# Beyond Single Models: Mitigating Multimodal Hallucinations via Adaptive Token Ensemble Decoding

Anonymous ACL submission

## Abstract

Large Vision-Language Models (LVLMs) have recently achieved impressive results in multimodal tasks such as image captioning and visual question answering. However, they remain prone to **object hallucination**—generating descriptions of nonexistent or misidentified objects. Prior work have partially mitigated this via auxiliary training objectives or external modules, but often lacks scalability, adaptability, or model independence. To address these limitations, we propose **Adaptive Token Ensemble Decoding (ATED)**, a training-free, token-level ensemble framework that mitigates hallucination by aggregating predictions from multiple LVLMs during inference. ATED dynamically computes uncertainty-based weights for each model, reflecting their reliability at each decoding step. It also integrates diverse decoding paths to improve contextual grounding and semantic consistency. Experiments on standard hallucination detection benchmarks demonstrate that ATED significantly outperforms state-of-the-art methods, reducing hallucination without compromising fluency or relevance. Our findings highlight the benefits of adaptive ensembling and point to a promising direction for improving LVLM robustness in high-stakes vision-language applications. Code is available at https://anonymous.4open.science/r/ATED.

## 1 Introduction

In recent years, large language models (LLMs) have made significant breakthroughs in natural language processing (Touvron et al., 2023; Chiang et al., 2023; Achiam et al., 2023; Bai et al., 2023a) and have been increasingly extended to vision-language tasks, giving rise to large vision-language models (LVLMs) (Ye et al., 2023; Liu et al., 2023; Li et al., 2023a, 2024; Chen et al., 2024c,d; Bai et al., 2023b). These models have demonstrated strong capabilities in both understanding (Zhang et al., 2025; Lai et al., 2024) and generating (Geng et al., 2023) multimodal content.

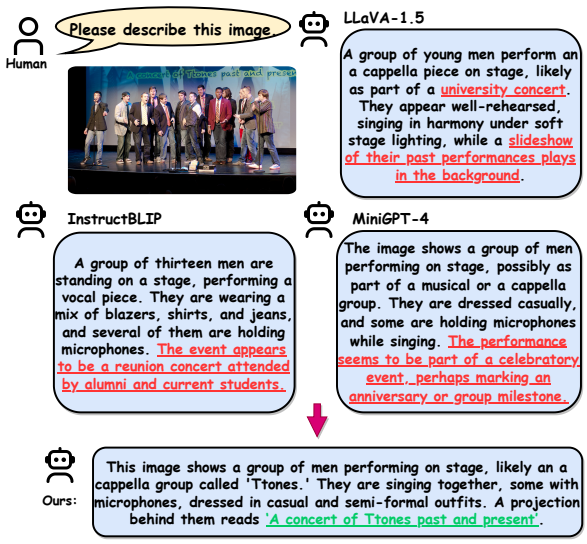


Figure 1: Comparison of image-description generation results from various LVLMs and our proposed *ATED* method. Red text indicates hallucination. Green text represents hallucination mitigating from *ATED*.

However, LVLMs often suffer from the problem of *object hallucination*, where the model generates details or objects that do not exist in the image (Li et al., 2023d; Wang et al., 2023; Gunjal et al., 2024; Liu et al., 2024b), significantly limiting their reliability in high-stakes applications, such as autonomous driving, medical image analysis, and remote sensing, where factual correctness and visual grounding are critical.

Early research on mitigating hallucinations primarily focused on enhancing data quality and training paradigms. Specifically, diverse instruction-tuning datasets and multi-task training approaches were introduced to reduce the models' tendency to hallucinate during generation (Li et al., 2023c; Liu et al., 2024a). Other methods adopted post-hoc strategies by implementing output-checking mechanisms to detect and correct hallucinated con-

tent (Yin et al., 2024; Zhou et al., 2024).

More recently, training-free approaches have emerged (Leng et al., 2023; Wang et al., 2024; Huang et al., 2024). However, many of these methods rely on additional annotations, large-scale fine-tuning, or complex inference, incurring substantial human and computational costs. Furthermore, single-model strategies are inherently limited by the knowledge scope of the underlying model, restricting generalization and adaptability. As shown in Figure 1, existing LVLMs exhibit notable levels of hallucination in image captioning tasks, highlighting the need for a more robust solution.

Ensemble learning (Polikar, 2012), which leverages the collective intelligence of multiple models, has proven highly effective at reducing errors and enhancing robustness in traditional classification and regression tasks (Mienye and Sun, 2022). More recently, it has been successfully extended to text generation tasks—particularly in LLMs—to improve output accuracy and mitigate issues such as hallucination (Jiang et al., 2023; Wan et al., 2024). Inspired by these advances, this paper presents a novel framework that integrates ensemble learning with diversified decoding strategies within the autoregressive generation process, harnessing the complementary strengths of different LVLMs. By aggregating outputs from multiple models or decoding paths, our approach not only improves the factual consistency and coherence of generated content, but also reduces the disparity in hallucination tendencies across models—ultimately enhancing the generalization and adaptability of LVLMs in multimodal tasks.

To this end, we introduce **Adaptive Token Ensemble Decoding (ATED)**—the first fine-grained, token-level ensemble strategy for multimodal LVLMs. ATED is a training-free framework that performs parallel inference across multiple LVLMs and aggregates their output logits at the token level. We estimate each model's hallucination tendency via output uncertainty, and use a greedy optimization algorithm to derive adaptive importance weights by minimizing overall uncertainty. Furthermore, ATED incorporates diverse decoding paths, substantially improving the factual accuracy and reliability of generated outputs while maintaining strong adaptability across diverse scenarios.

Our main contributions are as follows:

- We propose **ATED**, a training-free multimodal ensemble decoding method that mitigates hallucinations via fine-grained token-level fusion.
- We introduce an uncertainty-minimization weighting mechanism that dynamically assigns weights based on model confidence, improving the reliability of ensemble decoding.
- Extensive experiments show that ATED consistently outperforms existing methods across multiple multimodal benchmarks, achieving superior accuracy and robustness.

## 2 Related Work

**Hallucination in LVLMs** Hallucination was initially observed in LLMs, referring to generated content that deviates from factual knowledge or user intent (Jing et al., 2024; Liu et al., 2024b). Large vision-language models (LVLMs) (Bai et al., 2025; Zhang et al., 2023), which extend LLMs with visual inputs, also exhibit hallucinations—typically manifesting as mismatches between generated text and visual content. Existing studies categorize hallucinations in LVLMs into three main types: *object hallucination* (Biten et al., 2021; Li et al., 2023d; Rohrbach et al., 2019), *attribute hallucination*, and *relationship hallucination* (Wu et al., 2024; Zhou et al., 2024). Object hallucination refers to fabricated or omitted objects; attribute hallucination involves incorrect properties such as color or size; relationship hallucination describes inaccurate relations among objects. These errors may arise from visual misinterpretation, flawed reasoning, or over-reliance on language priors.

**Hallucination Mitigation in LVLMs** To address hallucination in LLMs and LVLMs, researchers have proposed a range of solutions, including improved instruction tuning (Jiang et al., 2024; Liu et al., 2024a; Yu et al., 2024a; Yue et al., 2024), reinforcement learning with human or AI feedback (Gunjal et al., 2024; Kim et al., 2024; Li et al., 2023b; Sun et al., 2023; Yu et al., 2024b,c), retrieval augmentation, and structural model enhancements (Zhai et al., 2024). More recently, several training-free decoding strategies have been developed to suppress hallucination in LVLMs. For example, conservative decoding methods applied to both original and perturbed inputs (Chen et al., 2024b; Favero et al., 2024; Huo et al., 2025; Leng et al., 2023; Wang et al., 2024; Woo et al., 2024) aim to reduce overreliance on language priors. Techniques such as input distortion—applied

to either visual content or instructions—amplify hallucinations to better identify and suppress them through contrastive decoding. Token-level pruning and related approaches (Favero et al., 2024; Woo et al., 2024) also manipulate visual inputs to mitigate hallucinations.

