ReViB: Mitigating Hallucinations in Large Vision-Language Models via Recovered Visual Bias Contrastive Decoding

Marian-Sergiu Nistor
Independent Researcher
sergiunistor.info@gmail.com

Abstract

Hallucination remains a central failure mode of large vision-language models (LVLMs), which generate descriptions of objects, attributes, or relations absent from the image. A leading explanation is that models fall back on language priors and statistical biases acquired during pretraining rather than grounding their output in the visual input. Contrastive decoding methods counter this by contrasting the model's prediction against one obtained from a perturbed input, but the perturbation is generic and only loosely aligned with the bias it aims to remove. We propose **ReViB** (**Re**covered **Vi**sual **B**ias), a training-free contrastive decoding method that recovers this bias directly and subtracts it. Because an LVLM's visual encoding captures coarse, prior-aligned semantics more than fine visual detail, inverting it back into image space through unCLIP reconstructs a scene dominated by the model's statistical bias. Contrasting the original prediction against this reconstruction removes the recovered bias and yields more grounded outputs. **ReViB** consistently reduces object and attribute hallucination on the POPE and MME benchmarks, with the largest gains on the popular and adversarial splits where language priors dominate. We further show that, contrary to recent work claiming contrastive decoding applies only a crude, unidirectional shift to the output distribution, **ReViB** revises predictions in both directions and can lower the yes-rate while improving accuracy.

1 Introduction

Hallucinations in LVLMs can have catastrophic consequences in high-stakes settings such as medical imaging, autonomous driving, and legal systems (Wang, 2024; Magesh et al., 2025). In vision-language models, hallucination manifests as fluent, coherent responses that misrepresent the visual input, and is commonly grouped into three types: non-existent objects, incorrect attributes,

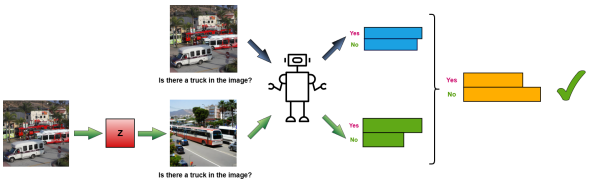


Figure 1: Illustration of the mechanism that drives **ReViB**. Standard inference yields an incorrect answer, driven by statistical biases the model acquired during pretraining. Inverting the visual encoding through unCLIP reconstructs an image that isolates this prior-driven content, exposing the hallucinated response. Contrasting the two removes the recovered bias and recovers the correct answer.

and wrong relations (Liu et al., 2024a; Jing et al., 2024; Gunjal et al., 2024; Zhai et al., 2023). Because such outputs are confident and fluent yet false, they are harder to identify than overt failures, making factuality a matter of AI safety rather than just output quality (Bender et al., 2021).

Research has identified several contributors to these errors. Building on hallucination studies in language modeling (Tonmoy et al., 2024; Wen et al., 2024), work on LVLMs points to statistical biases in multimodal training data (You et al., 2024), an over-reliance on language priors (Zhu et al., 2020; Ren et al., 2023), and multimodal misalignment between the visual and textual modalities (Jiang et al., 2024; Liu et al., 2024a; Wang et al., 2024a). A leading hypothesis, central to our work, is that hallucinations arise less from the prompt and more from the model reverting to these language priors and to the statistical associations and biases acquired during pretraining, rather than remaining grounded in the given visual context (Favero et al., 2024; Leng et al., 2024; Liu et al., 2024b; Wang et al., 2024c; Zhang et al., 2025; Zhu et al., 2025).

Various methods have been proposed to address this. One direction implies finetuning the model, either on curated instruction data such as LRV-

Instruction (Liu et al., 2024a) or through human and model feedback, as in RLHF-V (Yu et al., 2024) and factually augmented RLHF (Sun et al., 2024). Instead, training-free methods correct behavior directly. For example, LURE revises the generated descriptions to remove hallucinated objects after decoding (Zhou et al., 2024). OPERA introduces an over-trust penalty together with a retrospection step during beam search (Huang et al., 2024). PAI reinforces visual grounding by upweighting attention to image tokens (Liu et al., 2024d). MARINE provides object-grounding features drawn from external vision models (Zhao et al., 2024). RITUAL instead aggregates predictions across multiple randomly transformed views of the image, relying on the low chance that a hallucination persists across transformations (Woo et al., 2024).

Contrastive decoding represents an additional inference-time route, which amplifies the disagreement between two next-token distributions to suppress unfaithful predictions (Li et al., 2023a). Most LVLM variants follow the same recipe, contrasting the original distribution with one obtained from a perturbed input, where the perturbation may be visual (Leng et al., 2024), instruction-based (Wang et al., 2024c), internal (Huo et al., 2024), multi-view (Chen et al., 2024).

However, such perturbations are generic: they degrade the visual signal but do not specifically target the language prior that drives hallucination, so the recovered signal only loosely aligns with the bias being subtracted. We propose **ReViB** (**Re**covered **Vi**sual **B**ias), a contrastive decoding method that recovers and subtracts the model's visual bias. The image embedding produced by an LVLM's CLIP-based encoder (Radford et al., 2021) captures coarse, prior-aligned semantics more than precise visual detail (Tong et al., 2024), which makes it a natural carry of the model's statistical bias. We invert this embedding back into image space through unCLIP (Ramesh et al., 2022), where the generative prior reconstructs the scene from the embedding, amplifying prior-consistent content while discarding fine grounding detail. Contrasting the model's prediction on the original image against its prediction on this reconstruction subtracts the recovered visual bias and yields more grounded outputs.

