# Mitigating Object Hallucination in Large Vision-Language Models via False Discovery Controlled Visual Data Splitting

**Chang Liu**
School of Data, Mathematical, and Statistical Sciences
University of Central Florida
chang.liu@ucf.edu

**Yu Tian**
Institute of Artificial Intelligence
University of Central Florida
yu.tian2@ucf.edu

**Rui Xie***
School of Data, Mathematical, and Statistical Sciences
University of Central Florida
rui.xie@ucf.edu

## Abstract

Multiple object hallucination, where large vision-language models (LVLMs) generate objects not supported by the visual input, is a persistent challenge caused by visual uncertainty during decoding. Existing methods reduce hallucinations using contrastive signals, but they rely on heuristics and lack principled control of false positives at the image level. To address this, we propose False Discovery Rate-**CO**nt**R**ol of H**AL**lucination (CORAL), a training-free framework that models visual uncertainty using an uncertainty-aware visual data splitting strategy and leverages mirror statistic to quantify visual contrast during decoding. By computing mirror statistic from paired, symmetrically perturbed visual inputs, CORAL estimates spurious object predictions and sets a data-driven threshold to control the expected fraction of false discoveries per image, suppressing hallucinations while retaining high power for truly grounded objects. The framework is flexible, supports multiple LVLMs, and mitigates hallucinations without retraining or supervision. Extensive experiments on multiple benchmarks with several evaluation metrics demonstrate that CORAL consistently outperforms state-of-the-art methods, providing more reliable and robust hallucination control. Code is available at: https://changliu1993-cl.github.io/CORAL/

## 1 Introduction

Large Vision-Language Models (LVLMs) extend large language models with pretrained vision encoders to jointly reason over visual and textual inputs, enabling the generation of linguistically fluent and semantically aligned outputs grounded in visual content [1, 2]. This unified multimodal framework has led to strong performance across core vision–language tasks such as image captioning [3], visual question answering [4], and multimodal dialogue [5], further driven by advances in model architectures [6–8], multimodal alignment techniques [9, 10], and large-scale benchmarks [11, 12]. Despite this progress, LVLMs remain vulnerable to hallucinations, where models produce confident yet visually ungrounded descriptions of objects, attributes, or relationships absent from the input im-

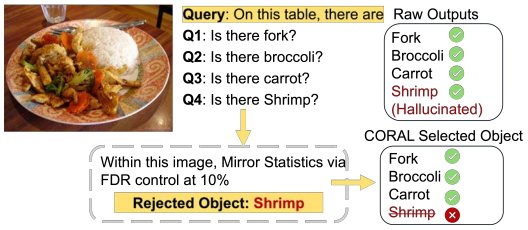



Figure 1: **Hallucination under multiple queries.** The model may generate plausible but hallucinated objects (e.g., *"shrimp"*, in red). CORAL performs image-level selection, rejecting hallucinations while retaining visually grounded objects under False Discovery Rate (FDR) control.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

age [1, 6–8, 13]. Such hallucinations are pervasive across tasks, including image captioning [14] and visual question answering [15], and persist under both closed-set and open-ended evaluation protocols [16–18]. As shown in Fig. 1, LVLMs may produce plausible yet hallucinated objects when answering multiple queries about the same image. Evaluating each query independently is insufficient, as hallucinated objects (e.g., *shrimp*) can be retained alongside real ones.

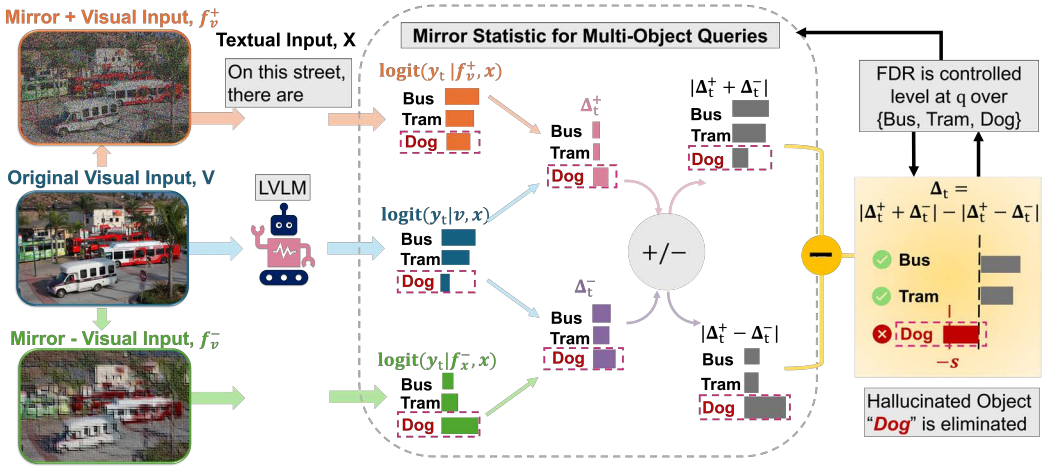



Figure 2: Illustration of CORAL, which applies visual uncertainty splitting and mirror statistic to control the FDR of hallucinated objects. The hallucinated object "Dog" (red) is removed under FDR control.

Object hallucination in LVLMs can be viewed as a failure to properly account for visual uncertainty during generation [18]. When visual inputs are reliable and unambiguous, token predictions are mainly driven by visual features extracted from the image [19]. In contrast, when visual evidence is weak or ambiguous, LVLMs tend to rely more on language priors, producing responses that are statistically plausible but not necessarily supported by the visual input [20]. As a result, objects with higher visual uncertainty are more likely to be hallucinated during the decoding process.

Existing object hallucination mitigation methods largely rely on closed-set object existence queries, evaluating each object independently [14, 15, 21]. Strategies such as image guidance [22] or global-local attention assembly [23] operate at the level of individual queries and require manually controlled signals. Contrastive approaches, such as Visual Contrastive Decoding (VCD) [24], probe visual uncertainty by comparing outputs from original and perturbed visual inputs, but still do not provide holistic, image-level control. From a statistical perspective, contrastive visual analysis can be understood through the lens of *data splitting*: the visual data are split into paired, symmetrically perturbed views from a shared noise and then used in separate stages of the analysis to quantify uncertainty and control errors [25]. As illustrated in Fig. 3, we introduce an **uncertainty-aware visual data splitting** strategy that constructs two paired visual inputs from each image via symmetric perturbations, by inducing controlled variability in the LVLM perception process. The splitting is instantiated in the visual domain by generating symmetric views from a shared noise source, allowing consistent visual signals to be distinguished from noise-induced variability.

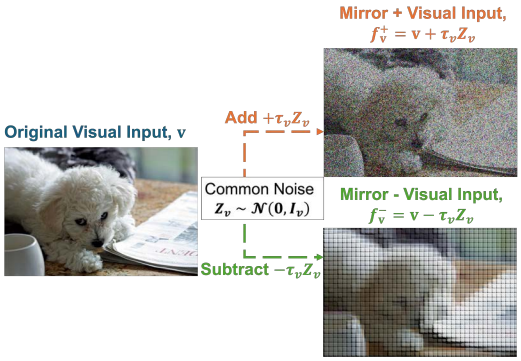



Figure 3: **Uncertainty-aware visual data splitting**. Paired, symmetrically perturbed visual views are generated from a shared noise source, producing contrastive views for the mirror statistic construction and thereby facilitating the separation of visual signal from noise.

Viewed in this way, visual hallucination detection can be cast as an uncertainty-aware objective selection problem, whose goal is to distinguish visually grounded outputs (e.g., *fork*, *broccoli*, and *carrot* in Fig. 1) from those that are not grounded in the visual input (e.g., the hallucinated *shrimp*).

Leveraging the symmetry of data splitting, we construct ***mirror statistic*** [26], a data-driven score for estimating false discoveries without access to ground truth. Built from two splitted visual data, mirror statistic reward objectives that exhibit consistent visual evidence and confidence across splits, while inconsistent or noisy objectives tend to cancel out due to symmetric positive and negative contributions (e.g., the hallucinated *Dog* in Fig.2). This property enables direct estimation of spurious visual signals and provides a principled mechanism to mitigate object hallucination by estimating and controlling the false discovery rate (FDR), defined here at the level of an entire image as the expected proportion of selected visual objectives that are not visually grounded, rather than at the level of individual object queries,

$$\text{FDR} = \mathbb{E}\left[\frac{\#\{\text{hallucinated objects identified as grounded}\}}{\#\{\text{all identified objects}\}}\right]. \tag{1}$$

e.g., four objects are predicted for the same Fig. 1, one of which (*shrimp*) is hallucinated, yielding an FDR of 25%.

As a consequence of leveraging mirror statistic for FDR control, we bounds the expected fraction of falsely selected visual objects while maintaining high power, the probability of retaining truly visually grounded objects:

$$\text{Power} = \mathbb{E}\left[\frac{\#\{\text{truly grounded objects correctly identified}\}}{\#\{\text{all truly grounded objects}\}}\right]. \tag{2}$$

In Fig. 1, our proposed method successfully retains all visually grounded objects (*fork*, *broccoli*, and *carrot*), achieving 100% power for this image. The same figure also illustrates a common failure mode of LVLMs, where a hallucinated object ("*shrimp*") is predicted alongside genuine objects within a single image, allowing visually unsupported and grounded predictions to co-exist.

In this setting, false discovery rate (FDR) control limits the proportion of hallucinated objects retained in the output, while high power naturally follows by preserving objects that are genuinely supported by visual evidence rather than suppressing them indiscriminately. This balance enables effective image-level control of object hallucination. Motivated by this observation, we propose `CORAL`, a false discovery rate **CO**nt**R**ol of object h**AL**lucination framework that controls the proportion of hallucinated objects at the image level while maintaining high power (Fig. 2). Our main contributions are summarized as follows:

- `CORAL` employs an **uncertainty-aware visual data splitting** strategy to construct two paired visual views via symmetric perturbations of each image, inducing controlled variability and enabling stochastic contrasts for object hallucination control.
- `CORAL` leverages **mirror statistic** constructed from split visual inputs to estimate and control the false discovery rate at the image level, enabling principled suppression of hallucinated objects while retaining high power.
- `CORAL` is a **training-free** approach that mitigates object hallucination through data-splitting–based FDR control, incurring low computational overhead compared to training-based alternatives.

## 2 Related Work and Preliminaries

Controlling the false discovery rate is a fundamental problem in multiple hypothesis testing [27]. The knockoff framework [28] introduces synthetic variables to construct feature-level test statistics with provable false discovery rate (FDR) control for variable selection, and has been extended to high-dimensional settings through Model-X knockoffs [29]. Similarly, Gaussian mirrors are developed for feature selection by constructing paired statistics that exhibit sign symmetry under the null hypothesis, such that positive and negative values are approximately balanced for null features [26]. This symmetry enables estimation of the number of false discoveries from the corresponding negative statistics, thereby facilitating FDR control [25, 26, 30]. Both knockoff-based and mirror-statistic approaches are specifically designed for variable or feature selection tasks, relying on paired variables with exchangeability properties to control the selection FDR while maintaining statistical power, particularly in linear models. Inspired by this line of work, we employ an uncertainty-aware visual data splitting strategy that constructs paired visual inputs, allowing mirror statistic to be applied for hallucination control in LVLMs.

Existing approaches to object hallucination in LVLMs focus on either model adaptation or post hoc correction. Fine-tuning-based methods improve grounding using curated datasets [31–33, 18, 34], while others rely on external language models to refine outputs [35–37]. Recent work further suggests that hallucination primarily arises from the language modeling head, which fails to fully leverage visually grounded representations despite their presence in intermediate features [37]. However, these approaches are often resource-intensive and difficult to scale, requiring large amounts of human annotation or reliance on proprietary models [38, 39]. This has motivated lightweight alternatives, particularly uncertainty-based methods that detect hallucinations from model confidence or variability [13, 40]. Several recent approaches mitigate object hallucination by enhancing visual grounding through contrastive or multi-view visual signals during generation [41]. Visual Contrastive Decoding (VCD) [24] leverages contrastive visual inputs to discourage text generation unsupported by the image. Along a similar line, the Assembly of Global and Local Attention (AGLA) [23] constructs complementary visual representations from multiple views and evaluates token-level consistency across visual perspectives using contrastive signals, thereby reducing hallucinated outputs. Mitigating Hallucination via Image-Grounded Guidance (MARINE) [22] further extends this paradigm by incorporating multi-image reasoning and cross-modal alignment mechanisms to refine visual–textual consistency and suppress hallucination. Recent work also shows that hallucination can be exacerbated under distribution shifts such as stylized images, where visual cues deviate from natural image statistics and become less reliable [42]. As this setting focuses on stylized distribution shifts, it is not directly comparable to standard natural-image benchmarks used in our evaluation.

While effective, existing approaches largely rely on logit comparisons during decoding to suppress visually unsupported outputs, without explicit control over hallucination. This limitation is amplified in multi-object scenes with heterogeneous visual evidence. In contrast, our `CORAL` formulates hallucination mitigation as an FDR-controlled selection problem, providing explicit image-level control over hallucinated objects while preserving power for grounded ones. We construct paired visual views via uncertainty splitting and apply mirror statistic to quantify consistent visual evidence, resulting in a principled, training-free framework for hallucination control.

**Preliminaries: Decoding of LVLMs** We consider a Large Vision-Language Model (LVLM) parameterized by $\theta$. The model takes a text prompt $\mathbf{x} = [x_1, \ldots, x_n]$, where each $x_i$ is a prompt token, and visual inputs $\mathbf{v}$, which provide contextual visual information. The model generates a response sequence $\mathbf{y} = [y_1, \ldots, y_m]$ autoregressively according to the conditional distribution $p_\theta(\mathbf{y} \mid \mathbf{v}, \mathbf{x})$, which factorizes as $p_\theta(\mathbf{y} \mid \mathbf{v}, \mathbf{x}) = \prod_{t=1}^{m} p_\theta(y_t \mid \mathbf{v}, \mathbf{x}, \mathbf{y}_{<t})$, where $\mathbf{y}_{<t} = [y_1, \ldots, y_{t-1}]$ for $t > 1$ and is empty for $t = 1$. At each step $t$, the next token is sampled from a categorical distribution defined by the model logits, $p_\theta(y_t \mid \mathbf{v}, \mathbf{x}, \mathbf{y}_{<t}) = \mathrm{softmax}(\mathrm{logit}_\theta(\cdot \mid \mathbf{v}, \mathbf{x}, \mathbf{y}_{<t}))$ . We can further view in the logit space, where the $t$-th token is sampled from the logit space by $\mathbf{y} \propto \exp \mathrm{logit}_\theta(\mathbf{y}|\mathbf{v}, \mathbf{x}, \mathbf{y}_{<t})$.

In the decoding phase of LVLMs, object hallucination usually appears when probabilities are erroneously allocated to tokens that do not align with the presented visual input $\mathbf{v}$. The main causes of this problem include statistical biases inherent in training data [43–45], and over-reliance on language priors embedded within large language models (LLMs) used as decoders [34, 16, 46, 47].

## 3 `CORAL` Methodology

To characterize and control object hallucination *globally* at the image level rather than on individual object queries, we introduce **uncertainty-aware visual data splitting**. An LVLM typically produces multiple semantic outputs from a single image, each supported by varying degrees of visual evidence, motivating the need for global regulation of hallucinated content within the same decoding context. Our approach constructs paired, symmetrically perturbed views of the same visual input, inducing controlled visual uncertainty while preserving semantic content. This design disentangles visual dependence from language priors and enables assessment of whether token generation is genuinely grounded in visual evidence. Building on this formulation, we cast hallucination mitigation as a false discovery rate (FDR) control problem using **mirror statistic**, a distribution-free procedure based on paired mirror comparisons that enables reliable image-level inference with principled error control over object hallucinations.

### 3.1 Uncertainty-Aware Visual Data Splitting

Visual uncertainty provides informative signals about the reliability of visual grounding in large vision language models rather than being merely a source of error [48]. To leverage this information, we adopt a data splitting strategy [25] that introduces symmetric perturbations to the visual input.

For each image, we apply an *uncertainty-aware visual data splitting* procedure to construct a paired set of mirror views that capture visual uncertainty during decoding. Let $Z_v \sim \mathcal{N}(0, \mathbf{I}_v)$ be a Gaussian noise vector matching the dimensionality of the visual input $\mathbf{v}$. We generate two symmetrically perturbed visual features via sign-symmetric randomization:

$$f_{\mathbf{v}}^{+} = \mathbf{v} + \tau_v Z_v, \qquad f_{\mathbf{v}}^{-} = \mathbf{v} - \tau_v Z_v, \tag{3}$$

where $\tau_v > 0$ controls the perturbation magnitude and can be calibrated on a held-out validation set for a target FDR. The calibrated $\tau_v$ is fixed for each visual backbone across all experiments, with sensitivity analyses provided in Appendix Fig. 10 and Fig. 11. The resulting perturbed views preserve the same underlying visual content while introducing symmetric stochastic deviations along a shared random direction. The uncertainty-aware visual data splitting is designed to create a paired, sign-symmetric contrast that (i) amplifies visually grounded signals while canceling injected noise, and (ii) produces symmetric evidence patterns in the absence of visual grounding, a property that underlies the mirror statistic introduced in the next subsection and enables principled control of hallucinated objects.

### 3.2 Mirror Statistic

Given the original visual input $\mathbf{v}$ and its two mirror views $f_{\mathbf{v}}^{+}$ and $f_{\mathbf{v}}^{-}$, we evaluate the LVLM under the same textual prompt $\mathbf{x}$. During autoregressive decoding, for the generated output token $y_t$ at step $t$, we compare the model's logits under the original and mirror view inputs. Specifically, we define the logit differences,

$$\Delta_t^{+} = \text{logit}_\theta(y_t \mid \mathbf{v}, \mathbf{x}, \mathbf{y}_{<t}) - \text{logit}_\theta\big(y_t \mid f_{\mathbf{v}}^{+}, \mathbf{x}, \mathbf{y}_{<t}\big)\,, \tag{4}$$

$$\Delta_t^{-} = \text{logit}_\theta(y_t \mid \mathbf{v}, \mathbf{x}, \mathbf{y}_{<t}) - \text{logit}_\theta\big(y_t \mid f_{\mathbf{v}}^{-}, \mathbf{x}, \mathbf{y}_{<t}\big)\,. \tag{5}$$

These quantities capture how the model's predictive behavior responds to sign-symmetric perturbations of the visual input along a shared noise realization. Since $f_{\mathbf{v}}^{+}$ and $f_{\mathbf{v}}^{-}$ preserve the same underlying visual content and differ only in perturbation sign, comparing $\Delta_t^{+}$ and $\Delta_t^{-}$ logit contrasts that reveal whether the model exhibits approximately symmetric responses to opposing visual perturbations.

Following the Gaussian mirrors construction [26], we define the mirror statistic for visual uncertainty

$$\Delta_t = \big|\Delta_t^{+} + \Delta_t^{-}\big| - \big|\Delta_t^{+} - \Delta_t^{-}\big|. \tag{6}$$

The mirror statistic $\Delta_t$ has two components. The first part, $\big|\Delta_t^{+} + \Delta_t^{-}\big|$, captures signal strength through consistent responses across symmetric perturbations, while the second part, $\big|\Delta_t^{+} - \Delta_t^{-}\big|$, reflects the noise cancellation effect. Their difference therefore provides a contrast between reliable visual evidence and perturbation-induced randomness.

Intuitively, when the generated token is not visually grounded, its logit response to the two sign-symmetric visual perturbations should not contain a stable visual-evidence component. In this null case, the paired differences $\Delta_t^{+}$ and $\Delta_t^{-}$ fluctuate primarily due to perturbation-induced noise, and the resulting mirror statistic is expected to have an approximately symmetric distribution around zero, rather than being always negative. In contrast, when the generated token is visually grounded, the two mirror-view differences are expected to share a stable visual-response component. This component is reinforced in $\big|\Delta_t^{+} + \Delta_t^{-}\big|$ and reduced in $\big|\Delta_t^{+} - \Delta_t^{-}\big|$, so visually grounded tokens tend to produce larger, more positive values of $\Delta_t$. As a result, $\Delta_t$ separates visually grounded outputs from noise-driven ones, enabling false discovery estimation from the negative tail of mirror statistic and principled FDR control with high power for true visual signals.

### 3.3 Controlling Hallucinations via FDR

We utilize the false discovery rate to control hallucinated tokens among the outputs generated by LVLMs. By leveraging the symmetric behavior of mirror statistic under sign-perturbed visual inputs,

we derive a data-driven threshold that uses negative mirror evidence to estimate spurious visual responses, enabling FDR-controlled selection while maintaining high power to retain truly visually grounded tokens.

Let $\mathcal{S}$ denote the set of decoding tokens generated for a given visual input, where each token $y_t \in \mathcal{S}$ is associated with mirror statistic $\Delta_t$, and let $\mathcal{S}_0 \subseteq \mathcal{S}$ denote the null subset whose generated token is not reliably grounded in the visual input (i.e., hallucinated tokens). Each token thus corresponds to a hypothesis testing whether it is visually grounded. A false discovery occurs when a token $y_t \in \mathcal{S}_0$ (*i.e.*, a hallucinated token not reliably grounded in the visual input) is incorrectly identified as visually grounded by rejecting its null hypothesis.

we state the required condition directly as approximate conditional sign-flip invariance under the null. Define

$$H_{0t} : \text{no systematic visual evidence supports token } y_t,$$

and let

$$\mathcal{G}_t = \sigma(\mathbf{v}, \mathbf{x}, \mathbf{y}_{<t})$$

denote the shared context consisting of the image, textual input, and decoding history, conditional on which $Z_v$ remains random.

