Scaling Test-Time Robustness of Vision-Language Models via Self-Critical Inference Framework

Kaihua Tang[1] Jiaxin Qi[2] Jinli Ou[4] Yuhua Zheng[3] Jianqiang Huang[2,3,4*]
[1] Tongji University, China [2] Computer Network Information Center, CAS, China
[3] HIAS, University of Chinese Academy of Sciences, China
[4] University of Chinese Academy of Sciences, China
tangkaihua@tongji.edu.cn, jxqi@cnic.cn, oujinli@zuaa.zju.edu.cn
zhengyuhua@ucas.ac.cn, jqhuang@cnic.cn

Abstract

The emergence of Large Language Models (LLMs) has driven rapid progress in multi-modal learning, particularly in the development of Large Vision-Language Models (LVLMs). However, existing LVLM training paradigms place excessive reliance on the LLM component, giving rise to two critical robustness challenges: language bias and language sensitivity. To address both issues simultaneously, we propose a novel Self-Critical Inference (SCI) framework that extends Visual Contrastive Decoding by conducting multi-round counterfactual reasoning through both textual and visual perturbations. This process further introduces a new strategy for improving robustness by scaling the number of counterfactual rounds. Moreover, we also observe that failure cases of LVLMs differ significantly across models, indicating that fixed robustness benchmarks may not be able to capture the true reliability of LVLMs. To this end, we propose the Dynamic Robustness Benchmark (DRBench), a model-specific evaluation framework targeting both language bias and sensitivity issues. Extensive experiments show that SCI consistently outperforms baseline methods on DRBench, and that increasing the number of inference rounds further boosts robustness beyond existing single-step counterfactual reasoning methods. [1]

1. Introduction

The recent advance in Large Language Models [1, 4, 6, 25, 38] (LLMs) has not only revolutionized the field of natural language processing but also catalyzed significant progress in multi-modal research, particularly in the vision-language domain [49, 51]. To better utilize the knowledge of LLMs, the prevalent training framework for Large Vision-Language Model (LVLM) integrates a visual encoder with a pretrained LLM and jointly fine-tunes the combined architecture, resulting in powerful and versatility LVLMs such as InstructBLIP [9], LLaVA series [26, 27] and Qwen-VL series [5, 39].

However, these LVLMs continue to suffer from robustness issues in two key aspects. First, the above-mentioned LLM-based vision-language framework inevitably inherits certain drawbacks of LLMs, such as sensitivity to language prompts [3, 18, 42]. Conventional VQA models lack the large-scale pretraining of LLMs and thus can only understand very limited textual information, failing to capture subtle prompt variations and thereby side-stepping this issue. As illustrated in Figure 1(a), simply requesting a LVLM to check image details without altering the question results in different outputs for the same input image. This language sensitivity undermines the consistency of LVLMs, reducing their reliability from the user's perspective. Second, vision-language models are also known to be susceptible to language bias. For example, conventional Visual Question Answering (VQA) models often rely heavily on language priors to answer questions, disregarding visual input [30, 41]. As shown in Figure 1(b), this problem also persists in LVLMs and can sometimes lead to generating non-existent contents, known as object hallucination [22, 24].

Recently, a growing number of research has focused on mitigating object hallucination in LVLMs [24, 52]. Among these efforts, Visual Contrastive Decoding (VCD) [22] and its variants [36, 43] have emerged as some of the most effective and widely adopted solutions. These methods typically perform a standard inference to obtain baseline logits and then estimate potential biases via a secondary inference with perturbed inputs. The final unbiased prediction is derived by weighted subtraction of the two logits. However, the object hallucination is merely a continuation of the lan-

*Corresponding author.
[1] Our code is publicly available on GitHub: https://github.com/KaihuaTang/Self-Critical-Inference-Framework

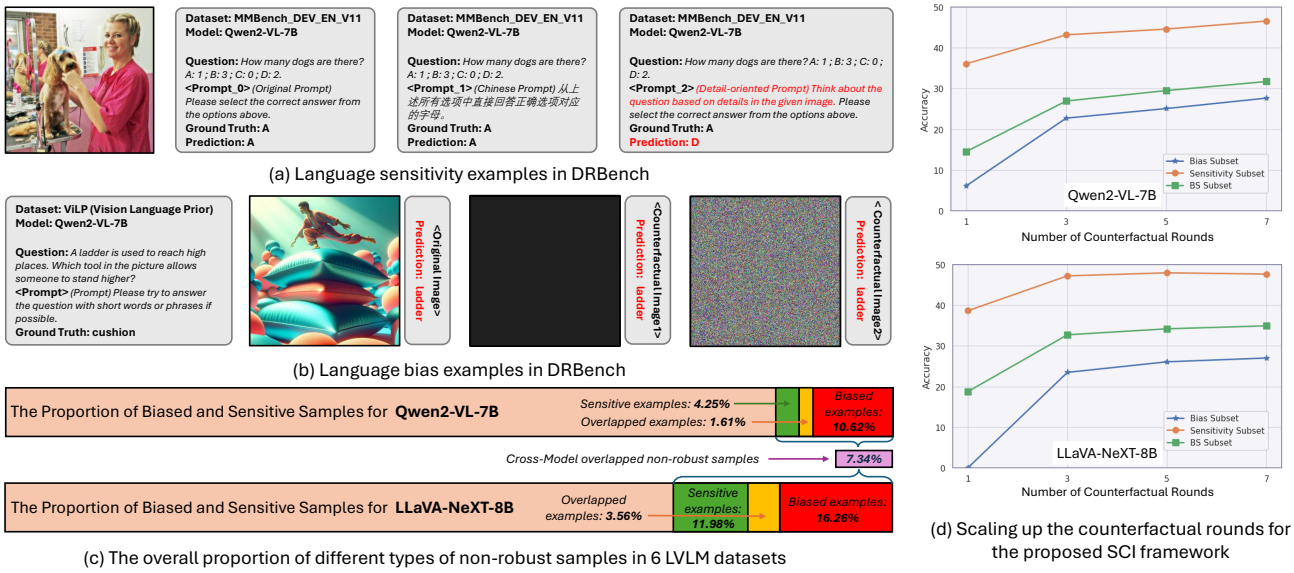

Figure 1. (a) and (b) are real DRBench examples suffering from language sensitivity and bias issues; (c) shows the overall proportion of different types of non-robust samples across all 6 datasets under two commonly used LVLMs; (d) demonstrates a novel test-time scaling strategy of robustness regarding the increased counterfactual rounds in the proposed SCI.

guage bias observed in conventional VLMs [30, 37], and it ignores the issue of language sensitivity that is newly introduced by LVLMs.

In this work, we first analyze the underlying principles of VCD, particularly the role of the trade-off parameter α, which is absent in the original Contrastive Decoding (CD) [23]. Through an in-depth mathematical analysis, we demonstrate that VCD is theoretically aligned with some debiasing algorithms used in previous vision-language tasks, such as TDE [37] and TIE [30]. Specifically, VCD leverages TIE logits to reweight the original logits, where $1/\alpha$ acts as the temperature parameter for logit scaling. Building on this insight, we propose a more comprehensive inference paradigm, termed Self-Critical Inference **(SCI)** framework, which unifies both Textual Counterfactual (TC) and Visual Counterfactual (VC) components. The final prediction is then derived from aggregating and comparing all multi-round counterfactual logits. This approach generalizes VCD and enables the simultaneous mitigation of both bias and sensitivity issues. We further examine three configurations: SCI_3, SCI_5, and SCI_7, with different numbers of input variations to investigate the effect of increasing counterfactual inference rounds. We argue that our approach establishes a new potential test-time scaling direction, distinct from prior methods that increase intermediate context token lengths within a single inference. Instead, robustness can be enhanced by performing increased round of counterfactual inference.

