T	*Hallucination Subset Total Scores*	*Perception Subset Total Scores*	*Recognition Subset Total Scores*
200	$586.67_{\pm 11.67}$	$1311.47_{\pm 4.33}$	$338.69_{\pm 19.87}$
500	$591.67_{\pm 36.06}$	$1340.89_{\pm 55.91}$	$323.45_{\pm 5.89}$
700	$578.89_{\pm 19.17}$	$1339.04_{\pm 40.40}$	$320.95_{\pm 14.18}$
999	$557.78_{\pm 1.92}$	$1345.81_{\pm 36.31}$	$321.90_{\pm 10.19}$

Table 4. An ablation study of total noise steps T on the MME benchmark.

α	*Hallucination Subset Total Scores*	*Perception Subset Total Scores*	*Recognition Subset Total Scores*
0.25	$583.89_{\pm 19.32}$	$1322.25_{\pm 32.58}$	$330.24_{\pm 13.60}$
0.5	$580.56_{\pm 17.11}$	$1315.49_{\pm 27.28}$	$333.45_{\pm 5.77}$
0.75	$578.33_{\pm 29.49}$	$1312.93_{\pm 39.31}$	$330.95_{\pm 13.58}$
1.0	$591.67_{\pm 36.06}$	$1340.89_{\pm 55.91}$	$323.45_{\pm 5.89}$

Table 5. An ablation study of α on the MME benchmark.

β	*Hallucination Subset Total Scores*	*Perception Subset Total Scores*	*Recognition Subset Total Scores*
0	$577.22_{\pm 11.10}$	$1299.04_{\pm 39.30}$	$302.98_{\pm 19.82}$
0.001	$574.44_{\pm 6.31}$	$1298.71_{\pm 40.24}$	$289.88_{\pm 14.02}$
0.01	$583.33_{\pm 18.78}$	$1324.44_{\pm 37.84}$	$327.38_{\pm 17.11}$
0.1	$591.67_{\pm 36.06}$	$1340.89_{\pm 55.91}$	$323.45_{\pm 5.89}$
0.2	$591.67_{\pm 7.26}$	$1343.06_{\pm 13.06}$	$328.57_{\pm 16.37}$
0.5	$635.00_{\pm 7.64}$	$1474.02_{\pm 15.53}$	$331.43_{\pm 13.03}$

Table 6. Ablation studies of β on the MME benchmark.

A. Detailed Experimental Settings

In all experimental setups, the hyper-parameters γ, α and β, as specified in Equations 2, 3 and 4, are fixed at values of 0.1, 1 and 0.1, respectively. For the total number of noise steps T delineated in Equation 2, we set a value of 500 for experiments involving the MME and LLaVA-Bench, while for those evaluating on POPE, the T value is set at 999.

B. Ablation Studies

For the Ablation Studies section, the default configuration for hyper-parameters α, β, and δ is set to 1, 0.1, and 500, respectively. These values are retained across all experiments unless an individual study specifies an alternative parameter adjustment for investigation. Across all the experiments, we use LLaVA-1.5 as the representative LVLM baseline to demonstrate the effect of tuning different hyper-parameters.

B.1. Effect of Total Noise Steps T

Figure 4 presents an ablation study examining the impact of varying noise levels, denoted as δ, using the LLaVA-1.5 model on the MME benchmark. In alignment with the experimental configuration, MME is subdivided into three subsets: hallucination, perception, and recognition. The hallucination subset includes tasks related to *Existence*, *Count*, *Position*, and *Color*, while the perception subset encompasses these and additional perception-focused tasks. The recognition subset, conversely, involves tasks that challenge LVLMs' cognitive reasoning abilities.

The study reveals a pronounced sensitivity to different δ values within the hallucination subset, where optimal noise levels correlate with substantially enhanced overall scores. In the realm of perception tasks, a surpassing of a specific noise threshold ($\delta > 500$) showcases VCD's capability to consistently yield improvements. For recognition tasks, VCD maintains steady performance across the spectrum of tested noise values.