**Theorem 3.1** (Conditional null symmetry)**.** *For $\Delta_t$ defined in Eq.* (6)*, if*

$$(\Delta_t^+, \Delta_t^-) \mid H_{0t}, \mathcal{G}_t \overset{d}{=} (\Delta_t^+, -\Delta_t^-) \mid H_{0t}, \mathcal{G}_t,$$

*then, for every $s > 0$,*

$$\Pr(\Delta_t \leq -s \mid H_{0t}, \mathcal{G}_t) = \Pr(\Delta_t \geq s \mid H_{0t}, \mathcal{G}_t). \tag{7}$$

The proof follows from mirror antisymmetry (Lemma B.1). For LVLMs, sign-flip invariance is an approximate working assumption, not guaranteed by Gaussian perturbations; uncorrelated contrasts are not required (Appendix B.4). With suitable tail-count concentration (Appendix B.6), approximate null symmetry motivates, for $s > 0$, $\#\{y_t \in \mathcal{S}_0 : \Delta_t \geq s\} \approx \#\{y_t \in \mathcal{S}_0 : \Delta_t \leq -s\} \leq \#\{y_t \in \mathcal{S} : \Delta_t \leq -s\}$.

The $\mathrm{FDR}(s)$ at threshold $s$ can be estimated by

$$\widehat{\mathrm{FDR}}(s) = \mathbb{E}\left[\frac{\#\{y_t \mid \Delta_t \leq -s\}}{\#\{y_t \mid \Delta_t \geq s\} \vee 1}\right]. \tag{8}$$

Intuitively, the negative tail provides a data-driven estimate of how many hallucinated tokens are expected among the selected ones.

For any designated target FDR level $q \in (0, 1)$, we choose a data-driven cutoff $T_q$ such that,

$$T_q = \min_s \left\{\widehat{\mathrm{FDR}}(s) \leq q\right\}, \tag{9}$$

and retain the set $\{y_t \in \mathcal{S} : \Delta_t \geq T_q\}$, suppressing the remaining tokens to mitigate hallucination.

Because mirror statistic are constructed independently for each decoding step, this design offers two advantages: (i) small, controllable perturbations reduce spurious correlations and improve power for identifying truly grounded tokens; and (ii) the computation is fully parallelizable across tokens, enabling scalability to outputs with many candidate tokens.

# 4 Experiments

## 4.1 Experimental Setting

**Models.** To demonstrate the broad applicability of our method across different LVLMs architectures, we apply and evaluate `CORAL` to widely used models, including LLaVA-OneVision-7B [49], Qwen2.5-VL-7B [50], and InternVL3-8B [51] (Table 1), as well as LLaVA-v1.5 [1], InstructBLIP-7B [52], and Qwen-VL-7B [53] (Appendix B).

**Dataset and Baselines.** In alignment with established evaluations from previous studies [24], we assess our method using the three dataset, MSCOCO [14], A-OKVQA [15], and GQA [21]. Each

dataset includes three negative sample settings, i.e. random, popular, and adversarial. We compare our CORAL with three state-of-the-art decoding methods and vanilla LVLMs without decoding techniques (regular), including VCD [24], which distorts the image inputs to impose penalties on logit outputs; MARINE [22], which introduces image-grounded guidance, and AGLA [23], which assembles global features for response generation and local features for visual discrimination simultaneously. Results for MSCOCO are presented here, whereas results for A-OKVQA and GQA appear in Appendix B.

**Metrics.** We evaluate all methods using FDR (Eq. (1)), power (Eq. (2)), and commonly used metrics including POPE and MME. We also report CHAIR metrics in the Appendix B. Details are provided in Appendix B.5.

Although CORAL is primarily designed for object hallucination, its token-level formulation, which operates directly on generated tokens without assuming a specific hallucination type, naturally generalizes to other types of hallucination, including attribute and relational hallucinations, as further validated empirically by MME results.

*Polling-based Object Probing Evaluation* (POPE) [16] evaluates closed-set choices. Multiple prompts with binary choices are formulated, each asking whether a specific object appears in the given image, such as "Is there a chair in this image?" to answer "yes" or "no".

Table 1: Evaluation of overall FDR control and overall power across multiple LVLM architectures on MSCOCO (3000 runs). Overall FDR and power denote averages of per-image FDR and power. **Bold** indicates the best result and underline indicates the second-best result.

| Method | LLaVA-OneVision-7B | | Qwen2.5-VL-7B | | InternVL3-8B | | Average | |
|---|---|---|---|---|---|---|---|---|
| | Overall FDR↓ | Overall Power↑ | Overall FDR↓ | Overall Power↑ | Overall FDR↓ | Overall Power↑ | Overall FDR↓ | Overall Power↑ |
| *Random* | | | | | | | | |
| Regular | 0.0935 (± 0.0133) | 75.36 (± 1.88) | 0.0944 (± 0.0114) | 77.43 (± 1.22) | 0.0946 (± 0.0122) | 73.52 (± 1.77) | 0.0942 (± 0.0111) | 75.44 (± 1.44) |
| VCD (2024) | 0.0892 (± 0.0188) | 78.50 (± 1.23) | 0.0908 (± 0.0132) | 79.14 (± 1.53) | 0.0899 (± 0.0109) | 78.86 (± 1.34) | 0.0900 (± 0.0145) | 78.83 (± 1.31) |
| MARINE (2025) | 0.0846 (± 0.0122) | 81.25 (± 1.17) | 0.0866 (± 0.0116) | 80.05 (± 1.60) | 0.0889 (± 0.0133) | 79.84 (± 1.33) | 0.0867 (± 0.0124) | 80.38 (± 1.21) |
| AGLA (2025) | 0.0799 (± 0.0223) | 84.92 (± 1.11) | 0.0821 (± 0.0114) | 84.67 (± 1.41) | 0.0810 (± 0.0144) | 80.92 (± 1.33) | 0.0810 (± 0.0123) | 83.50 (± 1.50) |
| CORAL(Ours) | **0.0691 (± 0.0113)** | **91.43 (± 1.42)** | **0.0718 (± 0.0332)** | **94.60 (± 1.49)** | **0.0799 (± 0.0122)** | **86.88 (± 1.12)** | **0.0736 (± 0.0149)** | **90.97 (± 1.44)** |
| *Popular* | | | | | | | | |
| Regular | 0.0971 (± 0.0205) | 80.11 (± 1.28) | 0.0903 (± 0.0111) | 77.01 (± 1.64) | 0.0937 (± 0.0106) | 50.80 (± 1.69) | 0.0937 (± 0.0141) | 69.31 (± 1.54) |
| VCD (2024) | 0.0893 (± 0.0187) | 83.11 (± 1.32) | 0.0911 (± 0.0133) | 75.10 (± 1.26) | 0.0875 (± 0.0211) | 67.22 (± 1.11) | 0.0893 (± 0.0177) | 75.14 (± 1.23) |
| MARINE (2025) | 0.0773 (± 0.0134) | 85.22 (± 1.55) | 0.0883 (± 0.0220) | 78.28 (± 1.25) | 0.0701 (± 0.0221) | 87.20 (± 1.11) | 0.0786 (± 0.0192) | 83.57 (± 1.30) |
| AGLA (2025) | 0.0733 (± 0.0177) | 88.13 (± 1.69) | 0.0861 (± 0.0233) | 82.22 (± 1.51) | 0.0741 (± 0.0188) | 81.14 (± 1.55) | 0.0778 (± 0.0199) | 83.83 (± 1.58) |
| CORAL(Ours) | **0.0552 (± 0.0122)** | **90.42 (± 0.24)** | **0.0802 (± 0.0200)** | **84.10 (± 1.15)** | **0.0675 (± 0.0155)** | **88.93 (± 1.21)** | **0.0676 (± 0.0159)** | **87.82 (± 0.87)** |
| *Adversarial* | | | | | | | | |
| Regular | 0.0988 (± 0.0114) | 51.22 (± 1.55) | 0.0973 (± 0.0108) | 52.19 (± 1.13) | 0.0953 (± 0.0133) | 54.88 (± 1.97) | 0.0971 (± 0.0118) | 52.76 (± 1.55) |
| VCD (2024) | 0.0903 (± 0.0115) | 56.32 (± 1.56) | 0.0935 (± 0.0109) | 53.16 (± 1.77) | 0.0901 (± 0.0131) | 60.59 (± 1.26) | 0.0913 (± 0.0118) | 56.69 (± 1.53) |
| MARINE (2025) | 0.0864 (± 0.0105) | 61.36 (± 1.99) | 0.0903 (± 0.0117) | 60.22 (± 1.53) | 0.0896 (± 0.0156) | 70.46 (± 1.59) | 0.0888 (± 0.0126) | 64.01 (± 1.70) |
| AGLA (2025) | 0.0775 (± 0.0198) | 75.77 (± 1.28) | 0.0899 (± 0.0155) | 73.99 (± 1.11) | 0.0826 (± 0.0211) | 81.36 (± 1.78) | 0.0833 (± 0.0188) | 77.04 (± 1.39) |
| CORAL(Ours) | **0.0694 (± 0.0155)** | **88.33 (± 1.45)** | **0.0839 (± 0.0111)** | **83.63 (± 1.55)** | **0.0733 (± 0.0122)** | **87.59 (± 1.33)** | **0.0755 (± 0.0129)** | **86.52 (± 1.44)** |

*MME* evaluates the perception and cognition abilities of LVLMs [54], including object existence, attributes, spatial position, and relations.

*MMBench* evaluates the fine-grained multimodal understanding capabilities of large vision-language models through multiple-choice questions spanning 20 ability dimensions [55].

**Implementation Details.** Throughout all experiments, we followed the recommended settings from the respective papers and used the released codes to ensure fair comparisons. All evaluations were repeated over 3000 randomized trials (details in Appendix B.7). For FDR control, we set the target level to $q = 0.1$, meaning that the expected proportion of hallucinated objects among all selected objects is controlled to be no greater than $10\%$.

### 4.2 Experimental Results

**Results on FDR & Power.** The FDR and power results in Table 1 demonstrate that CORAL achieves effective and consistent control of object hallucination under the target FDR level of $10\%$. Across most settings, CORAL attains the lowest or near-lowest false discovery rate, indicating strong control over hallucinated objects compared to existing baselines. Importantly, this improved FDR control does not sacrifice power; CORAL consistently achieves the highest or near-highest power, reflecting its ability to retain truly visually grounded objects while mitigating hallucinations. This favorable balance between multiple-testing error control (via FDR) and power is maintained on the more challenging

*popular* and *adversarial* subsets, where competing methods exhibit noticeable degradation in power. These results demonstrate that CORAL enables effective FDR control while maintaining high power, leading to robust hallucination mitigation.

Table 2: Evaluation with POPE score across multiple LVLM architectures on the MSCOCO dataset. We report individualized Accuracy and F1 score (mean ± std over 3000 runs). **Bold** indicates the best result and underline indicates the second-best result.

| Method | LLaVA-OneVision-7B | | Qwen2.5-VL-7B | | InternVL3-8B | | Average | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy↑ | F1 Score↑ | Accuracy↑ | F1 Score↑ | Accuracy↑ | F1 Score↑ | Accuracy↑ | F1 Score↑ |
| *Random* | | | | | | | | |
| Regular | 85.87 (± 1.77) | 85.72 (± 1.66) | 86.20 (± 1.30) | 86.98 (± 1.28) | 87.15 (± 1.88) | 86.26 (± 1.80) | 86.41 (± 1.65) | 86.32 (± 1.58) |
| VCD (2024) | 86.33 (± 1.23) | 88.86 (± 1.43) | 87.35 (± 1.83) | 90.06 (± 2.03) | 88.15 (± 1.53) | 88.05 (± 1.45) | 87.28 (± 1.53) | 88.99 (± 1.64) |
| MARINE (2025) | 87.21 (± 1.35) | 89.09 (± 1.09) | 89.72 (± 1.11) | 90.33 (± 2.09) | 89.23 (± 1.35) | 89.15 (± 1.19) | 88.72 (± 1.27) | 89.52 (± 1.46) |
| AGLA (2025) | 88.15 (± 1.65) | 89.97 (± 1.21) | 90.01 (± 1.23) | 91.17 (± 1.51) | 92.53 (± 1.42) | 93.54 (± 1.77) | 90.23 (± 1.43) | 91.56 (± 1.50) |
| CORAL (Ours) | **92.17 (± 0.89)** | **91.64 (± 1.83)** | **90.63 (± 1.54)** | **91.89 (± 1.83)** | **93.55 (± 1.52)** | **95.35 (± 1.87)** | **92.12 (± 1.32)** | **92.96 (± 1.84)** |
| *Popular* | | | | | | | | |
| Regular | 82.72 (± 1.99) | 83.16 (± 1.85) | 83.88 (± 1.48) | 84.79 (± 1.52) | 84.63 (± 1.48) | 85.26 (± 1.15) | 83.74 (± 1.65) | 84.40 (± 1.51) |
| VCD (2024) | 84.24 (± 1.46) | 86.16 (± 1.30) | 86.32 (± 1.83) | 89.34 (± 2.04) | 85.29 (± 1.14) | 86.25 (± 2.00) | 85.28 (± 1.48) | 87.25 (± 1.78) |
| MARINE (2025) | 86.81 (± 1.21) | 88.19 (± 2.09) | 88.72 (± 1.38) | 90.33 (± 1.18) | 87.67 (± 1.53) | 88.37 (± 1.81) | 87.73 (± 1.37) | 88.96 (± 1.69) |
| AGLA (2025) | 89.28 (± 1.11) | 89.34 (± 1.70) | 88.14 (± 1.23) | 90.79 (± 1.15) | 89.42 (± 1.32) | 90.17 (± 1.22) | 88.95 (± 1.22) | 90.10 (± 1.36) |
| CORAL (Ours) | **90.37 (± 1.89)** | **89.91 (± 1.33)** | **90.00 (± 1.45)** | **91.28 (± 1.83)** | **91.33 (± 1.35)** | **93.57 (± 1.44)** | **90.57 (± 1.56)** | **91.59 (± 1.53)** |
| *Adversarial* | | | | | | | | |
| Regular | 80.28 (± 2.21) | 81.02 (± 2.07) | 83.17 (± 1.37) | 83.46 (± 2.08) | 84.59 (± 1.33) | 84.96 (± 1.80) | 82.68 (± 1.64) | 83.15 (± 1.98) |
| VCD (2024) | 81.21 (± 1.89) | 82.96 (± 1.82) | 85.43 (± 1.88) | 85.24 (± 2.04) | 85.88 (± 1.65) | 86.99 (± 1.77) | 84.17 (± 1.81) | 85.06 (± 1.88) |
| MARINE (2025) | 85.26 (± 2.11) | 85.29 (± 2.09) | 87.42 (± 1.11) | 88.73 (± 1.91) | 87.25 (± 1.75) | 88.11 (± 1.16) | 86.64 (± 1.66) | 87.38 (± 1.72) |
| AGLA (2025) | 86.44 (± 1.61) | **88.14 (± 2.01)** | 91.39 (± 1.23) | 89.73 (± 2.15) | 90.11 (± 1.32) | 89.83 (± 1.10) | 89.31 (± 1.39) | 89.23 (± 1.75) |
| CORAL (Ours) | **88.27 (± 1.81)** | 87.99 (± 2.03) | **93.75 (± 1.54)** | **90.87 (± 1.83)** | **91.03 (± 1.00)** | **90.48 (± 1.83)** | **91.02 (± 1.45)** | **89.78 (± 1.90)** |

**Results on POPE.** POPE evaluates object-level grounding in LVLMs by testing their ability to answer *yes-or-no* questions about visual content. On the MSCOCO dataset, we report accuracy and F1 score. As shown in Table 2, CORAL consistently improves performance across all evaluated LVLMs and settings, demonstrating its effectiveness in mitigating object hallucinations. Notably, the improvements are particularly evident under the more challenging *adversarial* setting, where hallucination errors are more likely to occur. Additional results are provided in Tables 18, 19, and 20 in the Appendix. By explicitly controlling the false discovery rate (FDR), our method prioritizes limiting false positive predictions within each image. While this emphasis may introduce a mild trade-off with recall, it leads to more reliable predictions and improved overall performance in hallucination mitigation.

Table 3: Evaluation with MME score (mean ± std over 3000 runs) across multiple LVLM architectures on MSCOCO. **Bold** indicates the best result and underline indicates the second-best result.

| Model | Method | Object | | Attribute | Relation | | Total ↑ |
|---|---|---|---|---|---|---|---|
| | | Existence ↑ | Count ↑ | Color ↑ | Position ↑ | Commonsense ↑ | |
| LLaVA-OneVision-7B | Regular | 190.33 (± 6.50) | 145.53 (± 15.20) | 170.66 (± 9.10) | 160.25 (± 8.30) | 70.16 (± 6.50) | 736.93 (± 24.80) |
| | VCD (2024) | 186.25 (± 7.22) | 147.25 (± 11.44) | 175.35 (± 15.58) | 165.36 (± 2.55) | 79.42 (± 5.74) | 753.63 (± 18.76) |
| | MARINE (2025) | 191.26 (± 4.55) | 150.44 (± 10.15) | 177.45 (± 10.96) | 165.36 (± 7.19) | 82.33 (± 5.29) | 766.84 (± 17.47) |
| | AGLA (2025) | 190.54 (± 6.22) | 153.32 (± 15.37) | 178.42 (± 3.22) | 170.35 (± 4.22) | 84.36 (± 5.21) | 776.99 (± 15.96) |
| | CORAL (Ours) | **193.67 (± 3.21)** | **160.43 (± 10.11)** | **180.35 (± 8.46)** | **175.24 (± 10.58)** | **85.52 (± 4.22)** | **795.21 (± 18.12)** |
| Qwen2.5-VL-7B | Regular | 185.52 (± 5.80) | 140.46 (± 12.20) | 165.16 (± 8.60) | 155.22 (± 7.50) | 80.53 (± 6.20) | 726.89 (± 21.30) |
| | VCD (2024) | 187.35 (± 7.36) | 145.63 (± 10.45) | 173.35 (± 8.23) | 160.22 (± 11.34) | 82.55 (± 6.29) | 749.10 (± 11.61) |
| | MARINE (2025) | 189.83 (± 4.25) | 153.35 (± 10.24) | 176.25 (± 7.21) | 165.14 (± 10.44) | 84.21 (± 4.87) | 768.78 (± 11.88) |
| | AGLA (2025) | 190.15 (± 3.28) | 158.33 (± 11.38) | 179.35 (± 4.15) | 168.13 (± 9.83) | 87.28 (± 2.28) | 783.24 (± 13.42) |
| | CORAL (Ours) | **194.24 (± 3.18)** | **161.35 (± 10.01)** | **182.53 (± 6.98)** | **170.86 (± 8.14)** | **88.91 (± 3.71)** | **797.89 (± 12.67)** |
| InternVL3-8B | Regular | 192.25 (± 6.20) | 150.22 (± 10.80) | 175.48 (± 7.40) | 165.16 (± 6.50) | 95.38 (± 5.20) | 778.49 (± 19.80) |
| | VCD (2024) | 193.26 (± 3.28) | 159.25 (± 10.97) | 178.25 (± 9.25) | 175.35 (± 7.73) | 95.15 (± 1.54) | 801.26 (± 13.36) |
| | MARINE (2025) | **193.87 (± 2.82)** | 163.28 (± 3.29) | 180.93 (± 5.18) | 177.35 (± 4.26) | 96.18 (± 1.58) | 811.61 (± 13.42) |
| | AGLA (2025) | 193.17 (± 3.28) | 165.26 (± 7.77) | 183.85 (± 5.28) | 177.96 (± 4.26) | 95.15 (± 1.77) | 815.39 (± 12.81) |
| | CORAL (Ours) | 193.46 (± 4.26) | **168.36 (± 10.53)** | **185.29 (± 13.19)** | **178.54 (± 1.35)** | **96.28 (± 1.98)** | **821.93 (± 12.30)** |

**Results on MME.** The MME evaluation extends beyond POPE by covering a broader range of hallucination types, including object-, attribute-, and relation-level hallucinations. As shown in Table 3, CORAL consistently improves overall MME performance across all evaluated LVLM architectures, achieving the best total scores in all settings. These results suggest that, although CORAL

Table 4: Evaluation with MMBench score (mean ± std over 3000 runs) across multiple LVLM architectures on MSCOCO. **Bold** indicates the best result and underline indicates the second-best result.

| Method | LLaVA-OneVision-7B | Qwen2.5-VL-7B | InternVL3-8B |
|---|---|---|---|
| Regular | 80.77 (± 1.21) | 83.50 (± 1.01) | 83.40 (± 0.85) |
| VCD (2024) | 82.67 (± 1.00) | 85.14 (± 1.32) | 86.84 (± 0.99) |
| MARINE (2025) | 85.31 (± 1.09) | 87.44 (± 1.17) | 88.67 (± 1.03) |
| AGLA (2025) | 85.76 (± 1.11) | 87.02 (± 1.07) | 90.15 (± 1.00) |
| CORAL (Ours) | **86.70 (± 1.13)** | **87.90 (± 0.91)** | **94.50 (± 0.80)** |

is primarily designed for object-level hallucination control, it can also help mitigate attribute-level inconsistencies by suppressing noise-induced predictions that are not well supported by the visual input. This indicates that the benefits of CORAL extend beyond object presence to more fine-grained visual descriptions. Additional MME results for other models are reported in Table 14 in Appendix.