While existing methods mitigate hallucinations to some extent from various angles, they often suffer from limitations in scalability, generalizability, and practical deployment. This paper, therefore, focuses on reducing hallucinations in LVLMs without requiring additional training or complex reasoning pipelines, with an emphasis on adaptability to real-world applications.

## 3 Methodology

### 3.1 Preliminaries of LVLMs Generation

The generation mechanism of LVLMs can be deconstructed into three core modules: Vision Language Input, Model Forward Propagation, and Next Token Decoding.

**Vision Language Input.** LVLMs take both visual and textual inputs. Typically, raw images are processed by a vision encoder (e.g., a pretrained visual backbone), and the resulting features are projected into the input space of the language decoder via a cross-modal interface. These visual features are represented as visual tokens $V = \{v_1, v_2, ..., v_n\}$, where $n$ is the number of visual tokens. Similarly, the textual input is tokenized into text tokens $T = \{t_1, t_2, ..., t_m\}$ using a tokenizer, where $m$ is the number of text tokens. The visual and text tokens are then concatenated to form the final input sequence, denoted as $\{x_i, i \in [0, n+m-1]\}$.

**Model Forward Propagation.** A large language model (LLM) parameterized by $\phi$, such as Vicuna (Chiang et al., 2023), generates responses by conditioning on both the text and visual context. Following the autoregressive generation paradigm, LVLMs predict the probability of the next token $x_t$ at time step $t$ based on the previously generated tokens, the input text, and the visual features, over the vocabulary set $\boldsymbol{V}$. This process can be formally expressed as:

$$p(x_t \mid v, t, x_{<t}) = \text{softmax}(\text{logit}_\phi(x_t \mid v, t, x_{<t})), \quad x_t \in v \quad (1)$$

where $x_t$ denotes the token at time step $t$, $x_{<t}$ represents the sequence of previously generated tokens.

**Next Token Decoding.** Based on the predicted probabilities $p(x_t|v, t, x_{<t})$, various decoding strategies—such as greedy decoding, beam search, and contrastive decoding (e.g., VCD)—can be applied to generate output. While these strategies can marginally reduce hallucinations, they are typically restricted to single-model outputs, cannot leverage external knowledge, and fail to fully exploit complementary strengths across different models. As a result, they remain prone to errors, especially in open-domain scenarios. In contrast, our method adaptively fuses the token-level logits from multiple LVLMs that share the same vocabulary, immediately after the forward pass. By leveraging the diverse capabilities of different models, our approach more effectively mitigates hallucinations in both general-purpose and task-specific settings.

### 3.2 Adapative Token Ensemble Decoding

To leverage the complementary strengths of diverse LVLMs and enhance general task performance while reducing hallucinations, we propose a token-level ensemble decoding approach. Specifically, we introduce **Adaptive Token Ensemble Decoding (ATED)**, a training-free, uncertainty-guided fusion method that dynamically integrates multiple LVLMs at inference time. The overall framework is illustrated in Figure 2.

Given two or more LVLMs $\{M_i\}$ at test time, ATED fuses their output logits using adaptive importance weights $\{\lambda_1, \ldots, \lambda_M\}$, where each weight $\lambda_i$ reflects how well model $M_i$ interprets the visual and textual inputs. At each decoding step $t$, each model $M_i$ takes the visual token sequence $V = \{v_1, v_2, \ldots, v_n\}$ and the text token history $T = \{t_1, t_2, \ldots, t_{t-1}\}$ to generate a logit score $p_i = p_i(x_t \mid V, T)$ over the vocabulary. Assuming all models share the same vocabulary, the logits from the $i$-th model are computed as:

$$p(x_t|v, t, x_{<t})^{\text{orig},i} = LVLMs_i(X_{<t})\big|_t, \quad (2)$$

where $i = 1, \ldots, M$ indexes the models, and $p \in R^v$ with $v$ vocabulary size. ATED combines output logits to obtain the final probabilities as:

$$p(x_t|v, t, x_{<t}) = \Big(\sum_{i=1}^{M} \underbrace{\lambda_i}_{unknown} \; p_i\Big), \quad (3)$$

where $p_i$ is the decoding logits of model $M_i$. We assume that the weights $\lambda_i \in [0, 1]$ are normalized, i.e., $\sum_{i=1}^{M} \lambda_i = 1$. To further enhance robustness, we additionally introduce multiple perturbed

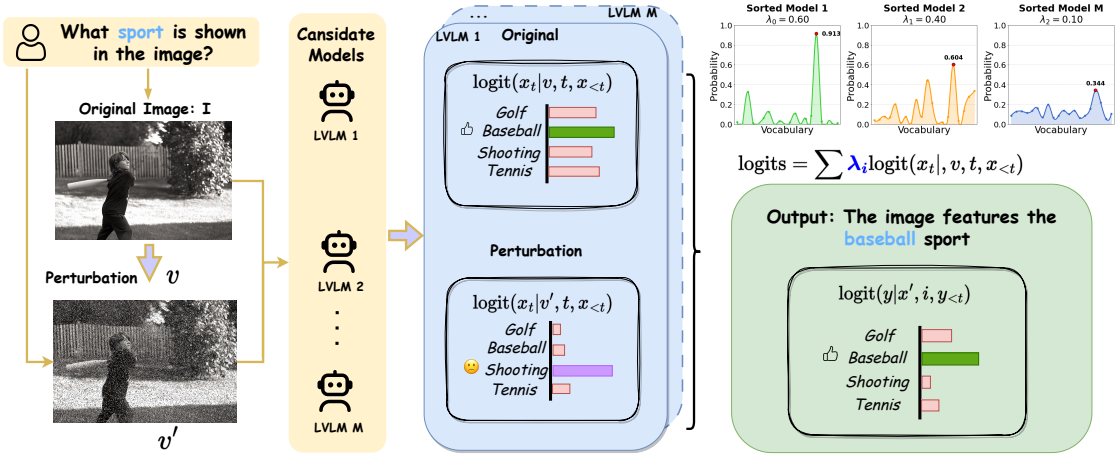


Figure 2: Overall pipeline of **Adapative Token Ensemble Decoding (ATED)**. Given a set of candidate LVLMs, a system instruction (inst), the original image, and its perturbed variants, the system produces multiple output streams. We ensemble the logits from each LVLM using uncertainty-guided weights and employ greedy uncertainty optimization to generate the final output. The entire process is dynamically repeated at each time step $t$ of token generation, ensuring high reliability and consistency of the generated results.

variants of the original image $I$ and employ contrastive decoding to effectively mitigate the backbone model's hallucinations. Formally, contrastive decoding can be expressed as:

$$p_{i,t} = \text{softmax}\Big[(1+\alpha)\ \text{logit}_\phi\big(x_t \mid v, t, x_{<t}\big) - \alpha\ \text{logit}_\phi\big(x_t \mid v_d, t, x_{<t}\big)\Big], \quad (4)$$

where $d$ and $\alpha$ indicate distortion operation and vision contrastive hyperparameter, respectively.

Inspired by Chen et al. (2024a); Qiu et al. (2025); Dey et al. (2025), we propose utilizing the entropy of probability distributions based on visual features as an uncertainty metric when LVLMs generate the next textual token. This metric reflects the model's token prediction confidence and associated weights $\lambda_i$ under current multimodal inputs. The metric aligns with the training objectives of causal language modeling. By analyzing prediction entropy under visual conditions, we can evaluate the model's depth of understanding of specific visual content, the quality of vision-language alignment, and the distributional differences between the visual input and the model's training data (Gonen et al., 2024). This cross-modal uncertainty analysis provides a novel perspective for assessing the generalization capabilities of multimodal large models in open-world scenarios.

### 3.3 Uncertainty-Guided Weight

**Uncertainty Minimization.** Given a tokenized input $X$ at time $t$, its uncertainty score is defined as the distribution entropy of $X$ at that time, with the following formula as:

$$H_{i,t}(X) = -\sum_{x_t \in \mathcal{V}} P_{i,t}\ \log P_{i,t}, \quad (5)$$

where $P_{i,t}$ represents the normalized probability for the $i$-th token corresponding to vocabulary $\boldsymbol{V}$, conditioned on the preceding tokens $x_{<t}$ according to model $M_i$.