We also introduce a new evaluation benchmark, termed Dynamic Robustness Benchmark **(DRBench)**, to adaptively assess the robustness improvements of individual models. The key motivation behind DRBench is that those non-robust data samples are not fixed. As shown in Figure 1(c), among all 24.68% hard samples for one LVLM (LLaVA-NeXT), there are only 7.34% shared with another LVLM (Qwen2-VL). This suggests that an LVLM may perform perfectly well on a fixed robustness dataset, yet still be vulnerable to other new samples. Besides, to enable a more precise analysis of algorithmic contributions, it is essential to disentangle the robustness gains from the confounding effect of base model performance. To this end, this benchmark is constructed by adaptively extracting non-robust subsets from existing LVLM datasets, based on the performance of a given LVLM. These model-specific subsets prevent newly introduced LVLMs from covering robustness issues by overfitting to existing datasets. Notably, the DRBench is easily scalable and can be seamlessly applied to widely used real datasets such as MMBench, MME, *etc.*, introducing more diverse and nature question types than previous datasets [24]. Furthermore, as illustrated in Figure 1(c), the additional statistical information itself from DRBench facilitates a more comprehensive diagnosis of the inherent vulnerabilities of each LVLM.

The main contributions of this paper are threefold: 1) We propose SCI, a counterfactual inference framework that simultaneously mitigates language bias and enforces language consistency. 2) We introduce the DRBench, a model-specific and dynamic benchmark designed to better assess the robustness of LVLMs under samples from real downstream tasks. 3) We demonstrate that SCI consistently im-

proves performance on both the DRBench and standard datasets, exhibiting strong generalizability. Furthermore, we reveal a previously underexplored potential for improving robustness by increasing the number of test-time counterfactual inference rounds.

2. Related Work

Large vision-language models. LVLMs integrate two of the most significant breakthroughs in recent years: the versatile image encoder CLIP [34] and LLMs for general-purpose question answering [31, 33, 38, 44]. The typical inference pipeline of a LVLM proceeds as follows: the input image is first encoded by CLIP or its more advanced successors [50] to extract patch-level visual features; an adapter then maps these features to the token embedding space of the LLM [5, 26]; finally, the visual and textual token embeddings are jointly fed into the LLM to generate the response. LVLMs have shown broad applicability in vision-language tasks such as image captioning [45–47] and Visual Question Answering (VQA) [2].

Language bias and sensitivity in vision-language models. Language bias has been a longstanding challenge for visual-language models. Previously, it was widely studied as the language prior in tasks like VQA[14, 30]. In today's LVLMs, it commonly manifests as object hallucination. Recent works have sought to mitigate it through targeted retraining and contrastive decoding strategies[15, 19, 22], which are parallel to earlier techniques such as rebalanced training and counterfactual inference [7, 30]. Meanwhile, sensitivity to language prompts has received considerably less attention in VL research. Early VQA systems sidestepping this issue by using a small language encoder. The emergence of LLMs has brought it to the forefront. Existing mitigation strategies can be broadly categorized into three groups: 1) prompt ensembling [32]; 2) RL-based prompt optimization [21]; 3) Chain-of-thought verification [40].

Test-time scaling. Scaling laws have always been central to understanding LLM behavior, particularly the positive correlation between the scale of model/dataset/compute and the performance [17, 20]. Recently, the attention has shifted toward test-time scaling, where increasing inference-time compute is also critical [35], such as adding demonstrations or decoding steps. In this work, we extend the notion of test-time scaling to the robustness: rather than increasing intermediate token length in a single inference, our method improves LVLM robustness by aggregating logits across more counterfactual inference rounds.

3. Methodology

3.1. Preliminaries

Counterfactual VQA: the use of counterfactual inference to mitigate language bias in vision-language tasks dates back to Unbiased SGG [37] and CF-VQA [30]. These works were the first to introduce the concepts of Total Direct Effect (TDE) and Total Indirect Effect (TIE) from the field of causality to achieve unbiased estimations via logit subtraction.

Since an LVLM can be regarded as a general VQA model, we take CF-VQA as an example. The TIE-based counterfactual logits can be formulated as:

$$TIE = Z(q, v, k) - Z(q, v^*, k^*), \tag{1}$$

where $Z(\cdot)$ denotes the model producing answer logits, q denotes the question feature, v is the visual feature, k is the multi-modal fusion feature, v^* and k^* are counterfactual dummy features agnostic to the inputs. In conventional VQA, which is formulated as a closed-set classification task, the unbiased answer is obtained by returning the candidate answer with the highest TIE logits.

Visual Contrastive Decoding (VCD): building upon the idea of Contrastive Decoding (CD) [23], VCD extends CD to mitigate object hallucination during LVLM inference, which can be formulated as follows:

$$\begin{aligned} p(y|v, v^*, q) = softmax((1+\alpha)\,&\text{logit}(y|v, q) \\ &- \alpha\,\text{logit}(y|v^*, q)), \end{aligned} \tag{2}$$

where y denotes the generated discrete token, α is a trade-off hyperparameter, q and v represent the input textual and visual tokens, respectively, and v^* corresponds to visual tokens obtained from a noisy image. The previously generated tokens are considered part of q for simplicity. The final VCD answer is therefore iteratively sampled from $p(y|v, v^*, q)$.

3.2. Self-Critical Inference Framework

In this paper, we observe that VCD essentially reweights the original logits using TIE logits from CF-VQA. Building on this insight, we propose a Self-Critical Inference (SCI) framework, which enhances model robustness through systematic logit-level reasoning over textual and visual counterfactual samples. The proposed SCI framework not only unifies the formulations of VCD and CF-VQA, but also provides a principled solution to both language bias and sensitivity.

We begin by revisiting VCD through the lens of CF-VQA. Specifically, we treat object hallucination in LVLMs as the consequence of iterative biased token generation and frame the decoding process as a sequence of biased classifications. This perspective highlights that LVLMs are fundamentally no different from conventional VQA models. At each generation step, the bias can be mitigated through reasoning over counterfactual logits. Based on this observation, we transform the probability expression in (2) into a logit-based formulation $Z_{vcd}(v, v^*, q)$ as follows:

$$Z_{vcd}(v, v^*, q) = (1+\alpha)\,Z(v, q) - \alpha\,Z(v^*, q), \tag{3}$$

where $Z(\cdot)$ denotes the LVLM that takes both textual tokens q and visual tokens (either v from real images or v^* from dummy ones) as input and output the logits for the next token. Since there are no explicit multi-modal fusion features in the LVLM inputs, we removes k or k^* in original TIE.

To better understand the relationship between VCD and TIE, we transform the above VCD logits into the $\exp(\cdot)$ domain. By explicitly expanding the $softmax$ function $exp(x_i)/(\sum_j exp(x_j))$ and omitting the normalization term, we approximate the probability using $p(y) \propto exp(\cdot)$. With this simplification, the VCD probability $p(y|v, v^*, q)$ in (2) can be rewritten as:

$$\begin{aligned} p(y|v, v^*, q) &= softmax(Z_{vcd}(v, v^*, q)) \\ &\propto exp(Z(v, q) + \alpha\,(Z(v, q) - Z(v^*, q))) \\ &= exp(Z(v, q)) \cdot exp(\alpha\,(Z(v, q) - Z(v^*, q))) \\ &= exp(Z(v, q)) \cdot exp(TIE/\tau). \end{aligned} \tag{4}$$

The above formulation bridges VCD and CF-VQA, showing that VCD essentially performs weighted token generation upon the original output token probability $p(y|v, q) \propto exp(Z(v, q))$, where TIE logits $exp(TIE/\tau)$ serves as a vocabulary-wise reweighting term, thus forcing the model to rely on visual difference. This formulation also clarifies the role of α in VCD. Neither vanilla CD nor TIE itself requires this additional parameter, because the logit difference itself captures the useful effect of real v over dummy v^*. Yet, as a reweighting term, it requires a temperature scaling factor to adjust the trade-off strength, so we further denote $\tau = 1/\alpha$.