B.2. Effect of α in Visual Contrastive Decoding

Table 5 demonstrates the outcomes of an ablation study focusing on α, which modulates the level of amplification between output distributions from original and distorted visual inputs, as formulated in Equation 3. The study observes minimal variance in the aggregate scores across the three MME subsets as α ranges from 0.25 to 1.0, showcasing a uniform improvement over regular decoding. This consistency evidences the efficacy and stability of the contrastive decoding strategy across a spectrum of α settings.

B.3. Effect of β in Adaptive Plausible Constraint

Table 6 presents the results of an ablation study on β, which controls the adaptive plausible constraint in Equation 4, where larger β indicates more aggressive truncation, keeping only high-probability tokens. The table illustrates that a β value of 0, implying no constraint, results in suboptimal performance, which validates our rationale for implementing this constraint: the output distribution with distorted visual inputs can still uphold fundamental linguistic standards and common sense reasoning. Indiscriminate penalization could inadvertently sanction these valid outputs and promote the generation of implausible tokens. As β increases, improvements in total scores across the hallucination and perception subsets are observed, highlighting the constraint's critical role in reducing hallucinations and improving LVLMs' perception capacities.

B.4. Effect of Different Sampling Strategies

Table 7 presents an ablation study on various sampling strategies conducted on the POPE-*Random* dataset using LLaVA-1.5. In addition to the direct sampling approach discussed in the main paper, this experiment includes four additional sampling strategies: Top P sampling (specifically, $p = 0.9$), Top K sampling (specifically, $k = 50$), Greedy decoding, and Top K sampling with temperature normalization ($k = 50, temp = 1.5/0.7$). The results indicate that applying VCD, irrespective of the sampling strategy employed, consistently contributes to hallucination mitigation

Sampling Strategy	w. VCD	Accuracy	Precision	Recall	F1 Score
Top P	No	$84.91_{\pm 0.25}$	$94.73_{\pm 0.30}$	$73.93_{\pm 0.52}$	$83.05_{\pm 0.32}$
	Yes	$\mathbf{87.82}_{\pm 0.66}$	$91.17_{\pm 0.57}$	$83.76_{\pm 0.87}$	$\mathbf{87.31}_{\pm 0.72}$
Top K	No	$83.04_{\pm 0.16}$	$91.84_{\pm 0.15}$	$72.53_{\pm 0.44}$	$81.05_{\pm 0.24}$
	Yes	$\mathbf{87.49}_{\pm 0.56}$	$91.09_{\pm 0.53}$	$83.11_{\pm 0.71}$	$\mathbf{86.92}_{\pm 0.60}$
Greedy	No	$87.10_{\pm 0.00}$	$97.33_{\pm 0.00}$	$76.29_{\pm 0.00}$	$85.54_{\pm 0.00}$
	Yes	$\mathbf{88.49}_{\pm 0.28}$	$91.78_{\pm 0.28}$	$84.56_{\pm 0.44}$	$\mathbf{88.02}_{\pm 0.30}$
Top K+Temperature 0.7	No	$85.17_{\pm 0.12}$	$94.82_{\pm 0.12}$	$74.40_{\pm 0.35}$	$83.38_{\pm 0.17}$
	Yes	$\mathbf{87.94}_{\pm 0.51}$	$91.21_{\pm 0.49}$	$83.98_{\pm 0.60}$	$\mathbf{87.45}_{\pm 0.54}$
Top K+Temperature 1.5	No	$79.28_{\pm 0.22}$	$86.48_{\pm 1.12}$	$69.42_{\pm 0.91}$	$77.01_{\pm 0.22}$
	Yes	$\mathbf{86.97}_{\pm 0.50}$	$90.96_{\pm 0.64}$	$82.09_{\pm 0.41}$	$\mathbf{86.30}_{\pm 0.51}$

Table 7. An ablation study of different sampling strategies.

