# Watermarking for Factuality: Guiding Vision-Language Models Toward Truth via Tri-layer Contrastive Decoding

**Anonymous ACL submission**


## Abstract

Large Vision-Language Models (LVLMs) have recently shown promising results on various multimodal tasks, even achieving human-comparable performance in certain cases. Nevertheless, LVLMs remain prone to hallucinations—they often rely heavily on a single modality or memorize training data without properly grounding their outputs. To address this, we propose a training-free, tri-layer contrastive decoding with watermarking, which proceeds in three steps: (1) select a mature layer and an amateur layer among the decoding layers, (2) identify a pivot layer using a watermark-related question to assess whether the layer is visually well-grounded, and (3) apply tri-layer contrastive decoding to generate the final output. Experiments on public benchmarks such as POPE, MME and AMBER demonstrate that our method achieves state-of-the-art performance in reducing hallucinations in LVLMs and generates more visually grounded responses. Our code will be publicly available upon publication.

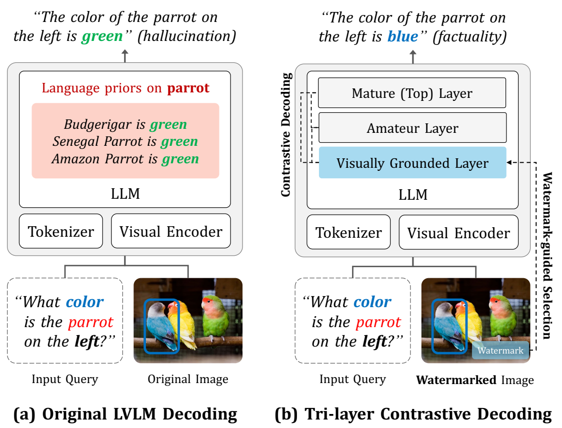




Figure 1: Architectural comparison between (a) the conventional decoding method of LVLMs and (b) our proposed watermark-based tri-layer contrastive decoding method. To mitigate hallucinations in LVLM, we leverage watermark for selecting visually grounded layer.

## 1 Introduction

Interest in Large Vision-Language Models (LVLMs) has surged recently, driven by integration of powerful large language models (LLMs) with visual encoders. This fusion enables a single model to interpret complex images and generate coherent descriptions. Recent LVLMs like LLaVA (Liu et al., 2023) and InstructBLIP (Dai et al., 2023) exemplify this trend: LLaVA connects a vision encoder to an LLM via a simple projection, while InstructBLIP uses a dedicated query transformer to bridge modalities. Such LVLMs have demonstrated impressive performance on tasks including image captioning, visual question answering, and other multimodal benchmarks.

A key limitation of LVLMs is their tendency to *hallucinate*—generating details absent from the image, such as naming non-existent objects or misattributing properties (see Fig. 1). Such hallucinations are often caused by the dominance of unimodal (language) priors. A lightweight vision module is often paired (and fine-tuned) with LLMs, which causes a modality imbalance where the language side can overwhelm the visual side (Han et al., 2022; Niu et al., 2021; Wu et al., 2022; Yan et al., 2023), outputting responses based mainly on LLMs' contextual or statistical biases. Thus, mitigating hallucinations is crucial for high-stakes applications, such as autonomous driving, medical imaging, and legal evidence analysis, where hallucinated responses could lead to severe consequences.

To mitigate such hallucinations, various approaches have been introduced. A straightforward approach is fine-tuning or specialized training: adjusting model weights on curated datasets that emphasize image-grounded truth (Gunjal et al., 2024; Yin et al., 2024a; Sarkar et al., 2025b), or employing Reinforcement Learning from Human Feedback (RLHF) or Direct Preference Optimization

(DPO) to penalize hallucinated outputs (Sun et al., 2023; Zhao et al., 2024). More recently, training-free inference-time contrastive decoding methods have emerged as efficient alternatives. For example, VCD (Leng et al., 2023) contrasts original and perturbed visual inputs to recalibrate the model's reliance on language priors. M3ID (Favero et al., 2024) boost visual relevance via mutual information, while AVISC (Woo et al., 2024) monitors and adjusts visual attention distributions. Octopus (Suo et al., 2025) combines these strategies by dynamically selecting contrastive approaches through DPO-trained controllers. However, existing methods often overlook how visual tokens interact with language across layers, assuming final outputs suffice for grounding. To address this, we embed lightweight visual watermarks into input images and evaluate layer-wise consistency via targeted visual queries. This enables the identification of the most visually grounded intermediate layer without retraining or architectural modifications, forming the basis of our tri-layer decoding strategy.

In this paper, we propose a novel training-free decoding strategy called Tri-layer Contrastive Decoding (TCD), which employs a watermark to guide the identification of the most visually grounded intermediate layer. To select this layer, we embed the watermark into the input image, query a corresponding ad-hoc question, and compare the probability distributions of an answer token across all layers. We explore *maximum probability gain search*, which identifies the layer based on the probability gain of the label token prompted by the watermark between adjacent layers. Given such visually grounded layer, we decode the model using tri-layer contrastive decoding with two additional layers, i.e., mature layer defined by top layer and amateur layer with the maximum Jensen-Shannon Divergence (JSD) compared to the mature layer, inspired by DoLa (Chuang et al., 2024). We evaluate our method on widely-used hallucination benchmarks—POPE (Li et al., 2023c), MME (Fu et al., 2024), and AMBER (Wang et al., 2023)—and show that the proposed approach achieves state-of-the-art performance across various models and settings. Detailed analyses further confirm the validity of our approach, demonstrating that watermark-guided TCD effectively mitigates hallucination. Our contributions are as follows:

- We propose Tri-layer Contrastive Decoding (TCD), a training-free inference framework that mitigates hallucination by contrasting three layer-wise outputs including mature, amateur, and visually grounded layer.
- We introduce a novel watermark-based approach to identify visually grounded layers in LVLMs by measuring visual information gain across intermediate outputs. Leveraging early-exit decoding with auxiliary visual prompts, our method enables interpretable and training-free layer selection.
- Extensive experiments on various benchmarks and models demonstrate the effectiveness of our proposed method, achieving state-of-the-art performance. Further analyses confirm that hallucinations are indeed alleviated, both quantitatively and qualitatively.

## 2 Related Work

**Hallucinations in LVLMs.** Various large vision-language models (LVLMs) have increasingly been introduced to improve the conventional multi-modal capabilities of traditional VLMs by leveraging and extending linguistic abilities of large language models (LLMs) (Liu et al., 2023; Li et al., 2023a; Bai et al., 2023a; Yang et al., 2024). Despite their promising performance in various multimodal tasks, LVLMs inherit the hallucination problem that is prevalent in LLMs. Among diverse types of hallucinations, object hallucination—where the model's descriptions of objects are not well-grounded in the input image—has drawn particular attention (Biten et al., 2022; Li et al., 2023c).

To mitigate hallucinations in LVLMs, several approaches have been proposed. Some frame hallucination as a binary classification task (Li et al., 2023c), while others design post-hoc correction modules (Zhou et al., 2023), or apply factually augmented reinforcement learning from human feedback (RLHF) (Sun et al., 2023) and Direct Preference Optimization (DPO) (Zhao et al., 2024). However, these methods typically require additional training stages and curated data.