To establish a more general robust inference framework, it is also necessary to address the overlooked language sensitivity issue as well. Therefore, we propose SCI framework to incorporates both a Visual Counterfactual (VC) component, which enhances visual cues similar to TIE, and a Textual Counterfactual (TC) component, which ensures prompt-consistent logits, as follows:

$$p_{\text{SCI}}(y|\boldsymbol{v}, \boldsymbol{q}) \propto exp(\text{TC}/\tau_1) \cdot exp(\text{VC}/\tau_2), \tag{5}$$

$$\text{TC}_k = max_i(Z_k(v^0, q^i)),\ i = 0, 1, 2, ..., N \tag{6}$$

$$\text{VC} = Z(v^0, q^0) - \mathbb{E}[Z(v^j, q^0)],\ j = 1, 2, ..., M \tag{7}$$

where $\boldsymbol{v} = \{v_j\}_{j=0}^{M}$ and $\boldsymbol{q} = \{q_i\}_{i=0}^{N}$ denote overall inputs; M and N are the number of visual and textual counterfactual variations, respectively; v^0 and q^0 stand for original visual and textual tokens; $\{v^j, j \neq 0\}$ and $\{q^i, i \neq 0\}$ represent counterfactual visual tokens generated from content-removed images and counterfactual textual tokens from semantically equivalent but lexically different prompts, respectively. The detailed implementation of these counterfactual samples will be explained in Experiments and Appendix. The operator $max_i(Z_k(\cdot))$ computes the element-wise maximum over $N+1$ samples on the $k-$th dimension of the logits for better consistency. VC enhances the original TIE by incorporating multiple counterfactual visual inputs to obtain a more stable estimation. τ_1 and τ_2 are temperature scaling factor for TC and VC logits, respectively. Following VCD, we also adopt Adaptive Plausibility Constraints as a post process before sampling from $p_{\text{SCI}}(y)$, details will be given in Appendix.

The overall SCI framework provides a generalized solution for robust LVLM inference. In this unified framework, prior works such as VCD and CF-VQA can be viewed as special cases. For VCD, there are no counterfactual prompt variations ($N = 0$) and only one counterfactual image ($M = 1$). For CF-VQA, the entire TC component is set to a constant and $M = 1$. As demonstrated in our experiments, increasing the number of counterfactual inference rounds, *i.e.* using larger M and N, leads to more robust final outputs, revealing a new potential test-time scaling strategy for robustness in LVLMs. We also believe that there remains a large opportunity to improve the effectiveness by developing more advanced TC and VC algorithms in future work.

Subset Size	B Subset	S Subset	BS Subset	Overlap
LLaVA-NeXT (MCQ)	1810	1005	2476	339
LLaVA-NeXT (Others)	345	582	794	133
LLaVA-NeXT (Overall)	2155	1587	3270	472
Qwen2-VL (MCQ)	1080	252	1243	89
Qwen2-VL (Others)	327	311	513	125
Qwen2-VL (Overall)	1407	563	1756	214

Table 1. The size of each subset in constructed DRBench. The overall number of samples across all 6 datasets is 13251, with MCQ and Others categories being 10632 and 2619, respectively.

3.3. Dynamic Robustness Benchmark

Collecting and constructing datasets tailored to specific robust issues is often cumbersome and costly. What's worse, once such datasets are publicly released, they may be inadvertently integrated into the web-crawled training corpus of subsequent LVLMs. To better evaluate language bias and sensitivity in real downstream tasks, we introduce the Dynamic Robustness Benchmark (DRBench), guided by two main motivations: 1) the evaluation benchmark should be model-specific and dynamic. Since different LVLMs may exhibit varying levels of robustness and their vulnerable samples are not the same, it is important to disentangle the confounding effect of the base model performance from the improvements brought by different inference strategies, so we can better understand the contribution of the inference algorithm itself; 2) existing LVLM bias evaluation datasets typically focus on a single question type and adopt formats that differ significantly from real-world LVLM tasks, *e.g.* exist-or-not questions [48]. Therefore, it is necessary to develop methods that can automatically adapt to diverse ques-

tion types and task formats.

Following the above two guiding principles, the proposed benchmark enables the transformation of any popular or newly released LVLM dataset regardless of its question formats into a robustness evaluation benchmark. Specifically, it will adaptively generate model-specific bias subset, sensitivity subset and their union BS Subset for any given LVLM dataset through a two-step process. First, we will evaluate the dataset using the given model. Then, we will adopt the following criteria for filtering the Bias Subset (BS) and the Sensitivity Subset (SS): BS $= \{(a_{gt}, v^0, q^0) \,|\, \forall j \neq 0, \arg\max_a p(a|v^0, q^0) = \arg\max_a p(a|v^j, q^0) \neq a_{gt}\}$ and SS $= \{(a_{gt}, v^0, q^0) \,|\, \forall i \neq 0, \arg\max_a p(a|v^0, q^0) \neq \arg\max_a p(a|v^0, q^i)\}$, where a and a_{gt} denote the predicted answer and the ground-truth answer, respectively, and $\arg\max_a p(a|\cdot)$ means the predicted answer is obtained via greedy decoding. The generation of counterfactual inputs v^j and q^i follows the same procedure as in SCI. In this paper, we fix $M = N = 2$ for all our subsets construction. In essence, for BS, we select samples that yield the same incorrect predictions under both the original and dummy visual inputs, indicating a reliance on spurious language priors; for SS, we identify samples whose predictions change in response to subtle, non-causal prompt variations. The final BS Subset is defined as the union of the above two subsets, enabling the investigation of both bias and sensitivity issues. We further split all samples into two groups based on their question types: MCQ for the dominant Multiple-Choice Question type and Others for Yes/No or general open-ended QA types.

In summary, the proposed DRBench offers three key advantages. First, robustness is a model-specific problem, samples that are biased or sensitive for one model may not be vulnerable for another, more evidence will be provided in Table 3. An adaptive and model-specific robustness benchmark can thus prevent newly developed LVLMs from being exposed to publicly released fixed datasets and misleading the evaluation of their real underline robustness. Second, as shown in Figure 1(c), different models exhibit varying levels of robustness, the size of each subset provides valuable insight into different models. For example, Table 1 indicates that: 1) Qwen2-VL is generally more robust than LLaVA-Next; 2) Qwen2-VL is more vulnerable to bias than to sensitivity; and 3) LLaVA-NeXT exhibits more sensitivity issues compared to Qwen2-VL. Third, DRBench enables the evaluation of robustness in various real-world tasks, rather than predefined questions such as a simple exist-or-not (Yes/No) assessment commonly used in previous work [48]. It also allows for the effortless conversion of any real-world LVLM dataset into the DRBench format, eliminating the need for labor-intensive sample collection and manual annotation.

4. Experiments

4.1. Benchmark Settings

In our experiments, we construct DRBench using 6 widely adopted LVLM benchmarks: MME [13], MMStar [8], CCBench [28], ViLP [29], MMBench-DEV-EN-V11 and MMBench-DEV-CN-V11 [28]. We begin by randomly splitting the datasets into 20% validation and 80% test sets, resulting in 3,315 and 13,251 samples, respectively. Detailed subset statistics are provided in Table 1. Note that the size of DRBench increases with larger number of M and N. For consistency and convenience, we fix $M = N = 2$ for all subsets constructions throughout our experiments. As we mentioned, to enable a more fine-grained analysis, we separately report performance for Multiple-Choice Question (MCQ) and Others (Open-ended QA for ViLP or Yes/No for MME) categories, in addition to the overall results. We use top-1 accuracy as the evaluation metric for all experiments. For the MME dataset, which adopts a different scoring metric, we convert its results to accuracy, so they can be integrated with samples from other datasets to get the final results.

4.2. Implementation Details

Model Zoo. We used Hugging Face versions of Qwen2-VL-7B-Instruct [39] and Llama3-LLaVa-NeXT-8B-hf [27] as our base models. Following their default configurations, the experiments were conducted using bfloat16 precision and top-k sampling decoding for Qwen2-VL, while LLaVa-NeXT used float16 and greedy decoding.