Dataset	**POPE**	**Model**	**Decoding**	Accuracy	Precision	Recall	F1 Score
MSCOCO	*Random*	LLaVA1.5(13B)	Regular	$83.31_{\pm 0.32}$	$91.46_{\pm 0.38}$	$73.48_{\pm 0.75}$	$81.49_{\pm 0.43}$
			VCD	$\mathbf{87.39}_{\pm 0.32}$	$92.68_{\pm 0.36}$	$81.19_{\pm 0.63}$	$\mathbf{86.55}_{\pm 0.41}$
		InstructBLIP(13B)	Regular	$82.36_{\pm 0.59}$	$86.93_{\pm 0.85}$	$76.19_{\pm 1.05}$	$81.20_{\pm 0.68}$
			VCD	$\mathbf{84.53}_{\pm 0.38}$	$88.55_{\pm 0.54}$	$79.32_{\pm 0.44}$	$\mathbf{83.68}_{\pm 0.40}$
	Popular	LLaVA1.5(13B)	Regular	$82.47_{\pm 0.55}$	$89.55_{\pm 0.92}$	$73.53_{\pm 0.78}$	$80.75_{\pm 0.61}$
			VCD	$\mathbf{85.74}_{\pm 0.25}$	$89.33_{\pm 0.52}$	$81.19_{\pm 0.63}$	$\mathbf{85.06}_{\pm 0.29}$
		InstructBLIP(13B)	Regular	$79.07_{\pm 0.66}$	$81.11_{\pm 0.70}$	$75.79_{\pm 1.27}$	$78.35_{\pm 0.78}$
			VCD	$\mathbf{81.47}_{\pm 0.42}$	$82.89_{\pm 0.64}$	$79.32_{\pm 0.44}$	$\mathbf{81.07}_{\pm 0.39}$
	Adversarial	LLaVA1.5(13B)	Regular	$80.00_{\pm 0.52}$	$84.46_{\pm 0.73}$	$73.53_{\pm 0.76}$	$78.62_{\pm 0.58}$
			VCD	$\mathbf{81.92}_{\pm 0.44}$	$82.40_{\pm 0.42}$	$81.17_{\pm 0.65}$	$\mathbf{81.78}_{\pm 0.47}$
		InstructBLIP(13B)	Regular	$76.57_{\pm 0.75}$	$77.00_{\pm 0.83}$	$75.79_{\pm 0.80}$	$76.39_{\pm 0.75}$
			VCD	$\mathbf{79.56}_{\pm 0.41}$	$79.67_{\pm 0.59}$	$79.39_{\pm 0.50}$	$\mathbf{79.52}_{\pm 0.38}$

Table 8. Results of 13B-sized LLaVA1.5 and InstructBLIP variants on the POPE metric. The best performance of each setting is **bolded**.

and an enhancement of the general performance capabilities of LVLMs. This consistency underscores the versatility and effectiveness of VCD across different sampling strategies in the context of LVLMs.

B.5. Effect of VCD when LVLMs Scale Up

Our evaluation extends to larger 13B variants of the LLaVA-1.5 and InstructBLIP models[8], assessing the scalability of our proposed VCD across different LVLM magnitudes. Table 8 reveals that the 7B and 13B variants of LLaVA-1.5 and InstructBLIP exhibit comparable performances across POPE settings (e.g., 81.33 and 81.49 F1 scores for LLaVA-1.5 7B and 13B in *Random* setting), suggesting that increasing the model parameters does not inherently resolve hallucination issues, thereby underscoring the pertinence of addressing this challenge. Crucially, VCD consistently boosts performance in all POPE configurations, reaffirming its robustness independent of model scale.

[8]Qwen-VL lacks larger variants.