**Results on MMBench.** We further evaluate CORAL on MMBench [55] to assess its effect on general multimodal understanding beyond object-hallucination benchmarks. As shown in Table 4, CORAL improves the MMBench score across all three evaluated backbones. For LLaVA-OneVision-7B, the score increases from 80.8 to 86.7, yielding a gain of 5.9 points. For Qwen2.5-VL-7B and InternVL3-8B, the scores improve from 83.5 to 87.9 and from 83.4 to 94.5, respectively, corresponding to gains of 4.4 and 11.1 points. These results suggest that CORAL mitigates object hallucination while improving broader multimodal understanding under the evaluated settings.

Table 5: Ablation study on POPE metrics using the MSCOCO dataset with LLaVA-OneVision-7B. Results are mean ± std over 10 runs. **Bold** indicates the best result.

| Method | Accuracy ↑ | Precision ↑ | Recall ↑ | F1 Score ↑ |
|---|---|---|---|---|
| Regular | 85.87 (± 1.77) | 83.41 (± 2.13) | 88.08 (± 1.47) | 85.72 (± 1.66) |
| w/o Visual Uncertainty Splitting | 88.12 (± 0.62) | 86.78 (± 0.74) | 89.07 (± 0.71) | 87.90 (± 0.58) |
| w/o Mirror Statistic | 88.76 (± 0.55) | 87.07 (± 0.49) | 90.33 (± 0.61) | 88.67 (± 0.53) |
| w/o Overall FDR Control | 89.36 (± 0.48) | 86.42 (± 0.92) | 92.37 (± 0.56) | 89.27 (± 0.41) |
| CORAL (Ours) | **92.17 (± 0.89)** | **88.25 (± 1.46)** | **94.87 (± 1.61)** | **91.64 (± 1.83)** |

### 4.3 Ablation Study

We conduct ablation studies to evaluate the contribution of visual uncertainty splitting, mirror statistic, and FDR control. As shown in Table 5, removing any component degrades performance, confirming their complementary roles. In particular, disabling FDR control increases recall but reduces precision and F1, indicating a higher tendency to retain hallucinated objects. Removing mirror statistic leads to reduced recall, reflecting weaker ability to preserve truly grounded objects, while removing visual uncertainty splitting results in overall performance degradation. The full CORAL achieves the best results across all metrics, demonstrating its effectiveness in balancing FDR control and power. Additional results are provided in Figures 10 and 11 in the Appendix.

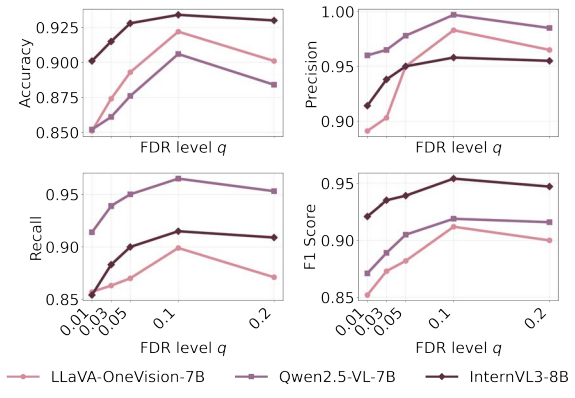



Figure 4: Ablation study on the effect of FDR target level ($q$) on the performance of LLaVA-OneVision-7B, Qwen2.5-VL-7B, InternVL3-8B using POPE metrics with $q = \{0.01, 0.03, 0.05, 0.1, 0.2\}$.

**Effect of FDR Control at Different Levels on Object Hallucinations.** Figure 4 shows the impact of the FDR target level $q$ on performance across LVLMs. As $q$ increases from small values, both accuracy and F1 score improve, indicating that overly strict control suppresses not only hallucinated objects but also true positives. Performance peaks at intermediate values of $q$, where a favorable balance between hallucination mitigation and retention of grounded objects is achieved. Further increasing $q$ leads to diminishing returns or slight degradation, as more spurious predictions are admitted. These results highlight the fundamental FDR-power trade-off, with consistent trends observed across architectures. This behavior aligns with the theoretical role of FDR control in regulating false discoveries while preserving statistical power.

**Latency Analysis.** We evaluate inference efficiency by measuring the average latency per generated token. Detailed results are provided in Figure 8 in the Appendix B.7. Among all approaches, CORAL incurs the smallest latency increase. Among all approaches, CORAL incurs the smallest latency increase (50.26 ms/token). While AGLA [23] (51.42 ms/token) and MARINE [22] (52.21 ms/token) rely on iterative decoding or repeated sampling, and VCD [24] (53.42 ms/token) performs contrastive decoding within the autoregressive loop, these methods introduce step-wise overhead that accumulates over the sequence. Although CORAL evaluates three visual inputs, these forward passes are used only for post-hoc mirror statistic computation. As a result, CORAL avoids decoding-time logit reweighting and achieves lower latency than VCD despite the additional view.

## 5 Conclusion, Limitations, and Future Work

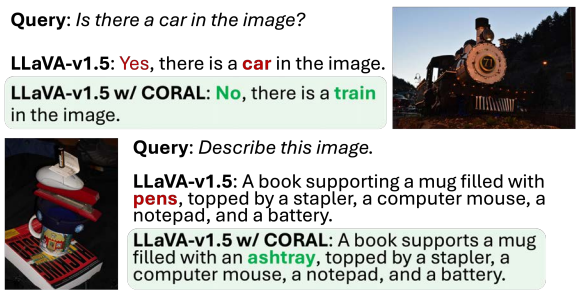



Figure 5: Hallucination mitigation examples by our method CORAL across multiple tasks. Hallucinated objects are highlighted in red.

We propose CORAL, a training-free framework that introduces visual uncertainty via data splitting and leverages mirror statistic to control the FDR of hallucinated objects during decoding. By explicitly controlling false positives at the image level while maintaining high power, CORAL suppresses hallucinations without discarding informative visual evidence. Extensive experiments across multiple LVLM architectures demonstrate consistent reductions in FDR and improvements in power and accuracy, particularly under challenging *popular* and *adversarial* settings.

**Limitations and Future Work.** Although effective, CORAL cannot be applied directly to black-box commercial APIs that do not expose token-level logits. Future work includes extending the FDR-based framework to handle more complex hallucinations, such as relational and attribute errors, and to open-ended generation tasks beyond object existence queries. More broadly, integrating CORAL with adaptive decoding strategies and applying FDR control to multimodal reasoning tasks are promising directions.

## References


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

[2] Zongxia Li, Xiyang Wu, Hongyang Du, Huy Nghiem, and Guangyao Shi. Benchmark evaluations, applications, and challenges of large vision language models: A survey. *arXiv preprint arXiv:2501.02189*, 1, 2025.

[3] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[4] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

[5] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[6] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[7] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[8] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[9] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 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, 2024.

[10] Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Y Zou, Kai-Wei Chang, and Wei Wang. Enhancing large vision language models with self-training on image comprehension. *Advances in Neural Information Processing Systems*, 37:131369–131397, 2024.

[11] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

[12] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[13] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[14] 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*, pages 740–755. Springer, 2014.

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

[16] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL `https://openreview.net/forum?id=xozJw0kZXF`.

[17] Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023.

[18] 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. *arXiv preprint arXiv:2310.00754*, 2023.

[19] Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G Shapiro, and Ranjay Krishna. Perception tokens enhance visual reasoning in multimodal language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3836–3845, 2025.

[20] Hongyan Chang, Ali Shahin Shamsabadi, Kleomenis Katevas, Hamed Haddadi, and Reza Shokri. Context-aware membership inference attacks against pre-trained large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7299–7321, 2025.

[21] Drew A Hudson and Christopher D Manning. 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, 2019.

[22] 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*. PMLR, 2025.

[23] Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29915–29926, 2025.

[24] 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*, pages 13872–13882, 2024.

[25] Chenguang Dai, Buyu Lin, Xin Xing, and Jun S Liu. False discovery rate control via data splitting. *Journal of the American Statistical Association*, 118(544):2503–2520, 2023.

[26] Xin Xing, Zhigen Zhao, and Jun S Liu. Controlling false discovery rate using gaussian mirrors. *Journal of the American Statistical Association*, 118(541):222–241, 2023.

[27] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[28] Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. 2015.

[29] Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):551–577, 2018.

[30] Chenguang Dai, Buyu Lin, Xin Xing, and Jun S Liu. A scale-free approach for false discovery rate control in generalized linear models. *Journal of the American Statistical Association*, 118 (543):1551–1565, 2023.

[31] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

[32] Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. *arXiv preprint arXiv:2411.02712*, 2024.

[33] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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

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

[36] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

[37] Yifan Li, Kun Zhou, Xin Zhao, Lei Fang, and Jirong Wen. Analyzing and mitigating object hallucination: A training bias perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 6636–6643, 2026.

[38] SMTI Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6, 2024.

[39] Zichao Lin, Shuyan Guan, Wending Zhang, Huiyan Zhang, Yugang Li, and Huaping Zhang. Towards trustworthy llms: a review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review*, 57(9):243, 2024.

[40] Qi Yang, Shreya Ravikumar, Fynn Schmitt-Ulms, Satvik Lolla, Ege Demir, Iaroslav Elistratov, Alex Lavaee, Sadhana Lolla, Elaheh Ahmadi, Daniela Rus, et al. Uncertainty-aware language modeling for selective question answering. *arXiv preprint arXiv:2311.15451*, 2023.

[41] Lixin Xiu, Xufang Luo, and Hideki Nakayama. A comprehensive information-decomposition analysis of large vision-language models. *arXiv preprint arXiv:2603.29676*, 2026.

[42] Zhaoxu Li, Chenqi Kong, Yi Yu, Qiangqiang Wu, Xinghao Jiang, Ngai-Man Cheung, Bihan Wen, Alex Kot, and Xudong Jiang. Saver: Mitigating hallucinations in large vision-language models via style-aware visual early revision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35617–35625, 2026.

[43] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698, 2020.

[44] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*, 2016.

[45] 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*, pages 6904–6913, 2017.

[46] Hong Yan, Lijun Liu, Xupeng Feng, and Qingsong Huang. Overcoming language priors with self-contrastive learning for visual question answering. *Multimedia Tools and Applications*, 82 (11):16343–16358, 2023.

[47] Zhibo Ren, Huizhen Wang, Muhua Zhu, Yichao Wang, Tong Xiao, and Jingbo Zhu. Overcoming language priors with counterfactual inference for visual question answering. In *China National Conference on Chinese Computational Linguistics*, pages 58–71. Springer, 2023.

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

[49] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[50] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. Technical report, Qwen Team, 2025.

[51] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[52] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

[53] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

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

[55] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[56] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

[58] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

[59] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228, 2018.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contribution, and these claim directly supported by the proposed method (in Sec. 3) and consistent experimental results demonstrating reduced hallucination without sacrificing utility (in Sec. 4.2).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses the limitations of the proposed methods in conclusion section (in Sec. 5).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: This paper provides the theory proof in Sec. 3 (Subsec. 3.3), with formal proofs including Lemma B.1 in App. B.4.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides sufficient details to reproduce the main experimental results, including model configurations, datasets, decoding procedures, and evaluation metrics in Sec. 4.2 and App. B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides an anonymous GitHub repository link (included in the abstract) with code and instructions to reproduce the main results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental settings, including datasets, evaluation protocols, and key hyperparameters, are described in Sec. 4.2, with additional implementation details provided in App. B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports experimental results with mean $\pm$ standard deviation across multiple runs, capturing variability due to random sampling and experimental conditions (Sec. 4.2 and App. B).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides details on the computational resources used for the experiments, including hardware configuration and execution settings, in Sec. 4.2, with additional details in the appendix (App. B).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics, and all experiments are conducted using publicly available datasets without involving human subjects or sensitive personal data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: This work improves the reliability of large vision-language models by reducing hallucinated outputs through FDR-controlled decoding, enhancing visual grounding and trustworthiness in Section 4.2 and Appendix B. However, improved output consistency may increase the risk of generating convincing misleading content, and the method does not address broader concerns such as bias or harmful language inherited from pretrained models in Appendix A.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release new datasets or pretrained models with high risk for misuse; the proposed method is a decoding method applied to existing models, and thus no additional safeguards are required.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper uses publicly available datasets and pretrained models, which are properly cited, and their terms of use are respected as described in the corresponding references.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce new datasets or pretrained models; the proposed method is implemented as code, which is provided via an anonymous repository for reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects; all experiments are conducted using publicly available datasets.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects; therefore, Institutional Review Board (IRB) approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core methodology does not involve LLMs as an important, original, or non-standard component; existing LVLMs are only used as evaluation platforms to assess hallucination control.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

# A Impact Statement

This paper presents research aimed at improving the reliability of large vision-language models by reducing hallucinated outputs during generation. The proposed method, CORAL, operates directly at decoding time and does not require additional training, external data, or model fine-tuning. This makes it easy to apply to existing models and practical for real-world use. By improving the alignment between generated outputs and visual input, CORAL helps reduce visually ungrounded content while preserving useful and relevant information. This targeted reduction of hallucinations can improve the quality and dependability of model responses, especially in applications where incorrect outputs may mislead users. At the same time, CORAL focuses specifically on hallucinations related to visual grounding and does not address other issues such as harmful language or biases inherited from model pretraining. These limitations are outside the scope of this work and may be explored in future research. To the best of our knowledge, this work does not introduce negative societal impacts associated with our research that merit highlighting in these discussions.

# B Experiment Setup

We conduct all of the experiments using a cluster with four NVIDIA Quadro RTX 6000 (24GB) GPUs using CUDA 11.7. Each single experiment can be run on a single RTX 6000 GPU.

## B.1 Model Architecture

In Table 6, we provide detailed descriptions of the LVLM architectures used in our experiments. These LVLMs respectively leverage the pre-trained vision encoder of the models we listed, which are all based on the Vision Transformer (ViT) [56] architecture.

Table 6: Details of the LVLM architectures used in our experiments.

| Model | Vision encoder | LLM |
|---|---|---|
| LLaVA-v1.5 [1] | CLIP-L-336px [57] | Vicuna-v1.5-7B [58] |
| Qwen-VL [53] | ViT-based visual encoder | Qwen-7B [53] |
| InstructBLIP [52] | BLIP-2 [3] | Vicuna-v1.1-7B [58] |
| LLaVA-OneVision-7B [49] | CLIP ViT-L/14 (336px) [57] | Vicuna-7B [58] |
| Qwen2.5-VL-7B [50] | ViT-based visual encoder | Qwen2.5-7B [50] |
| InternVL3-8B [51] | InternViT (high-resolution ViT) | InternLM2-8B [51] |

## B.2 Prompt Template

For each query, we randomly select a prompt template from the available template list, as shown in Table 7.

## B.3 Implementation Details of Visual Perturbations

In Eq. (3), the visual input $\mathbf{v}$ refers to the patch-level visual embeddings produced by the vision encoder. Specifically, given a patch-level visual embeddings $\mathbf{v}$, we construct mirror views as $\mathbf{v}^{\pm} = \mathbf{v} \pm \tau_v \mathbf{Z}_v$, where $\mathbf{Z}_v \sim \mathcal{N}(0, \mathbf{I})$ is sampled independently per image. This design ensures that both mirror views share the same visual semantics while inducing controlled uncertainty in the visual signal.

## B.4 Empirical Validation of the Mirror-Symmetry Property

Our FDR estimator in Eq. (8) relies on approximate null symmetry of the mirror statistic. The paired visual views $f_v^{\pm} = \mathbf{v} \pm \tau_v Z_v$ share the same noise realization with opposite signs. Consequently, the induced logit contrasts $\Delta_t^+$ and $\Delta_t^-$ are coupled and are not necessarily uncorrelated. Conditioning on $Z_v$ fixes the perturbations and does not justify an uncorrelatedness claim.

Instead, we state the required condition directly as approximate conditional sign-flip invariance under the null. Define

$$H_{0t} : \text{no systematic visual evidence supports token } y_t,$$

Table 7: Details of the LVLM architectures that we used in our paper.

| **Template Type** | **Prompt Template** |
|---|---|
| POPE task | This image contains only the following objects: <OBJECT_GROUNDING>. Do not assume any objects beyond this list. Based solely on this information, <QUERY><br>The detected objects in the image are: <OBJECT_GROUNDING>. Answer the question using only these objects. <QUERY><br>This image shows the following objects: <OBJECT_GROUNDING>. You must answer using only the objects in this list.Given these detected objects, <QUERY><br>The objects found in this image are limited to: <OBJECT_GROUNDING>. You should rely strictly on this list of objects and make no other guesses. Based on this, <QUERY> |
| CORAL grounded | This image contains the following visually grounded objects: <OBJECT_GROUNDING>. Based on the image, <QUERY><br>The following objects are visible in the image: <OBJECT_GROUNDING>. Using only the information from the image, <QUERY><br>This image shows: <OBJECT_GROUNDING>. Please answer the following question based on the image content: <QUERY> |
| CORAL-restricted | The objects visible in this image are limited to: <OBJECT_GROUNDING>. Do not assume any objects beyond this list. <QUERY><br>Only the following objects appear in the image: <OBJECT_GROUNDING>. Answer the question using only visual evidence from the image. <QUERY><br>Based strictly on the objects shown in the image: <OBJECT_GROUNDING>. Do not infer any additional objects. <QUERY> |
| CORAL-complementary | The same image is analyzed using multiple internally constructed visual representations, including the original view and two mirrored views.<br>Original view detects the following objects: <OBJECT_GROUNDING><br>Mirror view (+) detects the following objects: <OBJECT_GROUNDING_A><br>Mirror view (-) detects the following objects: <OBJECT_GROUNDING_B><br>Using the visual information above from the same image, <QUERY><br>Multiple complementary visual representations are derived from the same image.<br>Original view objects: <OBJECT_GROUNDING><br>Mirrored view (+) objects: <OBJECT_GROUNDING_A><br>Mirrored view (-) objects: <OBJECT_GROUNDING_B><br>Based on the image, <QUERY> |

and let

$$\mathcal{G}_t = \sigma(\mathbf{v}, \mathbf{x}, \mathbf{y}_{<t})$$

denote the shared context consisting of the image, textual input, and decoding history.

**Lemma B.1** (Conditional null symmetry)**.** *Suppose that, under $H_{0t}$,*

$$(\Delta_t^+, \Delta_t^-) \big| \, H_{0t}, \mathcal{G}_t \overset{d}{\approx} (\Delta_t^+, -\Delta_t^-) \big| \, H_{0t}, \mathcal{G}_t, \tag{10}$$

*where $\overset{d}{\approx}$ denotes approximate equality in distribution. Then the mirror statistic satisfies*

$$\Delta_t | \, H_{0t}, \mathcal{G}_t \overset{d}{\approx} -\Delta_t | \, H_{0t}, \mathcal{G}_t. \tag{11}$$

*Consequently, for every $s > 0$,*

$$\Pr(\Delta_t \geq s \mid H_{0t}, \mathcal{G}_t) \approx \Pr(\Delta_t \leq -s \mid H_{0t}, \mathcal{G}_t). \tag{12}$$

*Under exact conditional sign-flip invariance, these relations hold exactly.*

*Proof.* Define

$$M(a, b) = |a + b| - |a - b|.$$

The mirror functional is antisymmetric under a sign flip of its second argument:

$$M(a, -b) = |a - b| - |a + b| = -M(a, b).$$

Applying the assumed conditional sign-flip invariance gives

$$\begin{aligned} \Delta_t \big| H_{0t}, \mathcal{G}_t &= M(\Delta_t^+, \Delta_t^-) \big| H_{0t}, \mathcal{G}_t \\ &\overset{d}{\approx} M(\Delta_t^+, -\Delta_t^-) \big| H_{0t}, \mathcal{G}_t \\ &= -\Delta_t \big| H_{0t}, \mathcal{G}_t. \end{aligned} \tag{13}$$

This yields the corresponding approximate positive- and negative-tail equality. Exact sign-flip invariance gives exact equalities. □

The conditional sign-flip condition is a working assumption for nonlinear LVLM responses; it is not guaranteed by Gaussian input perturbations alone. It replaces the unsupported claim that the two logit contrasts are uncorrelated by construction. The mirror statistic and empirical thresholding procedure remain unchanged.