More recently, training-free, inference-time methods have emerged to re-balance models during decoding. OPERA (Huang et al., 2024) penalizes over-aggregated anchor tokens in beam search. VCD (Leng et al., 2023) contrasts outputs from original and distorted visual inputs to reduce over-reliance on unimodal priors and statistical biases. ICD (Wang et al., 2024) suppress hallucinatiosn

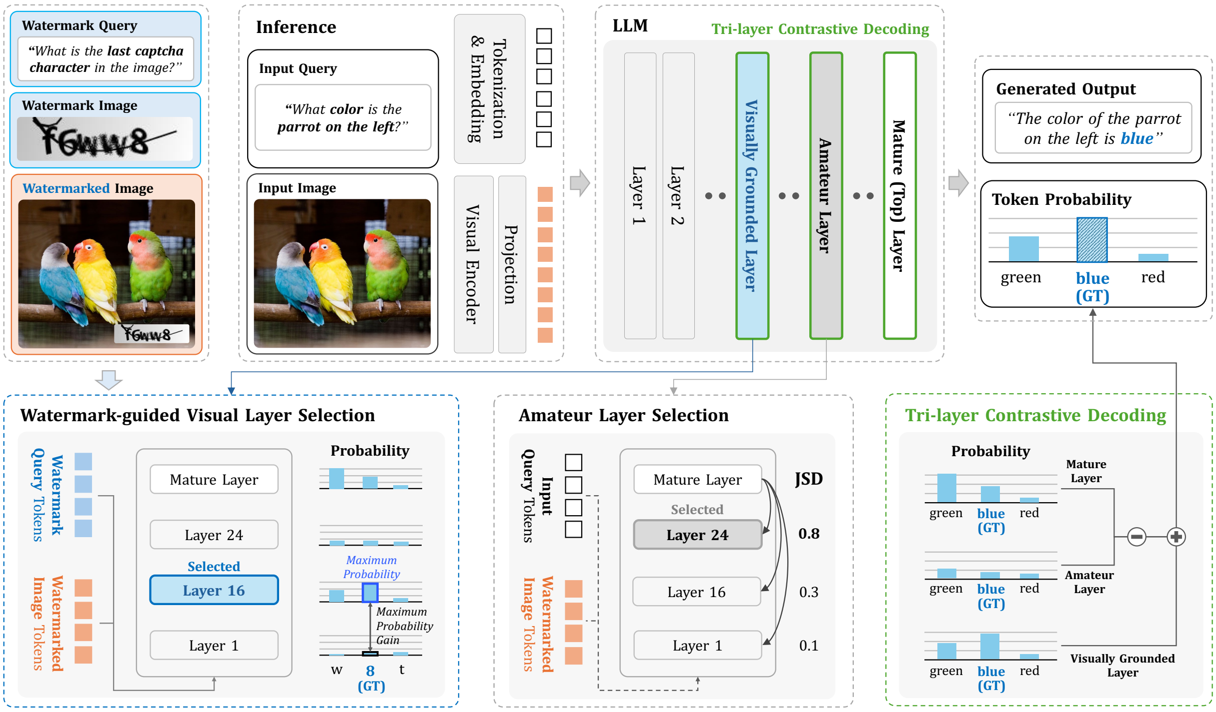



Figure 2: An overview of TCD, which leverages a tri-layer contrastive decoding approach by dynamically selecting and comparing following three decoding layers: (i) mature layer, (ii) amateur layer, and (iii) visually well-grounded layer. The process involves embedding a watermark into the input image, posing an ad-hoc question (e.g., "What is the last captcha character in the image?"), and selecting the visually well-grounded layer. Note that the top layer is chosen as the mature layer, while the amateur layer is selected based on the highest JSD from the mature layer.

by contrasting responses to perturbed instructions. M3ID (Favero et al., 2024) upweights image features during token sampling, and AVISC (Woo et al., 2024) reduces attention to blind tokens by monitoring visual focus. Octopus (Suo et al., 2025) dynamically selects contrastive decoding strategies using a controller trained via DPO.

All of these methods share a common philosophy: adjusting model behavior post hoc at inference time without retraining. Our proposed method aligns with this direction, but uniquely explores intermediate layers of the LVLM decoder. Instead of modifying inputs or attention distributions, we leverage the transformer's hierarchical representations to identify and utilize visually grounded layers for more reliable decoding.

**Layer-wise Contrastive Decoding.** Contrastive decoding (CD) is originally introduced in LLMs to improve fluency and coherence by contrasting the outputs of a strong expert model and a weaker amateur model (Li et al., 2022). Building on this idea, CAD (Shi et al., 2024) leverages surrounding context to guide generation more effectively, while ACD (Gera et al., 2023) enhances diversity and coherence in small LMs by fine-tuning early-layer prediction heads. Notably, DoLa (Chuang et al., 2024) introduces a layer-wise contrastive decoding framework that dynamically selects early layers based on token complexity to reduce hallucinations.

While these studies primarily focus on LLMs, applying CD to LVLMs poses new challenges, as models must incorporate both visual and linguistic modalities. Interestingly, we observe that intermediate layers in LVLMs often generate outputs that are more visually well-grounded than those from the final decoding layer. This observation motivates our use of layer-wise contrastive decoding as a potential solution for mitigating hallucinations.

However, identifying visually grounded layers in a training-free setting remains difficult. To address this, we propose leveraging watermarks—perturbations embedded into the input image that do not alter the final output but serve as cues for judging whether an intermediate layer is visually grounded.

## 3 Method

Given a visual context $v$ (e.g., an image) and a textual query $x$, LVLMs generate a textual response $y$. The response $y = \{y_1, y_2, \ldots, y_T\}$ is calculated in an auto-regressive manner, where each token is predicted sequentially based on the preceding tokens, and $T$ represents the total number of tokens in the generated response. Formally, the token probability distribution at each time step $t \in [1, T]$ can be formulated as follows:

$$p_\theta(y_t \mid x, v, y_{1:t-1}) = \frac{\exp(z_\theta(y_t \mid x, v, y_{1:t-1})/\tau)}{\Sigma_{y'_t \in \mathcal{Y}} \exp(z_\theta(y'_t \mid x, v, y_{1:t-1})/\tau)}, \tag{1}$$

where $\theta$ denotes model parameters, $z$ represents the logit of a layer, $\tau$ is a temperature for logit scaling, and $y'_t$ is a token in vocabulary set $\mathcal{Y}$. Output token selection, or decoding, determines the final generated response $y$ by selecting tokens from the probability distribution in Eq. (1). Common decoding strategies include greedy decoding (Sutskever et al., 2014), beam search (Bahdanau et al., 2014), and top-$k$ sampling (Fan et al., 2018).

Despite the effectiveness of these decoding strategies, a critical challenge remains: *hallucination*. In the context of LVLMs, even if the probability distribution $p_\theta$ assigns a high likelihood, a token $y_t$ is considered hallucinated if it lacks sufficient grounding in the provided textual query $x$ or visual context $v$. To this end, we propose a novel tri-layer contrastive decoding with a watermark-guided visual layer selection scheme. This approach aims to realign the model's token probability distribution with the factual constraints in $x$ and $v$, thereby reducing the incidence of hallucinations in the generated output. An overview of our proposed method is shown in Fig. 2.

### 3.1 Watermark-Guided Layer Selection

To mitigate hallucinations in LVLMs, we first select the most visually representative layer through watermark-based verification. The key intuition is that the visual information in LVLMs evolves across layers, which aligns with observations from prior work on LLMs (Chuang et al., 2024).

**Watermark Integration.** To identify a visually informative layer, a novel question emerges: *how can we identify a layer as visually informative, while preserving the visual representations of an input image?* This motivates us to design a watermark-based verification approach that can be seamlessly integrated with an input image and simultaneously provides a cue about the information in each layer. Specifically, we embed a watermark image into the input image and prepend a watermark question to the textual query. The watermark serves to examine each layer's representation in the model by leveraging image data related to vision-language tasks, such as CAPTCHAs. Formally, given a watermark image $\mathcal{I}_{\text{wm}}$ and a watermark textual query $x_{\text{wm}}$, the visual context $v$ and the textual query $x$ are generated as follows:

$$v = f_{\text{visual}}(\mathcal{I}_{\text{org}} + \alpha \mathcal{I}_{\text{wm}}), \tag{2}$$
$$x = \text{concat}(x_{\text{wm}}, x_{\text{org}}), \tag{3}$$

where $f_{\text{visual}}$ is a visual encoder, $\mathcal{I}_{\text{org}}$ is the input image, $x_{\text{org}}$ is the input text query, and $\alpha$ is the opacity hyperparameter for the watermark. For clarity, we construct a watermark question that has a fixed length and a clear answer (e.g., *"What is the last number in the CAPTCHA image?"*). In this section, we assume that $\mathcal{I}_{\text{wm}}$ is appropriately preprocessed (e.g., in terms of size and position) for the integration. For further details and analyses of watermark preprocessing, please see Section 4.1, as well as Algorithm 1 and Fig. 6, both located in the Appendix.

