is an image of a red and yellow train on the tracks. The train is parked at a train station and there are people standing on the platform looking at it. There are also other trains in the background, parked on the tracks. The sky is clear and there are no clouds in the sky.

Nullu ×1 ✓2/2 rank k = 8
This is an image of a red and yellow train at a train station. The train is made up of two cars, one in front of the other, with the front car being red and the back car being yellow. The train is parked at a platform at a train station, with people boarding and alight…

Placebo ×0 ✓2/2 ★ beam 2
The image is of a red and yellow train on the tracks. The train is a passenger train with two cars, one in front of the other. The front car has a large window on the side and a small window on the front. The back car has a large window on the side and a small window on the…



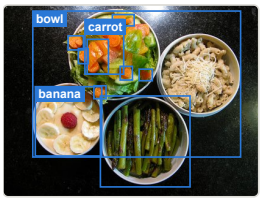
COCO #510122 · COCO boxes

annotated objects (2): person kite

Unedited ×1 ✓1/2 beam 3
The image shows a group of people standing on a wooden platform overlooking a body of water. They are all looking up at a bird flying in the sky. The sky is clear and blue, with a few clouds scattered about. The people are dressed in casual clothing, such as t-shirts and…

Nullu ×0 ✓2/2 ★ rank k = 8
This is an image of a group of people standing on a rooftop, looking up at a kite flying in the sky. The kite is made of white and red fabric and has a long tail. The people are all wearing casual clothing and are standing in a semicircle around…

Placebo ×1 ✓2/2 beam 1
The image shows a group of people standing on a wooden platform overlooking a body of water. They are all looking up at something in the sky, possibly a bird or a kite. The sky is clear and blue with some clouds scattered across it. The people are dressed in casual clothing and appear to be…



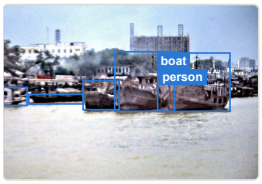
COCO #364399 · COCO boxes

annotated objects (4): carrot banana bowl dining table

Unedited ×1 ✓3/4 beam 3
The image shows a table with four bowls of food on it. The bowls are filled with different types of vegetables, such as broccoli, carrots, and green beans. There are also some pieces of chicken and rice in the bowls. The table is made of wood and has…

Nullu ×1 ✓2/4 rank k = 8
This is an image of a tray with four bowls of food on it. The bowls contain a variety of fruits and vegetables, including carrots, broccoli, and asparagus. The tray is made of stainless steel and has a smooth, glossy finish…

Placebo ×0 ✓3/4 ★ 32 tokens
The image shows a table with four bowls of food on it. The bowls are filled with different types of vegetables, such as carrots,…



COCO #383119 · COCO boxes

annotated objects (2): boat person

Unedited ×1 ✓2/2 beam 3
This is a black and white photograph of a river with several boats docked along the shore. There are several buildings visible in the background, including a large industrial building with smokestacks and a smaller building with a clock tower. The water is calm and there are no other boats or people visible in the image.

Nullu ×0 ✓1/2 rank k = 8
This is a black and white photograph of a group of boats moored in a river. The boats are old and rusted, with their hulls covered in barnacles and seaweed. The water is murky and choppy, with waves crashing against the shore. In the background, there is…

Placebo ×0 ✓2/2 ★ beam 2
This is a black and white photograph of a river with several boats docked along the shore. The boats are old and rusted, and there are several people standing on the shore. The sky is cloudy and there are some buildings visible in the background.



COCO #42276 · COCO boxes

annotated objects (3): skis toilet potted plant

Unedited ×2 ✓2/3 beam 3
The image appears to be a bathroom with a toilet, sink, and shower. There are several pairs of skis and snowboards leaning against the wall, as well as a pair of gloves and a helmet on the floor. The walls are painted a light blue color and there is a…

Nullu ×1 ✓2/3 ★ rank k = 8
This is a bathroom with a toilet, shower, and sink. The walls are painted white and the floor is made of ceramic tiles. There are several pairs of shoes lined up against the wall, including skis, snowboarding boots, and hiking boots…

Placebo ×1 ✓1/3 beam 2
The image shows a bathroom with a toilet, sink, and shower. There are several pairs of shoes and a pair of gloves on the floor. The walls are painted white and there is a window on one side of the room with curtains. The floor is made of tile and there…

Figure 15: **MiniGPT-4, COCO (2/3).** Rows: unedited model; Nullu at its published rank (k=8); best placebo for the image. Starred: the placebo on 4 of 6 images, a published method on 2. Legend and selection: App. G.