***Proof of Theorem 3.1***. The conditional sign-flip assumption in Theorem 3.1 is the exact-invariance case of Lemma B.1. Applying that lemma to the mirror statistic defined in Eq. (6) yields, for every $s > 0$,

$$\Pr(\Delta_t \leq -s \mid H_{0t}, \mathcal{G}_t) = \Pr(\Delta_t \geq s \mid H_{0t}, \mathcal{G}_t). \tag{14}$$

which completes the proof. □

The null symmetry in Theorem 3.1, together with suitable tail-count concentration conditions, motivates using negative-tail counts to estimate false discoveries among positively selected outputs.

To empirically assess whether this assumption holds in practice for LVLMs, we conduct a dedicated diagnostic analysis on *negative POPE queries*, where the queried object is guaranteed to be absent from the image. In this setting, all object-level statistic correspond to null hypotheses, providing a controlled environment for evaluating symmetry.

**Interpretation of the mirror statistic.** The mirror statistic admits the equivalent expression

$$\Delta_t = 2 \, \mathrm{sign}(\Delta_t^+ \Delta_t^-) \min\{|\Delta_t^+|, |\Delta_t^-|\}. \tag{15}$$

Thus, $\Delta_t$ is positive when the two contrasts have the same nonzero sign, negative when they have opposite signs, and zero when either contrast is zero. Its magnitude is twice the smaller absolute contrast. A large positive value therefore indicates a strong, directionally consistent response across the paired views, rather than establishing visual grounding by itself.

In particular, both $\Delta_t^+ > 0, \Delta_t^- > 0$ and $\Delta_t^+ < 0, \Delta_t^- < 0$ produce positive mirror statistics. In the latter regime, the token's logit is higher under both perturbed views than under the clean input. The current threshold-based selection rule does not explicitly distinguish these two regimes. Consequently, interpreting selected outputs as visually grounded depends on the validity of the null-symmetry assumption and the enrichment of grounded outputs in the positive tail.

To empirically assess whether this assumption holds in practice for LVLMs, we conduct a dedicated diagnostic analysis on *negative POPE queries*, where the queried object is guaranteed to be absent from the image. In this setting, all object-level statistic correspond to null hypotheses, providing a controlled environment for evaluating symmetry.

**Setup.** We use the POPE random split on LLaVA-v1.5 and extract object-level mirror statistic $\Delta_t$ from 1500 negative (answer = "no") queries. Each $\Delta_t$ is computed from mirrored visual features as described in Sec. 3.2, using the logit-margin difference between the positive and negative mirror views. No filtering or thresholding is applied in this analysis.

**Distributional symmetry.** Figure 6 visualizes the empirical distribution of $\Delta_t$ and its sign-flipped counterpart $-\Delta_t$. The overlaid histograms show strong overlap, and the QQ plot (in Figure 7) of $\Delta_t$ versus $-\Delta_t$ aligns closely with the identity line, indicating approximate symmetry across the full range of quantiles.

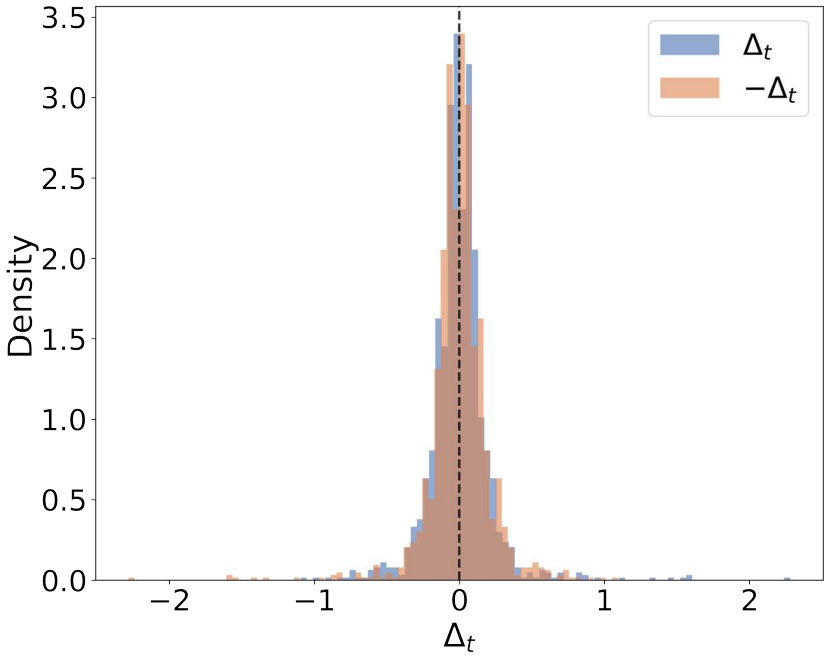



Figure 6: Empirical validation of mirror symmetry for mirror statistic $\Delta_t$. The figure shows overlaid histograms of $\Delta_t$ and its sign-flipped counterpart $-\Delta_t$ computed from negative POPE queries, where the queried object is guaranteed to be absent. The strong overlap between the two distributions and their near symmetry around zero indicate that mirror statistic for visually ungrounded objects are approximately symmetric, supporting their use as reference quantities for FDR estimation.

**Quantitative diagnostics.** We further report numerical symmetry diagnostics. The empirical mean of $\Delta_t$ is close to zero ($\text{mean}(\Delta_t) = 0.0012$), and the Kolmogorov-Smirnov test between $\Delta_t$ and $-\Delta_t$ yields a statistic of $0.0167$ with $p = 0.99$, failing to reject the null hypothesis that the two distributions are identical. These results indicate no detectable asymmetry at conventional significance levels.

**Implications for FDR estimation.** While LVLMs are highly nonlinear and do not strictly satisfy classical assumptions of mirror-based multiple testing, the above results suggest that, under symmetric feature perturbations, the induced mirror statistic for visually ungrounded objects are empirically well-approximated by a symmetric distribution. This empirical symmetry supports the use of Eq. (8) as a reliable plug-in estimator for the false discovery rate in our decoding-time selection procedure.

### B.5 Implementation Details: Object Sets and Token Span Construction

This appendix provides a precise description of how object sets and token spans are defined across tasks, and how object-level mirror statistic are constructed for false discovery rate (FDR) control.

**Object set definition.** For a given image and prompt, `CORAL` performs statistical testing at the object level. Let $\mathcal{S}$ denote the set of candidate objects associated with the input. The definition of

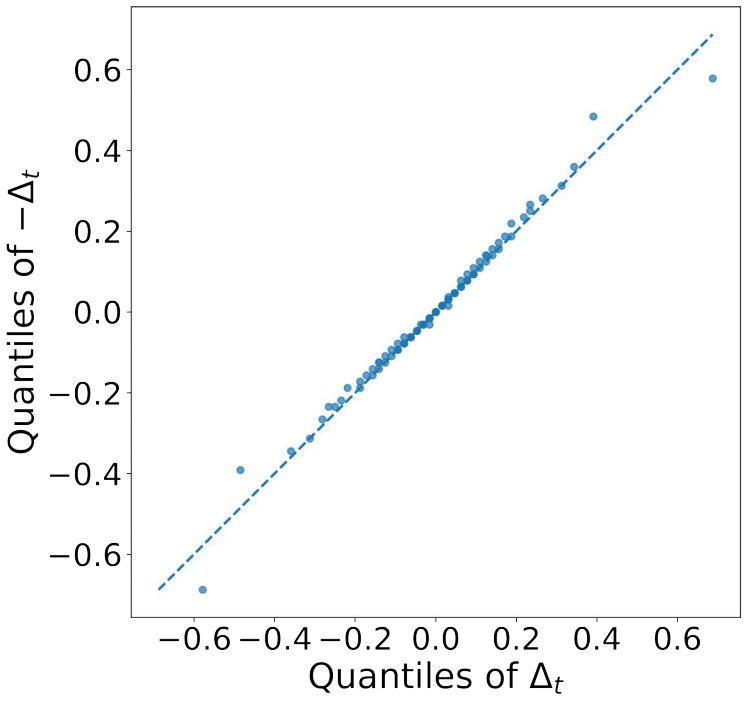



Figure 7: Quantile-Quantile (QQ) plot comparing the empirical quantiles of $\Delta_t$ and $-\Delta_t$ for negative POPE queries. The close alignment with the identity line across the full range of quantiles indicates approximate distributional symmetry of the mirror statistic, providing empirical support for the mirror-symmetry assumption used in Eq. (8).

$\mathcal{S}$ is task-dependent. For closed-set object existence benchmarks such as POPE and MME, $\mathcal{S}$ is directly given by the queried object categories provided by the benchmark (e.g., *fork*, *bus*, *zebra*). For captioning evaluation with CHAIR, $\mathcal{S}$ consists of object categories extracted from the generated caption using the standard CHAIR evaluation pipeline, which matches noun phrases against the MSCOCO object vocabulary and its synonym list.

**Object mention identification.** For each object $y_t \in \mathcal{S}$, we identify its occurrences in the generated text by matching the canonical object name and its associated synonyms to the generated sequence. All matching is performed at the tokenizer level of the evaluated LVLM, ensuring a deterministic and reproducible mapping. Each matched occurrence corresponds to a contiguous span of subword tokens.

**Token span construction.** For captioning, let $\mathcal{T}_i$ denote the set of token indices corresponding to all occurrences of object category $i$ in the generated text. If an object appears multiple times, $\mathcal{T}_i$ is the union of its corresponding token spans. Token-level mirror statistics $\Delta_t$ are computed for each associated token position $t$. The constituent tokens are not treated as independent object discoveries; instead, they are grouped into a single object-level decision.

**Object-level aggregation.** For objects represented by multiple tokens, we summarize the evidence using the least-supported token:

$$\Delta_i = \min_{t \in \mathcal{T}_i} \Delta_t. \tag{16}$$

At the selection threshold $T_q$, object $i$ is retained only if

$$\Delta_i \geq T_q \quad \Longleftrightarrow \quad \Delta_t \geq T_q \ \text{ for all } t \in \mathcal{T}_i. \tag{17}$$

This rule prevents an object from being retained solely because one token has strong visual evidence while another token in the associated span is weakly supported. The grouped tokens contribute a single object-level decision.

**Scope of FDR control.** Token-level FDR control does not, in general, automatically imply object-level FDR control under an arbitrary token-to-object mapping. For POPE, the one-object-one-decision structure closely aligns the tested decision with the evaluated object-level hypothesis. For multi-token objects and CHAIR-style caption evaluation, a general provable mapping requires additional assumptions on the parser, grouping function, and object-level null.

Accordingly, the guarantee is stated at the level at which the testing procedure is applied, under its required assumptions. The object-level improvements observed in our experiments provide empirical evidence of transfer to standard object-hallucination metrics, rather than a universal object-level FDR theorem for arbitrary structured outputs. A hierarchical procedure that combines within-span error control with FDR control across object groups is a possible direction for future work.

**Mirror statistic.** For each token $y_t \in \mathcal{S}$, we define a token-level mirror statistic $\Delta_t$ based on the logit differences under the original and perturbed visual inputs, as described in Eq. (6). This formulation operates directly at the token level and avoids the need for additional aggregation. As a result, the mirror-symmetry property required for valid FDR control is preserved. Object-level decisions can be derived from token-level statistic during evaluation by grouping tokens corresponding to the same object.

**FDR and power computation.** False discovery rate and power are computed at the image level over the object set $\mathcal{S}$ using $\{\Delta_t\}_{y_t \in \mathcal{S}}$, rather than over individual tokens. This guarantees that the statistical testing procedure is directly aligned with object-level hallucination evaluation protocols such as POPE, CHAIR, and MME.

### B.6 Statistical Dependence across Tokens and Objects

Mirror statistics are computed separately for each candidate decision, but are not assumed to be mutually statistically independent. Shared images, scene context, and decoding histories can induce dependence among the statistics.

**Autoregressive dependence.** At decoding step $t$, the clean and mirror-view logits are evaluated using the same fixed history $\mathbf{y}_{<t}$. The paired comparison is therefore made conditional on a shared decoding history. However, this does not imply independence of the mirror statistics across decoding steps.

**Null-tail symmetry and weak dependence.** Let $\mathcal{S}_0$ index the null testing units. For a candidate threshold $s > 0$, define the null-tail indicators

$$I_t^+(s) = \mathbf{1}\{\Delta_t \geq s\}, \qquad I_t^-(s) = \mathbf{1}\{\Delta_t \leq -s\}. \tag{18}$$

The analysis requires approximate marginal symmetry,

$$\mathbb{E}[I_t^+(s)] \approx \mathbb{E}[I_t^-(s)], \qquad t \in \mathcal{S}_0, \tag{19}$$

together with a weak-dependence condition such as

$$\sum_{t,u \in \mathcal{S}_0} \left|\mathrm{Cov}\big(I_t^{\pm}(s), I_u^{\pm}(s)\big)\right| = o\big(|\mathcal{S}_0|^2\big), \tag{20}$$

uniformly over the candidate thresholds. Here, the condition applies separately to the positive and negative tail indicators as the number of null testing units increases.

This condition allows correlations induced by shared visual inputs and contextual information, provided that their aggregate contribution satisfies Eq. (20). Local or block dependence can therefore be compatible with the condition, whereas pervasive dependence under which most null statistics move together can violate it.

**Dependence among object-level decisions.** Hallucinated objects such as a dog, a car, and a bicycle may be correlated through the same scene context. Such correlation does not, by itself, violate the weak-dependence condition. The relevant requirement concerns concentration of aggregate null-tail counts, rather than independence of every pair of object statistics.

When testing is performed on object-level statistics, the symmetry and dependence conditions must hold for those statistics. Thus, separate computation should not be interpreted as statistical independence. Within-image dependence diagnostics and end-to-end empirical FDR calibration assess the plausibility of these conditions in the evaluated settings.

### B.7 Implementation Details for Hallucination Evaluations

We evaluate the effectiveness of our methods on several state-of-the-art LVLMs, including LLaVA-v1.5 (7B and 13B) [1], InstructBLIP (7B and 13B) [52], Qwen-VL (7B) [53], LLaVA-OneVision-7B [49], Qwen2.5-VL-7B [50], and InternVL3-8B [51]. We compare our methods with three state-of-the-art decoding methods, including VCD [24], MARINE [22], and AGLA [23]. All evaluations were repeated over 3000 randomized trials. In each trial, we fix the model outputs and compute the evaluation metrics using random sampling procedures (e.g., POPE query sampling), rather than re-running full model inference. This allows efficient and stable estimation of mean and variance. We followed the suggested settings in their respective papers and released codes to ensure fair comparison. For the POPE [16] datasets, $\alpha$ is set to 2 and $\beta$ is set to 0.5. For the proposed methods, the target level for FDR control is set to 0.1. For the CHAIR [4], $\alpha$ is set to 2 and $\beta$ is set to 0.5. For the MME [54] dataset, we set $\alpha$ to 2 and $\beta$ to 0.5 for LLaVA-v1.5, and $\alpha$ and $\beta$ are set to 0.1 for InstructBLIP.In addition, the hyperparameter for VCD [24], MARINE [22], and AGLA [23] are reported in Table 8, Table 9, and Table 10 respectively. We strictly followed the original implementations and default hyperparameters described in their papers to reproduce each baseline's results.

Table 8: VCD [24] Hyperparameter Settings.

| **Parameters** | **Value** |
|---|---|
| Amplification Factor $\alpha$ | 1 |
| Adaptive Plausibility Threshold $\beta$ | 0.1 |
| Diffusion Noise Step | 500 |

Table 9: MARINE [22] Hyperparameter Settings.

| **Parameters** | **Value** |
|---|---|
| Guidance Strength | 0.7 |
| Score Threshold for DERT | 0.95 |
| Detect Threshold for RAM++ | 0.68 |

Table 10: AGLA [23] Hyperparameter Settings.

| **Parameters** | **Value** |
|---|---|
| Weighting Factor $\alpha$ | 2 |
| Adaptive Plausibility Constraint Factor $\beta$ | 0.5 |

Key factors that potentially affect the hallucination evaluation outcomes, including the evaluation dataset and prompt template, LVLM's sampling strategy and batched generation techniques, data splitting $\tau$, and FDR target level $q$. The hyperparameter settings for `CORAL` and overall experiment settings are shown in Table 11 and Table 12.

**Experiments for POPE Evaluations** POPE is a flexible approach to evaluating hallucinations in LVLMs, which formulates a binary classification task by prompting LVLMs with questions such as "Is there a keyboard in this image?" to answer "yes" or "no". Following [24], the POPE benchmark aggregates data from three distinct sources: MSCOCO [14], A-OKVQA [15], and GQA [21]. It involves 500 images from each dataset under each sampling setting and formulates 6 questions per

Table 11: CORAL Hyperparameter Settings. The settings are fixed depending on the question-answer tasks.

| Parameters | Value |
|---|---|
| Data Splitting Factor $\tau$ | 0.1 |
| FDR Thresholding $q$ (false positive level in an image) | 0.1 |

Table 12: Batch size settings for LVLM generation across different models. Unless otherwise noted, the batch size is fixed for each model throughout all experiments. To improve evaluation efficiency, we adopt batched generation. When the LVLM does not explicitly specify a padding strategy for inference, we apply left padding to avoid potential negative effects of batched generation.

| Model | Batch Size |
|---|---|
| LLaVA-v1.5 | 4 |
| Qwen-VL | 16 |
| InstructBLIP | 16 |
| LLaVA-OneVision-7B | 4 |
| Qwen2.5-VL-7B | 16 |
| InternVL3-8B | 8 |

image, culminating in a total of 27,000 query answer pairs from the development sets of these datasets. We reported the results on the MSCOCO dataset in Table 18, results on the A-OKVQA dataset in Table 19, and results on the A-OKVQA dataset in Table 20.

**Experiments for CHAIR Evaluations** *Caption Hallucination Assessment with Image Relevance* (CHAIR) [4] quantifies object hallucinations in image captions by comparing generated objects to ground-truth ones. We adopt the same prompt "Generate a short caption of the image." as utilized by Li et al. (2023). The maximum token length is 64, and the sampling approach with random seed of 242. For the calculation of CHAIR metrics, we referenced the 80 object categories annotated in the MSCOCO dataset, following [4].

$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}, \quad \text{CHAIR}_S = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}$$

Besides, we employed the synonym list from [59] to align synonymous words in the generated text with MSCOCO object categories. And we reported the result in Table 13. Following previous work [22], we randomly select 500 images from MSCOCO [14] and use $\text{CHAIR}_I$ and $\text{CHAIR}_S$. Additionally, we incorporate the overall power (Eq. (2)) to evaluate whether the descriptions accurately include the necessary visual content from the image. CORAL consistently reduces both sentence-level and instance-level hallucination rates while maintaining strong power across all models.

Table 13: Evaluation with CHAIR score across multiple LVLM architectures. Lower $C_S$ and $C_I$ indicate fewer hallucinated objects, while higher power indicates better retention of visually grounded objects. **Bold** indicates the best result.

| Method | LLaVA-v1.5 | | | Qwen-VL | | | InstructBLIP | | | LLaVA-OneVision-7B | | | Qwen2.5-VL-7B | | | InternVL3-8B | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $C_S \downarrow$ | $C_I \downarrow$ | Power↑ | $C_S \downarrow$ | $C_I \downarrow$ | Power↑ | $C_S \downarrow$ | $C_I \downarrow$ | Power↑ | $C_S \downarrow$ | $C_I \downarrow$ | Power↑ | $C_S \downarrow$ | $C_I \downarrow$ | Power↑ | $C_S \downarrow$ | $C_I \downarrow$ | Power↑ |
| Regular | 9.2 | 5.1 | 94.8 | 9.5 | 19.2 | 80.7 | 8.7 | 4.9 | 95.2 | 8.8 | 4.6 | 95.4 | 8.2 | 18.1 | 81.9 | 5.0 | 3.2 | 96.8 |
| VCD (2024) | 7.8 | 4.5 | 95.3 | 7.4 | 18.5 | 81.3 | 7.3 | 4.1 | 95.9 | 7.3 | 4.1 | 95.9 | 6.8 | 17.4 | 82.6 | 2.4 | 1.5 | 98.5 |
| MARINE (2025) | 6.9 | 3.8 | 96.1 | 6.3 | 14.5 | 84.7 | 6.2 | 3.0 | 97.0 | 6.2 | 3.0 | 97.0 | 5.9 | 13.8 | 86.2 | 2.2 | **1.3** | **98.7** |
| AGLA (2025) | 7.5 | 4.2 | 95.8 | 6.1 | 12.4 | 87.2 | 7.0 | 3.8 | 96.2 | 7.0 | 3.8 | 96.2 | 5.6 | 11.2 | 88.8 | 2.3 | 1.6 | 98.4 |
| CORAL (Ours) | **5.8** | **3.1** | **96.9** | **4.5** | **11.0** | **88.9** | **5.0** | **2.6** | **97.4** | **5.0** | **2.6** | **97.4** | **3.8** | **10.8** | **89.2** | **1.8** | **1.3** | **98.7** |

**Experiments for MME Evaluations** Similar to the POPE dataset, the MME dataset [54] contains only two types of answers (i.e., Yes or No). For hallucination-related tasks, MME likewise relies on *Yes-or-No* judgments, yielding instance-level correctness outcomes that can be aggregated for performance comparison. Following the setting in their original paper, we use the sum of accuracy and accuracy+ as the final score, where accuracy is calculated based on each question, and accuracy+

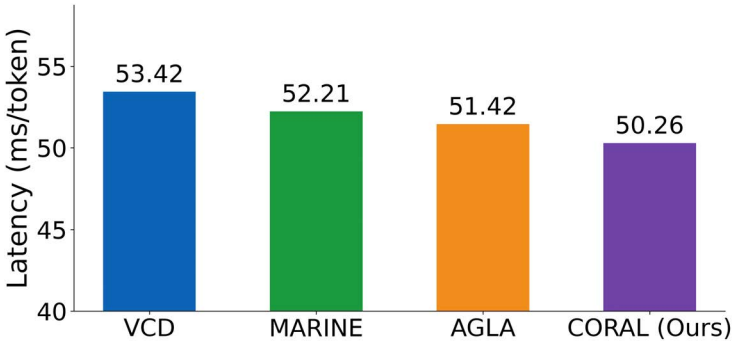



Figure 8: Inference latency comparisons. All measurements are conducted on a single NVIDIA RTX 6000 GPU with batch size 1, using identical input images and prompts.

is calculated based on each image where both of the two questions need to be answered correctly. So accuracy+ is a stricter measurement that can better reflect the comprehensive understanding degree of the model. We reported our results in Table 14.