**Layer Selection in LVLMs.** Our goal is to identify the decoding layer $l_v$ that contains visually informative representations using the watermark-integrated inputs $x$ and $v$. We select a layer based on the probability distribution $p_\theta$ in Eq. (1), where the logit $z$ is computed using the hidden representation $h_{t-1}$ and the vocabulary head $g$, i.e., $z = g(h_{t-1})$. Although $z$ is often computed using the last layer representation for final output generation (i.e., $z = g(h_{t-1}^{(L)})$), it is also possible to apply the language head $g$ to intermediate layers—an approach known as early exit (Teerapittayanon et al., 2016; Schuster et al., 2022; Chuang et al., 2024)—to leverage a model's implicit factual knowledge.

Given the watermark-integrated textual query $x$ and visual context $v$, the hidden representation of layer $l$, $h_{t-1}^{(l)}$, is generated by first processing the input through the embedding layer $f_{\text{embed}}$ and then through a series of transformer layers $f_{\text{trans}}^{(l)}$:

$$h_{t-1}^{(0)} = f_{\text{embed}}(x, v, y_{1:t-1},), \tag{4}$$
$$h_{t-1}^{(l)} = f_{\text{trans}}^{(l)}(h_{t-1}^{(l-1)}), \;\; l \in \{1, 2, \ldots, L\}, \tag{5}$$

where $L$ is the total number of transformer layers. Using these hidden representations, we compute the layer-wise token probability distribution $p_\theta^{(l)}$:

$$p_\theta^{(l)} = \text{softmax}(z_\theta^{(l)}) = \text{softmax}(g(h_{t-1}^{(l)})). \tag{6}$$

| LVLM | Method | MSCOCO | | OKVQA | | GQA | |
|---|---|---|---|---|---|---|---|
| | | Acc.(↑) | F1(↑) | Acc.(↑) | F1(↑) | Acc.(↑) | F1(↑) |
| *Referenced Results (Not Directly Comparable)* | | | | | | | |
| LLaVA-v1.5 | EOS | 86.80 | 86.00 | - | - | - | - |
| | HA-DPO | 86.63 | 86.87 | - | - | - | - |
| | Octopus | 85.79 | 83.44 | - | - | - | - |
| InstructBLIP | OPERA | 79.13 | 79.74 | - | - | - | - |
| | HA-DPO * | 85.43 | 85.64 | - | - | - | - |
| | Octopus | 84.79 | 83.43 | - | - | - | - |
| *Comparable Results (Training-Free Contrastive Decoding)* | | | | | | | |
| LLaVA-v1.5 | Base | 82.04 | 80.42 | 75.58 | 79.23 | 74.39 | 78.58 |
| | + ICD | 83.26 | 82.53 | - | - | - | - |
| | + VCD | 82.96 | 81.81 | 74.72 | 78.87 | 74.10 | 78.70 |
| | + M3ID | 82.57 | 80.26 | 76.16 | 79.91 | 74.60 | 78.99 |
| | + AVISC | 83.39 | 81.01 | 77.47 | 80.87 | 76.33 | 80.40 |
| | + TCD (Ours) | **87.00** | **86.65** | **86.46** | **87.07** | **85.47** | **85.44** |
| InstructBLIP | Base | 79.14 | 79.31 | 74.93 | 77.86 | 73.84 | 76.70 |
| | + ICD | 79.14 | 79.92 | - | - | - | - |
| | + VCD | 79.46 | 79.49 | 75.59 | 78.28 | 75.36 | 77.87 |
| | + M3ID | 80.59 | 80.15 | 75.83 | 78.80 | 74.68 | 77.62 |
| | + AVISC | 84.04 | 82.62 | 80.92 | 82.62 | 79.85 | 80.98 |
| | + TCD (Ours) | **84.10** | **83.88** | **82.88** | **84.33** | **80.96** | **82.39** |

Table 1: Performance comparison on discriminative tasks (ALL split) across the POPE-MSCOCO, A-OKVQA, and GQA datasets. The best results are shown in **bold** and the second-best is underlined. * Denotes InstructBLIP with the Vicuna-13B backbone; all other models are based on Vicuna-7B. Complete results for the Random, Popular, and Adversarial subsets are provided in Appendix Tables 9 to 11.

**Watermark-Guided Visual Layer Selection.** Given the layer-wise probability distribution of the watermark label $y_{\text{wm}}$, we identify the layer with the greatest probability increase compared to the previous layer—referred to as *maximum probability gain search*—as formulated as follows:

$$l_v = \operatorname{argmax}_l \; \Delta p_\theta^{(l)}(y_{\text{wm}} \mid x, v) \tag{7}$$

where $\Delta$ denotes the difference in probability between adjacent layers:

$$\Delta p_\theta^{(l)} = \begin{cases} p_\theta^{(l)} - p_\theta^{(l-1)}, & \text{(i)} \\ \log\left(\dfrac{p_\theta^{(l)}}{p_\theta^{(l-1)}}\right). & \text{(ii)} \end{cases} \tag{8}$$

Note that the watermark textual query $x_{\text{wm}}$ is prepended to $x$ (see Eq. (3)); therefore, $p_\theta^{(l)}$ is measured using the first sequence of generated tokens (for simplicity, we ignore the special tokens).

### 3.2 Tri-layer Contrastive Decoding

In our framework, we leverage the visual layer $l_v$ as a reference probability distribution for contrastive decoding. Following prior work (Chuang et al., 2024), we define the final layer $L$ as a mature layer and use it as an anchor distribution. The negative distribution, $l_a$ (referred to as an amateur layer), is selected based on the highest Jensen-Shannon Divergence (JSD) between the distributions of the intermediate layers and the anchor distribution:

$$l_a = \operatorname{argmax}_l \operatorname{JSD}\big(p_\theta^{(L)}, p_\theta^{(l)}\big), \tag{9}$$

where $l \in \{1, 2, \ldots, L-1\}$ is an intermediate layer index. Note that a high JSD implies that such a layer offers an alternative perspective prior to the final layer's information accumulation, making it a strong candidate for contrastive decoding.

**Constraints on Contrastive Decoding.** When a token exhibits high confidence in both the mature layer $L$ and the amateur layer $l_a$, the contrastive decoding process may reduce the relative difference between probabilities, making a previously certain decision ambiguous. To address this, we adopt the Adaptive Plausibility Constraint (APC), following prior works (Li et al., 2023b; Leng et al., 2023; Chuang et al., 2024). Formally, we define the set of viable tokens $\mathcal{V}$ as follows:

$$\mathcal{V}(x_t \mid x_{1:t-1}) = \left\{x_t \in X \mid p_\theta^{(L)}(x_t) \geq \beta \max_w p_\theta^{(L)}(w)\right\} \tag{10}$$

where $\beta \in [0, 1]$ is a hyperparameter that determines the threshold for plausible token selection.

**Final Output Generation.** To generate the final response $y$, we first define a constraint function

| LVLM | Method | Object-level | | Attribute-level | | Total(↑) |
|---|---|---|---|---|---|---|
| | | Existence(↑) | Count(↑) | Position(↑) | Color(↑) | |
| LLaVA-v1.5 | Base | 173.57 | 110.00 | 100.47 | 125.24 | 509.28 |
| | + VCD | 172.14 | 117.14 | 103.33 | 119.52 | 512.14 |
| | + M3ID | 178.33 | 107.22 | 96.39 | 127.50 | 509.44 |
| | + AVISC | **189.29** | 104.76 | 106.19 | 127.86 | 528.09 |
| | + TCD (Ours) | 185.00 | **158.3** | **135.0** | **175.0** | **653.30** |
| InstructBLIP | Base | 170.19 | 89.52 | 67.62 | 114.76 | 442.09 |
| | + VCD | 172.62 | 98.33 | 71.90 | 117.14 | 459.99 |
| | + M3ID | 173.89 | 89.72 | 72.72 | 110.56 | 446.88 |
| | + AVISC | **184.76** | 82.85 | 74.76 | 131.43 | 473.80 |
| | + TCD (Ours) | 180.00 | **116.67** | **76.66** | **158.33** | **531.67** |

Table 2: Performance comparison on the discriminative task using the coarse-grained perception subset of the MME (Fu et al., 2024) benchmark.