Our contributions are as follows:

- We introduce **ReViB** (**Re**covered **Vi**sual **B**ias), a training-free contrastive decoding method that constructs its contrastive distribution by inverting the visual encoding through unCLIP to recover the model's statistical bias, rather than applying generic input perturbations.
- We show that, contrary to Yin et al. (2025), **ReViB** does not apply a crude, unidirectional shift to the output distribution: it revises predictions in both directions, reducing rather than inflating positive responses while improving accuracy.
- We demonstrate that **ReViB** reduces object and attribute hallucination on the POPE (Li et al., 2023b) and MME (Fu et al., 2025) benchmarks, outperforming prior contrastive decoding approaches, with the largest gains on the popular and adversarial splits where language priors dominate.

2 Related Work

Contrastive decoding originated in text generation, contrasting an expert model against a weaker amateur to suppress the amateur's degenerate tendencies (Li et al., 2023a). In LVLMs the amateur becomes a copy of the same model conditioned on a hallucination-inducing input, and methods differ almost entirely in how that input is constructed. This yields a natural taxonomy of existing methods, organized by how the amateur distribution is produced.

Input perturbation. The most common approach perturbs the model's input to amplify hallucination, then subtracts the result. VCD (Leng et al., 2024) adds noise to the image, weakening visual grounding so the model leans on language priors that are then removed, though the corruption is indiscriminate and sensitive to the chosen noise level. ICD (Wang et al., 2024c) perturbs the textual side instead, prepending disturbance instructions that aggravate multimodal misalignment before contrasting them away.

Region and view resampling. HALC (Chen et al., 2024) grounds each generated token to an image region with a detector, then samples several overlapping fields of view at different scales around that region and contrasts the token distributions they induce, correcting hallucinated tokens locally on the fly. Globally, it adds a matching-based beam search that ranks candidate sequences by a visual matching score to preserve generation quality.

Internal intervention. SID (Huo et al., 2024) avoids external perturbation and intervenes inside the network. It observes that pretrained LVLMs can introspectively judge the importance of vision tokens, and its *Context and Text-aware Token Selection strategy* retains only the least important ones after the early decoder layers, adaptively amplifying vision-and-text association hallucinations that are then subtracted, keeping the contrastive signal fully self-generated. VaLiD (Wang et al., 2024b) partly attributes hallucination to distortions accumulated during visual encoding, and uses entropy-based uncertainty to select and fuse the most hallucination-prone visual encoder layers into a reference distribution that is contrasted against the standard output.

3 Method

3.1 Preliminaries

An LVLM parameterized by θ takes an image v and a textual query x, and generates a response autoregressively. At decoding step t, the next token is drawn from

$$\begin{aligned} y_t &\sim p_\theta(y_t \mid v, x, y_{<t}) \\ &= \mathrm{softmax}\left(\ell_\theta(y_t \mid v, x, y_{<t})\right) \end{aligned} \tag{1}$$

where ℓ_θ denotes the output logits and $y_{<t}$ the previously generated tokens.

Object hallucination is widely attributed to statistical biases absorbed during pretraining: because certain objects co-occur frequently in image-text corpora, the model learns to predict them from context rather than from the image, defaulting to language priors when visual evidence is weak (Leng et al., 2024; Zhou et al., 2024). Contrastive decoding approaches for LVLMs generally counteract this by constructing an amateur distribution $p_\theta(y_t \mid v', x, y_{<t})$ from a hallucination-inducing input v' and subtracting it from the original:

$$\begin{aligned} p_{\mathrm{cd}}(y_t) = \mathrm{softmax}\big[\\ (1+\alpha)\,\ell_\theta(y_t \mid v, x, y_{<t}) \\ -\,\alpha\,\ell_\theta(y_t \mid v', x, y_{<t})\big] \end{aligned} \tag{2}$$

where $\alpha \geq 0$ controls the strength of the contrast. Existing methods differ only in how v' is produced, and in every case it is a generic degradation of the input that is not aligned with the specific bias being removed.

3.2 Recovering Visual Bias via unCLIP Inversion

Our goal is to make v' carry the model's statistical bias directly, rather than approximate it through noise. We exploit a known property of the LVLM's visual front end. The image is encoded by a CLIP vision encoder into an embedding $z = E(v) \in \mathbb{R}^d$ that captures coarse, prior-aligned semantics far more than fine spatial detail, to the point that visually distinct images are mapped to nearly identical embeddings (Tong et al., 2024). The embedding therefore underdetermines the image: many scenes are consistent with the same z.

We turn this ambiguity into a bias estimator by inverting z back into image space through unCLIP (Ramesh et al., 2022), whose diffusion decoder reconstructs an image conditioned on a CLIP image embedding:

$$v_{\mathrm{b}} = D\big(E(v)\big) \tag{3}$$

where E is the CLIP vision encoder (Radford et al., 2021) and D is the unCLIP diffusion decoder (Ramesh et al., 2022). Because z leaves much of the scene unspecified, D fills in the missing content with whatever is most likely under its training distribution, reconstructing the prior-consistent, frequently co-occurring structure while discarding the fine grounding detail that distinguishes the actual image. The reconstruction v_{b} is thus an image-space rendering of the model's visual bias.