We formulate the assignment of importance weights across $M$ models as an optimization problem that requires no training or labeled data, and can be solved directly during next token prediction. Formally, our optimization framework is defined as follows:

$$\lambda_1^*, \ldots, \lambda_i^* = \arg\min_{\lambda_1,\ldots,\lambda_i} -\sum p(x_i|x_{<i}) \log p(x_i|x_{<i}),$$
$$p(x_i|x_{<i}) = \text{softmax} \sum_{i=0}^{M} \lambda_i p_i(x_t|v, t, x_{<t}), \quad (6)$$

where the weights $\lambda_i$ are inversely proportional to each model's normalized uncertainty score—i.e., models with lower uncertainty are assigned higher weights. All weights are constrained such that $\sum_{i=1}^{M} \lambda_i = 1$ and $\lambda_i \in [0, 1]$.

**Uncertainty Greedy Optimization.** To address the uncertainty minimization problem proposed in Equation 6, we introduce an efficient greedy optimization algorithm that incrementally ensembles LVLMs. Specifically, we first compute the uncertainty of each LVLM's next-token prediction using Equation 5, and then sort the LVLMs $M_1, \ldots, M_i$

based on their uncertainties scores. Let the sorted models be denoted as:

$$[M_1^*, M_2^*, \ldots, M_i^*] = \text{argsort}\big(H_{1,t}, \ldots, H_{i,t}\big), \tag{7}$$

where $[M_1^*, M_2^*, \ldots, M_i^*]$ are ordered by the lowest to highest uncertainty scores, and set the weight of the top-ranked model to $\lambda_1^* = 1$ and $\lambda_{i>1}^* = 0, i = \{2, \ldots, M\}$.

We then iteratively consider incorporating the next-ranked model $M_{i+1}^*$ into the current ensemble. A grid search is performed over interpolation weights $\lambda_i$ with the step values between the two models, and the fused logits are defined as:

$$p_\lambda\big(x_i \mid x_{<i}\big) = \lambda\, p^{(i)}\big(x_i \mid x_{<i}\big) + (1-\lambda)\, p^{(i+1)}\big(x_i \mid x_{<i}\big). \tag{8}$$

Finally, we compute the uncertainty score under each interpolation ratio with Equation 6, and select the weight that minimizes the uncertainty score. The same procedure is repeated to iterate through all LVLM for the next step and iteratively update the ensemble logits and weight allocation accordingly. We can perform early stopping when we find $\lambda(i) = 1$, i.e., the effect of the current LVLM is zero. The final generation is performed using the output probabilities from the adapative uncertainty-guided ensemble.

# 4 Experiments

## 4.1 Experimental Settings

### 4.1.1 Datasets

We evaluate our model on three datasets as listed below. More details are shown in the Appendix C.

**POPE (Probability of Object Presence Estimation).** Li et al. (2023d) is a benchmark dataset for evaluating object hallucination in LVLMs. POPE integrates the MSCOCO (Lin et al., 2015), A - OKVQA (Schwenk et al., 2022), and GQA (Hudson and Manning, 2019) datasets to form 27,000 query - answer pairs for evaluation. Performance is quantified using standard metrics, including *accuracy*, *precision*, *recall*, and *F1 score*.

**CHAIR(The Caption Hallucination Assessment with Image Relevance).** Rohrbach et al. (2019) is a dataset for evaluating object hallucination in image captioning. CHAIR has two main variants: $CHAIR_i$ and $CHAIR_s$, focusing on instance and sentence levels, respectively.

**MME (Multimodal Large Language Model Evaluation).** Fu et al. (2024) assesses the performance of LVLMs in terms of two core capabilities: perception and cognition. In our evaluation, we focus on four representative sub-tasks: object existence, counting, position, and color. Model performance is measured using the *accuracy+* metric.

### 4.1.2 Models

We integrate our proposed method with four popular LVLMs: InstructBLIP (Dai et al., 2023), MiniGPT-4 (Zhu et al., 2023), LLaVA-1.5 (Liu et al., 2024c), and LLaVA-Next (Liu et al., 2024d). All the LVLMs used have a language model size of 7 billion parameters (7B). InstructBLIP and MiniGPT-4 utilize a Q-former(Li et al., 2023a), which represents an image using only 32 tokens, effectively bridging the visual and textual modalities. LLaVA-1.5 and LLaVA-NeXT employ a linear projection layer to align features from the two modalities. All LVLMs adopt pre-trained vision encoders such as the CLIP vision encoder(Radford et al., 2021), along with pre-trained large language models (LLMs) as language decoders, such as LLaMA(Touvron et al., 2023) or Vicuna v1.1(Chiang et al., 2023). Complete experimental details are provided in the Appendix. A.1

### 4.1.3 Baselines

For the object hallucination evaluation, we employ several widely used decoding strategies, such as multinomial sampling (default), greedy decoding, and four state-of-the-art training-free decoding methods. Greedy decoding selects tokens step by step by always choosing the one with the highest probability from the language model's logits. Based on greedy decoding, beam search maintains a set of beams to explore a wider range of candidates and eventually selects the best one among them. OPERA(Huang et al., 2024) is an improved method based on beam search ,it alleviates hallucination by penalizing specific patterns of knowledge aggregation. VCD(Leng et al., 2023) reduces hallucination by decoding with noisy images in a contrastive manner. ICD(Wang et al., 2024) mitigates hallucination by designing negative prompts to interfere with the visual inputs during contrastive decoding. SID (Huo et al., 2025) mitigateas hallucinations by introspectively filtering low-relevance visual signals during generation. For all the baselines, we use the default hyperparameters provided by their original source code to ensure a fair com-

| Model | Method | Random | | Popular | | Adversarial | |
|---|---|---|---|---|---|---|---|
| | | Accuracy | F1 Score | Accuracy | F1 Score | Accuracy | F1 Score |
| **LLaVA-1.5** | default | 83.86 | 82.68 | 80.82 | 79.54 | 76.42 | 76.61 |
| | OPERA | 88.85 | 88.67 | 82.77 | 83.40 | 79.16 | 80.93 |
| | VCD | 87.2 | 87.17 | 83.08 | 83.07 | 77.70 | 79.14 |
| | ICD | 83.15 | 83.91 | 83.15 | 83.91 | 79.13 | 80.41 |
| | SID | 89.46 | 89.62 | 85.13 | **85.94** | **83.24** | **82.21** |
| **InstructBLIP** | default | 81.44 | 81.21 | 79.06 | 79.12 | 76.29 | 76.99 |
| | OPERA | 84.57 | 83.74 | 78.24 | 79.15 | 74.59 | 76.33 |
| | VCD | 84.91 | 84.08 | 81.89 | 81.46 | 79.97 | 79.90 |
| | ICD | 81.12 | 82.25 | 81.12 | 82.25 | 76.82 | 78.99 |
| | SID | 87.23 | 86.90 | 81.16 | 82.57 | 78.51 | 81.26 |
| **MiniGPT4** | default | 65.65 | 66.45 | 59.61 | 62.54 | 58.35 | 62.22 |
| | OPERA | 79.91 | 77.6 | 73.78 | 72.23 | 71.76 | 70.64 |
| | VCD | 67.79 | 68.54 | 62.42 | 65.24 | 60.17 | 63.94 |
| | ICD | 71.89 | 75.63 | 64.58 | 75.33 | 61.77 | 67.61 |
| | SID | 75.20 | 76.12 | 68.94 | 72.93 | 66.57 | 69.40 |
| **LLaVA-NeXT** | default | 84.83 | 81.78 | 81.00 | 79.72 | 76.01 | 75.83 |
| | OPERA | 88.41 | 87.33 | 82.69 | 83.48 | 79.22 | 79.40 |
| | VCD | 86.01 | 85.20 | 81.90 | 82.23 | 78.00 | 79.12 |
| | ICD | 82.14 | 82.09 | 81.95 | 81.87 | 79.24 | 78.89 |
| | SID | **89.54** | **89.67** | **85.24** | 85.67 | 82.43 | 81.51 |
| **Ensemble** | **ATED**$^{*}$ | 88.74 | 87.82 | 83.62 | 84.82 | 78.86 | 81.21 |
| | **ATED**$^{\&}$ | 89.21 | 89.39 | 85.32 | 85.66 | 81.51 | 82.32 |
| | **ATED**$^{\#}$ | **89.83** | **89.35** | **86.71** | **85.97** | **82.96** | **82.78** |

Table 1: **POPE evaluation results on different decoding strategies.** Results are from the papers or re-implemented based on official codes. ***Note:*** $*$ denotes ours ensemble method without vision contrastive decoding, $\&$ denotes ensemble with LLaVA-1.5 and InstructBLIP, $\#$ denotes ensemble with LLaVA-1.5, InstructBLIP and LLaVA-NeXT.

parison. We posit that our method, being LVLM-agnostic, can be easily integrated into various off-the-shelf LVLMs that share the same vocabulary.

