SELF-AUGMENTED VISUAL CONTRASTIVE DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision-Language Models (LVLMs) have demonstrated remarkable multimodal capabilities, but they inherit the tendency to hallucinate from their underlying language models. While visual contrastive decoding has been proposed to mitigate this issue, existing methods often apply generic visual augmentations that disregard the specific context provided by the text query, limiting their effectiveness. This study introduces a novel training-free decoding strategy that addresses these limitations, featuring two key contributions. First, a self-augmentation prompting strategy that leverages the intrinsic knowledge of the model to dynamically align semantics between the query and the visual augmentation. Second, an adaptive thresholding algorithm that adaptively adjusts next token candidate size based on the output sparsity, utilizing full information from the logit distribution. Extensive experiments across four LVLMs and seven benchmarks demonstrate that the proposed decoding significantly enhances factual consistency compared to state-of-the-art decoding methods. This work highlights the importance of integrating query-dependent augmentation and entropy-aware decoding for improving effective generation of LVLMs. The source code will be released upon acceptance.

1 INTRODUCTION

Large Language Models (LLMs) have achieved remarkable success in language comprehension, generation, and reasoning (Brown et al., 2020; Google, 2023; Touvron et al., 2023; Chiang et al., 2023; OpenAI, 2023). By integrating visual encoding and projection, Large Vision-Language Models (LVLMs) have extended these capabilities to multimodal applications such as visual perception and planning (Li et al., 2022; Yu et al., 2022; Li et al., 2023a; Maaz et al., 2023; Ye et al., 2023; Zhang et al., 2023a; Zhu et al., 2023; Huang et al., 2023). Despite their impressive performance, LVLMs inherit critical limitations from their foundational language models. One of the most significant issues is *hallucination*, a phenomenon of generating plausible but factually incorrect or nonsensical outputs. This behavior is largely a byproduct of the auto-regressive training objective of the model, a process that incentivizes a reliance on spurious correlations over a precise understanding of underlying facts by maximizing token likelihood based on surface-level statistical patterns (Bender & Koller, 2020; Huang et al., 2025).

Advanced decoding methods can significantly enhance the factual consistency by shaping how token sequences are selected from output distributions at each generation step (Van der Poel et al., 2022; Favero et al., 2024a). A prominent decoding strategy to reduce hallucination effect is Contrastive Decoding (CD) (Li et al., 2023c), a technique that improves factuality by contrasting the outputs of an expert model with those of a weaker, *amateur* counterpart (Zhang et al., 2023b; Chuang et al., 2023). Motivated by this principle, Visual Contrastive Decoding (VCD) (Leng et al., 2024) was introduced to improve the general perceptual capabilities of LVLMs by contrasting standard output with an amateur logit generated from an input image degraded by random noise.

Subsequent research in VCD has primarily focused on determining which visual modifications or hidden states with experimental heuristics can maximize the sample variance while maintaining the semantics (Li et al., 2023b; Huang et al., 2024). However, these methods often overlook the critical role of the *input text query*, which specifies which aspects of an image are relevant to the user request. For instance, asking to identify an object in the image and solving a handwritten math problem require entirely different capabilities and reasoning from the LVLM. While VACoDe (Kim et al., 2024b) addressed this by estimating the divergence between logit distributions among the predefined visual augmentation set at the first generation step in a brute-force manner, there are two

fundamental limitations. First, first-token divergence is an empirical measure that does not always assure a favorable augmentation choice for the entire generation sequence. Second, its dependence on a single token renders it suitable for short, multiple-choice style answers but fundamentally limits its effectiveness for complex tasks requiring open-ended generation and multi-step reasoning.

Moreover, a challenge in contrastive decoding arises from the subtraction of the amateur logit from the expert logit (Li et al., 2023c). This operation can cause undesired effects that amplify the scores of certain tokens; if the amateur model produces a negative logit value, it will have its final score erroneously increased (Lyu et al., 2024). To mitigate this amplification effect, existing methods (Jin et al., 2024) truncate the vocabulary set based on a threshold set proportionally to the maximum value of the expert logit distribution. However, while this approach is effective at penalizing false positives, its reliance on a single data point (*i.e.*, the maximum logit) hinders it from utilizing the rich information encoded in the full logit distribution, such as *model confidence*.

These aforementioned limitations lead us to two main research questions. (1) *How can the semantic intent of a text query guide the selection of a visual augmentation to elicit a maximally informative discrepancy for contrastive decoding?* (2) *Is there a correlation between a predictive confidence of the model and the plausibility of its next-token candidates?* To address these questions, this study introduces Self-Augmented Visual Contrastive Decoding (SAVCD), a novel decoding strategy that adaptively select which visual augmentation is best suited to be *contextually relevant*. Unlike prior works (Kim et al., 2024b), SAVCD utilizes the intrinsic model knowledge to determine an optimal visual modification out of the box. Furthermore, we introduce an improved thresholding algorithm, Sparsity Adaptive Truncation (SAT), to overcome the limitations of existing plausibility constraints. Where prior methods often fail to utilize full information from the logit, SAT dynamically determines a threshold by utilizing the entire logit distribution as a proxy for the confidence of the output. The proposed method integrates seamlessly into any LVLM without requiring any architectural modifications or additional training. Extensive experiments and analysis verify that the proposed methods significantly enhance factual consistency and reduce hallucinations across multiple models and benchmarks. The contributions of this study are summarized as follows:

1. This work introduces SAVCD, a prompting strategy that leverages parametric knowledge of the model to select a visual augmentation that is semantically relevant to the textual query, thereby extracting a more informative discrepancy.
2. The proposed SAT improves the existing adaptive plausibility constraint by leveraging the entropy of the expert logit and dynamically sets a threshold of token implausibilities.
3. Extensive experiments validate the effectiveness of the proposed method across 4 LVLMs and 7 benchmarks. The results demonstrate that SAVCD significantly reduces hallucinations while amplifying the relevance and informativeness in the response.

2 PRELIMINARIES

Auto-regressive Generation of LVLMs Suppose that f_θ is an LVLM (Gong et al., 2023; Maaz et al., 2023; Li et al., 2025a), parameterized by θ. The model operates on a vocabulary set $\mathcal{V}$, and the set of all possible token sequences can be denoted by its Kleene closure, $\mathcal{V}^* = \bigcup_{i \geq 0} \mathcal{V}^i$, where i indicates the timestamp of the LVLM output. The function $f_\theta : \mathcal{V}^* \times \mathbb{R}^{h \times w \times 3} \to \mathcal{V}^*$ auto-regressively generates a response from a given text query $x \in \mathcal{V}^*$ and a visual input $v \in \mathbb{R}^{h \times w \times 3}$. At each timestep t, the LVLM computes a logit distribution over the vocabulary for the next token y_t, conditioned on the inputs (x, v) and the sequence of previously generated tokens $y_{<t}$. This yields the probability distribution over the next token:

$$p_\theta(y_t | v, x, y_{<t}) \propto \exp\left(\text{logit}_\theta(y_t | v, x, y_{<t})\right). \tag{1}$$

The next token is then selected from this distribution according to a chosen decoding method. Decoding methods are broadly categorized into two families: deterministic search, including greedy and beam search (Graves, 2012), and stochastic sampling, such as top-k, Nucleus (Holtzman et al., 2019), Mirostat (Basu et al., 2020), and typical (Meister et al., 2023) sampling.

Hallucination Ideally, the generated response y should be factually accurate, relevant to the query x, and faithful to the visual content v. However, current LVLMs often fail to meet these criteria, suffering

from a critical issue known as hallucination (Rohrbach et al., 2018). This phenomenon stems from multiple reasons, including imperfect learning and decoding (Ji et al., 2023), misalignment of vision and language modalities (Tong et al., 2024), and failure of understanding the context (Daunhawer et al., 2021). To address this issue, recent studies have suggested scaling the input image resolution (Liu et al., 2024b; Chen et al., 2024b), combining another inductive bias of visual encoders (Li et al., 2025b), post-hoc rectifying (Zhou et al., 2023), self-correction after generation (Yin et al., 2024), and advanced decoding methods (Shi et al., 2024; Favero et al., 2024b). Among those approaches, decoding-based methods are particularly promising since they enable real-time control, do not require additional training, and are compatible with other hallucination mitigation strategies.

Contrastive Decoding CD (Li et al., 2023c) tackled hallucination problems in the NLP domain by contrasting the predictions of two different language models with different capacities. VCD (Leng et al., 2024) extended the idea of CD with vision modality and introduced the contrastive counterpart v' by degrading visual content with random noise to v. It sequentially treats the logit from v' as an output of the amateur model, sampling the next token from:

$$p_{\mathrm{CD}}(y|v, v', x) = \mathrm{softmax}\left((1+\alpha)\cdot \mathrm{logit}_\theta(y|v, x) - \alpha \cdot \mathrm{logit}_{\theta'}(y|v', x)\right), \tag{2}$$

where α denotes an amplification parameter. Recent studies have focused on curating a better selection of the degradation to achieve maximal differentiation while preserving semantic integrity. For instance, cropping the patch which is likely to cause hallucinations (Chen et al., 2024a), caption substitute (Kim et al., 2024a), and visualization of the textual output (Park et al., 2025). While most VCD methods rely on a shared underlying principle of query-agnostic input modifications, VACoDe (Kim et al., 2024b) has introduced a dynamic visual augmentation strategy. This approach attempts to be query-aware by using L_2 distance between the expert and amateur logit distribution at the first token generation as a score function to select an augmentation.

However, this reliance on a first-generated token has fundamental limitations. The overall semantics of a task are not universally guaranteed to be reflected by first-token divergence, which is an empirical proxy. One example of failure is when two logits have the same argmax but are distinct in terms of overall entropy. In this case, the query could not be invalidated because, despite the distortion, the model could still answer correctly, although the divergence remained large. This may be effective for short-answer and multiple-choice questions, but it can be ineffective for other scenarios including open-ended questions and multi-step reasoning.