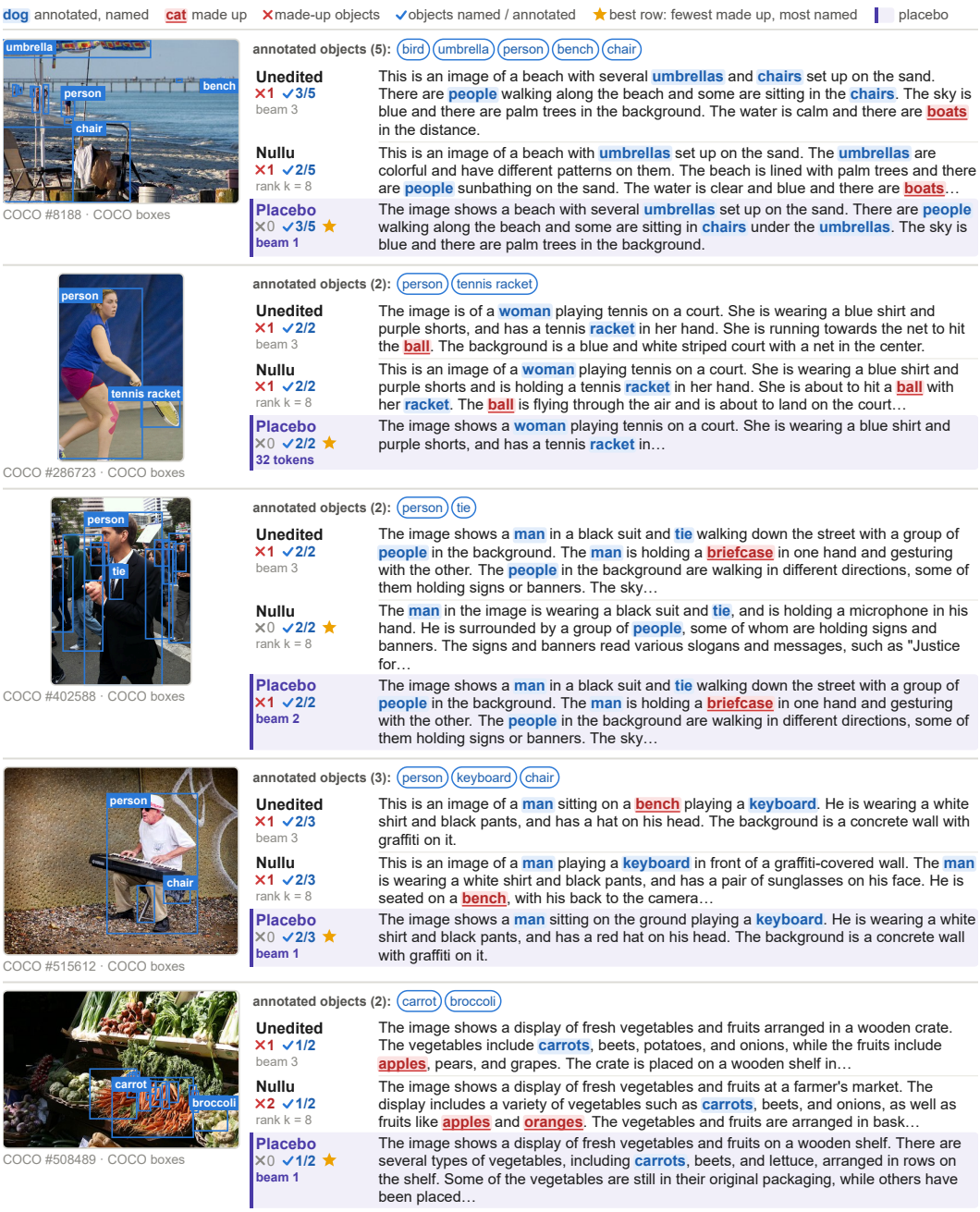


Figure 16: **MiniGPT-4, COCO (3/3).** Rows: unedited model; Nullu at its published rank (k=8); best placebo for the image. Starred: the placebo on 4 of 5 images, a published method on 1. Legend and selection: App. G.