### 4.2 Experimental Results

**Results on POPE.** We begin with the most widely adopted benchmark for evaluating object hallucination. Table 1 reports the average performance across three evaluation settings—*random*, *popular*, and *adversarial*—on various datasets, where *Default* refers to the unmodified backbone model. Our evaluation of ATED includes three configurations: two distinct LVLM ensemble variants ($\mathbf{ATED}^{\&}$ and $\mathbf{ATED}^{\#}$), as well as a version ($\mathbf{ATED}^{*}$) based on the $\mathbf{ATED}^{\#}$ that excludes vision-contrastive decoding ($\mathbf{ATED}^{*}$). For clarity, we highlight the best baseline results for each backbone in bold. Compared to each respective backbone, ATED achieves improvements of 4.20%–6.29% in *Accuracy* and 6.29%–6.97% in F1-*score*. Furthermore, on both LLaVA-1.5 and InstructBLIP, ATED consistently surpasses state-of-the-art methods ICD and VCD, attaining additional gains ranging from 0.89% to 5.10% in *Accuracy* and 0.80% to 2.94% in F1-*score*, thereby effectively mitigating hallucination issues.

**Results on CHAIR.** Beyond the binary "yes" or "no" evaluations on the POPE benchmark, we further validate the effectiveness of TADE in open-ended image captioning using the CHAIR metric. Specifically, we randomly sample 500 images from the validation split of the MSCOCO dataset and query various LVLMs with the prompt, "Please describe this image in detail." As shown in Table , when setting the maximum new token length to 64, our proposed ATED method significantly outperforms all baseline decoding approaches on the CHAIR metric, achieving improvements of 81.57% and 1.23% over the strongest baseline, respectively. Notably, when increasing the generation length to 512 tokens, ATED still attains the best performance on the CHAIR metric, with an improvement of approximately 30.0%. More detailed result are

| Type | METHOD | LLaVA-1.5 | | InstructBLIP | | MiniGPT4 | | LLaVA-NeXT | |
|---|---|---|---|---|---|---|---|---|---|
| | | $CHAIR_S$ | $CHAIR_I$ | $CHAIR_S$ | $CHAIR_I$ | $CHAIR_S$ | $CHAIR_I$ | $CHAIR_S$ | $CHAIR_I$ |
| Single | Default | 24.8 | 8.9 | 30.3 | 13.9 | 19.8 | 8.5 | 24.26 | 8.51 |
| | OPERA | 21.8 | 8.2 | 28.4 | 9.7 | 22.6 | 8.2 | 21.33 | 7.73 |
| | VCD | 23.6 | 8.6 | 30.0 | 11.2 | 22.0 | 10.6 | 23.27 | 8.34 |
| | ICD | 21.0 | 8.7 | 21.8 | 8.2 | 20.0 | 8.7 | 20.59 | 8.54 |
| | SID | 20.7 | 8.4 | 20.7 | 8.4 | 23.1 | 10.7 | 19.37 | 7.83 |
| Ensemble | Ours | ATED$^{\&}$ | | | | ATED$^{\#}$ | | | |
| | | $CHAIR_S$ | | $CHAIR_I$ | | $CHAIR_S$ | | $CHAIR_I$ | |
| | | **15.3** | | **10.9** | | **11.4** | | **8.1** | |

Table 2: **CHAIR evaluation results on different decoding strategies**. Results are from the papers or re-implemented based on official codes. ***Note:*** & denotes ensemble with LLaVA-1.5 and InstructBLIP, # denotes ensemble with LLaVA-1.5, InstructBLIP and LLaVA-NeXT.

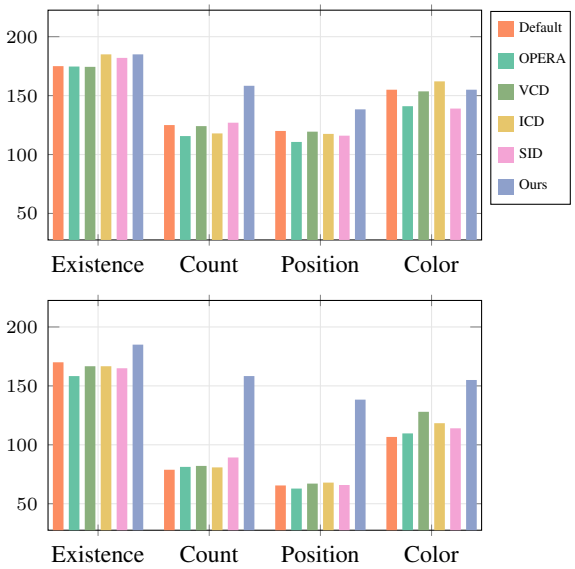


Figure 3: MME evaluation on hallucination subset with decoding strategies in LLaVA-1.5(up) and InstructBLIP(down).

| Setting | Method | Accuracy | F1 Score |
|---|---|---|---|
| **Random** | Uniform | 87.03 | 86.65 |
| | Confidence-based | 88.13 | 86.76 |
| | Ours (w/o UGO) | 89.30 | 87.78 |
| | Ours | 88.97 | 88.29 |
| **Popular** | Uniform | 84.57 | 84.37 |
| | Confidence-based | 86.27 | 84.79 |
| | Ours (w/o UGO) | 87.10 | 85.98 |
| | Ours | 87.57 | 86.58 |
| **Adversarial** | Uniform | 81.07 | 81.65 |
| | Confidence-based | 84.77 | 83.40 |
| | Ours (w/o UGO) | 85.17 | 84.25 |
| | Ours | 85.37 | 84.64 |

Table 3: Average results on the POPE COCO benchmark comparing various weights fusion method.

provided in Appendix E.

**Results on MME.** We extend the evaluation to include hallucinations at the object attribute level. We further conducted a systematic and comprehensive evaluation of the ATED method on the MME hallucination subset, which includes object-level tasks (existence identification and quantity judgment) and attribute-level tasks (location identification and color classification). As shown in Figure 4, ATED achieves the highest performance on location questions, and attains nearly perfect accuracy on existence-related questions. Our ATED method significantly outperforms the default LVLMs and other baseline methods across all four tasks (with total score improvements of at least +61.7 and +54.2 respectively).

In addition, we provide several qualitative cases that proves ATED strong ability on mitigating hallucinations. These cases uses the instructions with "Please describe this image in detail.", and details are provided in Appendix E.3.

### 4.3 Ablation Studies

**Adapative Uncertainty-Guided Weight.** To further validate the effectiveness of the adaptive uncertainty-guided weighting strategy, we conduct an extensive comparative analysis of various weighting approaches on the POPE benchmark. Specifically, we consider uniform weighting, confidence-based weighting, and uncertainty-guided weighting without the uncertainty greedy optimization (UGO) module. As presented in Table 3, our proposed ATED method delivers average improvements of 3.66% in Accuracy and 2.71% in F1 Score over the uniform weighting baseline. Moreover, when the UGO module is removed, the performance of the model ensemble deteriorates to varying extents, indicating that the lack of uncertainty-aware optimization impairs the

effectiveness of the ensemble strategy. These results clearly highlight the crucial role of adaptive uncertainty-guided weighting—particularly when enhanced by greedy optimization—in maximizing the performance gains of multimodal model ensembles. Overall, our findings provide strong empirical evidence that the proposed adaptive weighting strategy is fundamental for robust and effective multimodal integration.