Table 14: Evaluation with MME score across multiple LVLM architectures on MSCOCO with 3000 random replications. We report the mean ± standard deviation. **Bold** indicates the best result and underline indicates the second-best result.

| Model | Method | Attribute | | | Relation | Total ↑ |
|---|---|---|---|---|---|---|
| | | Existence ↑ | Count ↑ | Color ↑ | Position ↑ | |
| LLaVA-v1.5 | Regular | 175.67 (± 7.51) | 124.67 (± 19.59) | 151.00 (± 10.45) | 114.00 (± 9.32) | 565.33 (± 33.92) |
| | VCD (2024) | 184.66 (± 6.81) | 138.33 (± 15.68) | 153.00 (± 7.58) | 128.67 (± 7.21) | 604.66 (± 18.76) |
| | MARINE (2025) | 190.53 (± 7.26) | 154.43 (± 16.01) | 166.34 (± 6.96) | 130.28 (± 8.01) | 641.58 (± 17.47) |
| | AGLA (2025) | **195.00 (± 7.32)** | 153.89 (± 16.32) | **167.67 (± 6.42)** | 129.44 (± 7.81) | 646.00 (± 15.96) |
| | CORAL (Ours) | 194.43 (± 8.38) | **157.41 (± 15.11)** | 167.34 (± 7.33) | **131.19 (± 7.58)** | **650.37 (± 18.12)** |
| Qwen-VL | Regular | 155.00 (± 3.54) | 127.67 (± 13.36) | 173.00 (± 9.75) | 131.67 (± 7.73) | 587.33 (± 31.06) |
| | VCD (2024) | 156.00 (± 6.25) | 131.00 (± 6.19) | 181.67 (± 5.14) | 128.00 (± 3.61) | 596.67 (± 11.61) |
| | MARINE (2025) | 164.20 (± 6.73) | **136.55 (± 6.24)** | 185.79 (± 6.21) | 132.37 (± 7.76) | 618.91 (± 11.88) |
| | AGLA (2025) | 165.78 (± 5.28) | 134.18 (± 7.14) | 187.12 (± 5.33) | 133.17 (± 7.77) | 620.25 (± 13.42) |
| | CORAL (Ours) | **166.00 (± 6.58)** | 135.21 (± 7.09) | **188.22 (± 7.76)** | **134.00 (± 5.49)** | **623.43 (± 12.67)** |
| InstructBLIP | Regular | 141.00 (± 13.97) | **75.33 (± 14.16)** | 97.33 (± 16.94) | **66.67 (± 3.91)** | 380.33 (± 40.20) |
| | VCD (2024) | 170.00 (± 11.55) | 61.67 (± 8.47) | 114.44 (± 11.27) | 57.22 (± 6.73) | 403.33 (± 13.36) |
| | MARINE (2025) | **189.43 (± 11.88)** | 63.77 (± 7.80) | 118.67 (± 12.15) | 66.00 (± 7.25) | **437.87 (± 13.42)** |
| | AGLA (2025) | 180.00 (± 12.11) | 63.33 (± 7.39) | **119.44 (± 12.11)** | 65.56 (± 8.48) | 428.33 (± 12.81) |
| | CORAL (Ours) | 182.03 (± 12.08) | 64.17 (± 8.53) | 118.25 (± 12.76) | 66.43 (± 7.83) | 431.18 (± 12.30) |

**Experiments for Latency Analysis** Experiment setting for latency analysis. We compared our method with existing baselines in terms of the trade-off between inference cost and the effectiveness of reducing object hallucinations, as shown in Figure 8. For decoding methods such as VCD, AGLA, MARINE and our method, we measured the latency of LLaVA-v1.5 generating captions directly. We prompted the models with "Generate a short caption of the image." on 500 MSCOCO images with a batch size of 1 and a maximum token length of 64, without any stopping criteria, using a single RXT 6000 GPU. Then latency was calculated as the ratio of the number of output tokens and encoding and generation time.

### B.8 Additional Experiment Results on FDR and Power

Additionally, we report additional experimental results on false discovery rate (FDR) and power under a fixed target level $q = 0.1$. Throughout all experiments, we apply the same setting across three datasets. The target level $q = 0.1$ corresponds to controlling the proportion of false positives per image, such that the expected fraction of falsely retained hallucinated objects does not exceed $0.1$ for each image. This setting is used consistently for all datasets to ensure a fair and comparable evaluation.

We reported the result for MSCOCO [14] in Table 15. Additional results on A-OKVQA [15] and GQA [21] under this setting are provided in Table 16 and Table 17.

Table 15: Evaluation of overall FDR control and overall power across multiple LVLM architectures on MSCOCO with 3000 random replications. Overall FDR and power denote averages of per-image FDR and power across images. **Bold** indicates the best result and underline indicates the second-best result in each column.

| Method | LLaVA-v1.5 | | Qwen-VL | | InstructBLIP | | Average | |
|---|---|---|---|---|---|---|---|---|
| | Overall FDR↓ | Overall Power↑ | Overall FDR↓ | Overall Power↑ | Overall FDR↓ | Overall Power↑ | Overall FDR↓ | Overall Power↑ |
| *Random* | | | | | | | | |
| Regular | 0.0985 (± 0.0144) | 77.32 (± 1.41) | 0.0913 (± 0.0122) | 76.08 (± 1.23) | 0.0957 (± 0.0102) | 74.35 (± 1.63) | 0.0952 (± 0.0122) | 75.92 (± 1.42) |
| VCD (2024) | 0.0965 (± 0.0200) | 82.50 (± 1.13) | 0.0898 (± 0.0111) | 78.10 (± 1.12) | 0.0941 (± 0.0119) | 76.80 (± 1.22) | 0.0935 (± 0.0143) | 79.13 (± 1.16) |
| MARINE (2025) | 0.0935 (± 0.0112) | 89.20 (± 1.07) | 0.0852 (± 0.0119) | 79.05 (± 1.13) | 0.0944 (± 0.0114) | 77.14 (± 1.35) | 0.0910 (± 0.0115) | 81.80 (± 1.18) |
| AGLA (2025) | 0.0918 (± 0.0107) | 92.80 (± 1.15) | 0.0818 (± 0.0132) | 82.67 (± 1.10) | 0.0921 (± 0.0122) | 78.92 (± 1.21) | 0.0886 (± 0.0120) | 84.80 (± 1.15) |
| CORAL(Ours) | **0.0910 (± 0.0101)** | **97.62 (± 1.02)** | **0.0755 (± 0.0113)** | **84.90 (± 1.09)** | **0.0836 (± 0.0100)** | **84.00 (± 1.34)** | **0.0834 (± 0.0105)** | **88.84 (± 1.15)** |
| *Popular* | | | | | | | | |
| Regular | 0.0971 (± 0.0205) | 80.11 (± 1.28) | 0.0903 (± 0.0111) | 77.01 (± 1.64) | 0.0937 (± 0.0106) | 50.80 (± 1.69) | 0.0937 (± 0.0141) | 69.31 (± 1.54) |
| VCD (2024) | 0.0979 (± 0.0110) | 87.20 (± 1.18) | 0.0904 (± 0.0109) | 76.10 (± 1.31) | 0.0917 (± 0.0108) | 54.50 (± 1.29) | 0.0933 (± 0.0109) | 72.60 (± 1.26) |
| MARINE (2025) | 0.0963 (± 0.0108) | 90.21 (± 1.24) | 0.0831 (± 0.0118) | 79.38 (± 1.30) | 0.0835 (± 0.0138) | 59.20 (± 1.49) | 0.0876 (± 0.0121) | 76.26 (± 1.34) |
| AGLA (2025) | 0.0930 (± 0.0104) | 92.49 (± 1.21) | 0.0791 (± 0.0119) | 83.97 (± 1.28) | 0.0898 (± 0.0170) | 61.93 (± 1.35) | 0.0873 (± 0.0131) | 79.46 (± 1.28) |
| CORAL(Ours) | **0.0886 (± 0.0112)** | **96.49 (± 0.98)** | **0.0779 (± 0.0201)** | **85.10 (± 1.18)** | **0.0809 (± 0.0113)** | **64.30 (± 1.31)** | **0.0825 (± 0.0142)** | **81.96 (± 1.16)** |
| *Adversarial* | | | | | | | | |
| Regular | 0.0976 (± 0.0221) | 50.66 (± 1.42) | 0.0955 (± 0.0189) | 54.37 (± 1.13) | 0.0956 (± 0.0107) | 53.97 (± 1.85) | 0.0962 (± 0.0169) | 53.00 (± 1.47) |
| VCD (2024) | 0.0988 (± 0.0121) | 48.70 (± 1.35) | 0.0956 (± 0.0099) | 58.40 (± 1.23) | 0.0993 (± 0.0157) | 53.20 (± 1.33) | 0.0979 (± 0.0126) | 53.43 (± 1.30) |
| MARINE (2025) | 0.0915 (± 0.0111) | 60.24 (± 1.09) | 0.0915 (± 0.0119) | 79.96 (± 1.43) | 0.0999 (± 0.0134) | 80.20 (± 1.25) | 0.0943 (± 0.0121) | 73.47 (± 1.26) |
| AGLA (2025) | 0.0835 (± 0.0122) | 70.87 (± 1.19) | 0.0812 (± 0.0109) | 81.01 (± 1.23) | 0.0953 (± 0.0200) | 84.21 (± 1.21) | 0.0867 (± 0.0144) | 78.70 (± 1.21) |
| CORAL(Ours) | **0.0826 (± 0.0115)** | **71.46 (± 1.23)** | **0.0803 (± 0.0113)** | **84.50 (± 1.05)** | **0.0942 (± 0.0116)** | **88.80 (± 1.19)** | **0.0857 (± 0.0115)** | **81.59 (± 1.16)** |

# C Experiments for Ablation Study

## C.1 Effect of FDR Threshold

We also examine the role of the false discovery rate (FDR) control procedure on the MSCOCO dataset across three models. In this ablation study, we remove the adaptive threshold selection based on the estimated FDR and instead apply fixed thresholds chosen on a validation set or select a fixed proportion of tokens. This variant preserves the mirror statistic but disables explicit error control. The results show that without FDR-based thresholding, the empirical FDR varies substantially across experimental settings on MSCOCO, often exceeding the target level. In contrast, the full method consistently maintains empirical FDR close to the desired rate while achieving comparable or better hallucination reduction. These findings indicate that the performance gains of our approach are not solely attributable to the mirror statistic but critically rely on the FDR control procedure, which provides explicit and stable error control. The corresponding results are reported in Figure 9.

## C.2 Calibration of the Perturbation Scale $\tau_v$

As described in the main text, the perturbation scale $\tau_v$ is calibrated using a lightweight sensitivity analysis rather than fine-grained hyperparameter optimization. Specifically, $\tau_v$ is selected based on a coarse sweep over a representative range of values on a small subset disjoint from the reported evaluation images, with the target false discovery rate (FDR) level fixed at $q = 0.1$. The goal of this calibration is to identify a stable operating regime that maintains empirical FDR close to the target level while yielding robust power, rather than to optimize any task-specific metric.

Once selected, the perturbation scale $\tau_v$ is kept fixed across all datasets and experiments for each backbone, and no further tuning is performed on the test benchmarks. Figures 10 and 11 report the corresponding sensitivity analysis. The results show that CORAL exhibits smooth and consistent behavior across a wide range of $\tau_v$ values, indicating that the method is not sensitive to precise tuning of the perturbation scale and mitigating the risk of overfitting to any specific dataset or evaluation protocol.

Table 16: Evaluation on overall FDR control and overall power across multiple LVLM architectures on **A-OKVQA** with 3000 random replications. **Bold** indicates the best result and underline indicates the second-best result in each column.

| Method | LLaVA-v1.5 | | Qwen-VL | | InstructBLIP | | Average | |
|---|---|---|---|---|---|---|---|---|
| | FDR↓ | Power↑ | FDR↓ | Power↑ | FDR↓ | Power↑ | FDR↓ | Power↑ |
| *Random* | | | | | | | | |
| Regular | 0.0957 (± 0.0131) | 77.57 (± 1.28) | 0.0953 (± 0.0133) | 79.53 (± 1.52) | 0.0987 (± 0.0107) | 75.34 (± 1.26) | 0.0966 (± 0.0124) | 77.48 (± 1.35) |
| VCD (2024) | 0.0930 (± 0.0100) | 84.10 (± 1.22) | 0.0892 (± 0.0121) | 82.30 (± 1.13) | 0.0967 (± 0.0112) | 80.40 (± 1.15) | 0.0930 (± 0.0111) | 82.27 (± 1.17) |
| MARINE (2025) | 0.0920 (± 0.0101) | 85.02 (± 1.35) | 0.0835 (± 0.0202) | 86.64 (± 1.05) | 0.0920 (± 0.0100) | 84.93 (± 1.08) | 0.0892 (± 0.0134) | 85.53 (± 1.16) |
| AGLA (2025) | 0.0892 (± 0.0112) | 86.42 (± 1.42) | 0.0832 (± 0.0158) | 87.01 (± 1.11) | 0.0923 (± 0.0123) | 87.43 (± 1.11) | 0.0882 (± 0.0131) | 86.95 (± 1.21) |
| CORAL (Ours) | **0.0781 (± 0.0201)** | **92.40 (± 1.22)** | **0.0745 (± 0.0133)** | **90.85 (± 1.08)** | **0.0831 (± 0.0101)** | **88.95 (± 1.14)** | **0.0786 (± 0.0145)** | **90.73 (± 1.15)** |
| *Popular* | | | | | | | | |
| Regular | 0.0938 (± 0.0142) | 79.34 (± 1.66) | 0.0896 (± 0.0201) | 81.75 (± 1.41) | 0.0959 (± 0.0159) | 72.02 (± 1.94) | 0.0931 (± 0.0167) | 77.70 (± 1.67) |
| VCD (2024) | 0.0933 (± 0.0034) | 83.25 (± 1.05) | 0.0913 (± 0.0101) | 80.90 (± 1.02) | 0.0977 (± 0.0103) | 70.10 (± 1.14) | 0.0941 (± 0.0089) | 78.08 (± 1.07) |
| MARINE (2025) | 0.0902 (± 0.0102) | 84.59 (± 1.09) | 0.0875 (± 0.0214) | 87.22 (± 0.97) | 0.0941 (± 0.0104) | 79.51 (± 1.11) | 0.0906 (± 0.0140) | 83.77 (± 1.06) |
| AGLA (2025) | 0.0874 (± 0.0142) | 90.87 (± 1.04) | 0.0841 (± 0.0125) | 88.91 (± 1.00) | 0.0872 (± 0.0124) | 80.29 (± 1.11) | 0.0862 (± 0.0130) | 86.69 (± 1.05) |
| CORAL (Ours) | **0.0860 (± 0.0122)** | **91.80 (± 1.08)** | **0.0766 (± 0.0223)** | **89.40 (± 0.98)** | **0.0814 (± 0.0152)** | **82.75 (± 1.13)** | **0.0813 (± 0.0166)** | **87.98 (± 1.06)** |
| *Adversarial* | | | | | | | | |
| Regular | 0.0941 (± 0.0155) | 56.34 (± 1.62) | 0.0967 (± 0.0194) | 63.01 (± 1.21) | 0.0978 (± 0.0143) | 61.24 (± 1.44) | 0.0962 (± 0.0164) | 60.20 (± 1.42) |
| VCD (2024) | 0.0991 (± 0.0100) | 55.60 (± 1.19) | 0.0965 (± 0.0121) | 63.40 (± 1.01) | 0.0988 (± 0.0090) | 59.80 (± 1.30) | 0.0981 (± 0.0104) | 59.60 (± 1.17) |
| MARINE (2025) | 0.0924 (± 0.0101) | 56.91 (± 1.22) | 0.0940 (± 0.0132) | 69.92 (± 1.02) | 0.0943 (± 0.0071) | 66.01 (± 1.28) | 0.0936 (± 0.0101) | 64.28 (± 1.17) |
| AGLA (2025) | 0.0913 (± 0.0105) | 63.34 (± 1.20) | 0.0921 (± 0.0100) | 71.91 (± 1.01) | 0.0923 (± 0.0091) | 68.32 (± 1.28) | 0.0919 (± 0.0098) | 67.86 (± 1.16) |
| CORAL (Ours) | **0.0842 (± 0.0210)** | **68.30 (± 1.22)** | **0.0781 (± 0.0125)** | **73.10 (± 1.04)** | **0.0881 (± 0.0142)** | **70.25 (± 1.32)** | **0.0835 (± 0.0159)** | **70.55 (± 1.19)** |

| Method | LLaVA-OneVision-7B | | Qwen2.5-VL-7B | | InternVL3-8B | | Average | |
|---|---|---|---|---|---|---|---|---|
| | FDR↓ | Power↑ | FDR↓ | Power↑ | FDR↓ | Power↑ | FDR↓ | Power↑ |
| *Random* | | | | | | | | |
| Regular | 0.0972 (± 0.0110) | 70.46 (± 1.54) | 0.0965 (± 0.0102) | 71.43 (± 1.29) | 0.0980 (± 0.0118) | 73.28 (± 1.25) | 0.0972 (± 0.0110) | 71.72 (± 1.36) |
| VCD (2024) | 0.0942 (± 0.0115) | 75.37 (± 1.26) | 0.0901 (± 0.0110) | 77.29 (± 1.16) | 0.0921 (± 0.0114) | 76.25 (± 1.25) | 0.0921 (± 0.0113) | 76.30 (± 1.22) |
| MARINE (2025) | 0.0918 (± 0.0116) | 77.01 (± 1.66) | 0.0900 (± 0.0155) | 78.03 (± 1.13) | 0.0891 (± 0.0145) | 80.35 (± 1.56) | 0.0903 (± 0.0139) | 78.46 (± 1.45) |
| AGLA (2025) | 0.0874 (± 0.0203) | 84.22 (± 1.77) | 0.0868 (± 0.0117) | 85.38 (± 1.10) | 0.0877 (± 0.0201) | 83.89 (± 1.26) | 0.0873 (± 0.0174) | 84.50 (± 1.38) |
| CORAL (Ours) | **0.0772 (± 0.0189)** | **89.20 (± 1.65)** | **0.0714 (± 0.0200)** | **92.25 (± 1.26)** | **0.0733 (± 0.0177)** | **90.76 (± 1.41)** | **0.0740 (± 0.0189)** | **90.74 (± 1.44)** |
| *Popular* | | | | | | | | |
| Regular | 0.0981 (± 0.0091) | 69.25 (± 1.26) | 0.0979 (± 0.0100) | 70.67 (± 2.48) | 0.0970 (± 0.0143) | 71.35 (± 1.36) | 0.0977 (± 0.0111) | 70.42 (± 1.70) |
| VCD (2024) | 0.0964 (± 0.0133) | 70.41 (± 1.24) | 0.0955 (± 0.0102) | 71.43 (± 1.35) | 0.0961 (± 0.0100) | 70.22 (± 1.53) | 0.0960 (± 0.0112) | 70.69 (± 1.37) |
| MARINE (2025) | 0.0904 (± 0.0114) | 75.25 (± 1.52) | 0.0914 (± 0.0121) | 75.00 (± 1.31) | 0.0917 (± 0.0105) | 74.99 (± 1.64) | 0.0912 (± 0.0113) | 75.08 (± 1.49) |
| AGLA (2025) | 0.0826 (± 0.0132) | 80.76 (± 1.36) | 0.0814 (± 0.0110) | 82.84 (± 1.51) | 0.0820 (± 0.0130) | 82.01 (± 1.17) | 0.0820 (± 0.0124) | 81.87 (± 1.35) |
| CORAL (Ours) | **0.0774 (± 0.0118)** | **85.36 (± 1.36)** | **0.0751 (± 0.0209)** | **87.26 (± 0.99)** | **0.0757 (± 0.0116)** | **87.01 (± 1.04)** | **0.0761 (± 0.0148)** | **86.54 (± 1.13)** |
| *Adversarial* | | | | | | | | |
| Regular | 0.0973 (± 0.0099) | 70.40 (± 1.35) | 0.0969 (± 0.0104) | 73.01 (± 1.11) | 0.0977 (± 0.0106) | 69.46 (± 1.53) | 0.0973 (± 0.0103) | 70.96 (± 1.33) |
| VCD (2024) | 0.0903 (± 0.0099) | 77.26 (± 1.75) | 0.0915 (± 0.0104) | 76.49 (± 1.98) | 0.0899 (± 0.0100) | 80.01 (± 1.52) | 0.0906 (± 0.0101) | 77.92 (± 1.75) |
| MARINE (2025) | 0.0843 (± 0.0134) | 84.25 (± 1.63) | 0.0850 (± 0.0127) | 83.78 (± 1.05) | 0.0837 (± 0.0114) | 85.73 (± 1.11) | 0.0843 (± 0.0125) | 84.59 (± 1.26) |
| AGLA (2025) | 0.0776 (± 0.0143) | 88.92 (± 1.65) | 0.0768 (± 0.0107) | 89.01 (± 1.84) | 0.0763 (± 0.0136) | 89.57 (± 1.26) | 0.0769 (± 0.0129) | 89.17 (± 1.58) |
| CORAL (Ours) | **0.0687 (± 0.0221)** | **91.47 (± 2.63)** | **0.0680 (± 0.0178)** | **92.04 (± 1.87)** | **0.0678 (± 0.0129)** | **92.68 (± 1.99)** | **0.0682 (± 0.0176)** | **92.06 (± 2.16)** |

# D Boundary and Failure Cases.

Figure 12 presents representative failure cases of CORAL, including false negatives for present objects, confusion between visually similar categories such as a streetlamp and a traffic light, difficulties recognizing small or partially occluded objects such as a knife, and confusion between related activities. These examples illustrate that errors can remain when visual evidence is weak or semantic distinctions are fine-grained.