| LVLM | Method | CHAIR(↓) | Cover.(↑) | HalRate(↓) | Cog.(↓) |
|---|---|---|---|---|---|
| *Referenced Results (Not Directly Comparable)* | | | | | |
| LLaVA-v1.5 | EOS | 5.1 | 49.1 | 22.7 | 2.0 |
| | HA-DPO | 6.7 | 49.8 | 30.9 | 3.3 |
| | HALVA | 6.6 | 53.0 | 32.2 | 3.4 |
| | Octopus | 4.8 | 49.2 | 23.4 | 1.2 |
| *Comparable Results (Training free Contrastive Decoding)* | | | | | |
| LLaVA-v1.5 | Base | 8.0 | 44.5 | 31.0 | 2.2 |
| | + VCD | 6.7 | 46.5 | 27.8 | 2.0 |
| | + M3ID | 6.0 | **48.9** | 26.0 | **1.5** |
| | + AVISC | 6.3 | 46.6 | 25.6 | 2.0 |
| | + TCD (Ours) | **4.4** | 47.2 | **19.2** | 1.7 |
| InstructBLIP | Base | 8.4 | 46.4 | 31.1 | 2.6 |
| | + VCD | 7.6 | 47.7 | 29.9 | **2.2** |
| | + M3ID | 6.9 | 47.2 | 27.5 | **2.2** |
| | + AVISC | 6.7 | 46.7 | 28.0 | 2.6 |
| | + TCD (Ours) | **6.3** | **48.8** | **26.8** | 2.3 |
| *Appliance to a Stronger Backbone* | | | | | |
| DeepSeek-VL2-Tiny | Base | 3.8 | 56.8 | 18.2 | 1.0 |
| | + VCD* | 4.7 | **56.9** | 22.4 | 1.3 |
| | + TCD (Ours) | **3.6** | 56.3 | **16.5** | **0.8** |

Table 3: Performance comparison on the generative task using the AMBER (Wang et al., 2023) benchmark. * Indicates results implemented using the official code.

$F(\cdot)$ to leverage APC on the input tokens:

$$F(z_\theta(x_t)) = \begin{cases} z^{(L)} - z^{(l_a)} + \lambda z^{(l_v)} & \text{if } x_t \in \mathcal{V}(x_t \mid x_{<t}) \\ -\infty & \text{otherwise.} \end{cases} \tag{11}$$

This formulation ensures that contrastive decoding effectively integrates visual grounding while avoiding false positives (implausible tokens receiving disproportionately high scores) and false negatives (valid tokens being overlooked due to contrastive decoding effects) through the application of APC, thereby reducing hallucinations in generated responses. Finally, we generate the token sequence $y$ using the refined logits under the APC constraint:

$$y \sim \hat{p}_\theta = \text{softmax}(F(z_\theta(x_t))). \tag{12}$$

# 4 Experiments

## 4.1 Experimental Setup

**Benchmarks and LVLMs.** To evaluate LVLM's hallucination performance, we use three widely used benchmarks: POPE (Li et al., 2023c), a perception subset of MME (Fu et al., 2024), and AMBER (Wang et al., 2023). Following previous works (Leng et al., 2023; Woo et al., 2024; Suo et al., 2025), we evaluate the discriminative task on POPE, MME and generative task on AMBER. **POPE** is used to assess object hallucination by querying whether a specific object exists in an image, using a balanced set of positive and negative queries. It employs three sampling strategies—adversarial, popular, and random—across three datasets (i.e., MS-COCO (Lin et al., 2014), A-OKVQA (Schwenk et al., 2022), and GQA (Hudson and Manning, 2019)), thereby generating a total of 27,000 query-answer pairs. In addition, we use the **MME** benchmark to evaluate LVLMs on perception-related tasks. Following prior work (Yin et al., 2024b; Leng et al., 2023), we focus on object-level hallucination (existence and count) and attribute-level hallucination (position and color). For generative tasks, we utilize **AMBER**, an automated LLM-free multi-dimensional benchmark. Four metrics including Cover, Hal, Cog, and CHAIR (Rohrbach et al., 2018) are used to measure the generation quality of our method. Specifically, AMBER compares generated object mentions against human-annotated ground truth to evaluate object coverage (Cover), hallucination frequency (Hal), cognitively plausible hallucinations (Cog), and the proportion of hallucinated objects (CHAIR), providing a comprehensive and cost-efficient assessment of hallucination. In our experiments, we evaluate our method on two widely used LVLMs, LLaVA-1.5 (Liu et al., 2023) and InstructBLIP (Dai et al., 2023), both using Vicuna-7B as the backbone. We also apply our method to generative tasks using DeepSeek-VL2(Wu et al., 2024), a model with a Mixture of Expert (MoE) architecture, thereby demonstrating the robustness of TCD on a stronger backbone.

**Implementation Details.** Following prior work (Chuang et al., 2024; Leng et al., 2023), we set $\beta = 0.1$ for stable CD and use 20 candidate layers for both LVLMs, except in the case of MME evaluation for InstructBLIP. Other parameters such as $\lambda$ and question templates, are provided in the Appendix C. We leverage simple yet effective CAPTCHA (Wilhelmy and Rosas, 2013) dataset for watermark verification. Further, to seamlessly integrate a watermark into the input image, we apply light preprocessing (e.g., position, size, and opacity). The watermark is placed in the bottom-right corner with opacity $\alpha = 0.8$. Addi-

| Model | Setting | Decoding | Accuracy(↑) | F1(↑) |
|---|---|---|---|---|
| LLaVA1.5 (7B) | Random | Greedy | 85.87 | 84.37 |
| | | + AL | 87.70 (+1.83) | 86.37 (+2.00) |
| | | + AL+VL | 89.50 (+1.80) | 88.89 (+2.52) |
| | Popular | Greedy | 84.10 | 82.75 |
| | | + AL | 86.63 (+2.53) | 85.34 (+2.59) |
| | | + AL+VL | 87.60 (+0.97) | 87.14 (+1.80) |
| | Adversarial | Greedy | 81.03 | 80.10 |
| | | + AL | 84.27 (+3.24) | 83.18 (+3.08) |
| | | + AL+VL | 83.90 (-0.37) | 83.92 (+0.74) |
| LLaVA1.5 (13B) | Random | Greedy | 85.47 | 84.32 |
| | | + AL | 87.03 (+1.56) | 85.84 (+1.52) |
| | | + AL+VL | 90.23 (+3.20) | 89.20 (+3.36) |
| | Popular | Greedy | 84.07 | 82.89 |
| | | + AL | 87.03 (+2.96) | 85.84 (+2.95) |
| | | + AL+VL | 89.70 (+2.67) | 89.20 (+3.36) |
| | Adversarial | Greedy | 81.90 | 81.14 |
| | | + AL | 85.07 (+3.17) | 84.03 (+2.89) |
| | | + AL+VL | 85.87 (+0.80) | 85.79 (+1.76) |

Table 4: Effect of the components of the proposed contrastive decoding method: amateur layer (AM) and watermark-based visual layer (VL). We use the LLaVa-1.5 backbone on the POPE-MSCOCO benchmark. Performance gains are highlighted in red, and performance drops are highlighted in blue.

tional implementation details are provided in the Appendix A and Fig. 6.

## 4.2 Experimental Results

**Comparison with SOTA Approaches.** To validate the effectiveness of our method, we conduct evaluations using various benchmarks, models, and decoding methods. We use instruction fine-tuned LVLMs (referred to as "Base" in the tables), along with ICD, VCD, M3ID and AVISC, as our training-free contrastive decoding baselines. We additionally compare against EOS (Yue et al., 2024), HA-DPO (Zhao et al., 2024), HALVA (Sarkar et al., 2025a), and Octopus, which require additional training or external models, and serve as reference methods.

As shown in Table 1, TCD clearly outperforms the baselines and achieves state-of-the-art performance across all three subsets of POPE (Li et al., 2023c), in terms of both accuracy and F1 score. While Octopus combines all three baseline methods and requires additional DPO training, TCD still surpasses it—achieving higher performance for the LLaVA model and in F1 score for InstructBLIP.

The efficacy of our method in mitigating hallucinations is further confirmed in Table 2, while outperforming the baselines in object and attribute level. We provide full perception task score in the Appendix Table 6. For generative task, our method successfully mitigated hallucinations lowering the CHAIR score and Hallucination Rate of AMBER bench with huge margin as shown in Table 3. We also show our method's scalability using latest LVLM model, DeepSeek-VL-Tiny (Wu et al., 2024) with stronger backbone for both visual encoder and LLM compared with LLaVA-v1.5 and InstructBLIP.