Environments. All experiments were conducted using VLMEvalKit [11] on a single NVIDIA A800 GPU of 80GB memory with environment: Pytorch=2.6, Transformers=4.49, and Flash Attention=2.7 [10].

Algorithm details. We evaluated 4 inference strategies: TIE, VCD, M3ID, and the proposed SCI. We adapted Total Indirect Effect (TIE) from CF-VQA [30] to LVLMs by removing the multi-modal features k and k^* in Eq. (1). For fair comparison, we also incorporated the Adaptive Plausibility Constraints used in VCD and M3ID into TIE. VCD [22] and M3ID [12] share the same mathematical formulation as Eq. (4), except that the hyperparameter τ in M3ID varies depending on the position of the predicted token. For the proposed SCI, we added subscripts such as SCI_3, SCI_5, and SCI_7 to indicate the number of inference rounds. For example, SCI_5 means that the total number of counterfactual visual and textual variations, together with the original inputs is 5, *i.e.*, $M + N + 1 = 5$ In our experiments, we set $M = N = 1$, $M = N = 2$, and $M = N = 3$ for SCI_3, SCI_5, and SCI_7, respectively.

Counterfactual sample construction. We constructed up to 3 visual counterfactual variations and 3 prompt variations: 1) VC-Color0(C0) renders the input image into black;

Method	B Subset			S Subset			BS Subset		
	MCQ	Others	Overall	MCQ	Others	Overall	MCQ	Others	Overall
LLaVA-NeXT	0.0	0.0	0.0	39.2	37.63	38.63	15.91	27.58	18.75
LLaVA-NeXT-TIE	12.98	23.48	14.66	39.00	57.56	45.81	21.89	44.21	27.31
LLaVA-NeXT-VCD	12.65	25.51	14.71	40.50	56.53	46.38	22.54	44.58	27.89
LLaVA-NeXT-M3ID	16.91	25.22	18.24	39.90	56.36	45.94	24.15	44.33	29.05
LLaVA-NeXT-SCI$_3$ (ours)	21.22	35.36	23.48	39.60	60.31	47.20	27.14	50.13	32.72
LLaVA-NeXT-SCI$_5$ (ours)	23.81	37.97	26.08	**40.60**	**60.65**	**47.95**	28.80	51.01	34.19
LLaVA-NeXT-SCI$_7$ (ours)	**24.86**	**38.26**	**27.01**	40.10	**60.65**	47.64	**29.68**	**51.26**	**34.92**
Qwen2-VL	5.37	8.56	6.11	38.10	34.41	36.06	10.78	23.59	14.52
Qwen2-VL-TIE	16.20	16.82	16.35	45.63	36.66	40.67	20.27	27.29	22.32
Qwen2-VL-VCD	15.74	21.71	17.13	46.83	40.84	43.52	20.11	30.41	23.12
Qwen2-VL-M3ID	19.81	21.71	20.26	**47.22**	41.16	43.87	23.65	30.6	25.68
Qwen2-VL-SCI$_3$ (ours)	21.67	26.30	22.74	44.05	42.44	43.16	24.54	32.75	26.94
Qwen2-VL-SCI$_5$ (ours)	24.91	25.69	25.09	**47.22**	42.44	44.58	28.00	33.14	29.50
Qwen2-VL-SCI$_7$ (ours)	**27.04**	**29.66**	**27.65**	**47.22**	**45.98**	**46.54**	**29.61**	**36.84**	**31.72**

Table 2. Experiments on B(ias) Subset, S(ensitivity) Subset, and BS Subset of the proposed DRBench. **Bold texts** indicate the best result of each column and underline texts indicate the second best result.

Construction Model	Methods	MCQ	Others	Overall
LLaVA-NeXT	LLaVA-NeXT-Original	15.91	27.58	18.75
	LLaVA-NeXT-SCI$_5$	28.80	51.01	34.19
	Qwen2-VL-Original	59.29	63.48	60.31
	Qwen2-VL-SCI$_5$	61.15	67.88	62.78
Qwen2-VL	Qwen2-VL-Original	10.78	23.59	14.52
	Qwen2-VL-SCI$_5$	28.00	33.14	29.50
	LLaVA-NeXT-Original	30.25	39.18	32.86
	LLaVA-NeXT-SCI$_5$	34.59	41.33	36.56

Table 3. Ablation on cross-model BS Subset evaluation.

2) VC-Noise500(N500) and 3) VC-Noise400 apply the diffusion noise function from VCD, using noise steps of 500 and 400, respectively; 3) TC-V1 adds an additional system prompt instructing the model to focus on image details; 4) TC-V2 further modifies the system prompt's language from English to Chinese or vice versa; 5) TC-V3 that injects identity information by prompting the model to respond as a clever student. More detailed prompts will be given in the Appendix.

Hyperparameter settings. Based on the validation results, we set τ_1 to 1.5, 2, and 2.5 for SCI$_3$, SCI$_5$, and SCI$_7$, respectively. Since the TC component involves element-wise maximum over logits, its magnitude increases with the number of variations N. Therefore, the temperature scaling factor τ_1 should be increased accordingly to maintain a similar distribution of TC logits. The τ_2 is fixed at 0.2, because the averaging operation in the VC logits stabilizes the distribution and mitigates the need for the scaling change. For the Adaptive Plausibility Constraint [22] used in our experiments, the threshold parameter is set to 0.3 unless otherwise specified. More details about the constraint and hyperparameter ablation will be introduced in the Appendix.

4.3. Experimental Results

Experiments on the proposed DRBench. As shown in Table 2, we adopted two state-of-the-art LVLMs for our experiments: LLaVA-NeXT-8B and Qwen2-VL-7B. We compared the base model performances and three other algorithms: TIE, VCD, and M3ID that utilized counterfactual inference. The proposed methods, SCI$_3$, SCI$_5$, and SCI$_7$, consistently demonstrated superior performance across the B(ias), S(ensitivity), and combined BS Subsets. We further reported MCQ and Others results based on question types and saw that the improvements brought by SCI were consistent across both categories. Table 2 also reveals that the proposed DRBench can successfully disentangle the base model performance and focus on investigating the effectiveness of inference algorithms, as Qwen2-VL outperforms LLaVA-NeXT by 10.19% on the original datasets in Table 4, while their base and final overall performances on the DRBench are very close.

Experiments on real-world LVLM datasets. We further evaluated the proposed SCI on 6 popular LVLM datasets to verify its performance under real-world data distributions, in addition to the proposed subsets alone. Taking SCI$_5$ as an example in Table 4, it consistently outperformed the baseline models in all question types and almost all datasets. Meanwhile, TIE, VCD and M3ID decrease the performance on Others question type. Note that although the improvements appear relatively marginal, since vulnerable samples

Method	Single Dataset						Question Type		
	MMB-C	MMB-E	MME	CCB	MMS	ViLP	MCQ	Others	Overall
LLaVA-NeXT	78.0	79.72	79.57	47.0	44.75	51.53	70.12	71.86	70.46
LLaVA-NeXT-TIE	78.28	80.28	77.30	45.65	46.00	53.19	70.36	70.68	70.42
LLaVA-NeXT-VCD	78.38	80.28	78.09	46.63	45.00	54.31	70.44	71.55	70.66
LLaVA-NeXT-M3ID	78.31	80.18	78.62	45.89	45.92	54.03	70.36	71.86	70.66
LLaVA-NeXT-SCI$_5$ (ours)	78.21	80.08	80.15	46.20	45.75	53.06	70.32	72.70	70.79
Qwen2-VL	85.26	86.36	87.89	73.22	59.50	56.53	80.91	79.27	80.58
Qwen2-VL-TIE	86.00	86.59	86.52	73.84	59.00	57.08	81.30	78.43	80.73
Qwen2-VL-VCD	86.05	86.56	86.41	73.77	60.08	57.92	81.42	78.58	80.86
Qwen2-VL-M3ID	85.69	86.46	86.10	73.96	59.75	57.78	81.25	78.31	80.67
Qwen2-VL-SCI$_5$ (ours)	85.97	86.67	87.36	73.59	59.92	58.06	81.39	79.31	80.98

Table 4. Experiments on MMB(ench-Dev)-C/E(N-V11), MME, CCB(ench), MMS(tar), and ViLP indicate that **SCI has more consistent improvement** than TDE/VCD/M3ID on those real-world LVLM benchmarks (using 80% test splits). Blue texts indicate an improvement over the baseline.