One possible explanation is that perturbation responses for grounded and non-grounded concepts are insufficiently separated, limiting the discriminative ability of the mirror statistic. However, the qualitative examples alone do not establish this mechanism or imply that the corresponding statistics lie near the selection threshold. These cases complement the quantitative evaluation by highlighting the limitations of CORAL. The hallucination mitigation does not eliminate all unsupported predictions and may also reject genuinely present objects.

Table 17: Evaluation on overall FDR control and overall power across multiple LVLM architectures on **GQA** with 3000 random replications. **Bold** indicates the best result and underline indicates the second-best result in each column.

| Method | LLaVA-v1.5 | | Qwen-VL | | InstructBLIP | | Average | |
|---|---|---|---|---|---|---|---|---|
| | FDR↓ | Power↑ | FDR↓ | Power↑ | FDR↓ | Power↑ | FDR↓ | Power↑ |
| *Random* | | | | | | | | |
| Regular | 0.0933 (± 0.0132) | 76.55 (± 1.32) | 0.0914 (± 0.0188) | 81.66 (± 1.01) | 0.0889 (± 0.0188) | 83.88 (± 1.52) | 0.0912 (± 0.0169) | 80.69 (± 1.28) |
| VCD (2024) | 0.0924 (± 0.0098) | 80.92 (± 0.68) | 0.0924 (± 0.0090) | 80.24 (± 1.00) | 0.0872 (± 0.0145) | 85.47 (± 0.62) | 0.0907 (± 0.0111) | 82.21 (± 0.77) |
| MARINE (2025) | 0.0894 (± 0.0128) | 86.92 (± 0.40) | 0.0902 (± 0.0101) | 84.22 (± 0.93) | 0.0837 (± 0.0144) | 87.39 (± 0.55) | 0.0878 (± 0.0124) | 86.18 (± 0.63) |
| AGLA (2025) | 0.0802 (± 0.0164) | 90.42 (± 1.02) | 0.0872 (± 0.0100) | 85.11 (± 1.09) | 0.0855 (± 0.0148) | 89.43 (± 0.41) | 0.0843 (± 0.0137) | 88.99 (± 0.84) |
| CORAL (Ours) | **0.0719 (± 0.0223)** | **92.05 (± 0.32)** | **0.0825 (± 0.0121)** | **87.15 (± 0.44)** | **0.0813 (± 0.0170)** | **90.15 (± 0.67)** | **0.0786 (± 0.0171)** | **89.78 (± 0.48)** |
| *Popular* | | | | | | | | |
| Regular | 0.0911 (± 0.0122) | 80.62 (± 1.34) | 0.0900 (± 0.0133) | 79.98 (± 1.01) | 0.0977 (± 0.0142) | 80.01 (± 1.56) | 0.0929 (± 0.0132) | 80.20 (± 1.30) |
| VCD (2024) | 0.0901 (± 0.0100) | 86.42 (± 0.69) | 0.0861 (± 0.0122) | 84.92 (± 0.93) | 0.0982 (± 0.0022) | 79.23 (± 1.00) | 0.0915 (± 0.0081) | 83.52 (± 0.87) |
| MARINE (2025) | 0.0892 (± 0.0110) | 84.65 (± 1.11) | 0.0853 (± 0.0114) | 86.53 (± 0.88) | 0.0913 (± 0.0015) | 80.98 (± 1.14) | 0.0886 (± 0.0080) | 83.39 (± 1.04) |
| AGLA (2025) | 0.0852 (± 0.0139) | 89.07 (± 1.08) | 0.0793 (± 0.0142) | 87.34 (± 0.98) | 0.0892 (± 0.0109) | 81.32 (± 1.02) | 0.0846 (± 0.0130) | 85.91 (± 1.03) |
| CORAL (Ours) | **0.0816 (± 0.0142)** | **93.53 (± 0.25)** | **0.0714 (± 0.0133)** | **89.94 (± 1.21)** | **0.0823 (± 0.0110)** | **83.45 (± 1.05)** | **0.0784 (± 0.0128)** | **88.97 (± 0.84)** |
| *Adversarial* | | | | | | | | |
| Regular | 0.0945 (± 0.0146) | 63.66 (± 1.47) | 0.0980 (± 0.0149) | 55.89 (± 1.37) | 0.0983 (± 0.0120) | 55.16 (± 1.77) | 0.0969 (± 0.0138) | 58.24 (± 1.54) |
| VCD (2024) | 0.0934 (± 0.0115) | 57.09 (± 0.93) | 0.0973 (± 0.0132) | 58.93 (± 1.24) | 0.0987 (± 0.0010) | 66.92 (± 1.05) | 0.0965 (± 0.0086) | 60.98 (± 1.07) |
| MARINE (2025) | 0.0924 (± 0.0120) | 56.91 (± 0.86) | 0.0942 (± 0.0100) | 61.84 (± 1.28) | 0.0968 (± 0.0009) | 68.92 (± 1.03) | 0.0945 (± 0.0076) | 62.56 (± 1.06) |
| AGLA (2025) | 0.0953 (± 0.0111) | 62.42 (± 0.88) | 0.0921 (± 0.0112) | 65.21 (± 1.25) | 0.0988 (± 0.0100) | 70.87 (± 0.92) | 0.0954 (± 0.0108) | 66.17 (± 1.02) |
| CORAL (Ours) | **0.0892 (± 0.0101)** | **64.30 (± 1.00)** | **0.0881 (± 0.0200)** | **74.03 (± 1.32)** | **0.0941 (± 0.0044)** | **76.37 (± 1.00)** | **0.0905 (± 0.0115)** | **71.57 (± 1.11)** |

| Method | LLaVA-OneVision-7B | | Qwen2.5-VL-7B | | InternVL3-8B | | Average | |
|---|---|---|---|---|---|---|---|---|
| | FDR↓ | Power↑ | FDR↓ | Power↑ | FDR↓ | Power↑ | FDR↓ | Power↑ |
| *Random* | | | | | | | | |
| Regular | 0.0972 (± 0.0110) | 70.46 (± 1.54) | 0.0965 (± 0.0102) | 71.43 (± 1.29) | 0.0980 (± 0.0118) | 73.28 (± 1.25) | 0.0972 (± 0.0110) | 71.72 (± 1.36) |
| VCD (2024) | 0.0913 (± 0.0115) | 75.11 (± 1.62) | 0.0920 (± 0.0110) | 74.48 (± 1.15) | 0.0916 (± 0.0114) | 75.00 (± 1.01) | 0.0916 (± 0.0113) | 74.86 (± 1.26) |
| MARINE (2025) | 0.0876 (± 0.0136) | 77.89 (± 1.25) | 0.0869 (± 0.0133) | 78.03 (± 1.65) | 0.0866 (± 0.0151) | 78.58 (± 1.85) | 0.0870 (± 0.0140) | 78.17 (± 1.58) |
| AGLA (2025) | 0.0801 (± 0.0197) | 82.84 (± 1.54) | 0.0796 (± 0.0170) | 83.15 (± 2.07) | 0.0787 (± 0.0200) | 83.69 (± 1.96) | 0.0795 (± 0.0189) | 83.23 (± 1.86) |
| CORAL (Ours) | **0.0727 (± 0.0201)** | **89.28 (± 1.63)** | **0.0719 (± 0.0197)** | **90.79 (± 2.15)** | **0.0726 (± 0.0163)** | **90.76 (± 2.05)** | **0.0724 (± 0.0187)** | **90.28 (± 1.94)** |
| *Popular* | | | | | | | | |
| Regular | 0.0980 (± 0.0015) | 69.27 (± 1.25) | 0.0979 (± 0.0100) | 69.93 (± 1.73) | 0.0982 (± 0.0143) | 68.79 (± 1.55) | 0.0980 (± 0.0086) | 69.33 (± 1.51) |
| VCD (2024) | 0.0925 (± 0.0135) | 74.59 (± 1.48) | 0.0916 (± 0.0112) | 75.21 (± 1.62) | 0.0920 (± 0.0128) | 75.01 (± 1.72) | 0.0920 (± 0.0125) | 74.94 (± 1.61) |
| MARINE (2025) | 0.0872 (± 0.0143) | 78.29 (± 1.62) | 0.0863 (± 0.0138) | 79.15 (± 1.74) | 0.0860 (± 0.0125) | 79.59 (± 1.16) | 0.0865 (± 0.0135) | 79.01 (± 1.51) |
| AGLA (2025) | 0.0793 (± 0.0143) | 80.97 (± 1.64) | 0.0801 (± 0.0122) | 79.92 (± 1.01) | 0.0789 (± 0.0140) | 81.35 (± 1.26) | 0.0794 (± 0.0135) | 80.75 (± 1.30) |
| CORAL (Ours) | **0.0701 (± 0.0221)** | **86.92 (± 1.17)** | **0.0694 (± 0.0204)** | **86.00 (± 1.31)** | **0.0700 (± 0.0198)** | **86.90 (± 1.17)** | **0.0698 (± 0.0208)** | **86.61 (± 1.22)** |
| *Adversarial* | | | | | | | | |
| Regular | 0.0971 (± 0.0018) | 70.26 (± 1.26) | 0.0976 (± 0.0015) | 70.01 (± 1.37) | 0.0970 (± 0.0020) | 70.96 (± 1.29) | 0.0972 (± 0.0018) | 70.41 (± 1.31) |
| VCD (2024) | 0.0901 (± 0.0103) | 73.26 (± 1.37) | 0.0900 (± 0.0100) | 73.79 (± 2.00) | 0.0887 (± 0.0101) | 74.19 (± 1.22) | 0.0896 (± 0.0101) | 73.75 (± 1.53) |
| MARINE (2025) | 0.0825 (± 0.0113) | 76.98 (± 1.26) | 0.0816 (± 0.0117) | 77.21 (± 1.00) | 0.0820 (± 0.0110) | 77.07 (± 1.63) | 0.0820 (± 0.0113) | 77.09 (± 1.30) |
| AGLA (2025) | 0.0722 (± 0.0135) | 81.54 (± 1.76) | 0.0735 (± 0.0154) | 80.26 (± 1.27) | 0.0727 (± 0.0137) | 81.26 (± 1.75) | 0.0728 (± 0.0142) | 81.02 (± 1.59) |
| CORAL (Ours) | **0.0647 (± 0.0215)** | **89.26 (± 1.86)** | **0.0638 (± 0.0189)** | **90.05 (± 2.01)** | **0.0641 (± 0.0121)** | **89.78 (± 1.48)** | **0.0642 (± 0.0175)** | **89.70 (± 1.78)** |

Table 18: Evaluation with POPE score across multiple LVLM architectures on **MSCOCO** dataset comparing our method with several baselines. **Bold** indicates the best result and underline indicates the second-best result in each column.

| Decoding | LLaVA-v1.5 | | | | Qwen-VL | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy↑ | Precision↑ | Recall↑ | F1 Score↑ | Accuracy↑ | Precision↑ | Recall↑ | F1 Score↑ |
| *Random* | | | | | | | | |
| Regular | 83.29 (± 0.35) | 92.13 (± 0.54) | 72.80 (± 0.57) | 81.33 (± 0.41) | 84.73 (± 0.36) | 95.61 (± 0.45) | 72.81 (± 0.38) | 82.67 (± 0.41) |
| VCD (2024) | 87.73 (± 0.40) | 91.42 (± 0.55) | 83.28 (± 0.42) | 87.16 (± 0.41) | 88.63 (± 0.10) | 94.64 (± 0.25) | 81.91 (± 0.19) | 87.81 (± 0.11) |
| MARINE (2025) | 85.01 (± 0.24) | 88.27 (± 0.83) | 80.73 (± 0.12) | 84.33 (± 0.31) | 82.07 (± 0.14) | 89.27 (± 0.13) | **89.33 (± 0.24)** | 85.83 (± 0.87) |
| AGLA (2025) | 88.54 (± 0.64) | **94.41 (± 0.50)** | 82.08 (± 0.47) | 87.71 (± 0.51) | 84.60 (± 0.76) | **98.23 (± 0.29)** | 70.47 (± 0.41) | 82.07 (± 0.11) |
| CORAL (Ours) | **90.33 (± 0.54)** | 90.77 (± 0.76) | **89.80 (± 0.78)** | **90.28 (± 0.57)** | **91.33 (± 0.51)** | 96.20 (± 0.52) | 86.07 (± 0.90) | **90.85 (± 0.57)** |
| *Popular* | | | | | | | | |
| Regular | 81.88 (± 0.48) | **88.93 (± 0.60)** | 72.80 (± 0.57) | 80.06 (± 0.05) | 84.13 (± 0.18) | 94.31 (± 0.43) | 72.64 (± 0.45) | 82.06 (± 0.23) |
| VCD (2024) | 85.38 (± 0.38) | 86.92 (± 0.53) | 83.28 (± 0.42) | 85.06 (± 0.37) | 87.12 (± 0.07) | 91.49 (± 0.10) | 81.85 (± 0.19) | 86.40 (± 0.09) |
| MARINE (2025) | 71.32 (± 0.45) | 65.82 (± 0.11) | 88.79 (± 0.21) | 75.58 (± 0.12) | 87.92 (± 0.54) | 88.25 (± 0.93) | 80.75 (± 0.52) | 84.21 (± 0.39) |
| AGLA (2025) | 85.14 (± 0.87) | 87.88 (± 0.83) | 82.08 (± 0.83) | 84.68 (± 0.56) | 84.40 (± 0.21) | **97.16 (± 0.32)** | 70.87 (± 0.86) | 81.95 (± 0.32) |
| CORAL (Ours) | **87.20 (± 0.58)** | 85.36 (± 0.72) | **89.80 (± 0.71)** | **87.52 (± 0.62)** | **89.03 (± 0.57)** | 91.50 (± 0.51) | **86.07 (± 0.63)** | **88.70±0.41)** |
| *Adversarial* | | | | | | | | |
| Regular | 78.96 (± 0.52) | **83.06 (± 0.58)** | 72.75 (± 0.59) | 77.57 (± 0.57) | 82.26 (± 0.30) | 89.97 (± 0.33) | 72.61 (± 0.50) | 80.37 (± 0.37) |
| VCD (2024) | 80.88 (± 0.33) | 79.45 (± 0.29) | 83.29 (± 0.43) | 81.33 (± 0.34) | 84.26 (± 0.39) | 85.84 (± 0.45) | 82.05 (± 0.39) | 83.90 (± 0.39) |
| MARINE (2025) | 66.89 (± 0.14) | 61.73 (± 0.82) | 89.10 (± 0.10) | 79.24 (± 0.24) | 83.28 (± 0.21) | 85.37 (± 0.29) | 83.73 (± 0.23) | 84.21 (± 0.23) |
| AGLA (2025) | 81.13 (± 0.16) | 81.20 (± 0.78) | 82.10 (± 0.84) | 81.36 (± 0.70) | 82.70 (± 0.28) | **93.22 (± 0.41)** | 70.53 (± 0.72) | 80.30 (± 0.49) |
| CORAL (Ours) | **81.47 (± 0.74)** | 77.00 (± 1.03) | **89.73 (± 0.88)** | **82.88 (± 0.75)** | **86.53 (± 0.62)** | 87.03 (± 0.61) | **85.87 (± 0.63)** | **86.44 (± 0.44)** |

| Decoding | InstructBLIP | | | | LLaVA-OneVision-7B | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy↑ | Precision↑ | Recall↑ | F1 Score↑ | Accuracy↑ | Precision↑ | Recall↑ | F1 Score↑ |
| *Random* | | | | | | | | |
| Regular | 80.71 (± 0.73) | 81.67 (± 0.67) | 79.19 (± 1.14) | 80.41 (± 0.80) | 85.87 (± 1.77) | 83.41 (± 2.13) | 88.08 (± 1.47) | 85.72 (± 1.66) |
| VCD (2024) | 84.53 (± 0.38) | 88.55 (± 0.54) | 79.32 (± 0.44) | 83.68 (± 0.40) | 86.33 (± 1.23) | 89.01 (± 1.42) | 87.33 (± 1.78) | 88.86 (± 1.43) |
| MARINE (2025) | 86.72 (± 0.11) | 87.97 (± 0.53) | 80.03 (± 0.19) | 87.33 (± 0.91) | 87.21 (± 1.35) | 90.33 (± 1.79) | 88.11 (± 1.93) | 89.09 (± 1.09) |
| AGLA (2025) | 87.30 (± 0.32) | 88.83 (± 0.41) | 82.08 (± 0.47) | 87.71 (± 0.51) | 88.15 (± 1.65) | 93.26 (± 1.88) | **90.08 (± 1.47)** | 89.97 (± 1.21) |
| CORAL (Ours) | **90.13 (± 0.54)** | **92.63 (± 0.47)** | **87.20 (± 0.61)** | **89.84 (± 0.38)** | **92.17 (± 0.89)** | **98.25 (± 1.46)** | 89.87 (± 1.61) | **91.64 (± 1.83)** |
| *Popular* | | | | | | | | |
| Regular | 78.22 (± 0.84) | 77.87 (± 1.03) | 78.85 (± 0.52) | 78.36 (± 0.76) | 82.72 (± 1.99) | 79.81 (± 2.31) | 86.87 (± 1.70) | 83.16 (± 1.85) |
| VCD (2024) | 81.47 (± 0.42) | **82.89 (± 0.64)** | 79.32 (± 0.44) | 81.07 (± 0.39) | 84.24 (± 1.46) | 86.16 (± 1.69) | 86.53 (± 1.78) | 86.16 (± 1.30) |
| MARINE (2025) | 81.74 (± 0.88) | 80.27 (± 0.35) | 81.73 (± 0.92) | 79.39 (± 0.12) | 86.81 (± 1.21) | 89.15 (± 1.55) | 87.20 (± 1.91) | 88.19 (± 2.09) |
| AGLA (2025) | 81.86 (± 0.14) | 80.17 (± 0.88) | 85.68 (± 0.64) | 82.58 (± 0.52) | 89.28 (± 1.11) | 91.63 (± 2.09) | 89.08 (± 1.31) | 89.34 (± 1.70) |
| CORAL (Ours) | **83.43 (± 0.68)** | 81.09 (± 0.70) | **87.20 (± 0.61)** | **84.03 (± 0.47)** | **90.37 (± 1.89)** | **94.36 (± 1.46)** | **89.87 (± 2.04)** | **89.91 (± 1.33)** |
| *Adversarial* | | | | | | | | |
| Regular | 75.84 (± 0.45) | 74.30 (± 0.63) | 79.03 (± 0.68) | 76.59 (± 0.40) | 80.28 (± 2.21) | 77.14 (± 2.46) | 84.48 (± 1.93) | 81.02 (± 2.07) |
| VCD (2024) | 79.56 (± 0.41) | **79.67 (± 0.59)** | 79.39 (± 0.50) | 79.52 (± 0.38) | 81.21 (± 1.89) | 80.25 (± 1.71) | 84.61 (± 2.15) | 82.96 (± 1.82) |
| MARINE (2025) | 75.23 (± 0.24) | 78.27 (± 0.52) | 78.72 (± 0.68) | 80.13 (± 0.43) | 85.26 (± 2.11) | 84.93 (± 1.97) | 85.42 (± 2.14) | 85.29 (± 2.09) |
| AGLA (2025) | 77.29 (± 0.13) | 74.09 (± 0.14) | 85.67 (± 0.21) | 79.16 (± 0.61) | 86.44 (± 1.61) | 89.26 (± 1.51) | **90.01 (± 1.11)** | **88.14 (± 2.01)** |
| CORAL (Ours) | **80.63 (± 0.72)** | 77.14 (± 0.77) | **87.07 (± 0.61)** | **81.80 (± 0.50)** | **88.27 (± 1.81)** | **90.14 (± 1.21)** | 86.93 (± 1.87) | 87.99 (± 2.03) |