3.3 Recovered Visual Bias Contrastive Decoding

ReViB instantiates the contrastive decoding of Eq. (2) with the recovered-bias reconstruction as the amateur input, setting $v' = v_{\mathrm{b}}$. Tokens favored by the bias reconstruction but not by the original image are suppressed, while tokens supported by genuine visual evidence are retained.

To avoid penalizing valid tokens when the model is confident, we apply an adaptive plausibility constraint (Li et al., 2023a; Leng et al., 2024), restricting the output to a truncated set $\mathcal{V}_{\mathrm{head}}$ computed from the original-image distribution before contrasting:

$$\begin{aligned} \mathcal{V}_{\mathrm{head}} = \big\{ y_t \in \mathcal{V} : p_\theta(y_t \mid v, x, y_{<t}) \geq \\ \beta \max_w p_\theta(w \mid v, x, y_{<t}) \big\} \end{aligned} \tag{4}$$

and setting the contrastive score to $-\infty$ for $y_t \notin \mathcal{V}_{\mathrm{head}}$, where $\beta \in [0, 1]$ sets the truncation threshold. The next token is then selected from the contrasted distribution, either greedily or by sampling.

Table 1: POPE accuracy per subset and dataset total on LLaVA-1.5-7B. Every contrastive method lifts accuracy well above regular sampling (+5-6 points overall). **ReViB** (ViT-L, 600) gives the largest overall gain (+0.063 vs. VCD's +0.058); VCD leads only on the easier MSCOCO random split, whereas **ReViB** leads on the harder A-OKVQA and GQA adversarial subsets (e.g. GQA-adv +0.079).

Method	MSCOCO				A-OKVQA				GQA				Overall
	rand	pop	adv	Total	rand	pop	adv	Total	rand	pop	adv	Total	
Regular	0.814	0.793	0.761	0.789	0.798	0.769	0.710	0.759	0.813	0.745	0.704	0.754	0.767
VCD	**+0.076**	+0.066	+0.058	**+0.067**	+0.068	+0.050	+0.037	+0.052	+0.054	+0.051	+0.057	+0.054	+0.058
ReViB (ViT-H, 300)	+0.063	+0.065	+0.053	+0.061	+0.071	+0.052	+0.040	+0.054	+0.052	+0.044	+0.056	+0.051	+0.056
ReViB (ViT-H, 600)	+0.062	+0.068	+0.059	+0.063	+0.070	+0.059	+0.048	+0.059	**+0.058**	+0.055	+0.065	+0.059	+0.061
ReViB (ViT-L, 300)	+0.064	+0.064	**+0.063**	+0.064	+0.067	**+0.062**	+0.047	+0.059	+0.056	+0.054	**+0.079**	+0.063	+0.062
ReViB (ViT-L, 600)	+0.063	**+0.069**	+0.062	+0.065	**+0.073**	+0.057	**+0.052**	**+0.061**	+0.054	**+0.067**	+0.069	**+0.064**	**+0.063**

Table 2: POPE precision per subset and dataset total on LLaVA-1.5-7B. **ReViB** yields clearly larger precision gains than VCD (+0.054 vs. +0.037 overall), and **ReViB** (ViT-L) is best in almost every column. Its added positive answers are more often correct, i.e. the gain is not bought with a yes-bias.

Method	MSCOCO				A-OKVQA				GQA				Overall
	rand	pop	adv	Total	rand	pop	adv	Total	rand	pop	adv	Total	
Regular	0.852	0.810	0.761	0.806	0.794	0.746	0.673	0.734	0.807	0.715	0.669	0.726	0.753
VCD	+0.073	+0.064	+0.041	+0.058	+0.049	+0.030	+0.015	+0.030	+0.030	+0.025	+0.032	+0.029	+0.037
ReViB (ViT-H, 300)	+0.073	+0.078	+0.054	+0.068	+0.066	+0.039	+0.025	+0.041	+0.039	+0.027	+0.038	+0.035	+0.046
ReViB (ViT-H, 600)	+0.074	+0.079	+0.060	+0.070	+0.059	+0.046	+0.029	+0.043	+0.041	+0.035	+0.044	+0.040	+0.049
ReViB (ViT-L, 300)	**+0.076**	+0.074	**+0.065**	+0.071	+0.057	**+0.052**	+0.033	+0.047	+0.041	+0.037	**+0.061**	**+0.047**	+0.053
ReViB (ViT-L, 600)	+0.075	**+0.081**	+0.062	**+0.072**	**+0.068**	+0.048	**+0.036**	**+0.049**	**+0.043**	**+0.049**	+0.049	**+0.047**	**+0.054**

ReViB requires no training and adds a single un-CLIP inversion per image, which can be cached and reused across all prompts that make use of the given image.

4 Experimental Setup

Models and implementation. We evaluate **ReViB** on LLaVA-1.5-7B (Liu et al., 2024c), a widely used LVLM that pairs a CLIP vision encoder with a language-model decoder and is the model on which VCD (Leng et al., 2024) was originally developed, making it a natural testbed for a controlled comparison. All experiments run in `bfloat16` with SDPA attention.