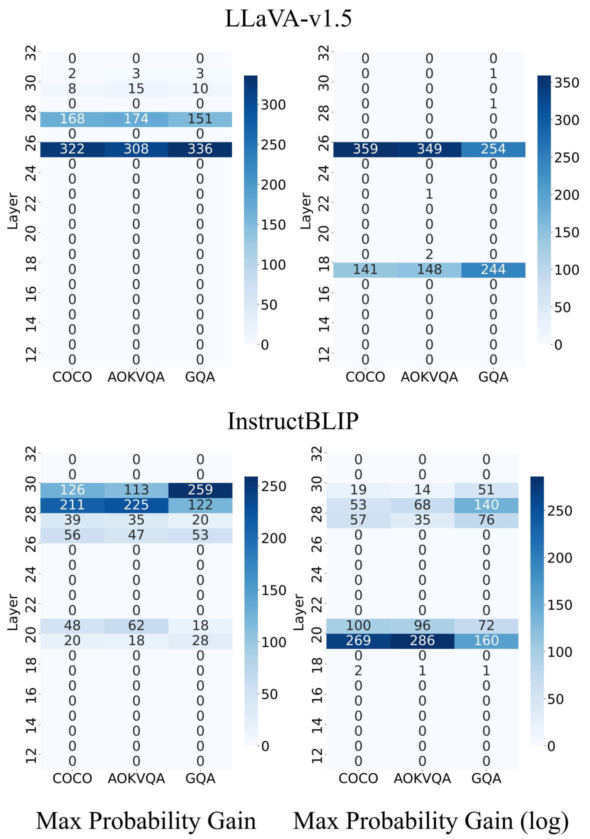
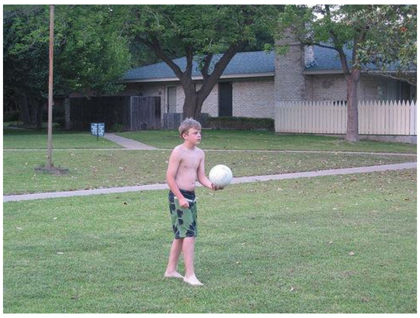



Figure 3: Heatmaps of each LVLMs' layers selected with POPE datasets. (COCO, AOKVQA and GQA). As shown, the proposed methods showed consistent layer selection with low variance.

**Visual Layer Selection Analysis.** Fig. 3 shows the results of our proposed visually grounded layer search, demonstrating its robustness through consistent layer selection layer selection across the model. We also observe noticeable variation across different models, implying representational differences among LVLMs.

**Qualitative Analysis.** As depicted in the Fig. 4, our method successfully mitigated hallucinations in the original model, increase in factuality, describing number of people correctly where other methods suffer. While other models generates similar hallucinations (i.e., cars in the background), likely to be triggered from memorization of the training data, our method distinguishes the house

**Question:** Describe this image.



**Base:**
The image features a young boy standing in a grassy field, holding a soccer ball. He appears to be enjoying his time outdoors, possibly playing a game of soccer. In the background, there are two cars parked, one on the left side and the other on the right side of the field. Additionally, there is a bench located near the right edge of the field, providing a place for spectators to sit and watch the game.

**VCD:**
The image captures a young man wearing a green and blue swimsuit standing on a grassy lawn in front of a house. The scene takes place near a driveway, with a car visible in the background on the right side of the image. There is a bench situated on the left side of the lawn, providing a place to sit and relax. The overall atmosphere suggests a fun and leisurely moment outdoors.

**AVISC:**
A young and shirtless boy is standing in a grassy yard, playing with a soccer ball. He is in the process of kicking it, demonstrating his enthusiasm for the game. There is a sports ball lying on the ground in the field, possibly related to his play or practice. A few cars can be seen in the background, one parked to the right of the boy and another one further to the left.

**TCD (Ours):**
The image features a young boy standing in a grassy field, holding a soccer ball in his hands. He appears to be enjoying his time outdoors, possibly preparing to play a game of soccer or simply practicing his skills. The boy is wearing swim trunks, which suggests that he might be at a beach or near a body of water. In the background, there is a house visible, adding to the outdoor setting.

Figure 4: Example generated by Base, VCD (Leng et al., 2023), AVISC (Woo et al., 2024) and our method, given the question: "Describe this image." We randomly sample from Amber (Wang et al., 2023) and all results are based on LLaVA-v1.5-7b model. We highlight phrases with red if they are *not* well visually grounded, and with blue if they are well visually grounded. We observe that our model successfully mitigates hallucinations compared to the other three baselines. Additional examples are provided in Appendix Fig. 8.

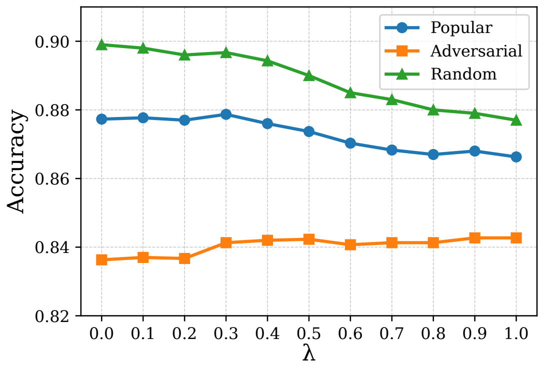



Figure 5: Comparison of accuracy across subsets of POPE-MSCOCO under varying $\lambda$ in the ablation setup, based on Eq. (13). While the Random and Popular subsets show improved performance when the visual layer dominates (i.e., lower $\lambda$), the Adversarial subset benefits from a greater contributions of the amateur layer (i.e., larger $\lambda$), highlighting the distinct roles of the visual and amateur layers in mitigating different forms of hallucination.

visible in the background.

**Tri-layer Selection Analysis.** Table 4 shows that contrasting the visual layer (+VL) with amateur layer (+AL) consistently boosts F1, except in the adversarial split. To isolate each layer's role, we interpolate the logits as follows:

$$z^{(L)} - \lambda z^{(l_a)} + (1-\lambda) z^{(l_v)}, \tag{13}$$

and sweep $\lambda$. Fig. 5 highlight the distinct roles played by each layer in our tri-layer decoding framework. In Random and Popular subsets, accuracy increases as $\lambda$ decreases, emphasizing the importance of the visually grounded layer $l_v$ in typical scenarios. Conversely, the Adversarial subset benefits from larger $\lambda$, as the amateur layer $l_a$ injects a complementary distribution less biased by co-occurrence patterns learned during pretraining(Chuang et al., 2024). This helps mitigate hallucinations triggered by visually plausible yet incorrect objects. These results suggest that our tri-layer formulation effectively addresses two major sources of hallucination commonly discussed in LVLMs: (i) internal linguistic biases and (ii) weak visual grounding. The JSD-guided selection of $l_a$ helps counteract the former, especially in adversarial contexts, while the watermark guided $l_v$ enhances visual alignment in standard inputs. While we fix $\lambda$ for simplicity in our main results, the ablation findings suggest promising directions for adaptive weighting strategies based on input characteristics.

## 5 Conclusion

In this paper, we introduce Tri-layer Contrastive Decoding (TCD), a training-free framework for reducing hallucinations in Large Vision-Language Models (LVLMs). Rather than assuming the final model output always provides the best visual grounding, we propose a principled approach that embeds lightweight visual watermarks into input images and leverages targeted visual queries to probe layer-wise consistency. By combining this watermark-guided visual layer selection with contrastive decoding across mature, amateur, and visually grounded layers, TCD dynamically recalibrates the model's reliance on vision and language, significantly improving factuality.

## 6 Limitations

While our method demonstrates consistent improvements across multiple benchmarks and models, several limitations remain. First, our layer selection mechanism is intentionally simple and interpretable, relying on fixed, rule-based comparisons of intermediate logits. This choice benefits reproducibility and transparency, but more sophisticated or learned strategies—such as attention-based routing or score aggregation—could further enhance flexibility and robustness, especially for models with more complex encoder-decoder architectures. Additionally, extending interpretability beyond decoder layers to the visual encoder itself remains an open and promising direction.

Second, our current implementation requires multiple decoding passes to evaluate candidate layers. Although inference can be reduced to a single pass if the preferred layer is predefined or learned, developing a seamless and fully dynamic layer selection mechanism without multi-pass exploration is still an open challenge.