| Decoding | Qwen2.5-VL-7B | | | | InternVL3-8B | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy↑ | Precision↑ | Recall↑ | F1 Score↑ | Accuracy↑ | Precision↑ | Recall↑ | F1 Score↑ |
| *Random* | | | | | | | | |
| Regular | 86.20 (± 1.30) | 88.68 (± 1.54) | 87.62 (± 1.24) | 86.98 (± 1.28) | 87.15 (± 1.88) | 86.87 (± 1.67) | 87.24 (± 1.76) | 86.26 (± 1.80) |
| VCD (2024) | 87.35 (± 1.83) | 89.55 (± 1.45) | 89.23 (± 1.35) | 90.06 (± 2.03) | 88.15 (± 1.53) | 89.01 (± 1.17) | 88.24 (± 1.44) | 88.05 (± 1.45) |
| MARINE (2025) | 89.72 (± 1.11) | 91.79 (± 1.35) | 91.31 (± 1.91) | 90.33 (± 2.09) | 89.23 (± 1.35) | 90.24 (± 1.66) | 88.52 (± 1.15) | 89.15 (± 1.19) |
| AGLA (2025) | 90.01 (± 1.23) | 93.38 (± 1.14) | 92.08 (± 1.74) | 91.17 (± 1.51) | 92.53 (± 1.42) | 93.86 (± 1.63) | **92.52 (± 1.11)** | 93.54 (± 1.77) |
| CORAL (Ours) | **90.63 (± 1.54)** | **99.71 (± 1.74)** | **96.47 (± 1.68)** | **91.89 (± 1.83)** | **93.55 (± 1.52)** | **95.83 (± 1.33)** | 91.50 (± 1.27) | **95.35 (± 1.87)** |
| *Popular* | | | | | | | | |
| Regular | 83.88 (± 1.48) | 85.62 (± 1.70) | 86.24 (± 1.44) | 84.79 (± 1.52) | 84.63 (± 1.48) | 86.17 (± 1.28) | 86.33 (± 1.14) | 85.26 (± 1.15) |
| VCD (2024) | 86.32 (± 1.83) | 90.15 (± 1.22) | 87.42 (± 1.44) | 89.34 (± 2.04) | 85.29 (± 1.14) | 88.01 (± 0.99) | 87.11 (± 1.32) | 86.25 (± 2.00) |
| MARINE (2025) | 88.72 (± 1.38) | 91.77 (± 1.33) | 88.24 (± 1.96) | 90.33 (± 1.18) | 87.67 (± 1.53) | 89.73 (± 1.11) | 88.33 (± 1.19) | 88.37 (± 1.81) |
| AGLA (2025) | 88.14 (± 1.23) | 91.88 (± 1.44) | **89.26 (± 1.73)** | 90.79 (± 1.15) | 89.42 (± 1.32) | 90.11 (± 1.53) | 88.91 (± 2.04) | 90.17 (± 1.22) |
| CORAL (Ours) | **90.00 (± 1.45)** | **97.93 (± 1.74)** | 89.20 (± 1.16) | **91.28 (± 1.83)** | **91.33 (± 1.35)** | **95.58 (± 1.73)** | **89.78 (± 1.17)** | **93.57 (± 1.44)** |
| *Adversarial* | | | | | | | | |
| Regular | 83.17 (± 1.37) | 80.62 (± 1.14) | 84.77 (± 1.43) | 83.46 (± 2.08) | 84.59 (± 1.33) | 81.31 (± 1.77) | 85.19 (± 1.17) | 84.96 (± 1.80) |
| VCD (2024) | 85.43 (± 1.88) | 87.13 (± 1.25) | 86.01 (± 2.01) | 85.24 (± 2.04) | 85.88 (± 1.65) | 86.16 (± 1.33) | 85.17 (± 1.12) | 86.99 (± 1.77) |
| MARINE (2025) | 87.42 (± 1.11) | 89.97 (± 1.41) | 88.41 (± 1.19) | 88.73 (± 1.91) | 87.25 (± 1.75) | 88.19 (± 1.37) | 87.70 (± 1.19) | 88.11 (± 1.16) |
| AGLA (2025) | 91.39 (± 1.23) | 92.28 (± 1.14) | **90.35 (± 1.75)** | 89.73 (± 2.15) | 90.11 (± 1.32) | 89.38 (± 1.14) | **90.08 (± 1.07)** | 89.83 (± 1.10) |
| CORAL (Ours) | **93.75 (± 1.54)** | **96.75 (± 1.74)** | 89.47 (± 2.16) | **90.87 (± 1.83)** | **91.03 (± 1.00)** | **92.33 (± 1.74)** | 89.02 (± 1.66) | **90.48 (± 1.83)** |

Table 19: Evaluation with POPE score across multiple LVLM architectures on **A-OKVQA** dataset comparing our method with several baselines with 3000 random replications. **Bold** indicates the best result and underline indicates the second-best result in each column.

| Decoding | LLaVA-v1.5 | | | | Qwen-VL | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy↑ | Precision↑ | Recall↑ | F1 Score↑ | Accuracy↑ | Precision↑ | Recall↑ | F1 Score↑ |
| *Random* | | | | | | | | |
| Regular | 83.45 (± 0.48) | 87.24 (± 0.68) | 78.36 (± 0.54) | 82.56 (± 0.50) | 86.67 (± 0.48) | 93.16 (± 0.55) | 79.16 (± 0.59) | 85.59 (± 0.53) |
| VCD (2024) | 86.15 (± 0.23) | 85.18 (± 0.34) | 87.53 (± 0.14) | 86.34 (± 0.21) | 89.22 (± 0.14) | 90.77 (± 0.04) | 87.32 (± 0.34) | 89.01 (± 0.16) |
| MARINE (2025) | 86.72 (± 0.14) | 87.71 (± 0.53) | 87.53 (± 0.34) | 86.34 (± 0.56) | 89.17 (± 0.45) | 89.87 (± 0.42) | **88.13 (± 0.54)** | 88.93 (± 0.18) |
| AGLA (2025) | **89.28 (± 0.33)** | **93.18 (± 0.43)** | 84.76 (± 0.42) | 88.77 (± 0.56) | 86.77 (± 0.43) | **95.02 (± 0.11)** | 77.60 (± 0.72) | 85.43 (± 0.98) |
| CORAL (Ours) | 88.90 (± 0.52) | 83.52 (± 0.49) | **96.93 (± 0.64)** | **89.73 (± 0.41)** | **89.53 (± 0.14)** | 90.92 (± 0.04) | 86.42 (± 0.58) | **89.10 (± 0.37)** |
| *Popular* | | | | | | | | |
| Regular | 79.90 (± 0.33) | 80.85 (± 0.31) | 78.36 (± 0.54) | 79.59 (± 0.37) | 85.56 (± 0.35) | 90.44 (± 0.56) | 79.53 (± 0.84) | 84.63 (± 0.42) |
| VCD (2024) | 81.85 (± 0.44) | 78.60 (± 0.58) | 87.53 (± 0.14) | 82.82 (± 0.36) | 87.85 (± 0.30) | 88.10 (± 0.36) | 87.53 (± 0.47) | 87.81 (± 0.31) |
| MARINE (2025) | 84.70 (± 0.59) | 86.02 (± 0.29) | 86.79 (± 0.24) | 85.13 (± 0.45) | 87.72 (± 0.42) | 88.73 (± 0.15) | **89.02 (± 0.54)** | 88.26 (± 0.24) |
| AGLA (2025) | 85.63 (± 0.78) | **86.27 (± 0.77)** | 84.67 (± 0.22) | 85.51 (± 0.25) | 86.30 (± 0.43) | **94.09 (± 0.11)** | 77.47 (± 0.22) | 84.97 (± 0.41) |
| CORAL (Ours) | **87.18 (± 0.54)** | 84.58 (± 0.70) | **89.65 (± 0.65)** | **87.52 (± 0.57)** | **89.02 (± 0.54)** | 91.48 (± 0.56) | 86.15 (± 0.61) | **88.60 (± 0.44)** |
| *Adversarial* | | | | | | | | |
| Regular | 75.84 (± 0.45) | 74.30 (± 0.63) | 79.03 (± 0.68) | 76.59 (± 0.40) | 80.28 (± 2.21) | 77.14 (± 2.46) | 84.48 (± 1.93) | 81.02 (± 2.07) |
| VCD (2024) | 79.56 (± 0.41) | **79.67 (± 0.59)** | 79.39 (± 0.50) | 79.52 (± 0.38) | 81.21 (± 1.89) | 80.25 (± 1.71) | 84.61 (± 2.15) | 82.96 (± 1.82) |
| MARINE (2025) | 75.23 (± 0.24) | 78.27 (± 0.52) | 78.72 (± 0.68) | 80.13 (± 0.43) | 85.26 (± 2.11) | 84.93 (± 1.97) | 85.42 (± 2.14) | 85.29 (± 2.09) |
| AGLA (2025) | 77.29 (± 0.13) | 74.09 (± 0.14) | 85.67 (± 0.21) | 79.16 (± 0.61) | 86.44 (± 1.61) | 89.26 (± 1.51) | **90.01 (± 1.11)** | **88.14 (± 2.01)** |
| CORAL (Ours) | **80.63 (± 0.72)** | 77.14 (± 0.77) | **87.07 (± 0.61)** | **81.80 (± 0.50)** | **88.27 (± 1.81)** | **90.14 (± 1.21)** | 86.93 (± 1.87) | 87.99 (± 2.03) |

| Decoding | InstructBLIP | | | | LLaVA-OneVision-7B | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy↑ | Precision↑ | Recall↑ | F1 Score↑ | Accuracy↑ | Precision↑ | Recall↑ | F1 Score↑ |
| *Random* | | | | | | | | |
| Regular | 80.71 (± 0.73) | 81.67 (± 0.67) | 79.19 (± 1.14) | 80.41 (± 0.80) | 85.87 (± 1.77) | 83.41 (± 2.13) | 88.08 (± 1.47) | 85.72 (± 1.66) |
| VCD (2024) | 84.53 (± 0.38) | 88.55 (± 0.54) | 79.32 (± 0.44) | 83.68 (± 0.40) | 86.33 (± 1.23) | 89.01 (± 1.42) | 87.33 (± 1.78) | 88.86 (± 1.43) |
| MARINE (2025) | 86.72 (± 0.11) | 87.97 (± 0.53) | 80.03 (± 0.19) | 87.33 (± 0.91) | 87.21 (± 1.35) | 90.33 (± 1.79) | 88.11 (± 1.93) | 89.09 (± 1.09) |
| AGLA (2025) | 87.30 (± 0.32) | 88.83 (± 0.41) | 82.08 (± 0.47) | 87.71 (± 0.51) | 88.15 (± 1.65) | 93.26 (± 1.88) | **90.08 (± 1.47)** | 89.97 (± 1.21) |
| CORAL (Ours) | **90.13 (± 0.54)** | **92.63 (± 0.47)** | **87.20 (± 0.61)** | **89.84 (± 0.38)** | **92.17 (± 0.89)** | **98.25 (± 1.46)** | 89.87 (± 1.61) | **91.64 (± 1.83)** |
| *Popular* | | | | | | | | |
| Regular | 78.22 (± 0.84) | 77.87 (± 1.03) | 78.85 (± 0.52) | 78.36 (± 0.76) | 82.72 (± 1.99) | 79.81 (± 2.31) | 86.87 (± 1.70) | 83.16 (± 1.85) |
| VCD (2024) | 81.47 (± 0.42) | **82.89 (± 0.64)** | 79.32 (± 0.44) | 81.07 (± 0.39) | 84.24 (± 1.46) | 86.16 (± 1.69) | 86.53 (± 1.78) | 86.16 (± 1.30) |
| MARINE (2025) | 81.74 (± 0.88) | 80.27 (± 0.35) | 81.73 (± 0.92) | 79.39 (± 0.12) | 86.81 (± 1.21) | 89.15 (± 1.55) | 87.20 (± 1.91) | 88.19 (± 2.09) |
| AGLA (2025) | 81.86 (± 0.14) | 80.17 (± 0.88) | 85.68 (± 0.64) | 82.58 (± 0.52) | 89.28 (± 1.11) | 91.63 (± 2.09) | 89.08 (± 1.31) | 89.34 (± 1.70) |
| CORAL (Ours) | **83.43 (± 0.68)** | 81.09 (± 0.70) | **87.20 (± 0.61)** | **84.03 (± 0.47)** | **90.37 (± 1.89)** | **94.36 (± 1.46)** | **89.87 (± 2.04)** | **89.91 (± 1.33)** |
| *Adversarial* | | | | | | | | |
| Regular | 75.84 (± 0.45) | 74.30 (± 0.63) | 79.03 (± 0.68) | 76.59 (± 0.40) | 80.28 (± 2.21) | 77.14 (± 2.46) | 84.48 (± 1.93) | 81.02 (± 2.07) |
| VCD (2024) | 79.56 (± 0.41) | **79.67 (± 0.59)** | 79.39 (± 0.50) | 79.52 (± 0.38) | 81.21 (± 1.89) | 80.25 (± 1.71) | 84.61 (± 2.15) | 82.96 (± 1.82) |
| MARINE (2025) | 75.23 (± 0.24) | 78.27 (± 0.52) | 78.72 (± 0.68) | 80.13 (± 0.43) | 85.26 (± 2.11) | 84.93 (± 1.97) | 85.42 (± 2.14) | 85.29 (± 2.09) |
| AGLA (2025) | 77.29 (± 0.13) | 74.09 (± 0.14) | 85.67 (± 0.21) | 79.16 (± 0.61) | 86.44 (± 1.61) | 89.26 (± 1.51) | **90.01 (± 1.11)** | **88.14 (± 2.01)** |
| CORAL (Ours) | **80.63 (± 0.72)** | 77.14 (± 0.77) | **87.07 (± 0.61)** | **81.80 (± 0.50)** | **88.27 (± 1.81)** | **90.14 (± 1.21)** | 86.93 (± 1.87) | 87.99 (± 2.03) |

| Decoding | Qwen2.5-VL-7B | | | | InternVL3-8B | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy↑ | Precision↑ | Recall↑ | F1 Score↑ | Accuracy↑ | Precision↑ | Recall↑ | F1 Score↑ |
| *Random* | | | | | | | | |
| Regular | 86.20 (± 1.30) | 88.68 (± 1.54) | 87.62 (± 1.24) | 86.98 (± 1.28) | 87.15 (± 1.88) | 86.87 (± 1.67) | 87.24 (± 1.76) | 86.26 (± 1.80) |
| VCD (2024) | 87.35 (± 1.83) | 89.55 (± 1.45) | 89.23 (± 1.35) | 90.06 (± 2.03) | 88.15 (± 1.53) | 89.01 (± 1.17) | 88.24 (± 1.44) | 88.05 (± 1.45) |
| MARINE (2025) | 89.72 (± 1.11) | 91.79 (± 1.35) | 91.31 (± 1.91) | 90.33 (± 2.09) | 89.23 (± 1.35) | 90.24 (± 1.66) | 88.52 (± 1.15) | 89.15 (± 1.19) |
| AGLA (2025) | 90.01 (± 1.23) | 93.38 (± 1.14) | 92.08 (± 1.74) | 91.17 (± 1.51) | 92.53 (± 1.42) | 93.86 (± 1.63) | **92.52 (± 1.11)** | 93.54 (± 1.77) |
| CORAL (Ours) | **90.63 (± 1.54)** | **99.71 (± 1.74)** | **96.47 (± 1.68)** | **91.89 (± 1.83)** | **93.55 (± 1.52)** | **95.83 (± 1.33)** | 91.50 (± 1.27) | **95.35 (± 1.87)** |
| *Popular* | | | | | | | | |
| Regular | 83.88 (± 1.48) | 85.62 (± 1.70) | 86.24 (± 1.44) | 84.79 (± 1.52) | 84.63 (± 1.48) | 86.17 (± 1.28) | 86.33 (± 1.14) | 85.26 (± 1.15) |
| VCD (2024) | 86.32 (± 1.83) | 90.15 (± 1.22) | 87.42 (± 1.44) | 89.34 (± 2.04) | 85.29 (± 1.14) | 88.01 (± 0.99) | 87.11 (± 1.32) | 86.25 (± 2.00) |
| MARINE (2025) | 88.72 (± 1.38) | 91.77 (± 1.33) | 88.24 (± 1.96) | 90.33 (± 1.18) | 87.67 (± 1.53) | 89.73 (± 1.11) | 88.33 (± 1.19) | 88.37 (± 1.81) |
| AGLA (2025) | 88.14 (± 1.23) | 91.88 (± 1.44) | **89.26 (± 1.73)** | 90.79 (± 1.15) | 89.42 (± 1.32) | 90.11 (± 1.53) | 88.91 (± 2.04) | 90.17 (± 1.22) |
| CORAL (Ours) | **90.00 (± 1.45)** | **97.93 (± 1.74)** | 89.20 (± 1.16) | **91.28 (± 1.83)** | **91.33 (± 1.35)** | **95.58 (± 1.73)** | **89.78 (± 1.17)** | **93.57 (± 1.44)** |
| *Adversarial* | | | | | | | | |
| Regular | 83.17 (± 1.37) | 80.62 (± 1.14) | 84.77 (± 1.43) | 83.46 (± 2.08) | 84.59 (± 1.33) | 81.31 (± 1.77) | 85.19 (± 1.17) | 84.96 (± 1.80) |
| VCD (2024) | 85.43 (± 1.88) | 87.13 (± 1.25) | 86.01 (± 2.01) | 85.24 (± 2.04) | 85.88 (± 1.65) | 86.16 (± 1.33) | 85.17 (± 1.12) | 86.99 (± 1.77) |
| MARINE (2025) | 87.42 (± 1.11) | 89.97 (± 1.41) | 88.41 (± 1.19) | 88.73 (± 1.91) | 87.25 (± 1.75) | 88.19 (± 1.37) | 87.70 (± 1.19) | 88.11 (± 1.16) |
| AGLA (2025) | 91.39 (± 1.23) | 92.28 (± 1.14) | **90.35 (± 1.75)** | 89.73 (± 2.15) | 90.11 (± 1.32) | 89.38 (± 1.14) | **90.08 (± 1.07)** | 89.83 (± 1.10) |
| CORAL (Ours) | **93.75 (± 1.54)** | **96.75 (± 1.74)** | 89.47 (± 2.16) | **90.87 (± 1.83)** | **91.03 (± 1.00)** | **92.33 (± 1.74)** | 89.02 (± 1.66) | **90.48 (± 1.83)** |

Table 20: Evaluation with POPE score across multiple LVLM architectures on **GQA** dataset, comparing our method with several baselines with 3000 random replications. **Bold** indicates the best result and underline indicates the second-best result in each column.