C. Detailed Experimental Results on MME

In Table 9, we present the performance of three LVLM baselines on the perception-related tasks of the MME benchmark. The baselines exhibit consistent performance patterns, and the deployment of VCD uniformly improves their perceptual competencies. This improvement is likely a consequence of VCD's capability to diminish statistical biases and language priors, thus recalibrating the LVLMs to favor visual information over pre-existing biases and priors.

Furthermore, Table 10 showcases the LVLMs' performances on recognition-related tasks within the MME benchmark. The results indicate that the application of VCD, while alleviating hallucination issues and augmenting perceptual capabilities, does not compromise the inherent reasoning abilities of LVLMs, as evidenced by the stable overall recognition scores.

D. Comparable Analysis with Prior Works

Our Visual Contrastive Decoding (VCD) stands out by requiring neither additional training nor tool usage, unlike

Model	Decoding	*Existence*	*Count*	*Position*	*Color*	*Posters*	*Celebrity*	Scene	Landmark	Artwork	OCR	***Percetion Total***
LLaVA1.5	Regular	175.67$_{\pm 7.51}$	124.67$_{\pm 19.59}$	114.00$_{\pm 9.32}$	151.00$_{\pm 10.45}$	127.82$_{\pm 7.13}$	113.59$_{\pm 3.43}$	148.30$_{\pm 3.49}$	129.95$_{\pm 5.33}$	102.20$_{\pm 4.70}$	92.00$_{\pm 31.29}$	1279.19$_{\pm 37.09}$
	VCD	**184.66**$_{\pm 6.81}$	**138.33**$_{\pm 15.68}$	**128.67**$_{\pm 7.21}$	**153.00**$_{\pm 7.58}$	**132.11**$_{\pm 6.53}$	**120.94**$_{\pm 7.57}$	**152.20**$_{\pm 0.21}$	**140.45**$_{\pm 6.73}$	**109.60**$_{\pm 2.66}$	**104.00**$_{\pm 30.96}$	**1363.96**$_{\pm 40.58}$
Qwen-VL	Regular	155.00$_{\pm 3.54}$	127.67$_{\pm 13.36}$	**131.67**$_{\pm 7.73}$	173.00$_{\pm 9.75}$	137.76$_{\pm 2.49}$	116.24$_{\pm 2.58}$	**150.17**$_{\pm 2.80}$	158.00$_{\pm 2.40}$	**125.75**$_{\pm 5.74}$	**89.50**$_{\pm 7.37}$	1364.74$_{\pm 30.78}$
	VCD	**156.00**$_{\pm 6.52}$	**131.00**$_{\pm 6.19}$	128.00$_{\pm 3.61}$	**181.67**$_{\pm 5.14}$	**142.45**$_{\pm 2.96}$	**137.35**$_{\pm 2.45}$	149.10$_{\pm 2.51}$	**163.95**$_{\pm 1.77}$	127.65$_{\pm 2.81}$	86.00$_{\pm 3.35}$	**1403.17**$_{\pm 14.57}$
InstructBLIP	Regular	141.00$_{\pm 13.97}$	75.33$_{\pm 14.16}$	**66.67**$_{\pm 3.91}$	97.33$_{\pm 16.94}$	109.66$_{\pm 6.21}$	87.50$_{\pm 6.80}$	128.74$_{\pm 3.13}$	100.55$_{\pm 3.33}$	94.10$_{\pm 5.05}$	83.50$_{\pm 19.25}$	1223.72$_{\pm 86.59}$
	VCD	**168.33**$_{\pm 11.55}$	**92.33**$_{\pm 8.47}$	64.00$_{\pm 6.73}$	**123.00**$_{\pm 11.27}$	**121.09**$_{\pm 5.12}$	**118.71**$_{\pm 3.93}$	**149.65**$_{\pm 1.46}$	**123.65**$_{\pm 1.89}$	**110.60**$_{\pm 2.89}$	**96.50**$_{\pm 8.94}$	**1447.19**$_{\pm 25.43}$

Table 9. Results on all MME perception-related tasks. The best performance of each setting is **bolded**.