Base	VC-C	VC-N	TC-V1	TC-V2	MCQ	Others	Overall
✓					10.78	23.59	14.52
	✓				8.77	18.52	11.62
		✓			10.62	25.15	14.86
			✓		10.38	24.37	14.46
				✓	12.07	23.00	15.26
✓	✓				21.72	29.43	23.97
✓			✓		10.54	23.39	14.29
✓	✓		✓		24.54	32.75	26.94
✓		✓		✓	26.67	30.97	27.93
✓	✓	✓			27.37	31.13	28.46
✓			✓	✓	11.58	23.21	14.98
✓	✓		✓	✓	26.07	32.55	27.96
✓	✓	✓	✓		26.71	33.33	28.64
✓	✓	✓	✓	✓	28.00	33.14	29.50

Table 5. Ablation experiments for different counterfactual logits combinations using Qwen2-VL on BS Subset.

comprise only a portion of the datasets. These results confirm that the gains observed on DRBench are not due to overfitting to specific data distributions, but rather reflect a general improvement in robustness.

Ablation study on test-time scaling effect with increasing inference rounds. To better understand the effect of each component in SCI framework, we conducted an ablation study on SCI$_5$. As shown in Table 5, we first evaluated the performance of the base inputs and four individual counterfactual inputs on the BS Subset. We then incrementally increased the number of counterfactual rounds to form progressively more complete versions of SCI to reach SCI$_5$. Note that experiments on VC component and TC component alone are also included. Together with the comprehensive results of SCI$_3$, SCI$_5$, and SCI$_7$ in Figure 2, the overall findings highlight the potential of test-time scaling: robustness of models can be improved with more incorporated counterfactual rounds.

Ablation study on cross-model DRBench evaluation. The ablation study in Table 3 provides additional insights: 1) non-robust samples vary significantly across different LVLMs. For instance, the BS Subset constructed by LLaVA-NeXT yields only 18.75% accuracy on its own model, while Qwen2-VL achieves 60.31% accuracy on the same subset, and vice versa. This demonstrates that even if an LVLM performs perfectly well on a fixed robustness benchmark, it may still fail on new vulnerable samples. These findings highlight the necessity of adopting a model-specific DRBench; 2) The performance gains achieved through SCI in one model are transferable to DRBenchs constructed by other models, thereby validating the generalization ability of SCI.

4.4. Discussions

We also provide some interesting discussions to shed lights on the proposed SCI framework and DRBench.

Q1: Why did the base models perform so poorly (*e.g.*, LLaVA-NeXT even got 0.0 on the Bias Subset) on the Bias, Sensitivity, and BS Subsets?

A1: The proposed DRBench are intentionally designed to probe samples particularly vulnerable to robust issues, *i.e.* they are hard examples for LVLMs. That's why the model performances on these subsets are sometimes even lower than random guessing, *e.g.*, MCQs have a 25% chance accuracy for random guess. In fact, according to the definition, the Bias Subset specifically collects samples for which the base model consistently produces incorrect predictions, so its accuracy is theoretically expected to be 0.0. The reason why Qwen2-VL does not yield exactly 0.0 is due to its use of top-k sampling for decoding by default. In contrast,

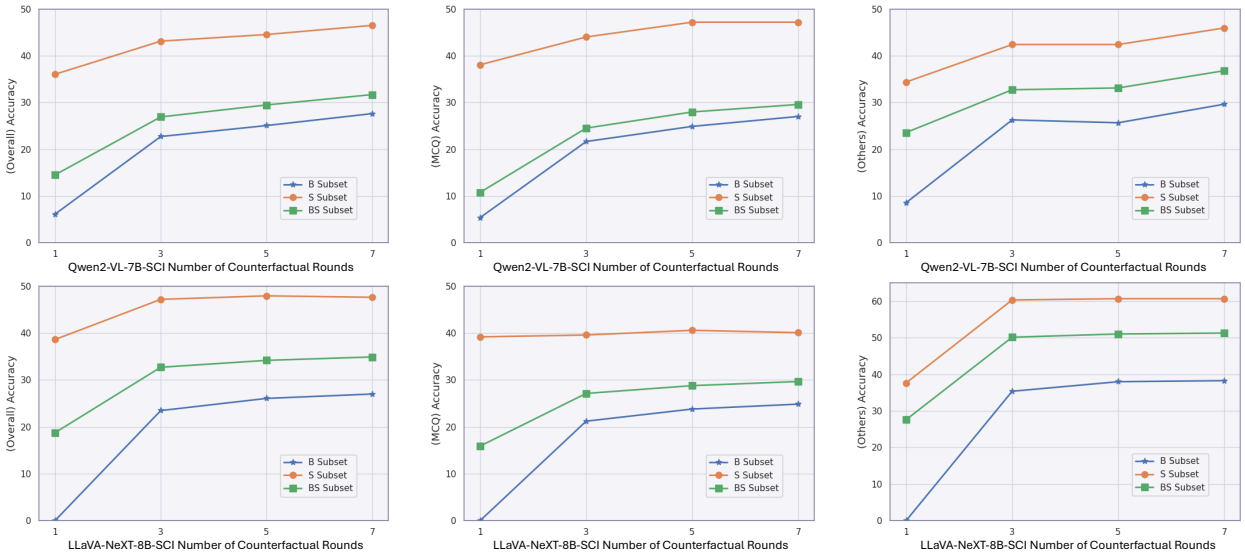

Figure 2. Investigating the test-time scaling effect on robustness with respect to the number of inference rounds on B/S/BS subsets across different question types and LVLMs.

LLaVA-NeXT uses greedy decoding, producing deterministic predictions, which explains its consistent 0.0 accuracy on the Bias Subset.

Q2: What's the computational overhead of SCI and are there potential solutions for acceleration?

A2: All test-time scaling strategies entail a trade-off between inference time and performance, which means that the proposed SCI will inevitably take more time. The most intuitive acceleration method for SCI is batch inference. Based on our experiments, the computational overhead of SCI_3, SCI_5, and SCI_7 using batch inference is approximately $1.29\times$, $1.81\times$, and $2.48\times$ that of the base model, respectively, which is much faster than the vanilla version, which costs $2.96\times$, $5.01\times$, and $6.68\times$, respectively. We also believe that KV Cache sharing for the visual and textual tokens that remain unchanged is a potential acceleration technique for SCI.

Q3: Why is SCI different from previous test-time scaling studies, and could it open up a new paradigm?

A3: Most of the existing test-time scaling studies [35] focus on increasing the length of intermediate thinking tokens. However, the prompt-level improvement only reveals whether the answer is correct or wrong. It provides no insight into whether, for instance, the input image increases the logit magnitude for a specific token such as "cushion". By introducing SCI, we go beyond discrete token outputs and analyze the underlying continuous logit distributions through comparison and aggregation of counterfactual logits. This approach provides significantly richer information than simply using final predicted tokens. Therefore, we believe that SCI opens up a promising new direction for test-time scaling studies.

Q4: Does the performance gains of the proposed SCI come from hacking the corresponding DRBench?