Third, for generation tasks, we follow AMBER's non-LLM-based evaluation protocol to reduce subjectivity and improve reproducibility. While this is consistent with prior literature, it limits direct comparison to studies that use LLM-based scoring. Developing a more robust evaluation framework—balancing reproducibility with semantic depth, for example via ensemble metrics or human-in-the-loop evaluation—would further strengthen future studies on hallucination mitigation.

Further discussions regarding baselines and experimental settings are provided in Appendix E.

## 7 Ethics Statement

All experiments are conducted using publicly available datasets (POPE, MME, AMBER), none of which contain personally identifiable or sensitive information. While our method aims to reduce object hallucinations by improving visual grounding, it does not address other potential biases—such as social, demographic, or ethical biases—that may already exist in the underlying LVLMs. In certain cases, stronger visual grounding could inadvertently reinforce existing biases by making them appear more factual. Future work may investigate the interaction between decoding-time visual grounding and bias.

## References


Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. 2022. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1381–1390.

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

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

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

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Ariel Gera, Roni Friedman, Ofir Arviv, Chulaka Gunasekara, Benjamin Sznajder, Noam Slonim, and Eyal Shnarch. 2023. The benefits of bad advice: Autocontrastive decoding across model layers. *arXiv preprint arXiv:2305.01628*.

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

Yudong Han, Liqiang Nie, Jianhua Yin, Jianlong Wu, and Yan Yan. 2022. Visual perturbation-aware collaborative learning for overcoming the language prior problem. *arXiv preprint arXiv:2207.11850*.

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

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

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *Preprint*, arXiv:2311.16922.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. 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.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

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

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

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

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024c. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710.

Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. 2025. Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6):6434–6442.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan O Arik, and Tomas Pfister. 2025a. Mitigating object hallucination in MLLMs via data-augmented phrase-level alignment. In *The Thirteenth International Conference on Learning Representations*.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö. Arık, and Tomas Pfister. 2025b. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. *Preprint*, arXiv:2405.18654.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. *Advances in Neural Information Processing Systems*, 35:17456–17472.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022.

A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. 2025. Octopus: Alleviating hallucination via dynamic contrastive decoding. *Preprint*, arXiv:2503.00361.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. 2016. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 2464–2469. IEEE.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

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

Rodrigo Wilhelmy and Horacio Rosas. 2013. captcha dataset.

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. 2024. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. *Preprint*, arXiv:2405.17820.

Yike Wu, Yu Zhao, Shiwan Zhao, Ying Zhang, Xiaojie Yuan, Guoqing Zhao, and Ning Jiang. 2022. Overcoming language priors in visual question answering via distinguishing superficially similar instances. *arXiv preprint arXiv:2209.08529*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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

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

Zihao Yue, Liang Zhang, and Qin Jin. 2024. Less is more: Mitigating multimodal hallucination from an EOS decision perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11766–11781, Bangkok, Thailand. Association for Computational Linguistics.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2024. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *Preprint*, arXiv:2311.16839.

Yiqi Zhong, Luming Liang, Ilya Zharkov, and Ulrich Neumann. 2023. Mmvp: Motion-matrix-based video prediction. *Preprint*, arXiv:2308.16154.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

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

**Algorithm 1** Embedding Visible Identifier (Watermarking)

**Input:** original image $\mathcal{I}_o$, watermark image $\mathcal{I}_w$, image dimensions $(x_o, y_o)$, $(x_w, y_w)$, and opacity $\alpha$

**Let** $(0,0)$ be the top-left pixel of $\mathcal{I}_o$, and $\mathcal{C}_w = (c_w^{(x)}, c_w^{(y)})$ be the center pixel of $\mathcal{I}_w$,

$\mathcal{P}_o \leftarrow (0.9x_o, 0.9y_o)$ ▷ bottom-right anchor pixel
$\mathcal{C}_w \leftarrow \mathcal{P}_o$ ▷ overlapping watermark
**while** $\mathcal{C}_w + (x_w/2, y_w/2) > (x_o, y_o)$ **do**
  **if** $c_w^{(x)} + x_w/2 > x_o$ **then** ▷ resize width
    $x_w \leftarrow \min(x_w/2, x_o - x_w)$
  **end if**
  **if** $c_w^{(y)} + y_w/2 > y_o$ **then** ▷ resize height
    $y_w \leftarrow \min(y_w/2, y_o - y_w)$
  **end if**
**end while**
$\mathcal{I} \leftarrow \mathcal{I}_o + \alpha\mathcal{I}_w$ ▷ watermark integration

**Output:** watermark-embedded image $\mathcal{I}$

## A Ablation Study on Watermark Parameters

**Visual Grounding Question and CAPTCHA selection.** Since the key of tri-layer contrastive decoding is to select a visually grounded pivot layer with early exit token prediction method, "a well designed question" that judges a layer robustly is crucial. Since the LVLM utilizes the LLM, it is sensitive to both the textual and visual input queries. If we design a task that is simple, the token probability may not be meaningful to choose a pivot layer. From this perspective, we chose CAPTCHA (Wilhelmy and Rosas, 2013) as a suitable complex visual input. Together with the visual query, we conducted a simple experiment with to fix both the image and text question. As shown in Fig. 6, we found that LVLM (i.e., LLaVA-1.5) tends to answer the last captcha character better. With some more finding such that LVLMs tend to have problems with recognizing numbers such as "0", "9" that may resemble the alphabet letters, we chose "f6ww8" as our experiment CAPTCHA. With these experiments, we fixed the question that select the visual-grounded layer as "What is the last captcha number in the image?".

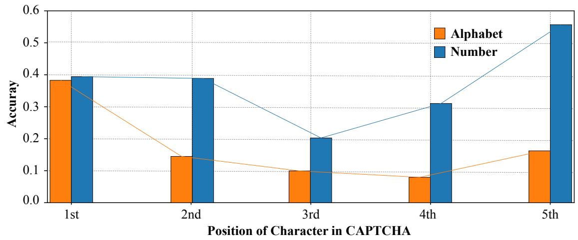



Figure 6: Qualitative result of CAPTCHA position. LVLM tends to answer numbers better than alphabet, last fifth character better than the other position.

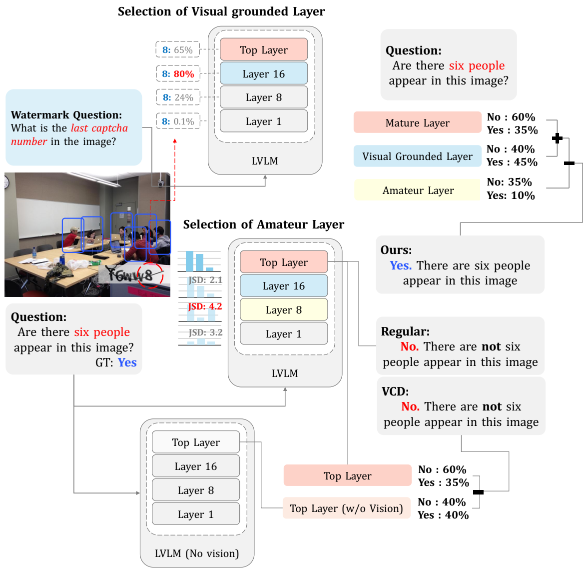



Figure 7: Examples of our tri-layer contrastive decoding approach on a sample from MME benchmark. We observe that our model outperforms the other alternatives, i.e., VCD (Leng et al., 2023) and regular LVLM model, successfully mitigating hallucinations while current models suffers. Note that an original image without watermark is used for all methods.