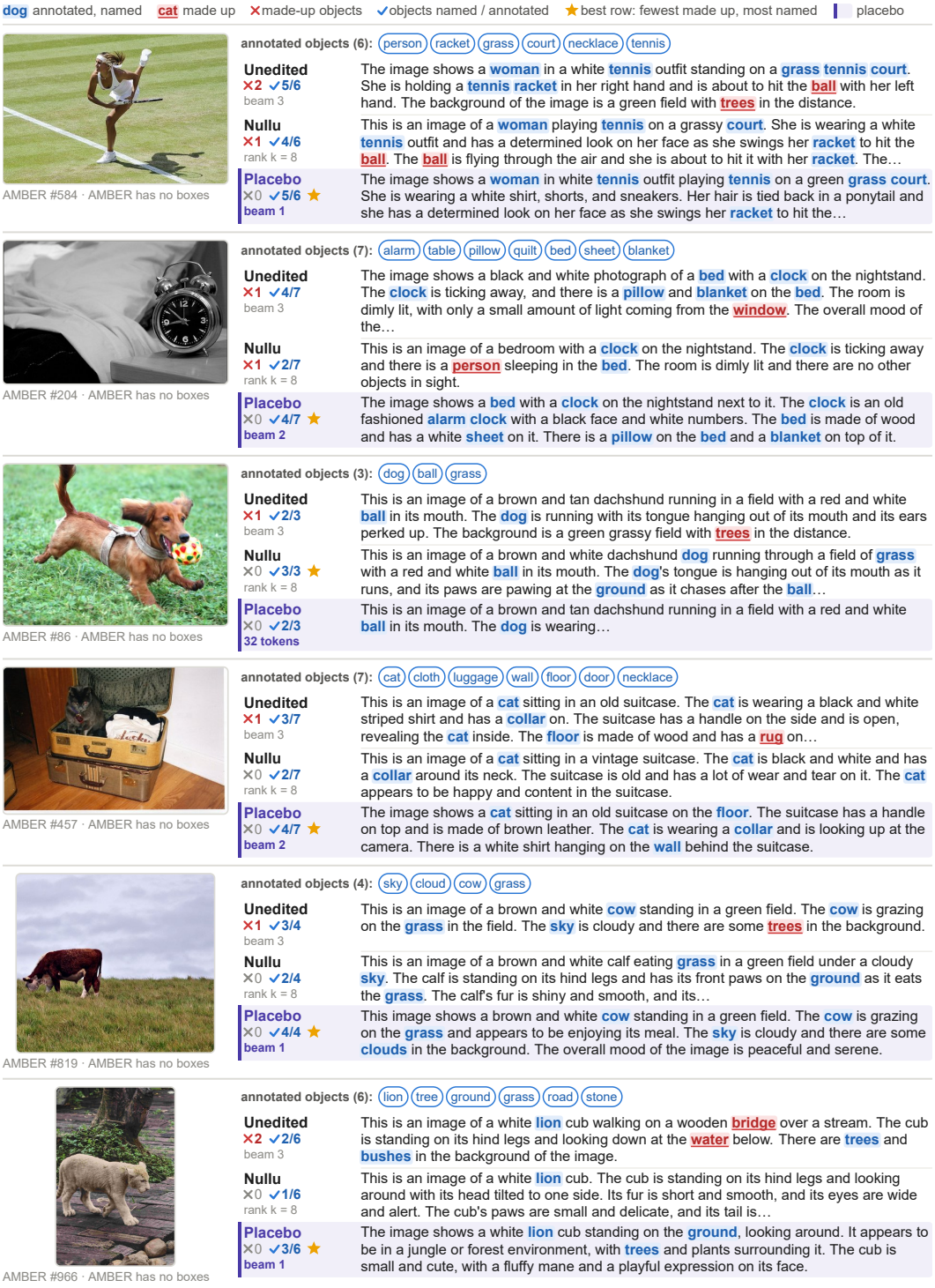


Figure 17: **MiniGPT-4, AMBER (1/3).** Rows: unedited model; Nullu at its published rank (k=8); best placebo for the image. Starred: the placebo on 5 of 6 images, a published method on 1. Legend and selection: App. G.