A4: Given the overlap process between the DRBench construction and the counterfactual sample construction of SCI, it is possible that the performance gains of SCI are tailored to its corresponding DRBench data. However, results in Table 3 demonstrate that SCI still yields consistent improvements when evaluated on the vulnerable test sets derived from other models, even though the relative improvements become smaller. This suggests that the proposed SCI possesses inherent generalization capability beyond its corresponding model-specific DRBench.

5. Conclusion

In this paper, we propose SCI, a generalized framework for robust inference in LVLMs that jointly addresses language bias and sensitivity through comprehensive logit-level counterfactual reasoning. Complemented by the DRBench, our contributions offer both a methodological advancement and an adaptive evaluation protocol for improving LVLM robustness. Extensive experiments further reveal a scalable pathway toward enhanced test-time robustness, which could be achieved by incorporating more counterfactual inference rounds and advanced logit-level reasoning algorithms. We hope that SCI and DRBench will serve as foundational paradigms and diagnostic tools for the development of future more reliable and trustworthy LVLMs.

Acknowledgments: This work was supported by Double First-Class Initiative Fund, Disciplinary Development Program of Institute of AI for Engineering, Tongji University.

References

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

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 3

[3] Simran Arora, Avanika Narayan, Mayee F Chen, Laurel J Orr, Neel Guha, Kush Bhatia, Ines Chami, and Christopher Ré. Ask me anything: A simple strategy for prompting language models. In *ICLR*, 2023. 1

[4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1

[5] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1, 3

[6] 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*, 33:1877–1901, 2020. 1

[7] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10800–10809, 2020. 3

[8] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 5

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1

[10] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023. 5

[11] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024. 5

[12] 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, 2024. 5, 1

[13] 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, 2023. 5

[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 3

[15] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18135–18143, 2024. 3

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[17] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. 3

[18] Mingjian Jiang, Yangjun Ruan, Sicong Huang, Saifei Liao, Silviu Pitis, Roger Baker Grosse, and Jimmy Ba. Calibrating language models via augmented prompt ensembles. *arXiv preprint*, 2023. 1

[19] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25004–25014, 2025. 3

[20] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. 3

[21] Minchan Kwon, Gaeun Kim, Jongsuk Kim, Haeil Lee, and Junmo Kim. Stableprompt: Automatic prompt tuning using reinforcement learning for large language models. *arXiv preprint arXiv:2410.07652*, 2024. 3

[22] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024. 1, 3, 5, 6

[23] 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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada, 2023. Association for Computational Linguistics. 2, 3

[24] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2024. 1, 2

[25] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 1

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 3

[27] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1, 5

[28] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 5

[29] Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. Probing visual language priors in vlms. *arXiv preprint arXiv:2501.00569*, 2024. 5

[30] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710, 2021. 1, 2, 3, 5

[31] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025. 3

[32] Silviu Pitis, Michael R Zhang, Andrew Wang, and Jimmy Ba. Boosted prompt ensembles for large language models. *arXiv preprint arXiv:2304.05970*, 2023. 3

[33] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *technical report*, 2019. 3

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3

[35] Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. 3, 8

[36] Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. Octopus: Alleviating hallucination via dynamic contrastive decoding, 2025. 1

[37] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3716–3725, 2020. 2, 3

[38] 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. 1, 3

[39] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 5

[40] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 3

[41] Zhiquan Wen, Guanghui Xu, Mingkui Tan, Qingyao Wu, and Qi Wu. Debiased visual question answering from feature and sample perspectives. *Advances in Neural Information Processing Systems*, 34:3784–3796, 2021. 1

[42] Gwenyth Portillo Wightman, Alexandra Delucia, and Mark Dredze. Strength in numbers: Estimating confidence of large language models by prompt agreement. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 326–362, 2023. 1

[43] Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. *arXiv preprint arXiv:2405.17820*, 2024. 1

[44] Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. On the generalization of sft: A reinforcement learning perspective with reward rectification. *arXiv preprint arXiv:2508.05629*, 2025. 3

[45] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 3

[46] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10685–10694, 2019.

[47] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. Exploring diverse in-context configurations for image captioning. *NeurIPS*, 36:40924–40943, 2023. 3

[48] Li Yifan, Du Yifan, Zhou Kun, Wang Jinpeng, Zhao Wayne Xin, and Wen Ji-Rong. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 4, 5

[49] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12), 2024. 1

[50] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 3

[51] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12401–12430, 2024. 1

[52] 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. In *12th International Conference on Learning Representations, ICLR 2024*, 2024. 1

Scaling Test-Time Robustness of Vision-Language Models via Self-Critical Inference Framework

Supplementary Material

A. Appendix

The following appendix contains supplementary details and experimental results excluded from the main paper due to space constraints. The overall appendix includes: B) adaptive plausibility constraint; C) generation of counterfactual inputs; D) additional experimental results and analyses.

B. Adaptive Plausibility Constraint

As mentioned in the main paper, we adopt adaptive plausibility constraint from VCD [22] and M3ID [12] as a post-processing step before sampling output tokens. This constraint masks tokens with low logit values under the original input, ensuring that low-confidence tokens are not sampled as final outputs. Specifically, the constraint can be formulated as:

$$Z_{vcd}(v, v^*, q)_k = -\infty, \tag{8}$$

$$\text{s.t. } Z(v, q)_k < \max_k(Z(v, q)) + \log(\beta), \tag{9}$$

where k is the token index for logits; the logit with value $-\infty$ ensures that $p_{vcd}(y|v, v^*, q)_k = 0$ for the masked tokens; β is the threshold; $\max_k(Z(v, q))$ is the largest logit value for original inputs.

The rationale behind the Adaptive Plausibility Constraint is that, although the output distribution under the original input may be biased, it can still serve as a valid filter to identify plausible candidate tokens. Only tokens with logits greater than $\max_k(Z(v, q)) + \log(\beta)$ are allowed to receive VCD logits and participate in final sampling. In contrast, low-confidence candidates with insufficient logits are directly masked out. As shown in Table 6, removing the adaptive plausibility constraint leads to a performance drop for SCI_5 on the B/S/BS subsets, and results in an even greater performance degradation on the original datasets as we expected.

For the proposed Self-Critical Inference (SCI) framework, we slightly change the constraint as follows:

$$p_{\text{SCI}}(y|\boldsymbol{v}, \boldsymbol{q}) = 0, \tag{10}$$

$$\text{s.t. } TC_k/\tau_1 < \max_k(TC/\tau_1) + \log(\beta), \tag{11}$$

where the key difference is that we use Textual Counterfactual (TC) logits, scaled by a temperature factor, to replace the original logits as the masking criterion, as we believe TC provides more consistent predictions. The final output tokens are then sampled from the unmasked candidates with non-zero probabilities.

Methods	Constraint	Original	B Subset	S Subset	BS Subset
Qwen2-VL	NA	81.12	6.10	37.59	15.46
Qwen2-VL-SCI_5	✗	68.93	26.16	34.04	27.63
Qwen2-VL-SCI_5	✓	81.03	29.65	40.43	32.55

Table 6. Ablation study for the adaptive plausibility constraint. To evaluate the effect of adaptive plausibility constraint, we conducted experiments on validation sets of original 6 datasets together with B(ias)/S(ensitive)/BS Subsets.

In our experiments, the default threshold β is set to 0.3 following the previous paper [12] for all DRBench experiments. We consider β as a trade-off parameter between relying on de-biased logits and original logits. When β approaches 1.0, the final output token closely resembles that produced by the original inputs. In contrast, when β approaches 0.0, the constraint becomes negligible, and the output behaves as if no filtering is applied. For experiments on original LVLM datasets, we increase β by 0.5 to 0.8, as these datasets exhibit less bias and the outputs are generally closer to those produced by the original inputs.