Benchmarks and metrics. We evaluate on POPE (Li et al., 2023b) and MME (Fu et al., 2025). POPE poses balanced yes/no object-existence questions over three datasets (MSCOCO (Lin et al., 2014), A-OKVQA (Schwenk et al., 2022), GQA (Hudson and Manning, 2019)), each with *random*, *popular*, and *adversarial* negative-sampling splits, where the adversarial split targets frequently co-occurring absent objects and most directly probes language priors. We report accuracy, precision, recall, F1, and the predicted-yes rate, and additionally track how many baseline answers each method flips (Yes→No and No→Yes) to characterize the direction of its correction. MME is a broader multimodal benchmark; we use its four hallucination-oriented categories, existence and count at the object level and position and color at the attribute level, and report the standard per-category score $100\,(\mathrm{acc} + \mathrm{acc}^{+})$ and their total.

Baselines. We compare against two baselines. The first is regular sampling, where the model answers directly with no contrastive intervention, establishing each LVLM's unmodified hallucination rate. The second is VCD (Leng et al., 2024), the canonical contrastive decoding method for LVLMs and the closest analogue to **ReViB**: both subtract a hallucination-inducing distribution from the original, and differ only in how that distribution is obtained. VCD constructs it by corrupting the image with forward diffusion noise, so the amateur input is a degraded version of the same image rather than a recovered-bias reconstruction. Following the original paper, we add 999 noise steps for POPE (Li et al., 2023b) and 500 for MME (Fu et al., 2025). Comparing directly against VCD isolates the effect of our reconstruction: because both methods share the same contrastive formulation, adaptive plausibility constraint, and hyperparameters, any difference in performance is attributable to the amateur input alone.

Table 3: POPE recall per subset and dataset total on LLaVA-1.5-7B. VCD attains the highest recall in every column (+0.088 overall), but this reflects its stronger push toward 'yes' rather than better perception; **ReViB**'s more moderate recall gains, paired with its higher precision, are the better-calibrated trade-off.

Method	MSCOCO				A-OKVQA				GQA				Overall
	rand	pop	adv	Total	rand	pop	adv	Total	rand	pop	adv	Total	
Regular	0.761	0.765	0.761	0.762	0.805	0.817	0.816	0.813	0.823	0.815	0.810	0.816	0.797
VCD	**+0.087**	**+0.074**	**+0.086**	**+0.083**	**+0.096**	**+0.082**	**+0.085**	**+0.087**	**+0.088**	**+0.097**	**+0.099**	**+0.095**	**+0.088**
ReViB (ViT-H, 300)	+0.060	+0.056	+0.051	+0.056	+0.076	+0.066	+0.068	+0.070	+0.068	+0.071	+0.077	+0.072	+0.066
ReViB (ViT-H, 600)	+0.057	+0.060	+0.057	+0.058	+0.083	+0.072	+0.083	+0.079	+0.081	+0.085	+0.090	+0.085	+0.074
ReViB (ViT-L, 300)	+0.059	+0.056	+0.060	+0.059	+0.080	+0.068	+0.065	+0.071	+0.076	+0.078	+0.090	+0.081	+0.070
ReViB (ViT-L, 600)	+0.058	+0.060	+0.062	+0.060	+0.079	+0.064	+0.072	+0.071	+0.069	+0.089	+0.090	+0.083	+0.071

Table 4: POPE F1 per subset and dataset total on LLaVA-1.5-7B. On F1, which balances precision and recall, **ReViB** (ViT-L, 600) attains the best overall gain (+0.063), surpassing VCD (+0.061); VCD leads only on MSCOCO while **ReViB** leads on A-OKVQA and GQA.

Method	MSCOCO				A-OKVQA				GQA				Overall
	rand	pop	adv	Total	rand	pop	adv	Total	rand	pop	adv	Total	
Regular	0.804	0.787	0.761	0.784	0.799	0.780	0.738	0.771	0.815	0.761	0.733	0.768	0.774
VCD	**+0.081**	**+0.069**	**+0.063**	**+0.070**	+0.072	+0.053	+0.043	+0.055	+0.058	+0.056	+0.059	+0.058	+0.061
ReViB (ViT-H, 300)	+0.066	+0.066	+0.053	+0.061	+0.071	+0.052	+0.042	+0.054	+0.053	+0.047	+0.054	+0.052	+0.055
ReViB (ViT-H, 600)	+0.065	**+0.069**	+0.058	+0.063	+0.071	+0.058	**+0.050**	**+0.060**	**+0.060**	+0.057	+0.063	+0.060	+0.061
ReViB (ViT-L, 300)	+0.066	+0.065	+0.062	+0.064	+0.069	**+0.059**	+0.046	+0.058	+0.057	+0.055	**+0.073**	**+0.063**	+0.061
ReViB (ViT-L, 600)	+0.065	**+0.069**	+0.062	+0.065	**+0.074**	+0.055	**+0.050**	**+0.060**	+0.056	**+0.067**	+0.066	**+0.063**	**+0.063**

ReViB configuration. We obtain the recovered-bias reconstruction with Stable unCLIP, which inverts a CLIP image embedding back into image space following the unCLIP formulation (Ramesh et al., 2022). To test sensitivity to the inversion backbone, we use two variants that differ in their CLIP image encoder (Radford et al., 2021): a ViT-H model and a ViT-L model. Reconstructions use 25 inference steps, guidance scale 10, an empty text prompt, and a fixed seed. We ablate the embedding noise level over values of 300 and 600 out of a maximum of 1000, which controls how much image-specific detail is discarded in favor of prior-consistent content. For all contrastive methods we set the contrast strength α=1.0 and the adaptive plausibility threshold β=0.5. We use sampling-based decoding (temperature 1.0, top-p 1.0), generating at most 8 new tokens per question.