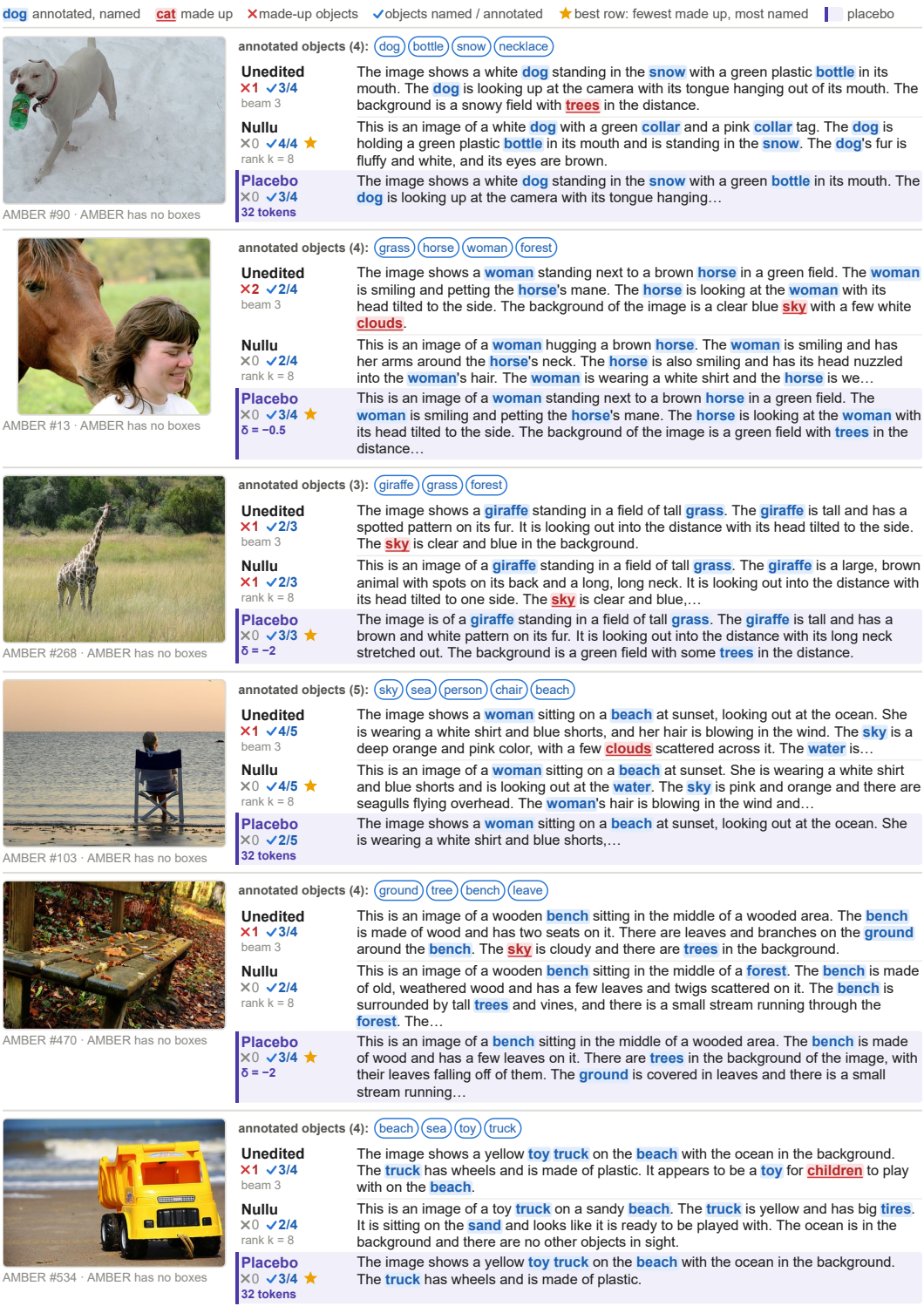


Figure 18: **MiniGPT-4, AMBER (2/3).** Rows: unedited model; Nullu at its published rank (k=8); best placebo for the image. Starred: the placebo on 4 of 6 images, a published method on 2. Legend and selection: App. G.

dog annotated, named cat made up ×made-up objects ✓objects named / annotated ★best row: fewest made up, most named placebo



AMBER #166 · AMBER has no boxes

annotated objects (2): cat grass

Unedited ×1 ✓1/2 beam 3
This is an image of a white and black cat standing in a green field. The cat is looking straight ahead with its ears perked up and its tail wagging. The background is a bright green field with some trees in the distance.