C. Generation of Counterfactual Inputs

In this section, we provide further details on the generation of counterfactual inputs. For the Visual Counterfactual input VC-Color0, we directly set the RGB values of all pixels in the input image to (0, 0, 0), resulting in a completely black image. For VC-Noise400 and VC-Noise500, we follow the method used in VCD [22], where Gaussian noise is added to simulate the forward diffusion process [16] at 400 and 500 time steps, respectively. The mathematical formulation of this forward process is as follows:

$$v_t = \sqrt{\bar{\alpha}_t} \cdot v_0 + \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon, \tag{12}$$

where v_t is the final noise image at at step t; v_0 is original image; $\epsilon \sim \mathcal{N}(0, 1)$ is random Gaussian noise; $\bar{\alpha}_t$ is cumulative product. The detailed implementation is available in the official GitHub repository of VCD.

For Textual Counterfactual input TC-V1, TC-V2, and TC-V3, as we can see from Figure 3, Figure 4, and Figure 5, each variations provide a semantically equivalent but lexically different prompts. Without change the meaning of instruction, TC-V1 adds an additional system prompt instructing the model to focus on image details, TC-V2 further modifies the system prompt's language from English to Chinese or vice versa, TC-V3 injects identity information by prompting the model to respond as a clever student.

Method	Qwen2-VL	Qwen2-VL-SCI$_3$	Qwen2-VL-SCI$_5$	Qwen2-VL-SCI$_7$
Inference Time (w/o batch inference)	540.47ms	1599.65ms	2707.16ms	3611.18ms
Inference Time (w/ batch inference)	540.47ms	697.24ms	978.14ms	1342.86ms

Table 7. We report the average inference time per sample on the MMStar dataset using one A800 GPU to illustrate the computational overhead introduced by SCI. Note that the baseline speed w/o batch inference sequentially conduct each counterfactual inference round, while w/ batch inference, all counterfactual inference rounds are conducted in one batch. Therefore, the later is significantly faster than the baseline speed.

Methods	Hyperparameters	MCQ	Others	Overall
Qwen2-VL	-	11.97	22.38	15.46
Qwen2-VL-SCI$_5$	$\tau_1 = 2.0$ $\tau_2 = 0.2$	33.45	30.77	32.55
Qwen2-VL-SCI$_5$	$\tau_1 = 2.0$ $\tau_2 = 2.0$	22.89	27.97	24.59
Qwen2-VL-SCI$_5$	$\tau_1 = 2.0$ $\tau_2 = 1.0$	26.06	29.37	27.17
Qwen2-VL-SCI$_5$	$\tau_1 = 2.0$ $\tau_2 = 0.5$	32.39	30.77	31.85
Qwen2-VL-SCI$_5$	$\tau_1 = 20$ $\tau_2 = 0.2$	3.87	18.88	8.89
Qwen2-VL-SCI$_5$	$\tau_1 = 10$ $\tau_2 = 0.2$	23.59	23.77	23.65
Qwen2-VL-SCI$_5$	$\tau_1 = 1.0$ $\tau_2 = 0.2$	28.17	26.57	27.63
Qwen2-VL-SCI$_5$	$\tau_1 = 0.2$ $\tau_2 = 0.2$	11.97	20.97	14.99

Table 8. Ablation study for temperature scaling hyperparameters τ_1 and τ_2 of SCI. Experiments are conducted under validation set of BS Subset.

D. Additional Experiments

This section will discuss some additional experiments, including ablation studies on hyperparameters, analysis of inference time for SCI, and other supplementary results.

Ablation study for hyperparameters. As shown in Table 8, we select the temperature scaling hyperparameters for the TC and VC logits based on validation performance on the BS Subset. For fair comparison, the hyperparameters were select on SCI$_5$ under base model Qwen2-VL and directly apply to LLaVA-NeXT. The temperature scaling τ_2 for VC is fixed as 0.2 across SCI$_3$, SCI$_5$, and SCI$_7$, because the logits distribution of VC would not change with the number of visual counterfactual inputs. As to the temperature scaling τ_1 for TC, since the calculation of TC involves maximum cross all outputs using different textual counterfactual inputs, the logits distribution of TC would change with number of textual counterfactual variations. Therefore, we decide to intuitively add 0.5 to τ_1 to prevent the distribution change when there is one more textual variation added to SCI.

Inference time and discussion about acceleration techniques. As shown in Table 7, we first evaluate the computational overhead of the vanilla implementation (sequential counterfactual inference) of SCI by measuring the average inference time per sample on the validation set of MMStar (Qwen2-VL BS Subset) using a single A800 GPU with Flash Attention 2.7. Specifically, we compare the original inference with SCI$_3$, SCI$_5$, and SCI$_7$. Since the vanilla implementation sequentially executes each counterfactual inference with different input variations, the computational overhead scales approximately linearly, resulting in $2.96\times$, $5.01\times$, and $6.68\times$ the base model's inference time, respectively. We then apply a straightforward acceleration technique, called batching inference to improve the efficiency. Since each counterfactual input variations together with the original input can be executed in the forward pass independently, we can put them into one batch and conduct batch parallel acceleration. The efficiency improvement after applying batch inference is significant, the computational overhead of SCI$_3$, SCI$_5$, and SCI$_7$ become $1.29\times$, $1.81\times$, and $2.48\times$, respectively. In future work, we believe that we can use KV cache sharing to further accelerate the SCI. Since each counterfactual input modifies only either the textual or visual modality, we can exploit shared components to reduce redundant calculations. For example, when the visual input is fixed and only textual prompts vary, we can prefill the visual tokens once and reuse the KV cache across all textual variations. While this approach requires additional engineering effort and potentially model fine-tuning, it offers significant theoretical efficiency gains.

The complete experiments on Bias/Sensitive/BS Subsets. Due to space constraints, the original paper only presented partial results for the Bias/Sensitive/BS Subsets experiments. The complete results are provided in Table 9. Experiments on all counterfactual inference settings with variant inputs are also included. Although LLaVA-NeXT shows 0.0 accuracy on the Bias Subset, as discussed in the main paper, variants such as LLaVA-NeXT-VCF-Color0, LLaVA-NeXT-VCF-Noise400, and LLaVA-NeXT-VCF-Noise500 may still achieve non-zero performance. This is because the Bias Subset is constructed from the combination of LLaVA-NeXT-VCF-Color0 and LLaVA-NeXT-VCF-Noise500 under our proposed setting. An incorrect prediction from one variant may coincidentally be correct in another (yet, it's still a blind guess), allowing for occasional non-zero accuracies in these counterfactual settings.

