Supplemental Material for "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[1,2,3,4*]
[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

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 [3] and M3ID [1] 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:

$$Z_{vcd}(v, v^*, q)_k = -\infty, \tag{1}$$
$$\text{s.t. } Z(v, q)_k < \max_k(Z(v, q)) + \log(\beta), \tag{2}$$

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 1, 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.

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 1. 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.

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

$$p_{\text{SCI}}(y|\boldsymbol{v}, \boldsymbol{q}) = 0, \tag{3}$$
$$\text{s.t. } TC_k/\tau_1 < \max_k(TC/\tau_1) + \log(\beta), \tag{4}$$

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.

In our experiments, the default threshold β is set to 0.3 following the previous paper [1] 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.

*Corresponding author.

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 2. 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.

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 [3], where Gaussian noise is added to simulate the forward diffusion process [2] at 400 and 500 time steps, respectively. The mathematical formulation of this forward process is as follows:

$$v_t = \sqrt{\bar{\alpha}_t} \cdot v_0 + \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon, \tag{5}$$

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 1, Figure 2, and Figure 3, 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.

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 3, 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.

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

Inference time and discussion about acceleration techniques. As shown in Table 2, 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

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 4. 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.

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 4. 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	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 5. 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-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 1. The list of all TC-V1 prompts that add an additional system prompt instructing the model to focus on image details.

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

References

[1] 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. 1

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[3] 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, 2