Nullu ×0 ✓2/2 ★ rank k = 8
This is an image of a white and black cat with a fluffy coat and green eyes. The cat is pawing at the grass with its front paws, as if it is about to pounce on something. The cat's ears are perked up and its tail is wagging. The…

Placebo ×0 ✓1/2 32 tokens
This is an image of a white and black cat standing in a green field. The cat is looking straight ahead with its ears perked up and its tail w…



AMBER #618 · AMBER has no boxes

annotated objects (7): wall cup toothbrush toothpaste sink mirror support

Unedited ×1 ✓3/7 beam 3
This is an image of a bathroom with a sink, toothbrush, and toothpaste on the counter. The sink is made of white ceramic and has a faucet on the right side. The toothbrush and toothpaste are on the left side of the sink. There…

Nullu ×0 ✓2/7 rank k = 8
This is an image of a bathroom with a toothbrush and toothbrush holder on the counter. The toothbrush holder is made of white plastic and has two slots for the toothbrushes to sit in. The toothbrushes are both blue and have the…

Placebo ×0 ✓3/7 ★ 32 tokens
This is an image of a bathroom with a sink, toothbrush, and toothpaste on the counter. The sink is made of white cer…



AMBER #44 · AMBER has no boxes

annotated objects (6): man grass road ground forest hat

Unedited ×1 ✓3/6 beam 3
The image shows a person walking down a dirt path in a park. The person is wearing black pants, a black shirt, and black sneakers. They are carrying a black backpack on their back. The path is lined with trees on both sides, and there is a small stream running…

Nullu ×0 ✓3/6 rank k = 8
This is an image of a young man wearing a black shirt and black pants, standing on a dirt path in a park. He is holding a skateboard in his hand and has a serious expression on his face. The path is lined with tall trees and there are no other people in sight…

Placebo ×0 ✓4/6 ★ beam 1
The image shows a person walking down a dirt path in a park. The person is wearing black pants and a black shirt with white letters on it. They are holding a skateboard in their right hand and have their left hand on their hip. The background is made up of trees and grass.



AMBER #324 · AMBER has no boxes

annotated objects (8): sky grass man forest fence road tie glass

Unedited ×1 ✓3/8 beam 3
The image shows a man in a white shirt and black pants standing in the middle of a grassy field. He is holding a ball in his hand and appears to be about to throw it. There are trees in the background and a clear blue sky above. The man is wearing sunglasses and…

Nullu ×0 ✓3/8 ★ rank k = 8
The man in the image is wearing a white shirt and black pants. He is standing on one leg, with the other leg stretched out in front of him. He is holding onto a tree branch with one hand and has a look of concentration on his face. The grass is green and the sky is…

Placebo ×0 ✓2/8 δ = −2
The image shows a man standing in the middle of a park, wearing a white shirt and black pants. He is holding a golf club in his right hand and is about to swing it. The background is a green field with trees in the distance. There are no other people or objects in the image.



AMBER #318 · AMBER has no boxes

annotated objects (4): grass woman football court

Unedited ×1 ✓3/4 beam 3
The image is of a young woman playing soccer on a grass field. She is wearing a white and black striped shirt and black shorts. She is holding a soccer ball in her right hand and kicking it with her left foot. The field is green and there are trees in the background.

Nullu ×1 ✓2/4 rank k = 8
This is an image of a young woman playing soccer on a grassy field. She is wearing a white and black striped shirt and black shorts, and has a soccer ball in her hand. She is running with the ball and appears to be about to kick it into the goal. The field…

Placebo ×0 ✓3/4 ★ 32 tokens
The image is of a young woman playing soccer on a grass field. She is wearing a white and black striped shirt and black shorts.



AMBER #221 · AMBER has no boxes

annotated objects (5): person grass football court glove

Unedited ×1 ✓2/5 beam 3
This is an image of a soccer player lying on the ground after making a save on a soccer ball. The player is wearing a green and white jersey and black shorts. The ball is on the ground next to the player. The background is a green grass field with trees in the distance.

Nullu ×0 ✓1/5 rank k = 8
This is an image of a soccer goalkeeper diving to save a soccer ball during a game. The goalkeeper is wearing a green and white jersey with the number 12 on the back and has long, curly blonde hair. She is lying on her stomach on the…

Placebo ×0 ✓2/5 ★ δ = −0.5
This image shows a female soccer player lying on the grass after making a save on a soccer ball. The player is wearing a green and white jersey with the number 12 on the back, and has long blonde hair tied back in a ponytail. She is holding her hands up…

Figure 19: **MiniGPT-4, AMBER (3/3).** Rows: unedited model; Nullu at its published rank (k=8); best placebo for the image. Starred: the placebo on 4 of 6 images, a published method on 2. Legend and selection: App. G.