3 SAVCD: Self-Augmented Visual Contrastive Decoding

To address the preceding limitations, we introduce SAVCD, a decoding method that identifies a query-specific visual augmentation to apply for visual contrastive decoding by utilizing the rich knowledge base of LVLM (Li et al., 2024). Unlike prior methods that rely on simple heuristics, SAVCD leverages the world knowledge and common sense embedded in the LVLM to achieve a *semantic alignment* between the query and the selected augmentation. This approach enables the model to reason the *underlying intent* of a query and make a choice that elicits a more targeted and informative discrepancy. Alg. 1 and Fig. 1 outline the proposed method.

3.1 Self-Augmentation Selection

SAS Prompting Self-Augmentation Selection (SAS) aims to employ parametric knowledge of the LVLM to dynamically select the best *task-optimal* visual augmentation on the fly that amplifies the output divergence. This is achieved through a structured SAS Prompt $\mathcal{P}$, which comprises three key components. First, the prompt contains explicit definitions of each visual augmentation and corresponding effects, providing the model with the necessary operational knowledge. Second, to minimize the risk of post hoc rationalization, the prompt is structured to elicit reasoning before the final selection is made (Zelikman et al., 2024). Finally, inspired by few-shot learning techniques (Brown et al., 2020; Patel et al., 2024), in-context learning (ICL) examples are included in the prompt $\mathcal{P}$ to further condition the contextual knowledge (Alayrac et al., 2022). The textual output is then processed by a parsing function $g(\cdot) : \mathcal{V}^* \to \mathcal{V}^* \times \mathcal{V}^*$, which separates the reasoning trace r and final augmentation choice c. The contrasted image is obtained by feeding the v and the final choice c to a predefined visual augmentation function $\mathcal{A}$.

$$(r, c) = g(f_\theta(\mathcal{P}, x)), \quad v' = \mathcal{A}(c, v). \tag{3}$$

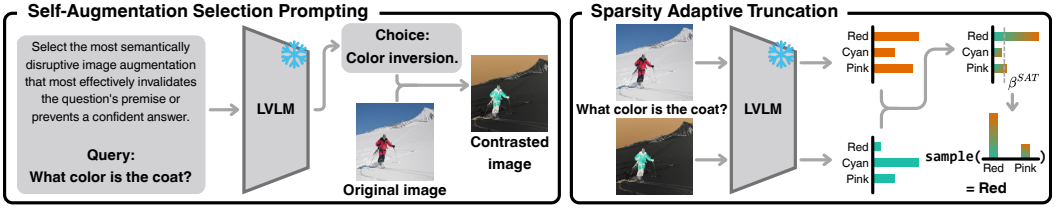


Figure 1: Overview of the proposed augmentation choice process and sparsity adaptive truncation.

Subsequently, contrasted logit distribution is calculated from expert logit $l = \text{logit}_\theta(y_t|v, x, y_{<t})$ and amateur logit $l' = \text{logit}_\theta(y_t|\mathcal{A}(c, v), x, y_{<t})$. The augmentation set is defined with random crop, random mask, noise addition, color inversion, horizontal flip, and vertical flip. Note that the generation configuration is set to greedy decoding for SAS Prompt $\mathcal{P}$ to ensure computational efficiency, determinism, and reproducibility. While further optimized prompting techniques (Manakul et al., 2023) and multiple combinations of different augmentations can be deployed, we limit the scope to two prompting features and the aforementioned augmentation set in this work. Full prompt is referred to the Appendix B.1.

3.2 Rethinking Adaptive Plausibility Constraint

CD-based methods encourage the generation of implausible tokens since the output distribution from contrasted visual input v' still involves the underlying semantics of v (Li et al., 2023c). This can cause the two distributions to not cooperate properly, resulting in the reward of undesirable tokens. Adaptive Plausibility Constraint (APC) (Li et al., 2023c; Leng et al., 2024) addresses this challenge with a controllable hyperparameter $\beta \in [0, 1]$, setting a threshold proportional to the logarithm of the maximum probability of the new token, formulated as:

$$\mathcal{V}_{\text{APC}} = \{y_t \in \mathcal{V} \mid p_\theta(y_t|v, x, y_{<t}) \geq \beta \cdot \max_{w \in \mathcal{V}} p_\theta(w|v, x, y_{<t})\}. \tag{4}$$

However, since this thresholding mechanism is based solely on the maximum logit value and the meaning of a logit value is *relative* to the other logits in the distribution, it is a *confidence-agnostic filter*. Although this approach penalizes false positives by truncating the sample space, it becomes unreliable in low-confidence states, when the risk of discarding the correct token from the candidate set is high (Guo et al., 2017; Wenkel et al., 2021). This deficiency arises because APC disregards the rich signal encoded in the full output distribution. The *entropy* of logit distribution provides a more robust and holistic measure of model uncertainty which can be leveraged for a more effective filtering of the candidate set.

Model uncertainty, characterized by the value of output entropy, is a recognized correlate of model errors (Manakul et al., 2023). When the logit distribution is highly entropic, a more lenient threshold is required to create a sufficiently inclusive candidate set and avoid erroneously discarding the context-relevant tokens. Conversely, in low-entropy scenarios where the model is confident (Tornetta, 2021) with a sparse output distribution, a more restrictive threshold is required to retain pivotal tokens with high probability and to refine the candidate set by taking over the probability mass from filtered tokens (Li et al., 2023c). This *inverse proportionality* heuristic improves generation fidelity by minimizing the risk of sampling erroneous, low-probability tokens on the tail of the distribution.

To enable the *confidence-aware* thresholding, the proposed method extends APC to SAT, a method that dynamically adjusts the plausibility constraint based on the *sparsity* of the output distribution. The method leverages the principle that a sparsity is inversely related to its uncertainty, quantified by Shannon Entropy $H : \mathbb{R}^d \to [0, \log_2 d]$ (Shannon, 1948), which maps a probability distribution p over its dimension d to its uncertainty, calculated as $H(p) = -\sum_{i=0}^{|\mathcal{V}|-1} p_i \log_2 p_i$. To implement an inversely proportional relationship where higher entropy yields a smaller threshold, a decayed entropy function $H_{\text{decay}} : \mathbb{R}^{|\mathcal{V}|} \to (0, 0.5]$ is formulated to compute the threshold value:

$$H_{\text{decay}}(p) = \sigma\left(-\gamma \sum_{i=0}^{|\mathcal{V}|-1} p_i \log_2 p_i\right), \tag{5}$$

where σ and $\gamma < 0$ denote a sigmoid function and a scaling parameter, respectively. The choice of a sigmoidal decay is deliberate, as other decaying functions, such as exponential or polynomial (Provencher, 1976; Borichev & Tomilov, 2010), could be potentially considered, but they lack

Algorithm 1 SAVCD: Self-Augment Visual Contrastive Decoding

Require: input image v, text query x, LVLM f_θ, augmentation function $\mathcal{A}$, SAS Prompt $\mathcal{P}$, vocabulary set $\mathcal{V}$, hyperparameter α.

$c \leftarrow f_\theta(\mathcal{P}, x)$ ▷ Identify augmentation c from given x
$t \leftarrow 0$ ▷ Initiate t
while $t < T$ **do**
  $l \leftarrow \text{logit}_\theta(y_t|v, x, y_{<t})$ ▷ Set expert logit l
  $l' \leftarrow \text{logit}_\theta(y_t|\mathcal{A}(c, v), x, y_{<t})$ ▷ Set amateur logit l'
  $l_{\text{CD}} \leftarrow (1+\alpha) \cdot l - \alpha \cdot l'$ ▷ Set contrasted logit
  $\beta_t^{\text{SAT}} \leftarrow H_{\text{decay}}\left(\text{softmax}(l)\right)$ ▷ Set SAT parameter β_t from Eq. 5
  $\mathcal{V}_{\text{SAT}} \leftarrow \{y_t \in \mathcal{V} \mid p_\theta(y_t|v, x, y_{<t}) \geq \beta_t^{\text{SAT}} \cdot \max_{w' \in \mathcal{V}} p_\theta(w'|v, x, y_{<t})\}$ ▷ Set threshold
  $l_{\text{CD}}[i] \leftarrow -\infty$ **for all** $i \notin \mathcal{V}_{\text{SAT}}$ ▷ Apply vocabulary truncation
  $y_t \sim \text{softmax}(l_{\text{CD}})$ ▷ Token sample
  $t \leftarrow t + 1$
end while
return $\{y_0, ..., y_{T-1}\}$

the versatility of a sigmoid. The curve of the sigmoid function is naturally bounded to $(0, 1)$, and its lower plateau creates a stable, consistent threshold for low confidence distributions, and precise controllability over the single steepness parameter γ of the decay for mid-range entropy. Furthermore, by ensuring the threshold remains strictly less than 1, SAT prevents the candidate set from collapsing to a single token, guaranteeing that the decoding process remains distinct from greedy decoding.