**Impact of Vision Perturbations.** We further investigate the impact of visual perturbations on hallucination reduction in LVLM ensemble decoding across different tasks. Specifically, we conduct systematic experiments on the POPE-MSCOCO and MME benchmarks to evaluate the performance of dynamic model ensembles under various conditions, including the absence of visual perturbations (**Ours(0)**) and different levels of perturbation intensity. Experimental results in Table 4 demonstrate that, without adaptation to visual perturbations, the performance of multimodal ensemble reasoning significantly degrades on both the POPE and MME datasets—for example, *Accuracy* decreases by 1.4%, F1-*score* drops by 2.7%, and *Accuracy+* decreases by 20. These findings further highlight that introducing multi-path contrastive decoding under visual perturbations can effectively mitigate hallucinations and enhance reasoning performance.

Additionally, we conduct an ablation study on the amplification factor between the output distributions of original and perturbed visual inputs, denoted as the hyperparameter $\alpha$ in Equation 4, to further validate the effect of visual contrast on model performance. Detailed experimental results are provided in Appendix E.2.

### 4.4 Performance Comparison of LVLMs Ensemble Strategy.

To investigate the performance of different LVLMs ensemble strategies across various tasks, we conducted ensemble experiments on the MME and POPE benchmarks usin LLaVA-1.5, InstructBLIP, and LLaVA-NeXT. The results are presented in Table 5. Our experiments reveal that when the performance gap between models is large (e.g., InstructBLIP and LLaVA-1.5 exhibit over a 10% difference on the MME benchmark), simple uniform(U) of token probabilities across models fails to improve results and may even degrade overall performance due to noise introduced by lower-performing models. Conversely, when the performance gap is small

| Noise | POPE | | MME |
|---|---|---|---|
| | Accuracy | F1 Score | Accuracy+ |
| **Ours(0)** | 85.90 | 84.19 | 616.67 |
| **Ours(200)** | 87.04 | 85.86 | **636.67** |
| **Ours(500)** | 86.73 | 85.44 | 608.33 |
| **Ours(700)** | 86.88 | 85.49 | 591.67 |
| **Ours(999)** | **87.17** | **86.59** | 576.67 |

Table 4: Evaluation results on POPE and MME with varying noise levels.

| Model | POPE | | MME |
|---|---|---|---|
| | Accuracy | F1 Score | Accuracy+ |
| InstructBLIP | 78.93 | 79.11 | 1385.87 |
| LLaVA-1.5 | 80.37 | 79.61 | 1715.40 |
| + InstructBLIP(U) | 83.90 | 85.14 | 1437.84 |
| + InstructBLIP | 85.35 | 85.79 | 1718.18 |
| + LLaVA-NeXT | **86.55** | **86.13** | **1788.09** |

Table 5: Ensemble performance on POPE and MME benchmarks

(for instance, LLaVA-1.5's F1 score on POPE exceeds that of LLaVA-NEXT by only about 5%), probability averaging can yield improvements over individual models. ATED overcomes these limitations. Unlike uniform methods, ATED employs an adaptive weighting strategy guided by uncertainty, effectively overcoming the aforementioned limitations and achieving stable performance improvements. This gives ATED greater robustness and broader applicability.

## 5 Conclusion

In this paper, we propose ATED, the first training-free multimodal ensemble decoding method that effectively mitigates hallucinations across diverse multimodal tasks. During inference, ATED performs parallel processing with multiple LVLMs and adaptively fuses token-level logits, enabling finer-grained semantic control and more consistent generation. By introducing an adaptive uncertainty-guided weighting mechanism, ATED dynamically adjusts model importance via uncertainty minimization, enhancing reliability of ensemble inference. Moreover, ATED supports diverse decoding paths, further improving the factual consistency and robustness of generated content. Extensive experiments show that ATED consistently outperforms previous methods on multiple benchmarks, achieving notable gains in both accuracy and robustness.

## Limitations

Despite ATED's strong performance in mitigating hallucinations and enhancing robustness, several limitations remain. (i) ATED cannot address all types of hallucinations in LVLMs, which is understandable since the method requires neither additional training nor modifications to model architectures—factors that may constrain its effectiveness in certain scenarios. (ii) The uncertainty minimization framework relies on estimating output logits over the vocabulary, but the accuracy of this estimation may fluctuate across different tasks and models, potentially affecting the reliability of the ensemble. (iii) Furthermore, our framework depends on the estimation of logits over a shared vocabulary among LVLMs; integrating models with differing vocabularies remains an underexplored area for future research.

## References


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. 2025. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. *Preprint*, arXiv:2406.12718.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

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 frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2025. Hallucination of multimodal large language models: A survey. *Preprint*, arXiv:2404.18930.

Ali Furkan Biten, Lluis Gomez, and Dimosthenis Karatzas. 2021. Let there be a clock on the beach: Reducing object hallucination in image captioning. *Preprint*, arXiv:2110.01705.

Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. 2024a. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. *Preprint*, arXiv:2403.01548.

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

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024d. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and 1 others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Yeongjae Cho, Keonwoo Kim, Taebaek Hwang, and Sungzoon Cho. 2025. Do you keep an eye on what i ask? mitigating multimodal hallucination via attention-guided ensemble decoding. In *Proceedings of the 2025 International Conference on Learning Representations (ICLR)*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Prasenjit Dey, Srujana Merugu, and Sivaramakrishnan Kaveri. 2025. Uncertainty-aware fusion: An ensemble framework for mitigating hallucinations in large language models. *arXiv preprint arXiv:2503.05757*.

Thomas G Dietterich. 2000. ensemble methods in machine learning. *International workshop on multiple classifer systems*.

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. *Preprint*, arXiv:2403.14003.

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.

Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Han Hu, Dong Chen, and Baining Guo. 2023. Instructdiffusion: A generalist modeling interface for vision tasks. *Preprint*, arXiv:2309.03895.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A. Smith, and Luke Zettlemoyer. 2024. Demystifying prompts in language models via perplexity estimation. *Preprint*, arXiv:2212.04037.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. *Preprint*, arXiv:2308.06394.

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. *Preprint*, arXiv:2311.17911.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Preprint*, arXiv:1902.09506.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2025. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *Preprint*, arXiv:2408.02032.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. *Preprint*, arXiv:2312.06968.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *Preprint*, arXiv:2306.02561.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. *Preprint*, arXiv:2311.01477.

Minchan Kim, Minyeong Kim, Junik Bae, Suhwan Choi, Sungkyung Kim, and Buru Chang. 2024. Exreal: Exploiting semantic reconstruction to mitigate hallucinations in vision-language models. *Preprint*, arXiv:2403.16167.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael S Bernstein, and Fei-Fei Li. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. Lisa: Reasoning segmentation via large language model. *Preprint*, arXiv:2308.00692.

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.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

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.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023b. Silkie: Preference distillation for large visual language models. *Preprint*, arXiv:2312.10665.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023c. $M^3$it: A large-scale dataset towards multi-modal multilingual instruction tuning. *Preprint*, arXiv:2306.04387.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023d. Evaluating object hallucination in large vision-language models. *Preprint*, arXiv:2305.10355.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. *Preprint*, arXiv:2306.14565.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024c. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024d. Llava-next: Improved reasoning, ocr, and world knowledge.

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

Ibomoiye Domor Mienye and Yanxia Sun. 2022. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access*, 10:99129–99149.

David Opitz and Richard Maclin. 1999. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*.

Robi Polikar. 2012. Ensemble learning. *Ensemble machine learning: Methods and applications*, pages 1–34.

Zexuan Qiu, Zijing Ou, Bin Wu, Jingjing Li, Aiwei Liu, and Irwin King. 2025. Entropy-based decoding for retrieval-augmented large language models. *Preprint*, arXiv:2406.17519.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. Object hallucination in image captioning. *Preprint*, arXiv:1809.02156.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. *Preprint*, arXiv:2206.01718.