5 Results

5.1 Hallucination Benchmark Results

Tables 1 and 5 report POPE accuracy and MME scores on LLaVA-1.5-7B, comparing regular sampling against VCD and the four **ReViB** configurations. Every contrastive method improves substantially over regular sampling, confirming that subtracting a hallucination-inducing distribution helps regardless of how that distribution is formed.

POPE Results

On POPE (Table 1), **ReViB** improves accuracy by up to 6.3 points overall, exceeding VCD's 5.8. The advantage is concentrated where language priors are strongest: VCD leads onlyon the MSCOCO random split, while **ReViB** is best on the harder A-OKVQA and GQA subsets, including the adversarial splits that target frequently co-occurring absent objects (e.g. +0.079 on GQA-adversarial). This is consistent with our hypothesis that **ReViB** removes prior-driven content specifically, so its benefit grows precisely where the prior does the most damage. The gap between VCD and **ReViB** on these splits is the effect of the reconstruction alone, since the two methods share the same contrastive formulation and differ only in the amateur input.

Precision improves more under **ReViB** than under VCD (+0.054 vs. +0.037 overall, as shown in Table 2), and the ViT-L variants lead in nearly every column. Because precision rises, the additional positive answers **ReViB** produces are more often correct, so its gains are not obtained by inflating the yes-rate.

Recall tells the complementary story (Table 3). VCD attains the highest recall in every column, but this follows directly from its stronger push toward "yes": answering "yes" more often mechanically

Table 5: MME scores per category on LLaVA-1.5-7B. **ReViB** improves MME far more than VCD (+121.7 vs. +98.3 total, about +23 points), with the noise-600 variants best across existence, count, position, and colour.

Method	Existence	Count	Position	Color	Total
Regular	158.33	100.00	90.00	136.67	485.00
VCD	+26.67	+28.33	+21.67	+21.66	+98.33
ReViB (ViT-H, 300)	**+31.67**	**+41.67**	+30.00	+16.66	+120.00
ReViB (ViT-H, 600)	**+31.67**	+31.67	**+36.67**	+21.66	**+121.67**
ReViB (ViT-L, 300)	+26.67	+28.33	+21.67	+23.33	+100.00
ReViB (ViT-L, 600)	**+31.67**	+31.67	+31.67	**+26.66**	**+121.67**

Table 6: POPE predicted-yes rate (%) per subset and dataset total on LLaVA-1.5-7B; results closer to 50 are better calibrated. Regular sampling is already yes-biased (53.0% overall, up to 60.6% on adversarial splits). VCD pushes the rate further from balance (+3.0 overall, to 56%), whereas **ReViB** shifts it far less (+0.8-1.3 overall), staying closest to the balanced 50%. On MSCOCO, **ReViB** reduces the rate by 0.5. **ReViB** corrects hallucination without amplifying the yes-bias.

Method	MSCOCO				A-OKVQA				GQA				Overall
	rand	pop	adv	Total	rand	pop	adv	Total	rand	pop	adv	Total	
Regular	44.7	47.2	50.0	47.3	50.7	54.8	60.6	55.4	51.0	57.0	60.6	56.2	53.0
VCD	+1.1	+0.8	+2.8	+1.6	+2.7	+3.1	+4.9	+3.5	+3.4	+4.6	+4.2	+4.1	+3.0
ReViB (ViT-H, 300)	-0.3	-1.0	-0.2	-0.5	+0.5	+1.4	+2.8	+1.5	+1.7	+2.7	+2.1	+2.1	+1.0
ReViB (ViT-H, 600)	-0.5	-0.8	-0.2	-0.5	+1.3	+1.3	+3.4	+2.0	+2.3	+3.0	+2.5	+2.6	+1.3
ReViB (ViT-L, 300)	-0.5	-0.8	-0.3	-0.5	+1.3	+0.6	+1.7	+1.2	+2.0	+2.3	+1.1	+1.8	+0.8
ReViB (ViT-L, 600)	-0.5	-0.9	+0.0	-0.5	+0.6	+0.7	+2.0	+1.1	+1.5	+2.2	+2.1	+1.9	+0.8

raises recall on the present-object questions while lowering precision. **ReViB**'s more moderate recall gains, combined with its higher precision, reflect a more balanced correction rather than weaker perception.

F1 summarizes this trade-off (Table 4). **ReViB** (ViT-L, 600) achieves the best overall F1 gain (+0.063), exceeding VCD (+0.061), with the same split-level pattern seen in accuracy: VCD leads on the easier MSCOCO random split, while **ReViB** leads on the harder A-OKVQA and GQA splits where language priors are strongest. **ReViB**'s gains come from a better precision-recall balance rather than from a one-sided change in answering behavior.