The proposed SAT method introduces a dynamic threshold β_t^{SAT}, which is calculated by incorporating the entropy of the logit distribution: $\beta_t^{\text{SAT}} = H_{\text{decay}}(\text{softmax}(\text{logit}_\theta(y_t|v, x, y_{<t})))$. The next-token candidate set, $\mathcal{V}_{\text{SAT}}$, is then constructed by filtering the vocabulary set with this adaptive threshold: $\mathcal{V}_{\text{SAT}} = \{y_t \in \mathcal{V} \mid p_\theta(y_t|v, x, y_{<t}) \geq \beta_t^{\text{SAT}} \cdot \max_{w \in \mathcal{V}} p_\theta(w|v, x, y_{<t})\}$. To exclude the implausible tokens, $-\infty$ is assigned to logit elements which are not involved in $\mathcal{V}_{\text{SAT}}$. Finally, the contrasted probability distribution is obtained by combining Equations 2 to 5:

$$p_{\text{CD}}(y_t|v, x, y_{<t}) = \begin{cases} \text{softmax}((1+\alpha) \cdot l - \alpha \cdot l'), & \text{if } y_t \in \mathcal{V}_{\text{SAT}} \\ 0. & \text{otherwise} \end{cases} \tag{6}$$

4 Experiments

4.1 Experimental Settings

Benchmark and Model Selection Following standard practices in the literature (Leng et al., 2024; Kim et al., 2024b), three foundation model families are selected to evaluate the effectiveness of SAVCD: LLaVA-1.5 (Liu et al., 2024a), Qwen-VL (Bai et al., 2023), and InstructBLIP (Dai et al., 2023) with vicuna-v1.1 (Chiang et al., 2023). 7B and 13B variants are selected for LLaVA-1.5, and 7B variants are chosen for the other model families. The evaluations are divided into two categories: discriminative and generative benchmarks. Discriminative benchmarks assess the factuality of visual recognition in the form of binary or multiple choice questions, while generative benchmarks evaluate broader capabilities by requiring open-ended responses and judge with proprietary models (Zheng et al., 2023; Gu et al., 2024; Ali et al., 2025). The selected discriminative benchmarks include POPE (Li et al., 2023d) constructed on MSCOCO (Lin et al., 2014), and A-OKVQA (Schwenk et al., 2022) dataset, MME-Perception (MME-P) (Fu et al., 2024), and MMVP (Tong et al., 2024). MMHal-Bench (Sun et al., 2023), LLaVA-Bench (In-the-Wild) (Liu et al., 2023), MM-Vet (Yu et al., 2023) are selected for generative benchmarks. The ablation studies were focused on the LLaVA-1.5 model family and the MME-P benchmark. This selection represents a methodological choice, as LLaVA-1.5 not only provides the strongest performance among the model families but also being one of the most widespread adoption within the open-source community, while MME-Perception offers the largest testbed among ones with diverse categories.

Implementation Details Unless explicitly stated otherwise, the CD hyperparameters are set to $\alpha = 1$, $\beta = 0.1$ for APC, and the SAT hyperparameter was set to $\gamma = -0.5$. Including automated

Table 1: Discriminative benchmark results on MME (Fu et al., 2024), MMVP (Tong et al., 2024), and POPE (Li et al., 2023d) constructed on COCO (Lin et al., 2014), and A-OKVQA (Schwenk et al., 2022). Avg. Δ denotes averaged gain against Multinomial sampling across benchmarks.

Model	Method	POPE-MSCOCO		POPE-AOKVQA		MME-P↑	MMVP↑	Avg. Δ
		Acc.↑	F1↑	Acc.↑	F1↑			
LLaVA-1.5-7B	Multinomial	$82.07_{\pm 1.83}$	$80.48_{\pm 1.66}$	$79.81_{\pm 4.12}$	$79.86_{\pm 3.26}$	$1278.42_{\pm 30.30}$	$32.40_{\pm 4.73}$	-
	VCD	$83.66_{\pm 1.97}$	$82.55_{\pm 1.76}$	$80.51_{\pm 4.49}$	$81.11_{\pm 3.56}$	$1323.67_{\pm 20.84}$	$34.00_{\pm 3.89}$	+10.86%
	VACoDe	$84.29_{\pm 2.41}$	**83.59**$_{\pm 2.11}$	$80.86_{\pm 4.97}$	$81.87_{\pm 3.90}$	$1372.50_{\pm 13.78}$	**36.67**$_{\pm 2.87}$	+9.52%
	SAVCD	**82.93**$_{\pm 1.77}$	$83.57_{\pm 1.63}$	**82.80**$_{\pm 4.75}$	**83.20**$_{\pm 3.86}$	**1431.30**$_{\pm 13.87}$	$36.00_{\pm 3.09}$	**+14.32%**
LLaVA-1.5-13B	Multinomial	$83.86_{\pm 1.51}$	$81.02_{\pm 1.33}$	$80.97_{\pm 3.51}$	$80.79_{\pm 2.84}$	$1351.69_{\pm 30.30}$	$31.60_{\pm 4.82}$	-
	VCD	$83.86_{\pm 1.72}$	$82.68_{\pm 1.53}$	$81.93_{\pm 3.65}$	$82.16_{\pm 2.94}$	$1372.77_{\pm 30.54}$	$31.60_{\pm 4.81}$	+6.33%
	VACoDe	$84.86_{\pm 1.90}$	**84.17**$_{\pm 1.68}$	$82.34_{\pm 3.90}$	$83.08_{\pm 3.02}$	$1434.09_{\pm 12.79}$	$32.13_{\pm 3.25}$	+8.03%
	SAVCD	**85.37**$_{\pm 1.42}$	$83.96_{\pm 1.32}$	**84.25**$_{\pm 3.64}$	**84.13**$_{\pm 3.04}$	**1462.18**$_{\pm 18.21}$	**34.80**$_{\pm 1.19}$	**+11.59%**
Qwen-VL	Multinomial	$75.72_{\pm 0.79}$	$72.07_{\pm 0.85}$	$76.64_{\pm 2.50}$	$74.68_{\pm 2.16}$	$1311.79_{\pm 23.42}$	$17.33_{\pm 2.54}$	-
	VCD	$77.98_{\pm 0.79}$	$75.42_{\pm 0.77}$	$78.85_{\pm 2.62}$	$77.73_{\pm 2.77}$	$1415.12_{\pm 21.31}$	$21.33_{\pm 2.62}$	+5.05%
	VACoDe	**78.35**$_{\pm 0.93}$	**76.08**$_{\pm 0.86}$	**78.98**$_{\pm 3.07}$	**78.02**$_{\pm 2.77}$	$1412.43_{\pm 10.27}$	$22.13_{\pm 3.75}$	**+7.49%**
	SAVCD	$77.58_{\pm 0.65}$	$74.71_{\pm 0.64}$	$78.23_{\pm 2.69}$	$76.78_{\pm 2.36}$	**1442.36**$_{\pm 11.87}$	**26.67**$_{\pm 1.63}$	+6.69%
InstructBLIP	Multinomial	$68.70_{\pm 1.74}$	$69.34_{\pm 1.42}$	$65.52_{\pm 3.00}$	$68.36_{\pm 1.91}$	$973.66_{\pm 41.81}$	$19.20_{\pm 1.52}$	-
	VCD	$71.99_{\pm 1.27}$	$72.77_{\pm 1.11}$	$69.26_{\pm 3.03}$	$72.23_{\pm 2.03}$	$1079.39_{\pm 46.30}$	$18.93_{\pm 2.77}$	+12.33%
	VACoDe	$73.29_{\pm 1.50}$	$74.26_{\pm 1.17}$	$70.01_{\pm 3.28}$	$73.40_{\pm 2.21}$	$1090.88_{\pm 33.01}$	**21.87**$_{\pm 1.10}$	+10.98%
	SAVCD	**82.86**$_{\pm 1.94}$	**82.34**$_{\pm 1.62}$	**72.09**$_{\pm 3.76}$	**75.37**$_{\pm 2.51}$	**1198.53**$_{\pm 17.95}$	$16.13_{\pm 3.18}$	**+18.78%**

Table 2: Generative benchmark results on LLaVA-Bench (In-the-Wild) (Liu et al., 2023), MM-Vet (Yu et al., 2023), and MMHal-Bench Sun et al. (2023).

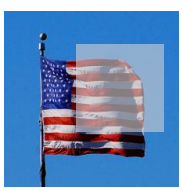
Model	Method	MMHal-Bench		MM-Vet↑	LLaVA-Bench↑	Avg. Δ
		Avg. Score↑	Hal. Rate↓			
LLaVA-1.5-7B	Multinomial	$2.27_{\pm 0.08}$	$0.65_{\pm 0.02}$	$27.74_{\pm 2.01}$	$58.48_{\pm 2.17}$	-
	VCD	$2.32_{\pm 0.09}$	$0.65_{\pm 0.02}$	$31.14_{\pm 1.15}$	$69.08_{\pm 2.07}$	+2.82%
	VACoDe	$2.32_{\pm 0.09}$	$0.64_{\pm 0.03}$	$29.88_{\pm 1.94}$	$69.12_{\pm 2.48}$	+6.14%
	SAVCD	**2.55**$_{\pm 0.11}$	$0.59_{\pm 0.03}$	**31.14**$_{\pm 0.95}$	**69.22**$_{\pm 1.80}$	**+6.97%**
LLaVA-1.5-13B	Multinomial	$2.35_{\pm 0.18}$	$0.65_{\pm 0.05}$	$31.20_{\pm 1.78}$	$69.48_{\pm 2.78}$	-
	VCD	$2.37_{\pm 0.24}$	$0.64_{\pm 0.07}$	$35.00_{\pm 1.61}$	$73.62_{\pm 1.40}$	+1.11%
	VACoDe	$2.52_{\pm 0.16}$	$0.61_{\pm 0.03}$	$34.18_{\pm 0.98}$	$74.56_{\pm 1.62}$	+4.78%
	SAVCD	**2.53**$_{\pm 0.09}$	**0.60**$_{\pm 0.03}$	**36.62**$_{\pm 1.33}$	**76.24**$_{\pm 0.83}$	**+6.04%**
Qwen-VL	Multinomial	$2.21_{\pm 0.12}$	**0.50**$_{\pm 0.02}$	$31.70_{\pm 1.76}$	$35.98_{\pm 1.56}$	-
	VCD	$2.17_{\pm 0.08}$	$0.51_{\pm 0.03}$	$34.04_{\pm 1.21}$	$38.78_{\pm 0.58}$	+9.21%
	VACoDe	**2.21**$_{\pm 0.13}$	$0.50_{\pm 0.03}$	$35.42_{\pm 1.36}$	$39.18_{\pm 1.75}$	+10.47%
	SAVCD	$2.15_{\pm 0.06}$	**0.50**$_{\pm 0.02}$	**35.98**$_{\pm 1.00}$	**39.84**$_{\pm 0.73}$	**+17.09%**
InstructBLIP	Multinomial	$1.89_{\pm 0.16}$	$0.69_{\pm 0.05}$	$23.06_{\pm 0.59}$	$53.24_{\pm 2.01}$	-
	VCD	$1.99_{\pm 0.14}$	$0.70_{\pm 0.04}$	$28.18_{\pm 1.35}$	**58.30**$_{\pm 1.38}$	+4.99%
	VACoDe	$2.03_{\pm 0.07}$	$0.68_{\pm 0.02}$	$27.40_{\pm 0.51}$	$56.82_{\pm 3.06}$	+9.17%
	SAVCD	**2.16**$_{\pm 0.13}$	**0.64**$_{\pm 0.05}$	**31.14**$_{\pm 0.95}$	$56.98_{\pm 2.13}$	**+17.08%**

judging of generative benchmarks, all API calls to proprietary models were made using OpenAI gpt-4o-mini with temperature 0 for deterministic results and reproducibility. All main experiments were conducted over five runs, and ablation studies over three runs with different random seeds, with results reported as the average and standard deviation to account for the inherent randomness from the augmentation process and multinomial sampling.