## B Artifacts

### B.1 Prompt Template

For each benchmark, we follow the official prompt template. For LLaVA-1.5, we adopt the POPE/MME instruction ending with *"Please answer the question using a single word or phrase."*, a commonly used template for short answer generation of LVLM model. For InstructBLIP, we follow its native *Short answer* scheme, which explicitly separates the image placeholder from the question. AMBER is designed as an open-ended description benchmark, so we keep its original single-sentence prompt. See Table 5 for detail.

| Dataset | Model | Template |
|---|---|---|
| POPE / MME | LLaVA-1.5 | `<question>\n Please answer the question using a single word or phrase.` |
| POPE / MME | InstructBLIP | `<ImageHere> <question> Short answer:` |
| AMBER | All | `Describe this image.` |

Table 5: Prompt templates used for each dataset–model pair. All baselines and our method use the identical text prompt.

| Model | Perception Score (↑) | | |
|---|---|---|---|
| | Regular | VCD | **Ours** |
| LLaVA1.5 | 1277.6 | 1338.2 | **1500.4** (+162.2) |
| InstructBLIP | 1050.9 | 1202.2 | **1240.73** (+38.53) |

Table 6: Evaluation of hallucination using various models and decoding methods on the coarse-grained perception subset of MME (Fu et al., 2024) benchmark. The best performances are **bolded**.

| Model | Dataset (Split) | $\lambda$ | Gain Search | Candidate $k$ |
|---|---|---|---|---|
| LLaVA-1.5 | MSCOCO (Random) | 1.0 | change | 20 |
| | MSCOCO (Popular) | 1.0 | change | 20 |
| | MSCOCO (Adversarial) | 1.0 | change | 20 |
| | AOK-VQA (Random) | 0.5 | log | 20 |
| | AOK-VQA (Popular) | 0.5 | log | 20 |
| | AOK-VQA (Adversarial) | 0.5 | log | 20 |
| | GQA (Random) | 0.1 | log | 20 |
| | GQA (Popular) | 0.1 | log | 20 |
| | GQA (Adversarial) | 0.1 | log | 20 |
| | MME (–) | 0.5 | change | 20 |
| | AMBER (–) | 0.5 | log | 20 |
| InstructBLIP | MSCOCO (Random) | 0.3 | change | 20 |
| | MSCOCO (Popular) | 0.3 | change | 20 |
| | MSCOCO (Adversarial) | 0.3 | change | 20 |
| | AOK-VQA (Random) | 0.3 | change | 20 |
| | AOK-VQA (Popular) | 0.3 | change | 20 |
| | AOK-VQA (Adversarial) | 0.3 | change | 20 |
| | GQA (Random) | 0.3 | change | 20 |
| | GQA (Popular) | 0.3 | change | 20 |
| | GQA (Adversarial) | 0.3 | change | 20 |
| | MME (–) | 1.0 | log | 10 |
| | AMBER (–) | 0.5 | log | 20 |

Table 7: Hyper-parameters for all dataset–scenario combinations. A single configuration per dataset is reused across splits to enable consistent comparison.

| Method | Latency (s) (↓) | Throughput (tokens/s) (↑) |
|---|---|---|
| LLaVA-1.5-7B | $0.17 \pm 0.06$ | $32.89 \pm 3.68$ |
| + VCD | $0.56 \pm 0.03$ | $17.97 \pm 0.83$ |
| + AVISC | $0.28 \pm 0.07$ | $15.93 \pm 1.45$ |
| + VCD (Ours) | $0.38 \pm 0.01$ | $26.88 \pm 0.58$ |

Table 8: Comparison with the baseline Contrastive Decoding methods for the Latency and Throughput.

# C Additional Implementation Details

## C.1 Hardware and Software Environment

All experiments with LLaVA v1.5 were conducted using PyTorch 2.1.2, CUDA 12.1, while InstructBLIP experiments relied on PyTorch 2.0.1, CUDA 11.7. The two configurations reflect the official code bases: LLaVA (Liu et al., 2024a) and OPERA (the reference implementation of InstructBLIP) (Huang et al., 2024). Unless otherwise noted, inference and evaluation were run on a single NVIDIA RTX A6000 (48 GB). Experiments with DeepSeek-VL2-Tiny were executed on an NVIDIA H100 NVL.

## C.2 Hyper-parameter Configuration

Table 7 lists the hyper-parameters used for every dataset–scenario–model combination. For each dataset we fix a single configuration and reuse it across the Random, Popular, and Adversarial splits to ensure a fair comparison. Although tuning the parameters per sample or subset can yield higher scores, our objective here is to show that **visually grounded** tri-layer selection is feasible; achieving optimal performance is left to future work.

## C.3 Implementation on stronger backbone

We additionally evaluate our method on the AMBER benchmark using DeepSeek-VL2-Tiny, a Mixture-of-Experts model with a substantially stronger backbone than Vicuna-7B despite its smaller parameter count (3.37 B). For the VCD baseline (Leng et al., 2023), we follow the authors' recommendations and sweep $\alpha = 1.0$ while varying $\beta \in [0.2, 0.5]$; we report the best AMBER score obtained. For TCD, we treat the last eight decoder layers (of twelve) as candidates and select layer 4 as the visually grounded pivot, based on a preliminary sweep with a small watermarking subset.

# D Latency

We report decoding latency (seconds) and throughput (tokens per second, t/s; mean $\pm$ standard deviation) on the AMBER generation task. Eleven samples were drawn at random, and the first sample in each run was discarded to avoid warm-up bias. All methods were executed with their *official* implementations on a single NVIDIA H100 GPU, using a batch size of one and a maximum generation length of ten tokens. Our method evaluates $k = 20$ candidate layers per decoding step.

## E Discussion of Baseline Selection

As discussed in Section 4.2, we selected VCD, M3ID, and AVISC as our primary training-free contrastive decoding baselines, and included ICD, EOS (Yue et al., 2024), HA-DPO (Zhao et al., 2024), HALVA (Sarkar et al., 2025a), and Octopus as reference methods that require additional training or external modules. Nonetheless, there exist other notable variations in decoding-based approaches for mitigating hallucinations in LVLMs. For example, PAI (Liu et al., 2024b) proposes a method similar to VCD, introducing visual perturbations to strengthen visual input, while ConVis (Park et al., 2025) leverages SDXL, a text-to-image model, to further ground LVLMs using generated images.

Given the diversity of possible experimental setups—such as model choices (e.g., LLaVA-1.5, InstructBLIP, QwenVL (Bai et al., 2023b), MiniGPT (Zhu et al., 2023), and Shikra (Chen et al., 2023)), benchmarks and evaluation metrics (e.g., POPE-MSCOCO, POPE-OKVQA, POPE-GQA, MME-Perception (example on Fig. 7), MME-Cognition, AMBER, CHAIR, MMVP (Zhong et al., 2023), and MMbench (Liu et al., 2024c)), we aimed to align our experimental design with the conventions established by recent works such as Octopus(Suo et al., 2025) and AVISC(Woo et al., 2024). For instance, although reporting results for each POPE subset independently could highlight the strengths of our method, we chose to aggregate all POPE subsets into a single evaluation to provide a fair and comprehensive comparison, as recommended by recent literature.

**Question:** Describe this image.

**Base:**
The image depicts a group of four people walking together in a grassy field. They are enjoying a leisurely stroll, possibly exploring the area or taking a break from their daily routine. The group is spread out, with one person on the left side, another in the middle, and the remaining two on the right side of the image.

**VCD:**
The image features a group of people walking through a lush, grassy field. There are five people in the scene, spread out across the field as they enjoy the outdoor activity. A frisbee can be seen on the ground near the middle of the scene, suggesting that they may have been playing with it earlier.

**AVISC:**
The image features a group of people strolling through a grassy field on a lovely day. All individuals are wearing shorts and enjoying their walk outdoors. One person toward the front is carrying a cell phone while at the same time leading the way.

**TCD(Ours):**
The image depicts a group of three people walking together in a grassy field, enjoying the outdoors. They are walking in a line, with one person slightly ahead of the others. The field is surrounded by trees, creating a serene and natural atmosphere.

**Question:** Describe this image.

**Base:**
The image features a woman running on a sandy beach, wearing a white shirt and a bikini. She is running with determination, possibly participating in a beach run or a fitness event. The woman's shadow is visible on the sand, indicating her movement. The beach setting is further emphasized by the presence of a cell phone in the scene, possibly belonging to the woman or someone else nearby.

**VCD:**
The image features a woman running on the beach, enjoying a sunny day. She is wearing a white tank top and appears to be running towards the water, likely enjoying the beautiful scenery around her. Her shadow can be seen on the sand, indicating the sun's position and her movement. In addition to the main woman, there is another person standing on the beach, slightly in the background and to the right side of the frame.

**AVISC:**
The image features a woman running on the beach wearing a pair of shorts and running shoes. She is in a line, jogging quickly, and engaging in a workout near the water. Additionally, there is a backpack in the scene, likely belonging to the woman or someone she is with. It is located to the side of her, possibly used for carrying her belongings or her running necessities while she enjoys her time on the beach.

**TCD(Ours):**
The image features a woman running on a sandy beach, enjoying her time outdoors. She is wearing a white top and appears to be in motion, possibly jogging or running. The beach is a wide-open space, providing ample room for her to run freely. The scene captures the essence of a sunny day at the beach, with the woman taking advantage of the pleasant weather to engage in physical activity.