Model	Decoding	*Common Sense Reasoning*	*Numerical Calculation*	*Text Translation*	*Code Reasoning*	***Recognition Total***
LLaVA1.5	Regular	106.43$_{\pm 9.04}$	**72.50**$_{\pm 15.51}$	**95.50**$_{\pm 12.80}$	78.50$_{\pm 22.12}$	352.93$_{\pm 27.98}$
	VCD	**111.29**$_{\pm 7.06}$	68.50$_{\pm 16.64}$	89.50$_{\pm 5.97}$	**84.00**$_{\pm 25.35}$	**353.29**$_{\pm 36.19}$
Qwen-VL	Regular	109.86$_{\pm 10.31}$	**60.00**$_{\pm 6.37}$	83.00$_{\pm 11.91}$	**67.50**$_{\pm 10.16}$	**320.36**$_{\pm 26.00}$
	VCD	**114.39**$_{\pm 5.83}$	54.00$_{\pm 9.62}$	**85.00**$_{\pm 7.29}$	64.50$_{\pm 7.37}$	317.89$_{\pm 11.59}$
InstructBLIP	Regular	79.57$_{\pm 6.81}$	62.86$_{\pm 11.23}$	55.00$_{\pm 10.75}$	70.00$_{\pm 10.75}$	267.43$_{\pm 10.42}$
	VCD	**109.71**$_{\pm 7.31}$	**66.00**$_{\pm 16.45}$	**69.00**$_{\pm 11.54}$	**74.50**$_{\pm 20.26}$	**319.21**$_{\pm 20.60}$

Table 10. Results on all MME recognition-related tasks. The best performance of each setting is **bolded**.

Method	Precision	Recall	F1	Accuracy
Baseline	81.89$_{(\pm 2.03)}$	71.07$_{(\pm 1.80)}$	76.09$_{(\pm 1.61)}$	77.67$_{(\pm 1.51)}$
WoodPecker[58]	86.96$_{(\pm 0.33)}$	72.00$_{(\pm 0.69)}$	78.77$_{(\pm 3.23)}$	80.60$_{(\pm 0.20)}$
VCD (Ours)	80.00$_{(\pm 0.54)}$	83.20$_{(\pm 1.20)}$	**81.57**$_{(\pm 0.86)}$	**81.20**$_{(\pm 0.80)}$

Table 11. POPE results on a 500-sample COCO Adversarial subset.

Method	Precision	Recall	F1	Accuracy
LLaVA-RLHF[60]	78.08$_{(\pm 0.20)}$	65.69$_{(\pm 0.30)}$	71.35$_{(\pm 0.23)}$	73.62$_{(\pm 0.18)}$
LLaVA-RLHF+VCD	73.87$_{(\pm 0.13)}$	81.24$_{(\pm 0.17)}$	**77.38**$_{(\pm 0.15)}$	**76.26**$_{(\pm 0.15)}$

Table 12. POPE results under the COCO Adversarial setting.

existing LVLM hallucination mitigation methods, which involve either post-hoc rewriting with external tools [58, 77], or further fine-tuning with specialized datasets or supervision [20, 43, 60].

To contextualize our method in prior work, we benchmarked VCD against the SOTA method WoodPecker[9], which we reproduced using the LLaVA1.5 model. Results in Table 11 showcase VCD's effectiveness in reducing hallucinations without additional resources[10].

Crucially, VCD's novel approach of contrasting logits from varying inputs within the same model is orthogonal to all existing methods. This orthogonality enables VCD to be integrated with both post-hoc rewriters and fine-tuned models, potentially enhancing their effectiveness. To illustrate this, we combined VCD with LLaVA-RLHF[11], and achieved improved performance as shown in Table 12.