4.2 Experimental Results

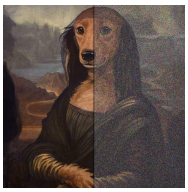
Main Results Tab. 1 and 2 summarize the averaged performance and standard deviations for all evaluated settings. The final column in each table, denoted as Avg. Δ, reports the average performance gain over the multinomial sampling baseline for each method and combination. For this calculation, the accuracy score is used for POPE and the average score is used for MMHal-Bench. For each configuration, the best-performing method is highlighted in bold, and ties are resolved in favor of the method exhibiting lower variance across runs. SAVCD achieves remarkable performance gains across both benchmark categories, ranging from 6.69% to 18.78% relative to the multinomial sampling.

To further probe the effectiveness of SAVCD, a token-level analysis was conducted to verify how the proposed method mitigates hallucinations by examining the output logits of LLaVA-1.5-7B. Fig. 2 illustrates two examples of logit values of LLaVA-1.5-7B with SAVCD on MM-Vet (Yu et al., 2023)



Query:
Is the wind blowing the flag?

Expert	_No
Amateur	_No
Final token	_Yes
Expert logit	24.14
	24.14
	23.56
Amateur logit	24.70
	24.70
	23.00
Final logit	23.58
	23.58
	24.12
β^{SAT}	0.265



Query:
Describe this photo in detail

Expert	_The	_image	_features	_a	_painting	_of	_a	_dog	_dog	...
Amateur	_The	_image	_features	_a	_painting	_of	_a	_blue	_dog	...
Final	_The	_image	_features	_a	_painting	_of	_a	_brown	_dog	...
Expert logit	26.69	22.98	22.36	25.06	16.25	21.34	23.09	15.13	19.33	...
	26.69	22.98	22.36	25.06	16.25	21.34	23.09	6.50	19.33	...
	26.69	22.98	22.36	25.06	16.25	21.34	23.09	14.14	19.33	...
Amateur logit	26.69	23.47	22.53	25.36	15.37	21.73	23.41	14.73	17.98	...
	26.69	23.47	22.53	25.36	15.37	21.73	23.41	15.36	17.98	...
	26.69	23.47	22.53	25.36	15.37	21.73	23.41	8.68	17.98	...
Final logit	26.69	22.50	22.19	24.77	17.13	20.95	22.78	15.52	20.67	...
	26.69	22.50	22.19	24.77	17.13	20.95	22.78	-inf	20.67	...
	26.69	22.50	22.19	24.77	17.13	20.95	22.78	19.59	20.67	...
β^{SAT}	0.42	0.39	0.24	0.46	0.12	0.35	0.46	0.08	0.30	...

Figure 2: Qualitative examples of SAVCD on MM-Vet (Yu et al., 2023) and LLaVA-Bench (Liu et al., 2023), and corresponding logit distributions and SAT thresholds by timestamp.

and LLaVA-Bench (Liu et al., 2023). Amateur and Expert logit indicate the selected token with and without augmentation, and the final token, highlighted with gray, is the token that corresponds to the argmax of the contrasted logit. Note that the applied augmentations are stylized for visual clarity.

These examples provide three important observations. (1) The example to the left shows a case of *failure correction* where the contrastive process between two logits successfully elevates the score for the correct `_Yes` token, making it the final answer. (2) The example on the right evidences *hallucination penalty*, where random noise triggered hallucination of generating `_blue` token from the amateur logit. It is penalized through subtraction, causing its final score to fall below the SAT threshold and be removed from the candidate set. (3) Adaptive nature of the SAT threshold β^{SAT} is observed, with a higher threshold applied to common tokens (*e.g.*, articles, prepositions) and a lower threshold applied to informative, lower-confidence tokens (*e.g.*, painting, red-boxed token). These findings highlight a clear validation of both core components of SAVCD, confirming that not only *contextually relevant* augmentation selection with model knowledge can effectively amplify the output divergence by invalidating the premise of the question, but also the efficacy of *confidence-aware* SAT.

Computational Overhead The computational cost of SAVCD was evaluated by comparing throughput (token/s) and latency (ms/token) against other CD-based methods. The analysis used the LLaVA-Bench with the LLaVA-1.5 family on an NVIDIA A100 GPU. Detailed results are presented in Tab. 3. The superscript $+$ on SAVCD denotes the full prompt configuration that includes reasoning and ICL, while $-$ denotes a lightweight configuration without both components. The results show that the primary computational bottleneck for both adaptive methods is the augmentation choice process. VACoDe is a *brute-force* that requires a separate forward pass for each predefined augmentation, which includes the full set of visual tokens, resulting in an overhead that scales linearly with the size of the augmentation set. On the other hand, SAVCD demonstrates architectural advantage by requesting a *single text-only generation pass*, bypassing the main computational expenses of visual tokens, which constitute the majority of the input. This architectural feature enables a flexible trade-off between performance and latency. While the full prompting incurs a cost comparable to VACoDe, the cost-optimized prompting exhibits substantially higher efficiency. This confirms that SAVCD provides a more scalable and controllable framework for query-aware contrastive decoding.

Table 3: Decoding throughput (token/s) and latency (ms/token). # Param and # tokens indicate model size and number of generated tokens for multimodal query, respectively.

Decoding	# Param	# tokens	token/s$^{\uparrow}$	ms/token$_{\downarrow}$
VCD	7B	9914	18.50	54.06
	13B	8785	14.01	71.38
VACoDe	7B	8418	16.97	58.93
	13B	8568	13.03	76.76
SAVCD^{-}	7B	8346	17.39	57.50
	13B	8793	11.37	87.92
SAVCD^{+}	7B	10163	15.08	66.32
	13B	8805	11.33	88.26

4.3 Ablation Study and Analyses

Augmentation Selection A detailed investigation of the augmentation choice made by the LLaVA-1.5 family is presented in Fig. 3 with different patterns by model capacities. For comparison, the choices from gpt-4o-mini are also included to provide a practical upper bound and will be denoted as the "Oracle" for notation convenience throughout the remainder of this paper. The results reveal that the distribution of selections varies significantly across different benchmarks. A notable contrast is found between the sparsest POPE and the most uniform MMVP. For POPE, random mask accounts for 87.6% of all selections. This strong preference arises because the queries related to object

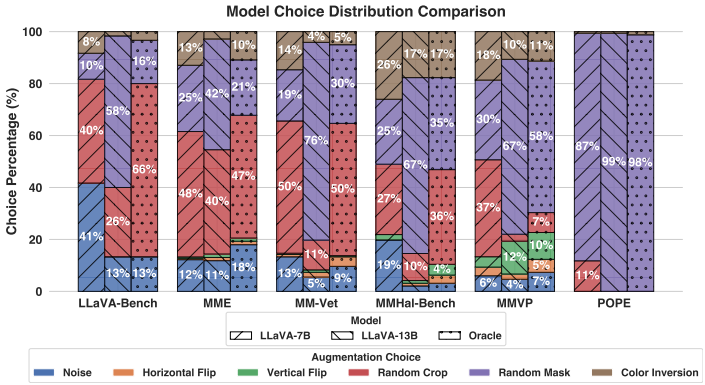


Figure 3: Distribution of self-augmentation choice across model size and benchmarks. Oracle indicates gpt-4o-mini decisions.

Table 4: Comparison with single augmentations with LLaVA-1.5-7B (Liu et al., 2024a) on MME-Perception (Fu et al., 2024). The compared single augmentations are the predefined augmentations in SAS Prompting.

Strategy	Augmentation	MME-P↑
Static	Noise	$1351.76_{\pm 7.59}$
	Horizontal flip	$1302.55_{\pm 47.62}$
	Vertical flip	$1354.42_{\pm 15.13}$
	Random crop	$1315.56_{\pm 41.75}$
	Random mask	$1302.50_{\pm 29.34}$
Adaptive	VACoDe	$1372.50_{\pm 13.78}$
	SAVCD	$1431.30_{\pm 13.87}$
	Oracle	$1435.07_{\pm 22.30}$

recognition in POPE are unified as `Is there a {object} in the image?`, which are directly addressed by the definition of random mask as an occluding operation within the SAS Prompt. In contrast, the uniform distribution of MMVP reflects the diverse nature of the benchmark itself, which queries nine different *visual pattern* categories and thus applies a wider range of augmentations. On the other hand, the infrequent selection of horizontal flip across all benchmarks is a direct result of the evaluated queries rarely testing for horizontal spatial relationships. These findings suggest a broader principle that the set of predefined augmentations must be sufficiently diverse to match the complexity of the visual patterns in a given task, while noting that the specific distribution of choices is also dependent on the SAS Prompt design.

Model Capacity The impact of model scale on the quality of augmentation selection was evaluated by comparing the LLaVA-1.5 7B and 13B models against the Oracle baseline using two primary metrics. First, selection accuracy was measured by calculating the agreement with the Oracle choice, where the 7B model achieved 64.15% and the 13B model achieved 66.19%. Second, the quality of the reasoning trace for each choice was assessed by gpt-4o-mini on a scale of 0 to 10, with the 13B model producing higher quality justifications with an average score of 9.04 compared to 8.28 for the 7B model across benchmarks. The results from both metrics confirm that larger model capacity leads to improved augmentation selection and reasoning quality. Full prompt for reasoning assessment and detailed breakdown of these agreements are provided in the Appendix B.2 and E, respectively.