**Question:** Describe this image.

**Base:**
The image features a young man walking down a path in a park. He is wearing a black shirt and a black hat, and he is smoking a cigarette. The man appears to be enjoying his time outdoors, possibly taking a break from his daily routine. The park setting is evident from the presence of a bench located near the path, providing a place for people to sit and relax.

**VCD:**
In the image, there is a man walking down a path in a park or along a street. He is wearing a black hat, black shirt, and jeans, and he is smoking a cigarette. The path is lined with green grass and trees, creating a pleasant outdoor atmosphere. Aside from the main subject, there are two other people in the background, but they are not the main focus of the image.

**AVISC:**
The scene captures an outdoor walking path where two men are walking. One man is wearing an all-black outfit with a beanie on; he appears to be closer to the camera in the foreground. Another man can be seen with a watch and is located further away from the camera. Additionally, there is a car parked on the side of the path.

**TCD(Ours):**
The image features a young man walking down a path in a park, wearing a black hat and black clothing. He appears to be enjoying his time outdoors, possibly taking a stroll or engaging in some form of exercise. The park is surrounded by trees, providing a serene and natural environment for the man to walk through.

Figure 8: More examples generated by Base, VCD (Leng et al., 2023), AVISC (Woo et al., 2024) and ours given a question, such as "Describe this image." We extract random samples from Amber (Wang et al., 2023) and all results are based on LLaVA v1.5 7b. We highlight phrases with red if it is *not* well visually grounded, and with blue if it is well visually grounded. We observe that our model successfully mitigates hallucinations compared to other three baselines.

| Method | Random | | Popular | | Adversarial | | ALL | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 |
| LLaVA-1.5-7B | 83.77 | 81.94 | 82.57 | 80.86 | 79.77 | 78.47 | 82.04 | 80.42 |
| +ICD | 87.51 | 83.28 | 83.15 | 83.91 | 79.13 | 80.41 | 83.26 | 82.53 |
| +ConVis | 84.70 | – | 83.20 | – | 81.10 | – | 83.00 | – |
| +OPERA | 84.40 | – | 83.40 | – | 81.20 | – | 83.00 | – |
| +VCD | 85.43 | 83.99 | 83.17 | 81.94 | 80.27 | 79.49 | 82.96 | 81.81 |
| +M3ID$^\dagger$ | 86.13 | 81.85 | 82.07 | 80.77 | 79.50 | 78.15 | 82.57 | 80.26 |
| +AVISC | 84.67 | 82.21 | 83.67 | 81.27 | 81.83 | 79.55 | 83.39 | 81.01 |
| +Octopus | 87.51 | 85.40 | 85.20 | 84.19 | 82.22 | 81.44 | 85.79 | 83.44 |
| TCD (Ours) | **89.50** | **88.89** | **87.60** | **87.14** | **83.90** | **83.92** | **87.00** | **86.65** |
| InstructBLIP | 81.53 | 81.19 | 78.47 | 78.75 | 77.43 | 78.00 | 79.14 | 79.31 |
| +ICD | 84.36 | 83.82 | 77.88 | 78.70 | 75.17 | 77.23 | 79.14 | 79.92 |
| +OPERA | 84.57 | 83.74 | 78.24 | 79.15 | 74.59 | 76.33 | 79.13 | 79.74 |
| +VCD | 82.03 | 81.56 | 79.13 | 79.20 | 77.23 | 77.72 | 79.46 | 79.49 |
| +M3ID$^\dagger$ | 82.33 | 81.53 | 80.90 | 80.42 | 78.53 | 78.49 | 80.59 | 80.15 |
| +AVISC | 86.03 | 84.41 | 84.27 | 82.77 | 81.83 | 80.67 | 84.04 | 82.62 |
| +Octopus | 86.63 | 85.30 | **84.90** | **83.55** | **82.83** | **81.43** | **84.79** | 83.43 |
| TCD (Ours) | **88.40** | **87.63** | 82.77 | 82.67 | 81.13 | 81.33 | 84.10 | **83.88** |

Table 9: Comparison with the state-of-the-art methods for the discriminative tasks on the POPE_MSCOCO dataset.

| Method | Random | | Popular | | Adversarial | | ALL (Avg.) | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 |
| LLaVA-1.5-7B | 82.73 | 84.26 | 76.10 | 79.34 | 67.90 | 74.09 | 75.58 | 79.23 |
| +ICD | - | - | - | - | - | - | | |
| +OPERA | - | - | - | - | - | - | | |
| +VCD | 81.30 | 83.23 | 75.43 | 79.26 | 67.43 | 74.11 | 74.72 | 78.87 |
| +M3ID$^\dagger$ | 83.57 | 85.09 | 76.80 | 80.06 | 68.10 | 74.58 | 76.16 | 79.91 |
| +AVISC | 84.60 | 85.88 | 78.83 | 81.63 | 68.97 | 75.11 | 77.47 | 80.87 |
| +Octopus | - | - | - | - | - | - | | |
| TCD (Ours) | **91.23** | **91.12** | **87.57** | **87.86** | **80.57** | **82.24** | **86.46** | **87.07** |
| InstructBLIP | 81.00 | 82.06 | 75.00 | 77.69 | 68.80 | 73.84 | 74.93 | 77.86 |
| +ICD | - | - | - | - | - | - | | |
| +OPERA | - | - | - | - | - | - | | |
| +VCD | 81.73 | 82.66 | 75.33 | 77.92 | 69.70 | 74.27 | 75.59 | 78.28 |
| +M3ID$^\dagger$ | 82.33 | 83.66 | 75.60 | 78.36 | 69.57 | 74.39 | 75.83 | 78.80 |
| +AVISC | **88.47** | **88.59** | 81.77 | 82.98 | 72.53 | 76.28 | 80.92 | 82.62 |
| +Octopus | - | - | - | - | - | - | | |
| TCD (Ours) | 88.00 | 88.36 | **84.03** | **85.08** | **76.60** | **79.56** | **82.88** | **84.33** |

Table 10: Comparison with the state-of-the-art methods for the discriminative tasks on the A-OKVQA dataset.

| Method | Random | | Popular | | Adversarial | | ALL (Avg.) | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc. | F1 |
| LLaVA-1.5-7B | 82.40 | 83.99 | 72.03 | 76.84 | 68.73 | 74.92 | 74.39 | 78.58 |
| +ICD | – | – | – | – | – | – | – | – |
| +OPERA | – | – | – | – | – | – | – | – |
| +VCD | 82.27 | 84.22 | 71.77 | 77.05 | 68.27 | 74.84 | 74.10 | 78.70 |
| +M3ID$^\dagger$ | 82.83 | 84.62 | 72.83 | 77.58 | 68.13 | 74.78 | 74.60 | 78.99 |
| +AVISC | <u>85.00</u> | <u>86.45</u> | <u>74.80</u> | <u>79.17</u> | <u>69.20</u> | <u>75.58</u> | <u>76.33</u> | <u>80.40</u> |
| +Octopus | – | – | – | – | – | – | – | – |
| TCD (Ours) | **88.90** | **88.43** | **85.57** | **85.46** | **81.93** | **82.44** | **85.47** | **85.44** |
| InstructBLIP | 80.00 | 81.02 | 73.53 | 76.49 | 68.00 | 72.59 | 73.84 | 76.70 |
| +ICD | – | – | – | – | – | – | – | – |
| +OPERA | – | – | – | – | – | – | – | – |
| +VCD | 81.73 | 82.45 | 74.10 | 76.87 | 70.27 | 74.29 | 75.36 | 77.87 |
| +M3ID$^\dagger$ | 80.57 | 81.85 | 74.57 | 77.53 | 68.90 | 73.47 | 74.68 | 77.62 |
| +AVISC | <u>86.47</u> | <u>86.57</u> | <u>78.00</u> | <u>79.84</u> | <u>73.07</u> | <u>76.54</u> | <u>79.85</u> | <u>80.98</u> |
| +Octopus | – | – | – | – | – | – | – | – |
| TCD (Ours) | **86.57** | **86.79** | **80.17** | **81.65** | **76.13** | **78.72** | **80.96** | **82.39** |

Table 11: Comparison with the state-of-the-art methods for the discriminative tasks on the GQA dataset.