MME Results

On MME (Table 5), the difference is much larger: **ReViB** improves the total score by up to 121.7 points against VCD's 98.3, a gain of roughly 23 points. The improvement is spread across all four categories, covering both object-level (existence, count) and attribute-level (position, color) hallucination, indicating that the recovered-bias contrast helps beyond simple object existence.

Across both benchmarks the noise-600 **ReViB** variants are strongest, and the ViT-L encoder is competitive with or better than the larger ViT-H, suggesting the method does not require a high-capacity inversion backbone.

5.2 Answer Calibration

Table 6 reports the predicted-yes rate, which exposes the direction of each method's correction. Regular sampling on LLaVA-1.5-7B is already yes-biased (53.0% overall). VCD raises the yes-rate on every split, by 3.0 points on average. **ReViB** behaves differently: it lowers the yes-rate on MSCOCO (by about 0.5 points across all splits) while raising it modestly on A-OKVQA and GQA. The same method, with identical settings, thus shifts the rate in opposite directions depending on the input.

This directly contradicts Yin et al. (2025), who characterize contrastive decoding as a unidirectional adjustment that simply biases the model toward more "yes" outputs. A shift that is unidirectional by construction would raise the yes-rate on every split; **ReViB** instead moves it up or down depending on the input, so its correction is not a

Table 7: POPE Yes→No shifts per subset and dataset total on LLaVA-1.5-7B. Yes→No shifts are similar across methods (VCD 2022, **ReViB** ∼2280).

Method	MSCOCO				A-OKVQA				GQA				Overall
	rand	pop	adv	Total	rand	pop	adv	Total	rand	pop	adv	Total	
VCD	185	214	211	610	220	242	237	699	205	248	260	713	2022
ReViB (ViT-H, 300)	208	242	247	697	245	254	256	755	245	279	302	826	2278
ReViB (ViT-H, 600)	214	248	250	712	230	252	245	727	226	277	292	795	2234
ReViB (ViT-L, 300)	210	244	247	701	235	264	271	770	221	279	309	809	2280
ReViB (ViT-L, 600)	208	243	248	699	247	267	265	779	230	285	295	810	2288

Table 8: POPE No→Yes flips per subset and dataset total on LLaVA-1.5-7B. No→Yes flips are far more numerous for VCD (2853) than for **ReViB** (∼2515). With the near-equal Yes→No counts in Table 7, this shows VCD applies a strongly asymmetric push toward 'yes', while **ReViB** flips labels in both directions more evenly: a balanced, better-calibrated correction.

Method	MSCOCO				A-OKVQA				GQA				Overall
	rand	pop	adv	Total	rand	pop	adv	Total	rand	pop	adv	Total	
VCD	219	238	294	751	301	337	382	1020	308	387	387	1082	2853
ReViB (ViT-H, 300)	198	212	241	651	259	298	338	895	295	359	365	1019	2565
ReViB (ViT-H, 600)	198	224	245	667	269	293	347	909	296	367	369	1032	2608
ReViB (ViT-L, 300)	195	220	238	653	273	283	322	878	281	349	342	972	2503
ReViB (ViT-L, 600)	192	215	248	655	263	289	325	877	274	351	358	983	2515

uniform push toward "yes." We motivate this further through per-answer decision-shift counts in Section 6.

6 Decision-Shift Analysis

Section 5.2 showed that **ReViB** moves the predicted-yes rate up or down depending on the input, unlike the uniform upward shift that Yin et al. (2025) attribute to contrastive decoding. Here we examine the underlying answer shifts directly, counting how many regular-sampling predictions each method changes from "yes" to "no" (Table 7) and from "no" to "yes" (Table 8). The direction and balance of these shifts reveal whether a method genuinely re-evaluates predictions or simply biases them one way.

The Yes→No shifts are comparable across methods (Table 7), with VCD and every **ReViB** variant converting roughly 2000 to 2300 previously-"yes" answers to "no".

The No→Yes shifts are where the methods diverge sharply (Table 8). VCD converts significantly more "no" answers to "yes" (2853) than **ReViB** (~ 2515), while their Yes→No counts are nearly equal. VCD therefore applies a strongly asymmetric correction, adding many more "yes" answers than it removes, which is exactly the unidirectional behavior Yin et al. (2025) describe. However, **ReViB** directly contradicts this account: its flips are far more balanced across the two directions. Furthermore, on MSCOCO, the No→Yes flips are lower than the Yes→No flips. Thus, its correction cannot be the crude unidirectional shift toward "yes" that Yin et al. (2025) claim contrastive decoding performs. Instead, **ReViB** re-grounds predictions on the image rather than nudging them toward one answer. This is the mechanism behind the near-flat yes-rate in Table 6: **ReViB** moves a large number of answers, but in both directions.

7 Conclusion

We introduced **ReViB**, a training-free contrastive decoding method that recovers an LVLM's visual bias directly rather than approximating it through generic input perturbation. By inverting the model's CLIP image embedding back into image space through unCLIP, **ReViB** reconstructs the prior-consistent content the embedding encodes and contrasts against it, subtracting the recovered bias from the model's predictions. On LLaVA-1.5-7B, **ReViB** reduces object and attribute hallucination on POPE and MME, outperforming VCD

and showing the largest gains on the harder splits where language priors dominate. Because **ReViB** and VCD share the same contrastive formulation and differ only in the amateur input, these gains are attributable to the recovered-bias reconstruction itself. We further found that **ReViB** does not act as a unidirectional shift toward "yes," but adjusts predictions in a direction that depends on the input, contradicting recent claims that contrastive decoding merely inflates positive responses. Together these results suggest that aligning the contrastive signal with the specific bias driving hallucination, rather than degrading the input at random, is a promising direction for grounding LVLM outputs.