Comparison with Single Augmentation A comparison between static and adaptive visual augmentation strategies is presented in Tab. 4. Static strategies apply a single, fixed augmentation across all inputs, while adaptive strategies utilize query-aware augmentations. There is a clear performance gap between the two approaches, underscoring the importance of context-optimal augmentation. The significant gap between the proposed method and the others underscores the importance of *query-augmentation semantic alignment* and architectural flexibility, opening the possibility of leveraging diverse knowledge sources from internal knowledge to external reasoning modules.

SAT Threshold The core inverse-entropy heuristic of SAT was validated through a comparison of the proposed H_{decay} against the APC baseline and a normalized scaled entropy H_{ns} : $(-\sum_{i=0}^{|\mathcal{V}|-1}(p)_i \log_2(p)_i / \log_2|\mathcal{V}|)^{1/\gamma}$. The H_{ns} function is a direct-proportional entropy function, *i.e.*, implements the opposing rule to H_{decay}, mapping high input entropy to a more confined threshold. As visualized in Fig. 4, the results confirm a performance hierarchy: H_{decay} consistently outperforms APC and H_{ns} with more stable outputs. A further observation arises from the performance trend within each entropy-based function with respect to the scaling parameter γ. A lower γ absolute value corresponds to a more restrictive threshold for H_{decay} but a more generous one for H_{ns}. These findings not only imply mature thresholding is required to properly penalize false positives, but also provide strong empirical support for the *inverse-entropy principle* in the design of SAT.

Furthermore, the generalizability of SAT was evaluated through a direct comparison with the APC baseline. Both thresholding algorithms were applied to VCD, VACoDe, and SAVCD, and the findings are presented in Tab. 5. The results show that SAT consistently outperforms APC across all decoding configurations, achieving an average performance gain of 4.94%. This performance gain is attributed

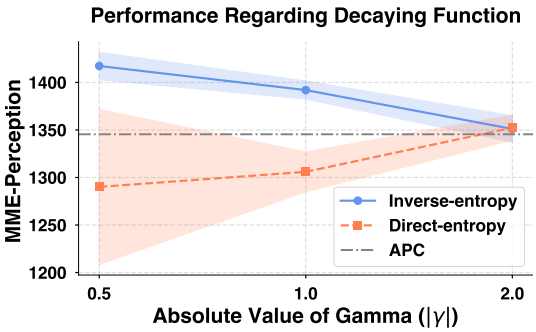


Figure 4: Comparison of the normalized entropy and proposed inverse-entropy function by γ.

Table 5: Plausibility constraint thresholding with APC and SAT.

Decoding	Thresholding	MME-P$^{\uparrow}$
VCD	APC ($\beta = 0.1$)	$1323.67_{\pm 20.84}$
	SAT	$1395.17_{\pm 17.09}$
VACoDe	APC ($\beta = 0.1$)	$1372.50_{\pm 13.78}$
	SAT	$1414.21_{\pm 26.85}$
SAVCD	APC ($\beta = 0.1$)	$1345.46_{\pm 4.65}$
	SAT	$1431.30_{\pm 13.87}$

to the foundational difference in the usage of model confidence. The consistency of this improvement suggests that SAT is broadly applicable to other CD-based methods.

SAS Prompting The individual contributions of the reasoning and ICL components within SAS Prompting were evaluated by selectively removing the reasoning instruction and in-context examples from the full prompt. The results presented in Tab. 6 indicate that removing either component has a minimal impact on performance. Even the weakest configuration, which omits both components, achieves a performance gain of 11.00% against regular sampling. Note that the reasoning instruction, however, is the most impactful factor affecting computational latency, as it requires the model to generate a full text sequence for the justification. Removing this instruction reduces the generation requirement to fewer than ten tokens for the final choice.

Table 6: SAS Prompting with and without reasoning steps and ICL.

Reasoning	ICL	MME-P$^{\uparrow}$
✗	✗	$1419.08_{\pm 11.39}$
✗	✓	$1428.63_{\pm 31.85}$
✓	✗	$1428.02_{\pm 7.92}$
✓	✓	$1431.30_{\pm 13.87}$

5 Limitations and Future Work

The proposed method also presents several branches for future work by addressing current limitations. First, the effectiveness of SAS Prompting depends on the reasoning and instruction-following ability of the base model. Less capable models might produce malformed outputs or poor augmentation choices. This dependency could be addressed in future work by developing more robust prompting methods (*e.g.*, Chain-of-Thoughts (Wei et al., 2022)) or utilizing a smaller, specialized model for the selection task. Second, the current method is limited to a predefined set of visual augmentations. While this set covers common scenarios, it may not contain the best augmentation for highly specialized visual reasoning tasks. A promising direction for future research involves developing methods that can dynamically select from a more diverse and larger library of transformations using external modules (*e.g.*, object detector), enhancing the approach's versatility. Finally, the inclusion of explicit reasoning creates a trade-off between performance and inference speed; this trade-off is flexible and can be controlled by simplifying the prompt as highlighted in the ablation studies. This offers a range of options to suit different application requirements, and further exploration into optimizing this balance is a valuable area for future investigation.

6 Conclusion

This work introduces SAVCD, a novel decoding strategy designed to mitigate hallucinations in LVLMs. The proposed method aligns the semantics between query and visual augmentation by leveraging the flexible intrinsic reasoning of the model without relying on predefined heuristics. In addition, the proposed sparsity adaptive truncation introduces a confidence-aware thresholding that dynamically adjusts candidate sets based on logit entropy, effectively penalizing false positives. Extensive experiments conducted across three LVLM families and seven benchmarks demonstrated that SAVCD consistently improves factual consistency over existing decoding strategies while maintaining practical computational efficiency. Beyond immediate performance gains, this study underlines the importance of the semantic coupling of query-augmentation with confidence-sensitive decoding as a principled approach for developing more robust multimodal generation.

REFERENCES

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

Muhammad Kashif Ali, Eun Woo Im, Dongjin Kim, Tae Hyun Kim, Vivek Gupta, Haonan Luo, and Tianrui Li. Harnessing meta-learning for controllable full-frame video stabilization. *arXiv preprint arXiv:2508.18859*, 2025.

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

Sourya Basu, Govardana Sachitanandam Ramachandran, Nitish Shirish Keskar, and Lav R Varshney. Mirostat: A neural text decoding algorithm that directly controls perplexity. *arXiv preprint arXiv:2007.14966*, 2020.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5185–5198, July 2020. doi: 10.18653/v1/2020.acl-main.463.

Alexander Borichev and Yuri Tomilov. Optimal polynomial decay of functions and operator semigroups. *Mathematische Annalen*, 347(2):455–478, 2010.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901, 2020.

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

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 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, 2024b.

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

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 49250–49267, 2023.

Imant Daunhawer, Thomas M Sutter, Kieran Chin-Cheong, Emanuele Palumbo, and Julia E Vogt. On the limitations of multimodal vaes. *arXiv preprint arXiv:2110.04121*, 2021.

Alessandro Favero, L. Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, A. Achille, Ashwin Swaminathan, and S. Soatto. Multi-modal hallucination control by visual information grounding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312, 2024a. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10655750.

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

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. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. URL https://arxiv.org/abs/2306.13394.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.

Google. Bard. https://bard.google.com/, 2023.

Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. 2019.

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

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

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:72096–72109, 2023.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. *arXiv preprint arXiv:2402.14409*, 2024.

Junho Kim, Hyunjun Kim, Kim Yeonju, and Yong Man Ro. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:133571–133599, 2024a.

Sihyeon Kim, Boryeong Cho, Sangmin Bae, Sumyeong Ahn, and Se-Young Yun. Vacode: Visual augmented contrastive decoding. *arXiv preprint arXiv:2408.05337*, 2024b.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13723–13733, 2025b.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, pp. 19730–19742. PMLR, 2023a.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023b.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. 2023c.

Yaoyiran Li, Anna Korhonen, and Ivan Vulić. Self-augmented in-context learning for unsupervised word translation. *arXiv preprint arXiv:2402.10024*, 2024.

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

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024b.

Xinyu Lyu, Beitao Chen, Lianli Gao, Hengtao Shen, and Jingkuan Song. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *Advances in Neural Information Processing Systems*, 37:122811–122832, 2024.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121, 2023.

OpenAI. ChatGPT. `https://openai.com/blog/chatgpt/`, 2023.

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

Maitreya Patel, Naga Sai Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, et al. Tripletclip: Improving compositional reasoning of clip via synthetic vision-language negatives. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:32731–32760, 2024.

SW Provencher. A fourier method for the analysis of exponential decay curves. *Biophysical journal*, 16(1):27–41, 1976.

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

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 146–162. Springer, 2022.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27 (3):379–423, 1948.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 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)*, pp. 783–791, 2024.

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

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, 2024.

Gabriele N Tornetta. Entropy methods for the confidence assessment of probabilistic classification models. *arXiv preprint arXiv:2103.15157*, 2021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Liam Van der Poel, Ryan Cotterell, and Clara Meister. Mutual information alleviates hallucinations in abstractive summarization. *arXiv preprint arXiv:2210.13210*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:24824–24837, 2022.

Simon Wenkel, Khaled Alhazmi, Tanel Liiv, Saud Alrshoud, and Martin Simon. Confidence score: The forgotten dimension of object detection performance evaluation. *Sensors*, 21(13):4350, 2021.

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

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

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D Goodman. Star: Self-taught reasoner bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 1126, 2024.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a.

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. Alleviating hallucinations of large language models through induced hallucinations. *arXiv preprint arXiv:2312.15710*, 2023b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:46595–46623, 2023.

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

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

SELF-AUGMENTED VISUAL CONTRASTIVE DECODING

APPENDIX

Due to space limitations in the main manuscript, we provide supplementary materials in this appendix that elaborate on the proposed design, experimental settings, and visualizations. This includes the complete prompt design for Self-Augmentation Selection (SAS) and LLM-as-a-Judge for reasoning quality, additional qualitative examples, extended experimental results, and a detailed breakdown of the model and benchmark information.