[9] SOTA with the LLaVA1 model on the POPE benchmark.

[10] We assess WoodPecker on a 500-sample COCO Adversarial subset, due to the cost consideration of GPT3.5-Turbo API calls for rewriting.

[11] The LLaVA1-based LLaVA-RLHF may have comparatively lower scores against other LLaVA1.5-based methods.

E. More Case Studies

Additional case studies on the LLaVA-bench are presented to illustrate the effectiveness of our approach across different LVLMs. Figure 7 provides further instances of hallucination corrections by our method. Meanwhile, Figure 8 offers supplemental examples of the enhancements brought by our proposed VCD in bolstering the general perception and recognition abilities of LVLMs.

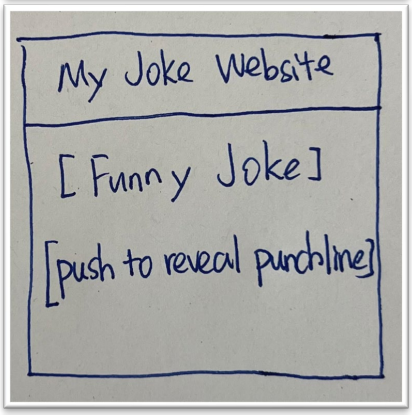
F. Prompt and Case for GPT-4V Aided Evaluation

To evaluate open-ended generation, we utilize GPT-4V to assess the accuracy and detailedness of LVLMs' responses. The specific configurations are detailed in Table 13. Additionally, an illustrative evaluation case is presented in Figure 9.

Prompt: How many coffee mugs are in the set?

Regular Decoding: Four.

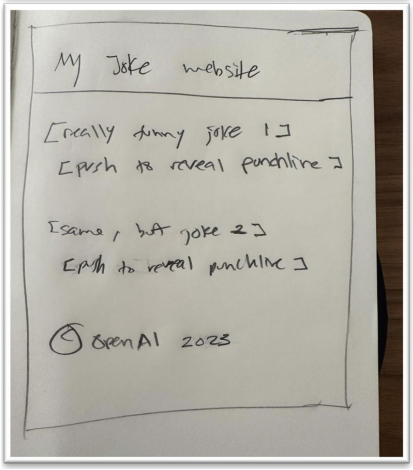
Visual Contrastive Decoding: Three.



Prompt: Describe this sketch in detail.

Regular Decoding: The sketch is a single line drawing that consists of two blocks, one labeled "my joke website" and the other labeled "funny joke". The drawing has a handwritten font that is important to its context as a sketch rather than a computer generated image. Additionally, there **are four ink splatters** surrounding the labeled sections, representing the various types of puns that the website might feature. The sketch is **part of a conversation between a user and an artificial intelligence assistant**, conveying the user's thoughts on humorous rhymes.

Visual Contrastive Decoding: The sketch depicts a hand-drawn note, with the phrase "My joke website" written in blue ink at the top. Below the phrase, the user has drawn a stick figure in the middle of the note. The stick figure is labeled with the caption "funny joke website," with a speech bubble coming out of its mouth that reads "push to reveal punch.“ The drawing has a sense of humor to it, with the joke revolving around the idea of a website that pushes people to reveal a punch.



Prompt: Describe this sketch in detail.

Regular Decoding: The image displays a handwriting diagram of a **joke sheet**, where one can see the **different steps** outlined to create a funny joke or story. These steps typically involve planning the punchline, practicing the setup, and working on the supporting material. The diagram begins with the setup, moves on to the punchline, and ends with the joke itself. The writing is captured on a book or piece of a notebook, with the handwriting clearly showing the structure of the **joke-making process**. The piece of paper appears to be placed on a wooden table or similar surface.