Rico Sennrich, Alexandra Birch, and Barry Haddow. 2016. Edinburgh neural machine translation system for wmt 16. In *Proceedings of the First Conference on Machine Translation*.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. 2024. Knowledge fusion of large language models. *Preprint*, arXiv:2401.10491.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. 2023. Evaluation and analysis of hallucination in large vision-language models. *Preprint*, arXiv:2308.15126.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *Preprint*, arXiv:2403.18715.

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.

Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. 2024. Evaluating and analyzing relationship hallucinations in large vision-language models. *Preprint*, arXiv:2406.16449.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, and 1 others. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

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

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2024a. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. *Preprint*, arXiv:2311.13614.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2024b. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *Preprint*, arXiv:2312.00849.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024c. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. *Preprint*, arXiv:2405.17220.

Zihao Yue, Liang Zhang, and Qin Jin. 2024. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *Preprint*, arXiv:2402.14545.

Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, Chunyuan Li, and Manling Li. 2024. Halle-control: Controlling object hallucination in large multimodal models. *Preprint*, arXiv:2310.01779.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023. How language model hallucinations can snowball. *Preprint*, arXiv:2305.13534.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. 2025. Gpt4roi: Instruction tuning large language model on region-of-interest. *Preprint*, arXiv:2307.03601.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *Preprint*, arXiv:2311.16839.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. *Preprint*, arXiv:2310.00754.

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

# A More Backgrounds

## A.1 Architecture of Vision-Language Models

Large Vision-Language Models (LVLMs) integrate pretrained image encoders and large-scale language models to support tasks such as image captioning and visual question answering. Typically, a frozen vision encoder such as CLIP extracts dense image embeddings (Radford et al., 2021), which are projected into the language space and fed into a decoder-only LLM like LLaMA. Architectures such as BLIP-2 (Li et al., 2023a) and InstructBLIP (Dai et al., 2023) adopt this two-tower design and align the modalities using lightweight adapters or learned commands.

**LLaVA-1.5** is a refinement over the original LLaVA model, featuring a simplified architecture and improved training pipeline. It employs CLIP-ViT-L as the vision encoder and a Vicuna-based decoder-only language model. The two modalities are connected via a trainable MLP projection layer, which maps visual tokens into the language embedding space. Trained with visual instruction tuning on synthetic datasets, it achieves strong results across various benchmarks. (Liu et al., 2023)

**InstructBLIP** builds on BLIP-2 by introducing an instruction-aware query transformer that conditions the vision encoder's output on task-specific prompts. It integrates a pretrained Vision Transformer (ViT) with a Q-Former, feeding the encoded visual queries to a language model such as Flan-T5 or Vicuna. It is trained using instruction tuning on a collection of 26 datasets. (Dai et al., 2023)

**MiniGPT-4** aims to replicate the capabilities of GPT-4-based vision-language systems using open-source components. It integrates a frozen ViT-based vision encoder with Vicuna, connected via a lightweight linear projection layer. Training involves two stages: pre-alignment on image-text pairs followed by fine-tunning on high quality image descriptions. Its minimal parameter count enables efficient multimodal alignment with strong performance in image captioning. (Zhu et al., 2023)

**LLaVA-NEXT** is an enhanced version of LLaVA-1.5, optimized for higher visual reasoning fidelity. It retains the MLP projection structure but augments training with improved instruction-following datasets and higher-resolution visual inputs. It achieves better performance in OCR, compositional reasoning, and world knowledge benchmakrs. (Liu et al., 2024d)

## A.2 Ensemble Learning in NLP

Ensemble learning has long been a reliable strategy in machine learning to improve robustness, reduce overfitting, and enhance generalization. By combining the predicitions of multiple models or decision rules, ensembles can correct individual biases and reduce the variance of outputs (Dietterich, 2000), Classical ensemble methods include bagging, boosting, and stacking, all of which have demonstrated strong performance in classification tasks such as sentiment analysis, topic classification, and named entity recognition (Opitz and Maclin, 1999).

In the domain of Natural Language Processing (NLP), ensemble methods have been applied extensively in both structured prediction and generation tasks. For example, ensemble decoding, which involves averaging or voting across multiple language models. has been shown to improve fluency and factuality in neural machine translation (Sennrich et al., 2016). Recent work has also explored

ensemble inference for large language models, aggregating outputs per-token-level logits from multiple sources to improve consistency and reduce hallucination

In multimodal learning, ensemble approaches are gaining traction as a decoding-level intervention. Rather than relying on a single model's output, ensembles constructed across different model variants or decoding stratigies can better capture complementary evidence, making them suitable for suppressing the hallucination in vision-language tasks.

## B Details about Baseline

To mitigate hallucination without retraining, a variety of decoding-time techniques have been proposed:

- **ICD:** (Wang et al., 2024) Instruction-Contrastice Decoding leverages instruction-level perturbations to reduce hallucination in multimodal large language models. It operates by introducing minimal semantic alterations to the input prompt, such as inserting irrelevant phrases or modifiying the question structure, and then comparing the model's output distributions under both the original and perturbed instructions. Tokens exhibiting instability across these variants are identified as potentially hallucinated and are downweighted during generation.

- **SID (Self-Introspective Decoding):** (Huo et al., 2025) Self-Introspective Decoding mitigateas hallucinations by introspectively filtering low-relevance visual signals during generation. It evaluates the contextual alignment of visual tokens with both the preceding textual context and the decoding history, retaining only those with strong semantic relevance. By pruning distractive or semantically weak visual features early in decoding, SID improves grounding accuracy, particularly in complex or visually dense scenarios.

- **VCD (Visual-Contrastive Decoding):** (Leng et al., 2023) Visual-Contrastive Decoding aims to improve visual consistency by introducing small-scale perturbations to the visual input and contrasting the model's responses. The approach applies controlled distortions, such as Gaussian blur, occlusion, or token masking, to the image embeddings and measures output divergence. Tokens highly sensitive to such perturbations are treated as visually fragile and are penalized during decoding.

- **OPERA (Overtrust Penalty with Retrospective Adjustment):** (Huang et al., 2024) Overtrust Penalty with Retrospective Adjustiment introduces a two-stage mechanism to address hallucination in multimodal generation: overtrust penalty and retrospective adjustment. During decoding, it applies a regularization term to suppress overconfident token predictions that exhibit weak visual grounding. After generation, a retrospective evaluation is performed to re-rank or adjust outputs based on their semantic agreement with the image.

- **Ensemble Decoding (ED):** (Cho et al., 2025) Ensemble Decoding combines multiple generation pathways to improve robustness and reduce hallucination. It operates by aggregating outputs from a set of models or decoding configurations, such as different random seed, visual crops, or temperature settings, and fusing them through majority voting, logit averaging, or response re-ranking. This ensemble process helps to mitigate the influence of unstable or outlier predicitions by emphasizing consensus across multiple decoders

All these methods operate without modifying model parameters, offering flexible, training-free solutions for enhancing visual faithfulness during inference.

## C Evaluation Metric Details

**The Polling-based Object Probing Evaluation (POPE)** benchmark is a systematic framework designed to assess object hallucination in Large Vision-Language Models (LVLMs) during image description tasks. POPE employs a binary question-answering format, using prompts such as "Does the image contain ___?" to evaluate a model's ability to accurately determine the presence or absence of specific objects within images. To construct negative samples—instances where the object is absent from the image—POPE utilizes three distinct strategies: random sampling involves selecting objects that do not appear in the image at random; popular sampling selects absent objects from a pool of frequently occurring objects across the dataset;

adversarial sampling prioritizes objects that commonly co-occur with present objects but are absent in the current image. The benchmark integrates three datasets: MSCOCO, A-OKVQA, and GQA. From each dataset, 500 images are selected, and six questions are generated per image, resulting in a total of 27,000 query-answer pairs for evaluation. Performance is measured using standard metrics, including accuracy, precision, recall, and F1 score, with higher values indicating a model's superior capability in mitigating hallucinations such as fabricated objects and erroneous descriptions.