A ADDITIONAL EXPERIMENTAL SETUP DETAILS

The visual augmentations were implemented on top of the official VCD (Leng et al., 2024) source code with the following specific parameters. The color inversion operation was performed using the PyTorch `torchvision.transforms.functional.invert` function. For the random crop and random mask augmentations, a ratio of 2.0 was used, which corresponds to applying the operation to a randomly placed square patch with side lengths equal to half the original image dimensions. For the noise augmentation, a diffusion noise step of 500 was applied from the official VCD random noise implementation.

B FULL PROMPT DESIGN

This section provides the verbatim prompts used for both the self-augmentation selection and the subsequent reasoning quality assessment. The full SAS Prompt, which leverages in-context learning and reasoning to achieve optimal query-augmentation semantic alignment, is presented first. This is followed by the prompt used to instruct the LLM-as-a-Judge for the evaluation of the SAS reasoning trace against the Oracle. The individual effects of the reasoning and in-context learning components within the SAS Prompt are quantified in the ablation study section of the main manuscript.

B.1 SAS PROMPTING

> You are an expert data augmentation analyst. Your task is to select the single most semantically disruptive image augmentation that most effectively invalidates the question's premise or prevents a confident answer. Provide a clear reason explaining why the augmentation is chosen, then state your final choice.
>
> ## Augmentations and Their Effects ##
> - Vertical flip: Flips image top-to-bottom. Disrupts questions about "above", "below", "under" or reading orientation.
> - Color inversion: Replaces each color with its complement. Disrupts questions relying on accurate color identification.
> - Random crop: Removes random parts of the image. Disrupts questions requiring global context or peripheral objects.
> - Random mask: Occludes portions of the image. Disrupts object presence, count, or attribute recognition.
> - Noise: Adds visual distortion. Disrupts questions requiring small details, texture, or text clarity.
> - Horizontal flip: Flips the image left-to-right. Disrupts questions about left/right positioning and left-to-right text reading.
>
> ## Examples ##
> Question: "Is the mirror above the TV?" Reason: The question focuses on vertical positioning. Vertical flip reverses top and bottom, making "above" mean "below," invalidating the question. Other augmentations don't affect vertical relationships. Choice: vertical flip
> Question: "Is this photo taken indoors?" Reason: The question requires identifying a specific environmental context. Random crop may exclude key background elements like trees, invalidating the question. Flips, color inversion, noise, and random mask don't directly affect scene context. Choice: random crop

Question: “Are there any green beans in the image?” Reason: The question requires identifying a specific color. Color inversion changes green to its complement, invalidating the question. Flips, noise, random mask, and random crop don’t target color directly. Choice: color inversion
Question: “How many people are in the image?” Reason: The question requires counting visible people. Random mask can completely obscure one or more people, making the exact count impossible. Noise obscures details but typically doesn’t hide entire objects, allowing approximate counting. Flips and color inversion don’t affect object visibility or count. Choice: random mask
Question: “Is the cat on the right side of the laptop?” Reason: The question relies on horizontal positioning. Horizontal flip reverses left and right, making “right” mean “left”, invalidating the question. Other augmentations don’t target horizontal positions. Choice: horizontal flip
Question: “Does this artwork exist in the form of painting?” Reason: The question requires identifying the texture of the artwork. Noise obscures fine details, making it hard to identify the medium. Other augmentations don’t target texture details. Choice: noise

Your Answer
If multiple augmentations could disrupt the question, select the one whose effect is most direct and unambiguous. You must choose one of the given augmentations following the “Reason:” and “Choice:” format.
Question: “{text}”

B.2 LLM-AS-A-JUDGE PROMPT FOR REASONING QUALITY

Your task is to evaluate a candidate model’s response against an expert-provided reference solution. The goal is to select the image augmentation that most effectively disrupts the premise of a given question.

Evaluation Rubric (Integer Scale 0-10)
- 10 (Excellent): The candidate’s choice is highly effective and the reasoning is clear, logically sound, and directly supports the choice. The response is of reference quality.
- 7-9 (Good): The choice is effective and the reasoning is logical, but may be slightly less specific or insightful than the reference.
- 4-6 (Acceptable): The choice is plausible but not optimal. The reasoning is generic, weak, or contains minor flaws.
- 1-3 (Poor): The choice is ineffective and the reasoning is flawed or irrelevant.
- 0 (Very Poor): The choice and reasoning are completely incorrect or nonsensical.

Reference Example
Question: “How many people are in the image?”
Reference Reason: “The question requires counting visible people. Random mask can completely obscure one or more people, making the exact count impossible.”
Reference Choice: “random_mask”
Candidate Reason: “Random crop might cut some people out of the frame.”
Candidate Choice: “random_crop”
Evaluation: Score: 7, Reason: The candidate’s choice is a valid strategy for disrupting a counting task, but it is less direct than the reference. The reasoning is correct but lacks specificity.

Task
Question: “{question}”
Reference Reason: “{oracle_reason}”
Reference Choice: “{oracle_choice}”
Candidate Reason: “{model_reason}”
Candidate Choice: “{model_choice}”

Evaluation:

C MODEL AND BENCHMARK DETAILS

Model Families

- LLaVA-1.5 (Liu et al., 2023) is a powerful open-source LVLM that establishes the effectiveness of visual instruct tuning for creating general-purpose visual assistants. Its

architecture is characterized by its simplicity, connecting a pretrained CLIP vision encoder to a Vicuna LLM using a single Multi-Layer Perceptron projection layer. The LLaVA-1.5 version improved upon the original by incorporating a more capable LLM and scaling the instruction-following data.

- Qwen-VL (Bai et al., 2023) is a series of highly performant, versatile vision-language models based on the Qwen language model family. A key feature of the Qwen-VL architecture is its support for multiple languages, the ability to process multi-image inputs, and its strong capabilities in fine-grained visual understanding, including text recognition and object localization.
- InstructBLIP (Dai et al., 2023) is a vision-language instruction tuning framework designed to enhance zero-shot generalization across a diverse set of tasks. Its central innovation is the use of an instruction-aware Query Transformer. This module is trained to extract visual features from the image encoder that are specifically relevant to the given text instruction, enabling more targeted and effective multimodal reasoning.

Discriminative Benchmarks

- **MME** (Fu et al., 2024) is a benchmark that provides a granular evaluation of multimodal tasks, spanning 10 perception and 4 cognition categories. The performance is measured on binary yes or no questions using an accuracy-based MME score. Following the standard practice (Leng et al., 2024; Kim et al., 2024), we consider the perception category for the experiments.
- **MMVP** (Tong et al., 2024) is designed to evaluate a model's understanding of fine-grained visual details. It achieves this by using 300 CLIP-blind image pairs, where models must capture subtle differences to perform paired classification accurately. These image pairs cover nine distinct visual patterns: orientation and direction, feature presence, state and condition, quantity and count, positional and relational context, color and appearance, structural and physical characteristics, text, and viewpoint and perspective. The evaluation follows a multiple-choice format, where final model responses are mapped to the answer options using GPT-4 as an automated judge.
- **POPE** (Li et al., 2023) serves as a dominant benchmark for assessing object hallucination by testing models with three distinct types of negative questions. These categories include queries about random non-existent objects, popular objects that are frequent in the dataset but absent from the image, and adversarial objects selected for their high co-occurrence. The dataset contains 9,000 question-image pairs built from 500 images, each evaluated against multiple questions across the three categories.

Generative Benchmarks

- **LLaVA-Bench (In-the-Wild)** (Liu et al., 2023) is a benchmark to evaluate the ability of Large Vision Language Models (LVLMs) to handle complex tasks and adapt to new domains. It features 24 images and 60 queries, which collapse into three categories: conversation, detailed description, and complex reasoning. The evaluation is conducted using GPT-4V as a judge to rate both the model response and a reference answer. The final performance is reported as a score ratio, calculated by dividing the total score of the reference answer.
- **MMHal-Bench** (Sun et al., 2023) evaluates and penalizes hallucinations across a diverse set of reasoning types. It is composed of 96 image-question pairs that cover eight distinct categories, including object attributes, comparison, and spatial relations. Evaluation is performed using GPT-4V as an automated judge to assess the severity of hallucination in the generated response. The responses are scored on a scale from 0 to 7, where a higher score indicates greater facutal consistency.
- **MM-Vet** (Yu et al., 2023) evaluates an LVLM to integrate multiple multimodal capabilities for complex reasoning. The benchmark defines six fundamental multimodal abilities: recognition, knowledge, OCR, spatial awareness, language generation, and mathematics. A key feature of MM-Vet is its focus on compositional tasks, where these six core abilities are combined to create 16 distinct capability integrations. The dataset itself is composed of 200 images and 218 questions, each requiring a specific combination of these integrated skills.

D ADDITIONAL QUALITATIVE RESULTS

To provide a more granular understanding of the behavior of the method, this section presents additional qualitative results from both discriminative and generative benchmarks. Each example provides a comprehensive analysis that includes the reasoning trace for the chosen augmentation, a stylized visualization of the augmentation, the logit values for the expert, amateur, and contrasted distributions, and the corresponding Sparsity Adaptive Truncation threshold. For improved visualization clarity, common punctuation tokens such as commas and periods have been omitted from the presented logit distributions.

Expert token	b
Amateur token	b
Final token	a
Expert logit	20.88
	20.88
	20.77
Amateur logit	20.89
	20.89
	20.50
Final logit	20.86
	20.86
	20.50
β^{SAT}	0.361

Query:
How many spots are on the animal in the image? (a) 5 (b) 6

Reason: The question requires counting the number of spots on the animal. Random mask can completely obscure the animal, making it impossible to count the spots. Noise and flips don't affect the visibility of the animal or the count.

Expert token	_Left
Amateur token	_Left
Final token	_Right
Expert logit	22.62
	22.62
	22.38
Amateur logit	22.69
	22.69
	22.03
Final logit	22.56
	22.56
	22.72
β^{SAT}	0.253

Query:
Is the sports car in the picture facing left or right?