Limitations

We validate **ReViB** on a single model, LLaVA-1.5-7B, so it remains to be shown that the recovered-bias contrast transfers across model families and across scales. Different LVLMs pair the vision encoder with the language model in different ways, and both the strength of the language prior and the fidelity of the unCLIP inversion may vary with architecture and size, so results on one 7B model may not predict behavior on larger, smaller, or differently structured LVLMs. A broader study spanning several families (for example the Qwen-VL and InstructBLIP lines) and a range of model sizes is needed to establish how general the effect is. **ReViB** also depends on the availability of a CLIP-based encoder that can be inverted through unCLIP; models whose visual embeddings lie outside the space unCLIP was trained to decode may require a different inversion pathway, which we have not explored. Finally, our experiments cover object and attribute hallucination on POPE and MME under short yes/no answers; whether the same gains hold for open-ended, long-form generation remains open.

References

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

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

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2024. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. 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*, pages 13872–13882.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023a. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 12286–12312.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *International Conference on Learning Representations*, volume 2024, pages 57689–57733.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024b. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

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

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

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2025. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of empirical legal studies*, 22(2):216–242.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

Zhibo Ren, Huizhen Wang, Muhua Zhu, Yichao Wang, Tong Xiao, and Jingbo Zhu. 2023. Overcoming language priors with counterfactual inference for visual question answering. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 600–610.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9568–9578.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Fei Wang, Liang Ding, Jun Rao, Ye Liu, Li Shen, and Changxing Ding. 2024a. Can linguistic knowledge improve multimodal alignment in vision-language pretraining? *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(12):1–22.

Jiaqi Wang, Yifei Gao, and Jitao Sang. 2024b. Valid: Mitigating the hallucination of large vision language models by visual layer fusion contrastive decoding. *arXiv preprint arXiv:2411.15839*.

Jue Wang. 2024. Hallucination reduction and optimization for large language model-based autonomous driving. *Symmetry*, 16(9):1196.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024c. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15840–15853.

Zhihua Wen, Zhiliang Tian, Zexin Jian, Zhen Huang, Pei Ke, Yifu Gao, Minlie Huang, and Dongsheng Li. 2024. Perception of knowledge boundary for large language models through semi-open-ended question answering. *Advances in Neural Information Processing Systems*, 37:88906–88931.

Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. 2024. Ritual: Random image transformations as a universal anti-hallucination lever

in large vision language models. *arXiv preprint arXiv:2405.17821*.

Hao Yin, Guangzong Si, and Zilei Wang. 2025. The mirage of performance gains: Why contrastive decoding fails to mitigate object hallucinations in mllms? In *Advances in Neural Information Processing Systems*, volume 38, pages 33056–33079. Curran Associates, Inc.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2024. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations*, volume 2024, pages 57153–57180.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024. 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*, pages 13807–13816.

Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, Chunyuan Li, and Manling Li. 2023. Halle-control: controlling object hallucination in large multimodal models. *arXiv preprint arXiv:2310.01779*.

YiFan Zhang, Yang Shi, Weichen Yu, Qingsong Wen, Xue Wang, Wenjing Yang, Zhang Zhang, Liang Wang, and Rong Jin. 2025. Debiasing multimodal large language models via penalization of language priors. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4232–4241.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. 2024. Mitigating object hallucination in large vision-language models via classifier-free guidance. *arXiv e-prints*, pages arXiv–2402.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, volume 2024, pages 56969–56998.

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

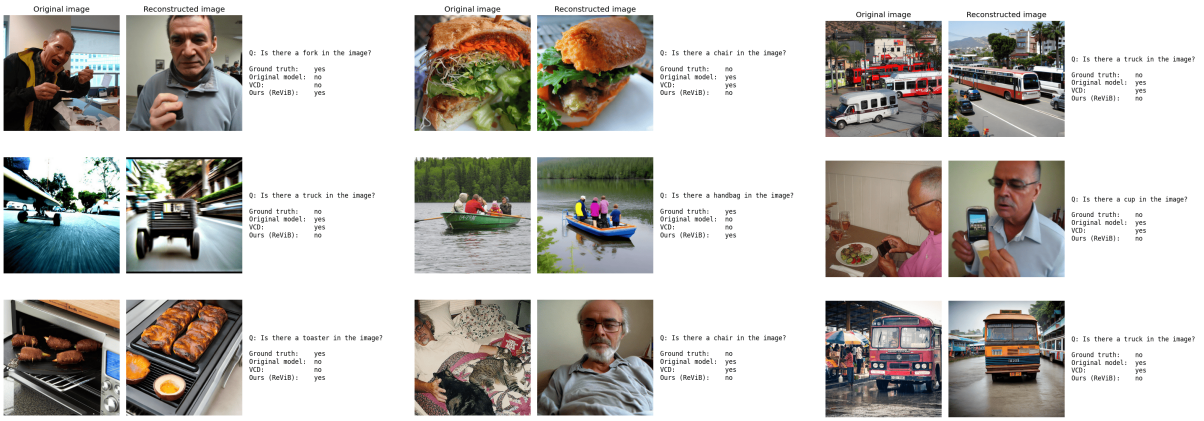
Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. 2020. Overcoming language priors with self-supervised learning for visual question answering. *arXiv preprint arXiv:2012.11528*.