**The Caption Hallucination Assessment with Image Relevance (CHAIR)** metric is a specialized evaluation framework designed to quantify object hallucination in image captioning models. CHAIR assesses the alignment between generated captions and the actual visual content by comparing the objects mentioned in the captions against ground-truth annotations from datasets like MSCOCO.

$$C_S = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|} \quad (9)$$

$$C_I = \frac{|\{\text{captions w/ hallucinated objects}\}|}{|\{\text{all captions}\}|} \quad (10)$$

The metric comprises two variants: CHAIRi (instance-level) and CHAIRs (sentence-level). CHAIRi calculates the proportion of hallucinated object mentions relative to all object mentions in the generated captions, while CHAIRs measures the fraction of sentences that contain at least one hallucinated object. Lower values in both metrics indicate better performance in mitigating object hallucinations .

**The Multimodal Model Evaluation (MME)** benchmark offers a comprehensive framework for assessing Large Vision-Language Models (LVLMs) across a spectrum of tasks, encompassing both perceptual and cognitive dimensions. Specifically, MME comprises ten perception-oriented subtasks and four cognition-focused ones, facilitating a holistic evaluation of LVLM capabilities .In the context of object-level hallucination evaluation, MME includes dedicated subsets targeting the "existence" and "count" tasks. The "existence" task assesses a model's ability to accurately identify the presence or absence of specific objects within an image, while the "count" task evaluates the model's proficiency in determining the correct number of instances of a given object.These tasks is quantified using a combined metric of accuracy and accuracy+. Accuracy measures the proportion of correct predictions, while accuracy+ accounts for near-correct responses.

# D Implementation Details

In all experimental settings, the hyper-parameter $\alpha$ is fixed at 1. For visual perturbations in the model ensemble, we adopt a noise-injection strategy, setting the noise steps $T$ to 200 for MME, 500 for LLaVA-Bench, and 999 for POPE. For OPERA, VCD, and SID, we use the default settings as specified in their original papers. Greedy decoding is used for comparison methods, while for open-ended generation tasks (such as CHAIR and LLaVA-Bench), we employ sampling with Top-p = 1. All experiments are conducted on Nvidia A40 GPUs.

# E More Detailed Comparison

## E.1 More Results on on CHAIR

The hyperparameter *max new tokens*, which controls the maximum length of generated responses, plays a critical role in CHAIR-based evaluation. In the main text, we report results using a setting of *max new tokens* = 64. Additional results under a relaxed constraint of *max new tokens* = 512 are provided in Table 6. As Table illustrates, the generation length limit has a substantial impact on LVLM performance under the CHAIR metric. when the token budget is increased from 64 to 512, our method consistently outperforms all baselines on the metric $\text{CHAIR}_S$, highlighting its robustness and adaptability under varying generation lengths. Furthermore, our model produces responses with an average length of 107.4 tokens as shown in Table 7 , indicating that the observed reduction in object hallucinations is achieved without compromising the richness of the generated descriptions.

## E.2 More Results on on MME

ATED is designed to integrate the expertise of multiple models, thereby bridging the hallucination gap that exists among different LVLMs during inference. To further investigate whether our approach not only preserves but also potentially enhances the fundamental perception and reasoning capabilities of LVLMs across a broader range of multimodal tasks, we also analyze the comprehensive performance on the MME benchmark, which consists of 14 sub-tasks for evaluating perception and recognition. As shown in Table 8, our method ($Ours_{\#}$) significantly outperforms all baseline approaches

| Type | METHOD | LLaVA-1.5 | | InstructBLIP | | MiniGPT4 | | LLaVA-NeXT | |
|---|---|---|---|---|---|---|---|---|---|
| | | $CHAIR_S$ | $CHAIR_I$ | $CHAIR_S$ | $CHAIR_I$ | $CHAIR_S$ | $CHAIR_I$ | $CHAIR_S$ | $CHAIR_I$ |
| Single | Default | 51.3 | 16.8 | 55.6 | 24.2 | 33.6 | 19.4 | 42.6 | 14.1 |
| | OPERA | 46.4 | 13.0 | 47.1 | 12.4 | 26.4 | 10.7 | 39.4 | 11.8 |
| | VCD | 51.7 | 15.6 | 51.0 | 16.7 | 30.4 | 14.2 | 41.1 | 12.9 |
| | ICD | 47.4 | 13.9 | 46.3 | 15.3 | 32.6 | 13.1 | 42.1 | 12.6 |
| | SID | 44.2 | 12.2 | 42.3 | 12.4 | 28.5 | 11.7 | 40.8 | 13.0 |
| Ensemble | Ours | ATED$^{\&}$ | | | | ATED$^{\#}$ | | | |
| | | $CHAIR_S$ | | $CHAIR_I$ | | $CHAIR_S$ | | $CHAIR_I$ | |
| | | - | | - | | **34.0** | | **17.1** | |

Table 6: **CHAIR evaluation results on different decoding strategies**. Results are from the papers or re-implemented based on official codes. ***Note:*** $\&$ denotes ensemble with InstructBLIP, $\#$ denotes ensemble with InstructBLIP and LLaVA-NeXT.

| Method | Length |
|---|---|
| Default | 100.6 |
| OPERA | 98.6 |
| VCD | 100.4 |
| ICD | 106.3 |
| **Ours**$^{\#}$ | 107.4 |

Table 7: Comparison of CHAIR performance across different methods in terms of output length on LLaVA-1.5.

| Method | Accuracy+ |
|---|---|
| Default | 1715.40 |
| OPERA | 1773.52 |
| VCD | 1756.02 |
| ICD | 1749.43 |
| SID | 1770.43 |
| **Ours**$^{\#}$ | 1788.09 |

Table 8: Comparison of total accuracy+ cacross different methods on LLaVA-1.5.

based on the LLaVA-1.5 backbone, surpassing both the original LVLMs and the best-performing baselines by a substantial margin (+18.34). These results indicate that our approach not only effectively manages hallucination during inference but also improves the accuracy of the underlying LVLMs on fundamental tasks.

Table 9 presents the quantitative evaluation results of the model under different $\alpha$ values on object-level metrics (Existence, Count), attribute-level metrics (Position, Color), and the overall accuracy (Total Accuracy+). As $\alpha$increases from 0.5 to 1.0, all metrics demonstrate varying degrees of improvement, with Color showing the most substantial gain—from 140 to 155. These improvements are reflected in the Total Accuracy+, which rises from 595.00 to 636.67 as $\alpha$ increases. Moreover, we observe that attribute-level metrics are more sensitive to changes in the intensity of vision-contrastive regularization compared to object-level metrics, resulting in greater improvements. This finding indicates that appropriately tuning the $\alpha$ parameter not only enhances the model's ability to confirm object information during adaptive ensemble inference but also significantly improves its capability to capture fine-grained attribute details. As a result, the overall prediction accuracy and robustness are further strengthened.

### E.3 Qualitative Analysis

To further evaluate whether ATED effectively mitigates hallucinations beyond quantitative metrics in open-ended generation tasks, we conducted a qualitative analysis on the MSCOCO dataset, using several decoding strategies as baselines. The LVLMs are prompt with "Please describe this image in detail", with the maximum token limit set to 150. As illustrated in Figure 8 and Figure 9, baseline methods including the default decoding, OPERA, and VCD often produce hallucinated content (highlighted in red). In contrast, ATED dynamically selects and weights token-level outputs from multiple models at each decoding step, guided by a greedy uncertainty-minimization strategy. This enables the model to better adapt to contextual environments and significantly improves the credibility and robustness of the generated content.

In addition, we perform GPT-assisted evaluation on the LLaVA-Bench benchmark (Liu et al., 2023). Following evaluation protocol proposed by (Yin et al., 2024; An et al., 2025), the model is presented with an image and two candidate descriptions, structured according to the prompt format

| $\alpha$ | Object-Level | | Attribute-Level | | Total Accuracy+ |
|---|---|---|---|---|---|
| | Existence | Count | Position | Color | |
| 0.5 | 180.00 | 143.33 | 131.67 | 140 | 595.00 |
| 0.7 | 180.00 | 143.33 | 136.67 | 140 | 600.00 |
| 1.0 | 185.00 | 158.33 | 138.33 | 155 | 636.67 |

Table 9: Quantitative results on Object-level (Existence, Count), Attribute-level (Position, Color), and Total Accuracy+ for using various noise steps.

shown in Figure 5. The GPT-4o API is employed to evaluate the generated responses in terms of factual accuracy (Accuracy) and descriptive richness (Detailedness).