Method	Bias Subset			Sensitivity Subset			BS Subset		
	MCQ	Others	Overall	MCQ	Others	Overall	MCQ	Others	Overall
LLaVA-NeXT	0.0	0.0	0.0	39.2	37.63	38.63	15.91	27.58	18.75
LLaVA-NeXT-TCF-V1	3.20	6.38	3.71	36.62	26.80	33.02	14.86	19.65	16.02
LLaVA-NeXT-TCF-V2	5.80	8.70	6.26	24.08	33.51	27.54	9.77	24.56	13.36
LLaVA-NeXT-TCF-V3	3.09	3.19	3.11	38.61	34.54	37.11	15.99	25.31	18.26
LLaVA-NeXT-VCF-Color0	4.59	4.06	4.50	27.26	23.54	25.90	13.85	18.01	14.86
LLaVA-NeXT-VCF-Noise400	6.63	3.19	6.08	27.96	23.71	26.40	14.98	17.51	15.60
LLaVA-NeXT-VCF-Noise500	6.30	3.48	5.85	27.16	23.02	25.65	14.54	17.63	15.29
LLaVA-NeXT-TIE	12.98	23.48	14.66	39.00	57.56	45.81	21.89	44.21	27.31
LLaVA-NeXT-VCD	12.65	25.51	14.71	40.50	56.53	46.38	22.54	44.58	27.89
LLaVA-NeXT-M3ID	16.91	25.22	18.24	39.90	56.36	45.94	24.15	44.33	29.05
LLaVA-NeXT-SCI$_3$ (ours)	21.22	35.36	23.48	39.60	60.31	47.20	27.14	50.13	32.72
LLaVA-NeXT-SCI$_5$ (ours)	23.81	37.97	26.08	**40.60**	**60.65**	**47.95**	28.80	51.01	34.19
LLaVA-NeXT-SCI$_7$ (ours)	**24.86**	**38.26**	**27.01**	40.10	**60.65**	47.64	**29.68**	**51.26**	**34.92**
Qwen2-VL	5.37	8.56	6.11	38.10	34.41	36.06	10.78	23.59	14.52
Qwen2-VL-TCF-V1	6.11	11.31	7.32	36.51	36.01	36.23	10.38	24.37	14.46
Qwen2-VL-TCF-V2	7.59	15.90	9.52	40.87	34.41	37.3	12.07	23.00	15.26
Qwen2-VL-TCF-V3	6.30	8.87	6.89	37.70	34.41	35.88	11.02	22.42	14.35
Qwen2-VL-VCF-Color0	5.83	6.73	6.04	20.24	28.94	25.04	8.77	18.52	11.62
Qwen2-VL-VCF-Noise400	7.59	21.41	10.80	21.03	25.72	23.62	10.22	24.17	14.29
Qwen2-VL-VCF-Noise500	7.59	21.71	10.87	20.63	27.33	24.33	10.62	25.15	14.86
Qwen2-VL-TIE	16.20	16.82	16.35	45.63	36.66	40.67	20.27	27.29	22.32
Qwen2-VL-VCD	15.74	21.71	17.13	46.83	40.84	43.52	20.11	30.41	23.12
Qwen2-VL-M3ID	19.81	21.71	20.26	**47.22**	41.16	43.87	23.65	30.6	25.68
Qwen2-VL-SCI$_3$ (ours)	21.67	26.30	22.74	44.05	42.44	43.16	24.54	32.75	26.94
Qwen2-VL-SCI$_5$ (ours)	24.91	25.69	25.09	**47.22**	42.44	44.58	28.00	33.14	29.50
Qwen2-VL-SCI$_7$ (ours)	**27.04**	**29.66**	**27.65**	**47.22**	**45.98**	**46.54**	**29.61**	**36.84**	**31.72**

Table 9. The complete experiments on Bias Subset, Sensitivity Subset, and BS Subset of the DRBench across two widely used base LVLMs demonstrate the effectiveness of the proposed SCI framework. **Bold texts** indicate the best result of each column.

Original Prompts	**TC-V1 Prompts**
Please select the correct answer from the options above.	Think about the question based on details in the given image. Please select the correct answer from the options above.
Please answer yes or no.	Think about the question based on details in the given image. Please answer yes or no.
Please try to answer the question with short words or phrases if possible.	Think about the question based on details in the given image. Please try to answer the question with short words or phrases if possible.
Answer the question directly using a single word or phrase.	Think about the question based on details in the given image. Answer the question directly using a single word or phrase.
Answer with the option's letter from the given choices directly.	Think about the question based on details in the given image. Answer with the option's letter from the given choices directly.
(Chinese Prompts) 请直接回答选项字母。	(Chinese Prompts) 结合问题与选项仔细观察图像中的信息，请直接回答选项字母。

Figure 3. The list of all TC-V1 prompts that add an additional system prompt instructing the model to focus on image details.

Method	Single Dataset						Question Type		
	MMB-C	MMB-E	MME	CCB	MMS	ViLP	MCQ	Others	Overall
LLaVA-NeXT	78.0	79.72	79.57	47.0	44.75	51.53	70.12	71.86	70.46
LLaVA-NeXT-TC-V1	77.46	79.95	76.20	46.75	43.92	51.53	69.87	69.42	69.78
LLaVA-NeXT-TC-V2	77.44	77.51	78.78	46.20	42.08	50.14	68.68	70.90	69.12
LLaVA-NeXT-VC-C0	29.97	31.85	50.29	27.02	25.08	28.47	29.66	44.29	32.55
LLaVA-NeXT-VC-N500	30.69	33.08	48.29	28.25	25.0	29.03	30.55	42.99	33.01
LLaVA-NeXT-TIE	78.28	80.28	77.30	45.65	46.00	53.19	70.36	70.68	70.42
LLaVA-NeXT-VCD	78.38	80.28	78.09	46.63	45.00	54.31	70.44	71.55	70.66
LLaVA-NeXT-M3ID	78.31	80.18	78.62	45.89	45.92	54.03	70.36	71.86	70.66
LLaVA-NeXT-SCI$_5$ (ours)	78.21	80.08	80.15	46.20	45.75	53.06	70.32	72.70	70.79
Qwen2-VL	85.26	86.36	87.89	73.22	59.50	56.53	80.91	79.27	80.58
Qwen2-VL-TC-V1	85.28	86.11	87.79	73.18	59.73	58.09	80.84	79.63	80.60
Qwen2-VL-TC-V2	85.26	86.39	87.96	72.92	59.53	56.37	80.88	79.27	80.56
Qwen2-VL-VC-C0	34.46	35.54	50.45	25.37	27.33	24.72	32.66	43.38	34.77
Qwen2-VL-VC-N500	31.33	31.82	50.13	25.43	28.50	26.81	30.29	43.72	32.94
Qwen2-VL-TIE	86.00	86.59	86.52	73.84	59.00	57.08	81.30	78.43	80.73
Qwen2-VL-VCD	86.05	86.56	86.41	73.77	60.08	57.92	81.42	78.58	80.86
Qwen2-VL-M3ID	85.69	86.46	86.10	73.96	59.75	57.78	81.25	78.31	80.67
Qwen2-VL-SCI$_5$ (ours)	85.97	86.67	87.36	73.59	59.92	58.06	81.39	79.31	80.98

Table 10. Experiments on MMB(ench-Dev)-C/E(N-V11), MME, CCB(ench), MMS(tar), and ViLP including all counterfactual inference results used by SCI$_5$. Blue texts indicate an improvement over the baseline.

Original Prompts	TC-V2 Prompts
Please select the correct answer from the options above.	(Chinese Prompts) 请仔细观察图像中的信息，然后结合问题与选项，从上述所有选项中直接回答正确选项对应的字母。
Please answer yes or no.	(Chinese Prompts) 观察给出的图片，请直接回答yes或no。
Please try to answer the question with short words or phrases if possible.	(Chinese Prompts) 请仔细观察图像中的细节，然后结合图像上的信息回答问题，请直接用一个简短的英语单词或数字回答。
Answer the question directly using a single word or phrase.	(Chinese Prompts) 请仔细观察图像中的细节，然后结合图像上的信息回答问题，请直接用一个简短的英语单词或数字回答。
Answer with the option's letter from the given choices directly.	(Chinese Prompts) 请仔细观察图像中的信息，然后结合问题与选项，从上述所有选项中直接回答正确选项对应的字母。
(Chinese Prompts) 请直接回答选项字母。	Please carefully examine the information in the image, then consider the question and options, and reply directly with the letter corresponding to the correct answer from the options above.

Figure 4. The list of all TC-V2 prompts that further modify the system prompt's language from English to Chinese or vice versa.

Original Prompts	TC-V3 Prompts
Please select the correct answer from the options above.	You are a smart student who is good at answering multiple-choice questions. Please select the correct answer from the options above.
Please answer yes or no.	You are a smart student who is good at answering yes or no questions. Please answer yes or no.
Please try to answer the question with short words or phrases if possible.	You are a smart student who is good at answering questions. Please try to answer the question with short words or phrases if possible.
Answer the question directly using a single word or phrase.	You are a smart student who is good at answering questions. Answer the question directly using a single word or phrase.
Answer with the option's letter from the given choices directly.	You are a smart student who is good at answering multiple-choice questions. Answer with the option's letter from the given choices directly.
(Chinese Prompts) 请直接回答选项字母。	(Chinese Prompts) 你是一名擅长回答选择题的聪明学生，请直接回答选项字母。

Figure 5. The list of all TC-V3 prompts that inject identity information by prompting the model to respond as a clever student.