A Qualitative Examples of Corrected Hallucinations

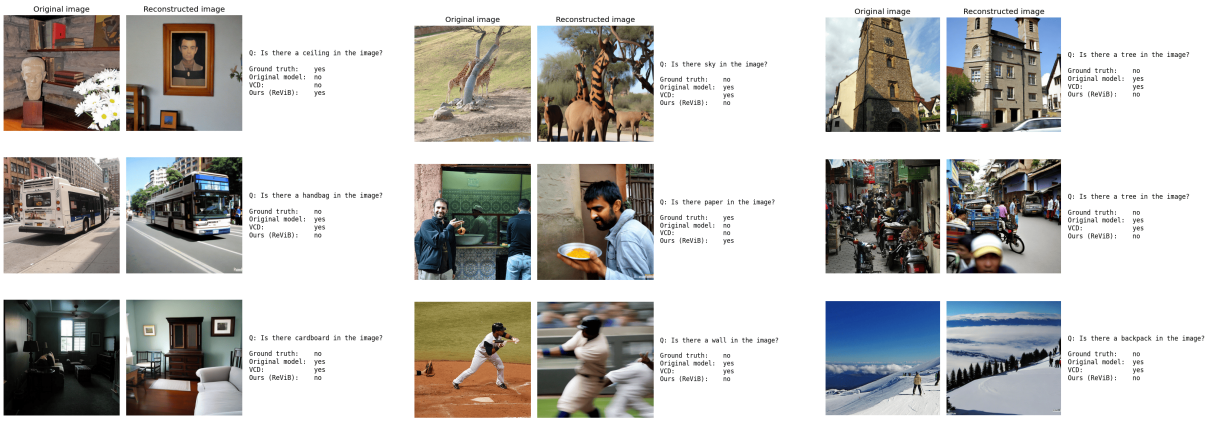
To make the mechanism behind **ReViB** concrete, we show qualitative examples drawn from each POPE split (Figure 2) and from MME (Figure 3). Each example presents the original image, its unCLIP reconstruction v_{b}, the prompt, and the answers produced by regular decoding, by VCD, and by **ReViB**, alongside the ground-truth label. These cases are selected to illustrate the qualitative behavior of the method rather than to estimate its success rate; aggregate performance is reported in the main text.

The reconstructions make the source of the correction visible. Passing the image through the CLIP encoder and back out through unCLIP produces a scene that keeps the coarse, prior-consistent layout while discarding the fine grounding detail that distinguishes the actual image. In the examples shown, this reconstruction visibly amplifies the model's priors: objects that frequently co-occur with the depicted scene are rendered more prominently or hallucinated outright, even when they are absent from the original, and conversely, objects that are genuinely present but atypical for the scene are often dropped or replaced by more expected content. Either way, the reconstruction pulls the image toward what the model expects rather than what it shows.

Therefore, in the reconstruction, the model commits to the hallucinated answer more strongly than it does on the real image, producing a sharper wrong-answer signal. It is precisely this amplified signal that contrastive decoding exploits: because the bias is exaggerated in the amateur branch, subtracting it removes the prior-driven token while leaving the genuinely grounded prediction intact, and the model arrives at the correct answer.



(a) MSCOCO - random (b) MSCOCO - popular (c) MSCOCO - adversarial



(d) A-OKVQA - random (e) A-OKVQA - popular (f) A-OKVQA - adversarial

(g) GQA - random (h) GQA - popular (i) GQA - adversarial

Figure 2: Qualitative examples from the POPE splits on LLaVA-1.5-7B, three per split. Each panel shows the original image, its unCLIP reconstruction, the prompt, and the answers from regular decoding, VCD, and **ReViB**, with the ground-truth label. In every case regular decoding and VCD answer incorrectly while **ReViB** answers correctly. The reconstruction amplifies the model's pretraining priors, pulling the scene toward expected co-occurring content and away from atypical detail, which sharpens the wrong-answer signal that the contrast then removes.

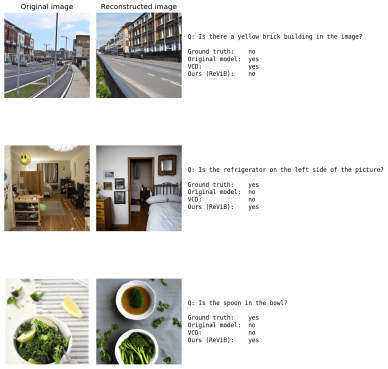


Figure 3: Qualitative examples from MME on LLaVA-1.5-7B. Each panel pairs the original image with its unCLIP reconstruction, the prompt, and the regular, VCD, and **ReViB** answers, together with the ground truth. In each case regular decoding and VCD answer incorrectly while **ReViB** answers correctly. Because the reconstruction amplifies the model's pretraining priors, it elicits a stronger biased response that the contrast then subtracts, restoring the grounded answer.