| Decoding | LLaVA-v1.5 | | | | Qwen-VL | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy↑ | Precision↑ | Recall↑ | F1 Score↑ | Accuracy↑ | Precision↑ | Recall↑ | F1 Score↑ |
| *Random* | | | | | | | | |
| Regular | 83.73 (± 0.27) | 87.16 (± 0.39) | 79.12 (± 0.35) | 82.95 (± 0.28) | 80.97 (± 0.32) | 88.07 (± 0.34) | 71.64 (± 0.57) | 79.01 (± 0.40) |
| VCD (2024) | 86.65 (± 0.45) | 84.58 (± 0.59) | 89.24 (± 0.34) | 86.99 (± 0.41) | 85.59 (± 0.38) | 86.88 (± 0.44) | 83.84 (± 0.36) | 85.33 (± 0.38) |
| MARINE (2025) | 86.33 (± 0.14) | 85.02 (± 0.63) | 88.23 (± 0.15) | 87.24 (± 0.11) | 85.54 (± 0.52) | 87.35 (± 0.25) | 87.26 (± 0.22) | 86.63 (± 0.22) |
| AGLA (2025) | 86.46 (± 0.34) | 85.84 (± 0.25) | 87.31 (± 0.15) | 86.57 (± 0.15) | 83.90 (± 0.43) | **93.05 (± 0.09)** | 73.26 (± 0.82) | 81.98 (± 0.13) |
| CORAL (Ours) | **88.02 (± 0.24)** | **87.25 (± 0.26)** | **90.02 (± 0.13)** | **89.11 (± 0.42)** | **89.13 (± 0.34)** | 89.96 (± 0.21) | **90.04 (± 0.23)** | **89.52 (± 0.63)** |
| *Popular* | | | | | | | | |
| Regular | 78.17 (± 0.17) | 77.64 (± 0.26) | 79.12 (± 0.35) | 78.37 (± 0.18) | 75.99 (± 0.33) | 78.62 (± 0.41) | 71.40 (± 0.38) | 74.84 (± 0.34) |
| VCD (2024) | 80.73 (± 0.47) | 76.26 (± 0.68) | **89.24 (± 0.34)** | 82.24 (± 0.35) | 81.83 (± 0.27) | 80.45 (± 0.47) | 84.09 (± 0.32) | 82.23 (± 0.22) |
| MARINE (2025) | 81.52 (± 0.63) | 82.12 (± 0.53) | 86.35 (± 0.24) | 84.35 (± 0.25) | 87.72 (± 0.42) | 88.73 (± 0.15) | 89.02 (± 0.54) | 88.26 (± 0.24) |
| AGLA (2025) | 83.67 (± 0.23) | 83.05 (± 0.24) | 84.60 (± 0.24) | 83.82 (± 0.15) | 80.80 (± 0.56) | 85.70 (± 0.35) | 73.93 (± 0.14) | 79.38 (± 0.51) |
| CORAL (Ours) | **85.26 (± 0.64)** | **85.25 (± 0.84)** | 87.25 (± 0.76) | **86.25 (± 0.74)** | **91.02 (± 0.15)** | **93.58 (± 0.09)** | **90.15 (± 0.11)** | **90.65 (± 0.15)** |
| *Adversarial* | | | | | | | | |
| Regular | 75.08 (± 0.33) | 73.19 (± 0.49) | 79.16 (± 0.35) | 76.06 (± 0.24) | 75.46 (± 0.63) | 77.92 (± 0.73) | 71.07 (± 0.97) | 74.33 (± 0.71) |
| VCD (2024) | 76.09 (± 0.43) | 70.83 (± 0.45) | **88.75 (± 0.56)** | 78.78 (± 0.36) | 80.01 (± 0.27) | 77.86 (± 0.24) | 83.85 (± 0.35) | 80.75 (± 0.27) |
| MARINE (2025) | 78.42 (± 0.62) | 77.22 (± 0.42) | 82.11 (± 0.62) | 79.93 (± 0.45) | 82.52 (± 0.13) | 79.22 (± 0.42) | 85.39 (± 0.14) | 82.11 (± 0.52) |
| AGLA (2025) | 80.66 (± 0.43) | 78.30 (± 0.23) | 84.82 (± 0.14) | 81.43 (± 0.42) | 78.73 (± 0.23) | 82.16 (± 0.53) | 73.40 (± 0.23) | 77.53 (± 0.53) |
| CORAL (Ours) | **82.42 (± 0.52)** | **80.12 (± 0.82)** | 85.52 (± 0.81) | **83.24 (± 0.53)** | **86.42 (± 0.54)** | **85.10 (± 0.52)** | **87.13 (± 0.13)** | **85.52 (± 0.53)** |

| Decoding | InstructBLIP | | | | LLaVA-OneVision-7B | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy↑ | Precision↑ | Recall↑ | F1 Score↑ | Accuracy↑ | Precision↑ | Recall↑ | F1 Score↑ |
| *Random* | | | | | | | | |
| Regular | 79.65 (± 0.24) | 77.14 (± 0.43) | 84.26 (± 0.36) | 80.56 (± 0.18) | 85.98 (± 0.14) | 86.53 (± 0.37) | 84.16 (± 0.42) | 85.27 (± 0.15) |
| VCD (2024) | 83.69 (± 0.11) | 81.84 (± 0.42) | 86.61 (± 0.48) | 84.16 (± 0.01) | 86.16 (± 0.13) | 85.26 (± 0.44) | 84.42 (± 0.17) | 85.37 (± 0.16) |
| MARINE (2025) | 85.25 (± 0.21) | 83.54 (± 0.14) | 86.92 (± 0.09) | 85.25 (± 0.53) | 88.33 (± 0.53) | 86.26 (± 0.14) | 87.33 (± 0.25) | 86.65 (± 0.15) |
| AGLA (2025) | 86.46 (± 0.35) | 85.84 (± 0.22) | 87.31 (± 0.54) | 86.57 (± 0.25) | 89.26 (± 0.42) | 89.72 (± 0.17) | 88.01 (± 0.35) | 88.53 (± 0.35) |
| CORAL (Ours) | **90.25 (± 0.24)** | **89.63 (± 0.54)** | **90.21 (± 0.20)** | **90.01 (± 0.25)** | **90.18 (± 0.16)** | **91.36 (± 0.35)** | **90.01 (± 0.35)** | **90.33 (± 0.31)** |
| *Popular* | | | | | | | | |
| Regular | 73.87 (± 0.58) | 69.63 (± 0.54) | 84.69 (± 0.68) | 76.42 (± 0.52) | 79.15 (± 0.35) | 80.53 (± 0.33) | 79.42 (± 0.17) | 80.11 (± 0.95) |
| VCD (2024) | 78.57 (± 0.14) | 74.62 (± 0.22) | 86.61 (± 0.48) | 80.17 (± 0.16) | 82.32 (± 0.14) | 81.93 (± 0.15) | 82.66 (± 0.16) | 81.97 (± 0.15) |
| MARINE (2025) | 78.11 (± 0.23) | 75.87 (± 0.12) | 87.01 (± 0.53) | 80.13 (± 1.05) | 84.15 (± 0.33) | 84.09 (± 0.15) | 84.11 (± 0.34) | 84.65 (± 0.22) |
| AGLA (2025) | 78.67 (± 0.11) | 77.44 (± 0.12) | **87.31 (± 0.34)** | 80.36 (± 0.12) | 87.15 (± 0.33) | 86.22 (± 0.35) | 87.01 (± 0.33) | 86.78 (± 0.25) |
| CORAL (Ours) | **79.23 (± 0.42)** | **78.03 (± 0.14)** | 87.21 (± 0.53) | **82.14 (± 0.34)** | **89.17 (± 0.42)** | **90.36 (± 0.14)** | **90.16 (± 0.32)** | **89.97 (± 0.27)** |
| *Adversarial* | | | | | | | | |
| Regular | 70.56 (± 0.53) | 66.12 (± 0.32) | 84.33 (± 1.05) | 74.12 (± 0.58) | 76.15 (± 0.63) | 77.35 (± 0.47) | 76.01 (± 0.28) | 76.25 (± 0.43) |
| VCD (2024) | 75.08 (± 0.13) | 70.59 (± 0.16) | 85.99 (± 0.10) | 77.53 (± 0.08) | 77.16 (± 0.24) | 78.24 (± 0.26) | 79.25 (± 0.11) | 78.41 (± 0.35) |
| MARINE (2025) | 74.97 (± 0.24) | 70.27 (± 0.12) | 86.13 (± 0.52) | 77.34 (± 0.42) | 79.25 (± 0.33) | 80.16 (± 0.23) | 80.04 (± 0.26) | 79.65 (± 0.26) |
| AGLA (2025) | 75.18 (± 0.52) | 70.19 (± 0.42) | **87.53 (± 0.43)** | 77.91 (± 0.29) | 80.15 (± 0.33) | 80.22 (± 0.17) | 79.15 (± 0.35) | 79.31 (± 0.37) |
| CORAL (Ours) | **79.62 (± 0.11)** | **72.32 (± 0.59)** | 86.92 (± 0.66) | **78.28 (± 0.53)** | **82.18 (± 0.17)** | **80.42 (± 0.16)** | **81.35 (± 0.33)** | **81.35 (± 0.33)** |

| Decoding | Qwen2.5-VL-7B | | | | InternVL3-8B | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy↑ | Precision↑ | Recall↑ | F1 Score↑ | Accuracy↑ | Precision↑ | Recall↑ | F1 Score↑ |
| *Random* | | | | | | | | |
| Regular | 86.19 (± 0.26) | 85.05 (± 0.24) | 85.37 (± 0.26) | 86.00 (± 0.26) | 85.74 (± 0.15) | 86.33 (± 0.33) | 85.56 (± 0.56) | 85.32 (± 0.27) |
| VCD (2024) | 87.14 (± 0.74) | 87.58 (± 0.11) | 86.74 (± 0.34) | 85.36 (± 0.15) | 85.14 (± 0.26) | 86.52 (± 0.16) | 86.16 (± 0.11) | 85.21 (± 0.26) |
| MARINE (2025) | 88.01 (± 0.15) | 89.79 (± 0.25) | 87.17 (± 0.13) | 88.10 (± 0.15) | 87.33 (± 0.14) | 89.53 (± 0.31) | 86.68 (± 0.13) | 87.42 (± 0.25) |
| AGLA (2025) | 90.22 (± 0.14) | 90.66 (± 0.32) | 89.71 (± 0.35) | 90.20 (± 0.45) | 90.04 (± 0.24) | 89.99 (± 0.32) | 90.79 (± 0.24) | 89.22 (± 0.42) |
| CORAL (Ours) | **91.03 (± 0.33)** | **91.39 (± 0.31)** | **90.40 (± 0.15)** | **90.65 (± 0.15)** | **92.15 (± 0.33)** | **93.36 (± 0.14)** | **91.36 (± 0.11)** | **91.16 (± 0.30)** |
| *Popular* | | | | | | | | |
| Regular | 85.19 (± 0.21) | 83.45 (± 0.14) | 86.63 (± 0.25) | 84.51 (± 0.23) | 83.74 (± 0.25) | 84.62 (± 0.33) | 83.65 (± 0.22) | 83.71 (± 0.27) |
| VCD (2024) | 86.37 (± 0.14) | 85.22 (± 0.17) | 86.91 (± 0.12) | 85.78 (± 0.15) | 84.26 (± 0.22) | 85.17 (± 0.26) | 84.33 (± 0.21) | 84.60 (± 0.23) |
| MARINE (2025) | 87.82 (± 0.11) | 86.76 (± 0.13) | 88.11 (± 0.21) | 87.43 (± 0.14) | 85.93 (± 0.19) | 86.88 (± 0.22) | 86.01 (± 0.18) | 85.97 (± 0.20) |
| AGLA (2025) | 87.31 (± 0.24) | 87.05 (± 0.18) | 87.60 (± 0.15) | 87.32 (± 0.21) | 87.08 (± 0.31) | 86.77 (± 0.28) | 87.15 (± 0.19) | 87.03 (± 0.24) |
| CORAL (Ours) | **89.14 (± 0.32)** | **88.37 (± 0.25)** | **89.65 (± 0.27)** | **89.01 (± 0.26)** | **88.92 (± 0.35)** | **89.75 (± 0.33)** | **88.90 (± 0.28)** | **89.32 (± 0.31)** |
| *Adversarial* | | | | | | | | |
| Regular | 83.02 (± 0.25) | 80.74 (± 0.18) | 84.88 (± 0.23) | 82.78 (± 0.21) | 82.45 (± 0.27) | 83.11 (± 0.22) | 82.66 (± 0.25) | 82.53 (± 0.24) |
| VCD (2024) | 84.61 (± 0.31) | 82.73 (± 0.27) | 85.75 (± 0.34) | 84.21 (± 0.29) | 83.16 (± 0.33) | 83.72 (± 0.29) | 83.95 (± 0.31) | 83.48 (± 0.30) |
| MARINE (2025) | 86.11 (± 0.28) | 84.36 (± 0.24) | 87.24 (± 0.30) | 85.78 (± 0.26) | 84.55 (± 0.30) | 85.21 (± 0.25) | 85.33 (± 0.27) | 85.02 (± 0.28) |
| AGLA (2025) | 87.42 (± 0.35) | 83.92 (± 0.31) | 88.53 (± 0.33) | 86.16 (± 0.30) | 86.32 (± 0.34) | 84.67 (± 0.28) | 86.55 (± 0.30) | 85.60 (± 0.31) |
| CORAL (Ours) | **88.73 (± 0.41)** | **85.94 (± 0.36)** | **89.92 (± 0.35)** | **87.89 (± 0.37)** | **87.94 (± 0.38)** | **86.52 (± 0.33)** | **87.36 (± 0.35)** | **87.02 (± 0.36)** |

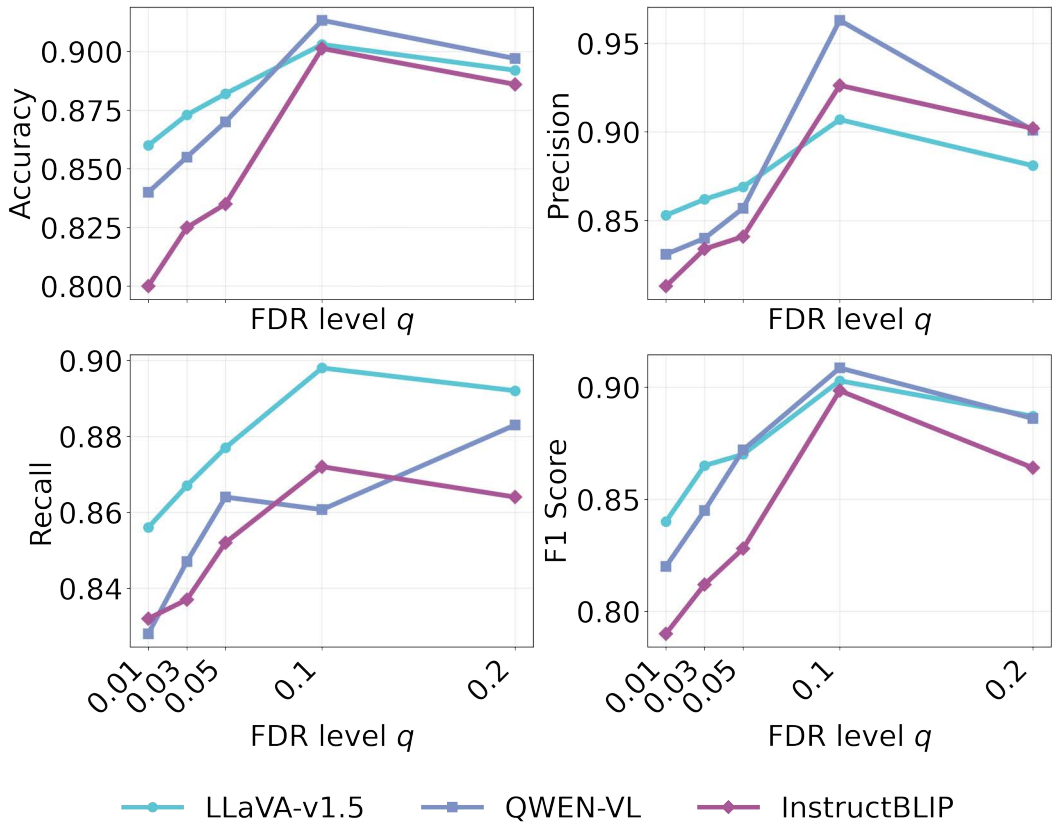



Figure 9: Ablation study on the effect of FDR target level ($q$) on the performance of LLaVA-v1.5, Qwen-VL, InstructBLIP using POPE metrics with $q = \{0.01, 0.03, 0.05, 0.1, 0.2\}$.

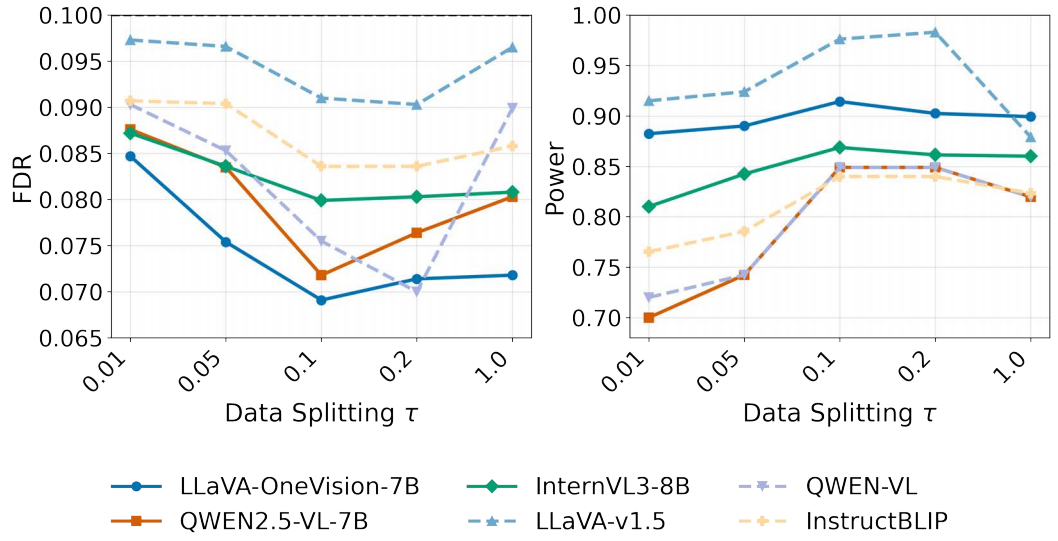



Figure 10: Sensitivity analysis of FDR and power with respect to the perturbation scale $\tau_v$. We report empirical FDR and power under different values of $\tau_v \in \{0.01, 0.05, 0.10, 0.20, 1.00\}$ while fixing the target FDR level at $q = 0.1$. The results show that CORAL achieves a favorable trade-off at intermediate perturbation scales and remains robust across a wide range of $\tau_v$ values.

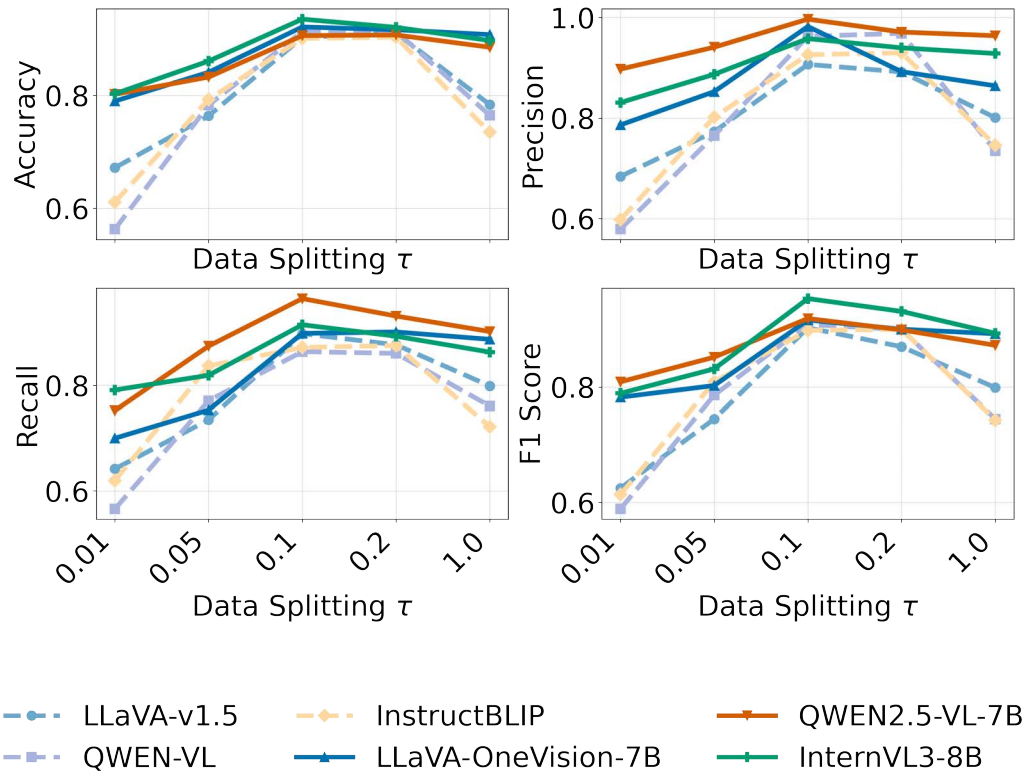



Figure 11: Sensitivity analysis of POPE accuracy and F1 with respect to the perturbation scale $\tau_v$. Performance is evaluated under different values of $\tau_v \in \{0.01, 0.05, 0.10, 0.20, 1.00\}$ while fixing the target FDR level at $q = 0.1$. The results indicate that CORAL is not sensitive to precise tuning of $\tau_v$ and achieves stable performance across a broad range of perturbation scales.

**(a) Over-suppression of grounded concepts with weak visual evidence**



**Query**: *Is there a cat in the image?*

**LLaVA-v1.5 w/ CORAL**: **No**, there is **not** a **cat** in the image.

**LLaVA-1.5 w/o CORAL**: **Yes**, there is a **cat** in the image.

**(b) Confusion between visually similar objects**



**Query**: *Is there a traffic light in the image?*

**LLaVA-v1.5 w/ CORAL**: **Yes**, there is a **traffic light** in the image.

**LLaVA-1.5 w/o CORAL**: **No**, there is a **streetlamp** in the image.

**(c) Missed small or partially occluded objects**



**Query**: *Is there a spoon in the image?*

**LLaVA-v1.5 w/ CORAL**: **No**, there is a **knife** in the image.

**LLaVA-1.5 w/o CORAL**: **Yes**, there is a **knife** in the image.

**(d) Fine-grained semantic ambiguity between related concepts or activities**



**Query**: *Is there a surfboard in the image?*

**LLaVA-v1.5 w/ CORAL**: **No**, there is a **parasailing** in the image.

**LLaVA-1.5 w/o CORAL**: **Yes**, there is a **surfboard** in the image.

Figure 12: Representative failure cases of CORAL on LLaVA-v1.5. Each example compares responses with and without CORAL to the same image and query. The cases illustrate remaining errors in object recognition and activity description, including failure to recognize a present cat, an unsupported traffic-light prediction, an inconsistent response concerning a knife, and confusion involving a surfboard and parasailing. Red and green text highlight incorrect and correct response components, respectively.