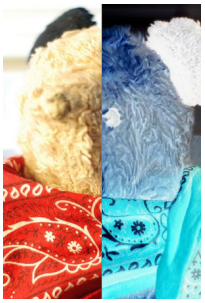
Reason: The question requires determining the direction the sports car is facing. Horizontal flip reverses left and right, making "left" mean "right" and vice versa, invalidating the question. Other augmentations don't target horizontal positions.

Figure 5: Qualitative results on MMVP (Tong et al., 2024).

Expert token	_No
Amateur token	_No
Final token	_Yes
Expert logit	27.56
	27.56
	27.22
Amateur logit	28.64
	28.64
	25.53
Final logit	26.48
	26.48
	28.91
β^{SAT}	0.380

Query:
Are there two horses in this image?

Reason: The question requires identifying the presence of two horses. Random mask can completely obscure one or both horses, making the exact presence impossible. noise and flips don't affect the number of horses or their visibility.

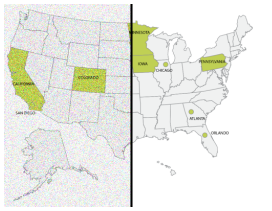


Expert token	_No
Amateur token	_No
Final token	_Yes
Expert logit	26.16
	26.16
	25.33
Amateur logit	27.36
	27.36
	23.19
Final logit	24.95
	24.95
	27.47
β^{SAT}	0.391

Query:
Is there a red scarf in the image?

Reason: The question requires identifying the color of an object. Color inversion changes red to its complement, making it impossible to determine if there is a red scarf in the image. Flips, noise, random mask, and random crop don't directly affect color identification.

Figure 6: Qualitative results on MME (Fu et al., 2024).



Query:
Which of these highlighted cities is the farthest north?

Expert token	_The	_farthest	_north	_highlight	ed	_city	_in	_the	_image	_is	_Min	...
Amateur token	_The	_farthest	_north	_highlight	ed	_city	_in	_the	_image	_is	_Min	...
Final token	_The	_farthest	_north	_highlight	ed	_city	_in	_the	_image	_is	_Chicago	...
Expert logit	17.89	18.28	23.14	19.75	24.11	24.66	21.52	25.17	18.83	23.50	15.25	...
	17.89	18.28	23.14	19.75	24.11	24.66	21.52	25.17	18.83	23.50	15.25	...
	17.89	18.28	23.14	19.75	24.11	24.66	21.52	25.17	18.83	23.50	15.25	...
Amateur logit	17.88	18.05	23.06	19.52	24.09	24.66	21.44	25.06	19.03	23.55	15.22	...
	17.88	18.05	23.06	19.52	24.09	24.66	21.44	25.06	19.03	23.55	15.22	...
	17.88	18.05	23.06	19.52	24.09	24.66	21.44	25.06	19.03	23.55	15.22	...
Final logit	17.91	18.52	23.22	19.98	24.12	24.66	21.59	25.28	18.62	23.45	15.28	...
	17.91	18.52	23.22	19.98	24.12	24.66	21.59	25.28	18.62	23.45	15.28	...
	17.91	18.52	23.22	19.98	24.12	24.66	21.59	25.28	18.62	23.45	15.28	...
β^{SAT}	0.196	0.341	0.475	0.342	0.500	0.494	0.298	0.444	0.351	0.488	0.125	...

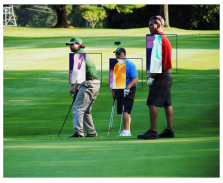
Reason: The question requires identifying the relative position of the cities. Noise obscures the map and makes it difficult to determine the distances between the cities. Other augmentations don't affect the relative positions of the cities.

Query:
What is located to the right of the shampoo?

Expert token	_To	_is	_a	_bott	er	_located	_to	_the	_sh	am	po	o	</s>
Amateur token	_To	_is	_a	_t	ing	_located	_to	_the	_sh	am	po	o	</s>
Final token	_There	_is	_a	_condition	er	_located	_to	_the	_sh	am	po	o	</s>
Expert logit	21.53	26.75	21.55	14.44	23.53	20.19	25.34	27.78	21.47	25.14	27.00	25.00	22.59
	21.53	26.75	21.55	13.80	22.75	20.19	25.34	27.78	21.47	25.14	27.00	25.00	22.59
	21.48	26.75	21.55	13.74	23.53	20.19	25.34	27.78	21.47	25.14	27.00	25.00	22.59
Amateur logit	21.55	26.39	20.62	11.69	22.88	19.44	25.00	27.11	20.28	24.72	26.66	24.09	22.27
	21.55	26.39	20.62	13.70	23.08	19.44	25.00	27.11	20.28	24.72	26.66	24.09	22.27
	21.14	26.39	20.62	10.05	22.88	19.44	25.00	27.11	20.28	24.72	26.66	24.09	22.27
Final logit	21.52	27.11	22.47	17.19	24.19	20.94	25.69	28.45	22.66	25.56	27.34	25.91	22.92
	21.52	27.11	22.47	13.90	22.42	20.94	25.69	28.45	22.66	25.56	27.34	25.91	22.92
	21.52	27.11	22.47	17.44	24.19	20.94	25.69	28.45	22.66	25.56	27.34	25.91	22.92
β^{SAT}	0.222	0.420	0.414	0.082	0.387	0.252	0.426	0.499	0.468	0.500	0.500	0.500	0.330

Reason: The question requires identifying the position of an object relative to another object. Random crop may remove the context of the objects' relationship, making it impossible to determine their position. Flips, color inversion, noise, and random mask don't directly affect relative positioning.

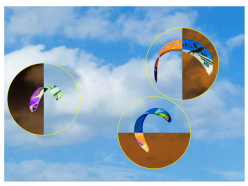
Figure 7: Qualitative results on MM-Vet (Yu et al., 2023).



Query:
What are the colors of the shirts worn by the three men from left to right in the image?

Expert token	_The	_colors	_of	_the	_shirts	_worn	_by	_the	_three	_men	_are	_green	_blue	_red	</s>
Amateur token	_The	_colors	_of	_the	_shirts	_worn	_by	_the	_three	_men	_are	_p	_orange	_p	</s>
Final token	_The	_colors	_of	_the	_shirts	_worn	_by	_the	_three	_men	_are	_green	_blue	_red	</s>
Expert logit	26.75	20.73	27.56	29.00	26.44	25.44	30.19	29.78	26.61	26.05	26.89	19.91	20.89	25.92	27.53
	26.75	20.73	27.56	29.00	26.44	25.44	30.19	29.78	26.61	26.05	26.89	12.73	14.61	17.27	27.53
	26.75	20.73	27.56	29.00	26.44	25.44	30.19	29.78	26.61	26.05	26.89	19.91	20.89	25.92	27.53
Amateur logit	26.66	21.00	27.36	28.94	26.12	25.20	30.02	29.67	26.52	25.97	26.72	16.20	17.41	19.91	27.19
	26.66	21.00	27.36	28.94	26.12	25.20	30.02	29.67	26.52	25.97	26.72	19.97	18.94	21.66	27.19
	26.66	21.00	27.36	28.94	26.12	25.20	30.02	29.67	26.52	25.97	26.72	16.20	17.41	19.91	27.19
Final logit	26.84	20.47	27.77	29.06	26.75	25.67	30.36	29.89	26.70	26.12	27.06	23.61	24.38	31.94	27.88
	26.84	20.47	27.77	29.06	26.75	25.67	30.36	29.89	26.70	26.12	27.06	-inf	-inf	-inf	27.88
	26.84	20.47	27.77	29.06	26.75	25.67	30.36	29.89	26.70	26.12	27.06	23.61	24.38	31.94	27.88
β^{SAT}	0.422	0.294	0.499	0.493	0.487	0.472	0.500	0.499	0.484	0.499	0.498	0.253	0.344	0.494	0.492

Reason: The question requires identifying the colors of the shirts. Color inversion changes each color to its complement, making it impossible to determine the original colors. Flips, noise, random mask, and random crop don't affect color perception.



Query:
What are the colors of the parachutes in the sky?

Expert token	_The	_par	ach	utes	_in	_the	_sky	_are	_blue	_orange	_yellow	</s>
Amateur token	_The	_par	ach	utes	_in	_the	_sky	_are	_blue	_yellow	_pur	</s>
Final token	_The	_par	ach	utes	_in	_the	_sky	_are	_blue	_orange	_green	</s>
Expert logit	27.91	23.41	25.34	25.12	23.72	29.25	25.98	24.62	19.25	19.05	20.59	26.77
	27.91	23.41	25.34	25.12	23.72	29.25	25.98	24.62	19.25	18.78	17.09	26.77
	27.91	23.41	25.34	25.12	23.72	29.25	25.98	24.62	19.25	19.05	20.20	26.77
Amateur logit	28.02	23.67	25.52	25.44	23.94	29.16	25.52	24.53	18.64	17.50	19.66	26.72
	28.02	23.67	25.52	25.44	23.94	29.16	25.52	24.53	18.64	18.64	19.81	26.72
	28.02	23.67	25.52	25.44	23.94	29.16	25.52	24.53	18.64	17.50	17.53	26.72
Final logit	27.80	23.14	25.17	24.81	23.50	29.34	26.45	24.72	19.86	20.59	21.53	26.81
	27.80	23.14	25.17	24.81	23.50	29.34	26.45	24.72	19.86	18.92	-inf	26.81
	27.80	23.14	25.17	24.81	23.50	29.34	26.45	24.72	19.86	20.59	22.88	26.81
β^{SAT}	0.452	0.375	0.499	0.488	0.441	0.500	0.497	0.362	0.128	0.220	0.272	0.475

Reason: The question requires identifying specific colors. color inversion changes each color to its complement, making it impossible to determine the original colors. Flips, noise, random mask, and random crop don't directly affect color perception.

Figure 8: Qualitative results on MMHal-Bench (Sun et al., 2023).