Furthermore, we conducted an additional evaluation based on GPT-4, following the methodology outlined in (Zhao et al., 2023). Specifically, we randomly sampled 200 images from the Visual Genome (VG-100K) dataset (Krishna et al., 2017) and assessed model performance by comparing the generated descriptions with the region descriptions associated with each image. This comparison allows for effective identification of hallucinated content based on semantic inconsistencies. We comprehensively analyzed five key metrics: sentences per image (SPI), words per image (WPI), hallucinated sentence ratio (HSR), hallucinated word ratio (HWR), and mean hallucination ratio (MHR). Notably, higher SPI and WPI values, as well as lower HSR, HWR, and MHR, indicate better model performance. In the radar charts, a larger area reflects superior performance. Multiple models and decoding strategies were included as baselines for comparison. The detailed results are presented in Figure 4. As shown, the proposed ATED method substantially reduces hallucination and effectively suppresses misleading content during generation.

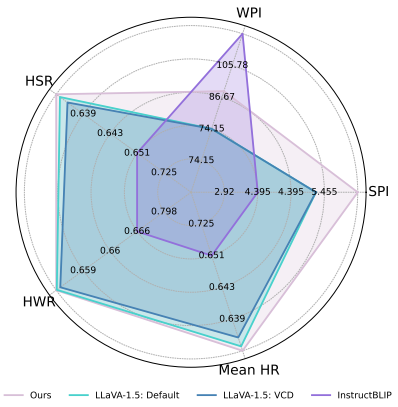


Figure 4: GPT-4 assisted hallucination evaluation.

**GPT-4o Prompt**

```
You are an AI designed to evaluate and score the performance of two AI
assistants in describing a given image. Your primary focus is on the
accuracy and detailedness of their descriptions. You will assess the
accuracy by checking for hallucinations—any part of the description
that is inconsistent with the image content. For detailedness, you
will consider how rich the response is in necessary details, excluding
any hallucinated parts. You will provide scores on a scale from 1 to
10 for each assistant separately, based on these criteria. After
scoring, you will offer an explanation for your evaluation, ensuring
it is free from bias and not influenced by the order of presentation
of the responses.

Input format:

[Assistant 1]

{}

[End of Assistant 1]

[Assistant 2]

{}

[End of Assistant 2]

Output format:

Accuracy:

Scores of the two answers:

Detailedness:

Scores of the two answers:
```

Figure 5: GPT-4o prompt format.

Instr: Please describe this image in detail

Max New Token = 150

LLaVA-1.5 OPERA

The image depicts a large, white skyscraper towering over a city with a clear blue sky in the background. The skyscraper is located in the center of the city, surrounded by other buildings of various heights. The city is bustling with activity, as evidenced by the presence of numerous cars and trucks on the streets below the skyscraper.

Ours

The image you've shared is a close-up of a tall, slender tower. The tower is predominantly white, with a few black accents that add a touch of contrast. The tower is situated in a city, as evidenced by the buildings and structures that surround it.

GPT-4o Evaluation: LLaVA-1.5 (OPERA)
Accuracy: 4/10
Detailedness: 5/10

GPT-4o Evaluation: Ours
Accuracy: 9/10
Detailedness: 7/10

*Accuracy:* Assistant 1: This description contains significant hallucinations. The structure in the image is the Space Needle in Seattle, which is not a "large, white skyscraper" and does not resemble one. It is a single, iconic observation tower with a distinct design. The claim of "numerous cars and trucks on the streets" is also inaccurate, as there are no visible streets bustling with traffic in the image. Assistant 2: This response is mostly accurate. The description identifies the subject as a "tall, slender tower" that is "predominantly white with a few black accents," which matches the appearance of the Space Needle. It also matches the setting by mentioning the city with surrounding buildings.

*Detailedness:* Assistant 1: While this response attempts to provide a more elaborate description of the setting, most of the details are fabricated and therefore detract from the evaluation. It does not mention defining features of the tower, such as its observation deck, tapered legs, or unique aesthetic. Assistant 2: This description is more concise but relatively to the point. The mention of black accents and the general shape of the tower improve its detailedness. However, it omits several crucial features, such as the observation deck, the flag, and the surrounding greenery in view.

Figure 6: Qualitative cases on LLaVA-Bench. The hallucinated content is highlighted in red.


Instr: Please describe this image in detail

Max New Token = 150

LLaVA-1.5 VCD: A cute little dog is laying down on a wooden floor, resting its head on its paws. The dog is white and brown in color and appears to be sleeping on a blue-colored surface.

Ours: The image features a small dog with a predominantly white coat, lying on its side on a wooden surface. The dog appears to be resting or sleeping, with its head comfortably resting on the wooden surface. The background is out of focus, but it seems to be an indoor setting.

GPT-4o Evaluation: LLaVA-1.5 (VCD)
Accuracy: 4/10
Detailedness: 6/10

***Accuracy:*** Assistant 1 incorrectly described the dog as "white and brown" and stated that it was "sleeping on a blue-colored surface" which is inaccurate. The dog's coat is light tan with slight white markings, and it is lying directly on a blue wooden floor, not "a blue-colored surface" like a blanket. Additionally, Assistant 1 claimed the dog was "resting its head on its paws" which is not the case—its head is resting on the floor. Assistant 2, on the other hand, provided an accurate description of the dog and its environment, correctly identifying that the dog is lying on its side on the blue wooden floor and avoiding any inaccurate claims.

GPT-4o Evaluation: Ours
Accuracy: 9/10
Detailedness: 7/10

***Detailedness:*** Assistant 1 included some descriptive details, such as identifying the dog's coloration (albeit inaccurately) and the mention of the "blue-colored surface". However, these inclusions were not meaningful due to inaccuracies. Assistant 2 provided sufficient and accurate details without adding unnecessary or hallucinated elements, such as describing the dog, its position, and the surrounding background accurately. Assistant 2's response is slightly more detailed due to a richer contextual hint about the background being "out of focus" and the scene being "an indoor setting."

Figure 7: Qualitative cases on LLaVA-Bench. The hallucinated content is highlighted in red.

Instr: Please describe this image in details

Max new token = 64

*LLaVA-1.5 (VCD):* The image features a woman lying on a bed, with her head resting on her hand. She appears to be looking at the camera, possibly posing for a picture. The bed is covered with a striped blanket, adding a cozy touch to the scene.

*LLaVA-1.5 (OPERA):* The image features a beautiful young woman sitting on a couch, wearing a white shirt. She appears to be posing for the camera, possibly for a magazine or advertisement. The couch she is sitting on has a striped pillow, adding a touch of color to the scene.

*InstructBLIP (OPERA):* The image features a young woman lying on a bed, wearing a white shirt and a red lipstick. She is positioned in the middle of the bed, with her hair cascading down her back. There are two pillows visible in the scene, one on the left side of the bed and

*Ours:* The image features a beautiful young woman with long, dark hair. She is lying down on a bed, dressed in a white shirt. The bed she is on is covered with a white comforter.

Figure 8: Qualitative cases on MSCOCO. The hallucinated content is highlighted in red.

Instr: Please describe this image in details

Max new token = 64

*LLaVA-1.5 (VCD):* The image features a woman sitting at a dining table, holding a large pizza in her hands. She is smiling and appears to be enjoying the moment. The pizza is placed in the center of the table, and the woman is positioned on the left side of the table."

*LLaVA-1.5 (OPERA):* The image features a woman sitting at a dining table, holding up a large pizza with a smile on her face. The pizza is placed in front of her, occupying a significant portion of the table. There are several other people in the scene, some sitting and others standing around the dining

*InstructBLIP (OPERA):* The image features a woman sitting at a dining table, holding a large pizza in her hands. She is smiling and appears to be eager to share the pizza with others. There are several chairs placed around the table, suggesting that there are other people present.

*Ours:* The image features a woman sitting in front of a dining table. She is holding a large pizza in her hands, which is placed on the table. The woman appears to be in the process of serving the pizza, as she is holding it up.

Figure 9: Qualitative cases on MSCOCO. The hallucinated content is highlighted in red.