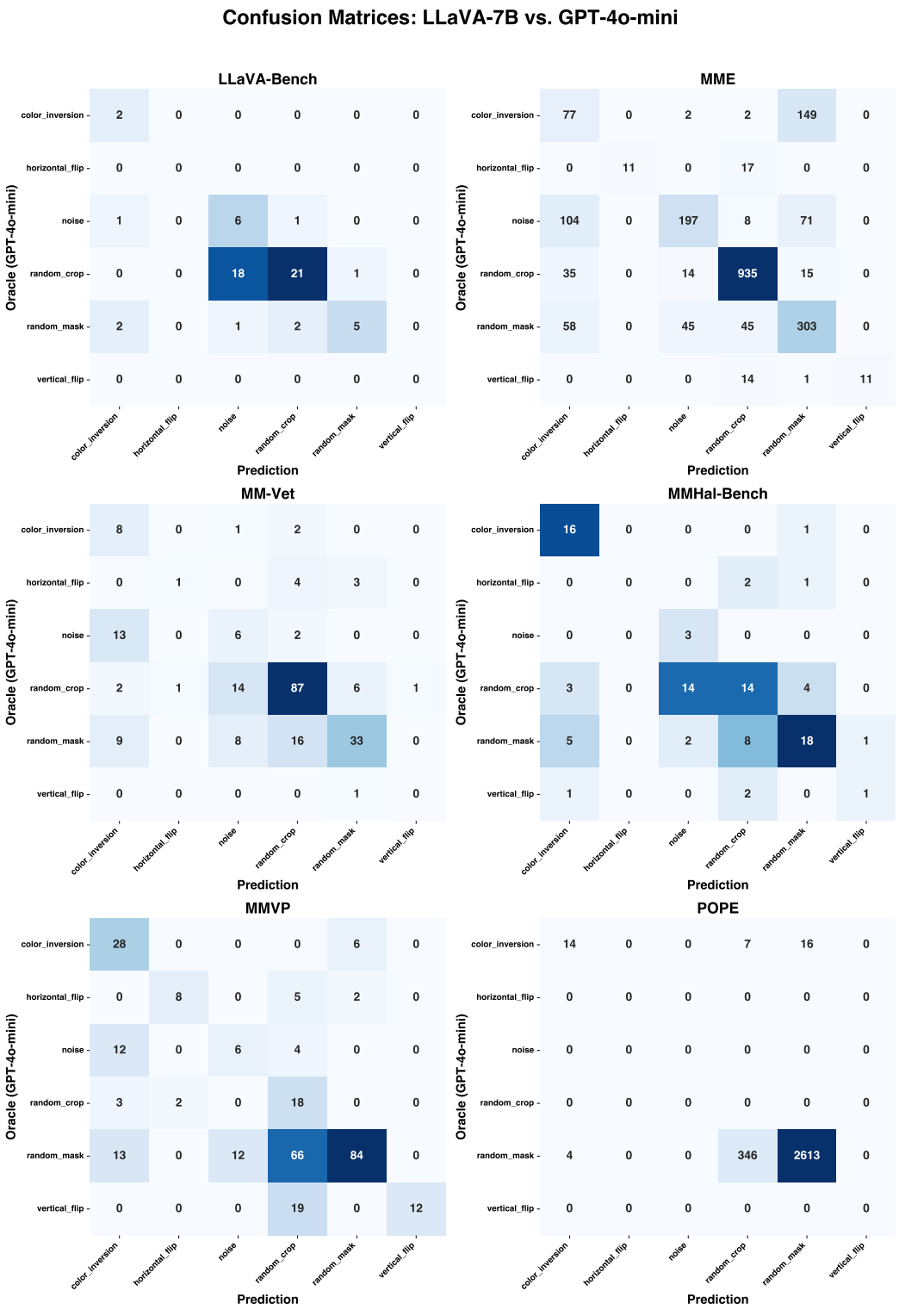


Figure 9: Confusion matrix for LLAVA-1.5 7B model.

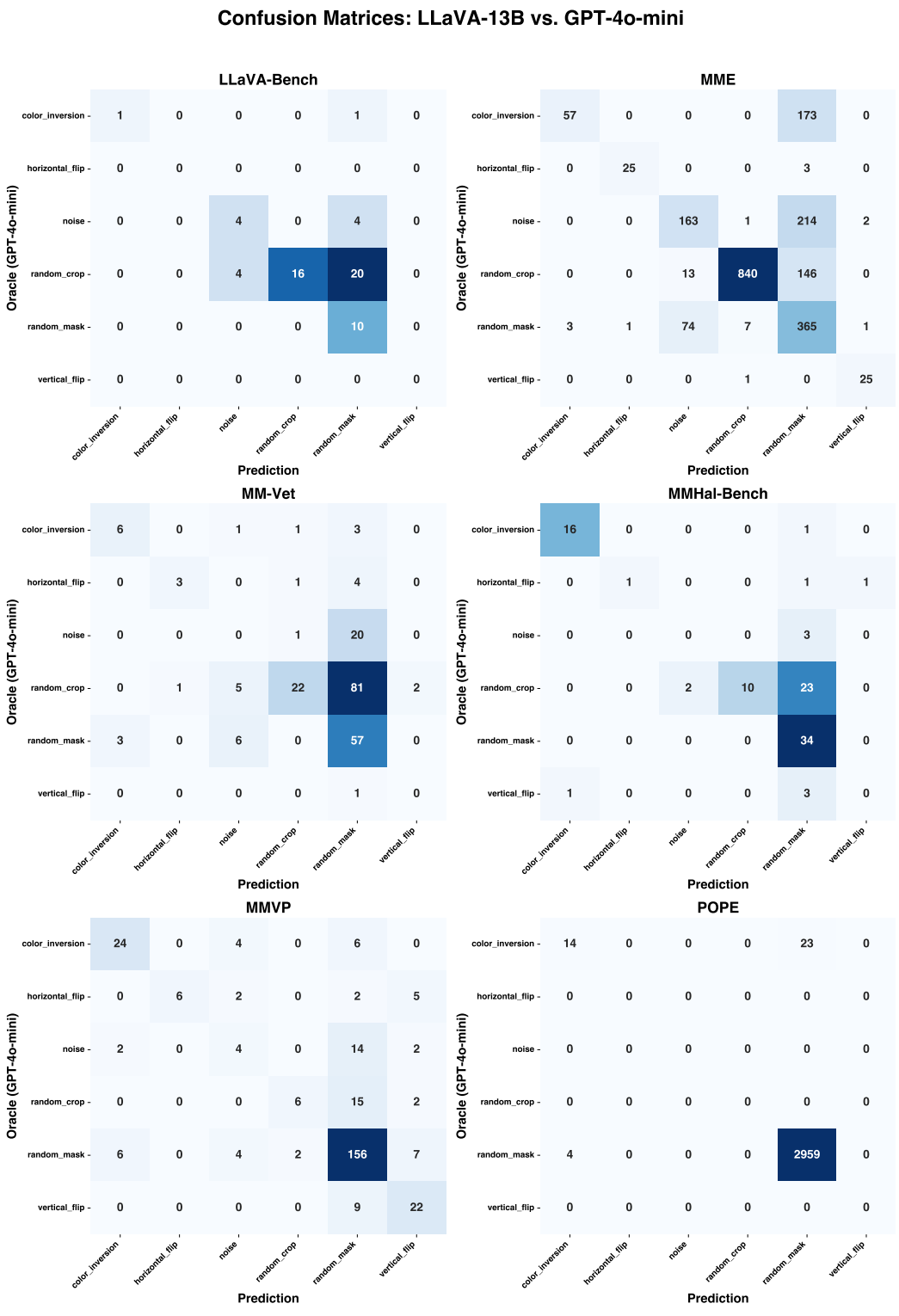


Figure 10: Confusion matrix for LLAVA-1.5 13B model.

Table 7: Comparison of model performance against the gpt-4o-mini oracle. Agreement measures the accuracy percentage (%) of augmentation choices, and Judge Score estimates the quality rating of the model reasoning on a 0 to 10 scale by gpt-4o-mini.

Model	Metric	LLaVA-Bench	MME	MM-Vet	MMHal	MMVP	POPE	Average
LLaVA-7B	Agreement (%)	56.67	72.56	61.93	54.17	52.00	87.57	64.15
	Judge Score	7.97	8.59	7.98	7.79	7.69	9.63	8.28
LLaVA-13B	Agreement (%)	51.67	69.77	40.37	63.54	72.67	99.10	**66.19**
	Judge Score	8.63	9.12	8.55	9.08	8.86	9.99	**9.04**

E Detailed Comparison Against Oracle

In the main script, experiments were conducted to evaluate the impact of the model scale on the quality of augmentation choice and reasoning. The agreement of each model's choice and the quality of its reasoning trace were measured against the Oracle, with the results summarized in Tab. 7. These results confirm that larger model capacity generally leads to better query-augmentation semantic alignment and higher reasoning quality.

A more granular analysis using the confusion matrices in Fig. 9 and Fig. 10, reveals a complex, task-dependent relationship. On uniform benchmarks such as POPE, the alignment between the 13B model and the Oracle is nearly optimal. In contrast, on more complex benchmarks such as MM-Vet, the 13B model exhibits a predictive bias, frequently selecting random crop when the Oracle chooses the functionally similar random mask. Note that this disagreement is not a critical failure, but rather a choice between two functionally similar occlusion-based augmentations.

This finding highlights a key strength of the proposed method. The fact that strong downstream performance is achieved without requiring a perfect, Oracle-level selection confirms that the framework is highly effective at leveraging the competent, albeit imperfect, reasoning of different model scales to significantly improve factual consistency.

References

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

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36: 49250–49267, 2023.

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. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. URL https://arxiv.org/abs/2306.13394.

Sihyeon Kim, Boryeong Cho, Sangmin Bae, Sumyeong Ahn, and Se-Young Yun. Vacode: Visual augmented contrastive decoding. *arXiv preprint arXiv:2408.05337*, 2024.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:34892–34916, 2023.

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

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.