Visual Contrastive Decoding: The sketch depicts a hand-drawn note, with the phrase "My joke website" written in blue ink at the top. Below the phrase, the user has drawn a stick figure in the middle of the note. The stick figure is labeled with the caption "funny joke website," with a speech bubble coming out of its mouth that reads "push to reveal punch.“ The drawing has a sense of humor to it, with the joke revolving around the idea of a website that pushes people to reveal a punch.

Figure 7. More examples from LLaVA-Bench of our proposed VCD for hallucination corrections. Hallucinated objects from LVLM's regular decoding are highlighted in red.

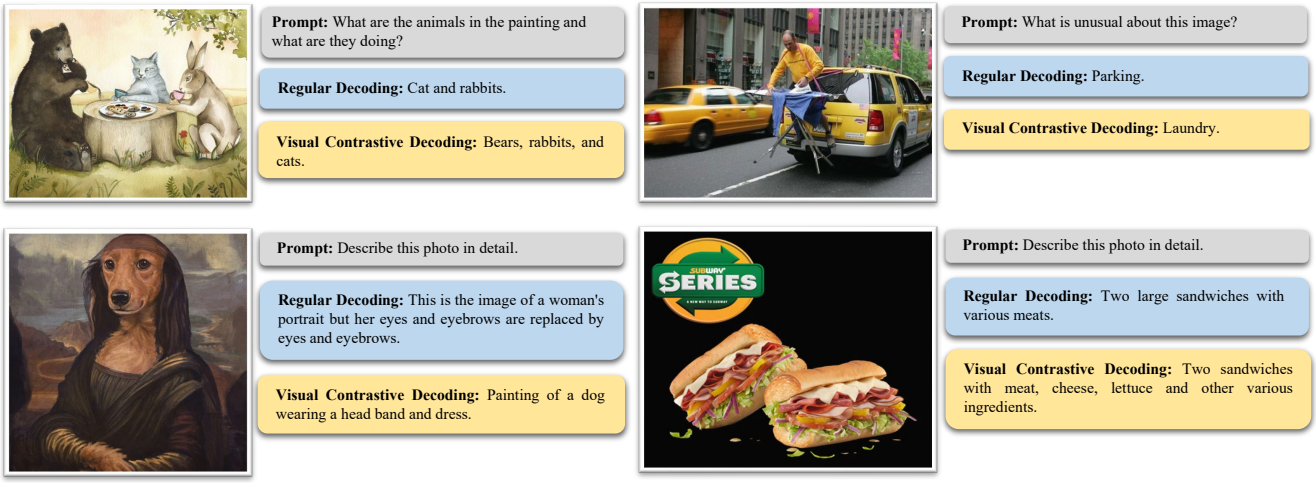


Figure 8. More examples from LLaVA-Bench of our proposed VCD for enhanced general perception and recognition capacities.

Description:
AI that scores image description accuracy and detailedness.

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

Input format:

[Assistant 1]
{Response 1}
[End of Assistant 1]

[Assistant 2]
{Response 2}
[End of Assistant 2]

Output format:

Accuracy:
Scores of the two answers:
Reason:

Detailedness:
Scores of the two answers:
Reason:

Table 13. The configuration to build an image-description evaluator with GPT-4V

Input:

[Assistant 1]

The image features a painting of a beautiful, smiling woman wearing a black veil. She appears to be a famous artwork known as the Mona Lisa, a portrait by the famous artist Leonardo da Vinci. The Mona Lisa is known for her enigmatic smile and captivating appearance. \n\nIn the background of the painting, there are mountains and a body of water, creating a serene and picturesque landscape. The woman, who is the main focus of the image, is wearing a headdress and has a veil over her face, adding to the allure and mystery of her portrait.“

[End of Assistant 1]

[Assistant 2]

The painting depicts a close-up of a classic Roman portrait style with a luxurious design. A woman is the main subject, wearing a brown squared veil adorned with intricate design. She is richly dressed and delicately painted, giving off an elegant feeling. Her profile is captured in the painting, emphasizing her facial features and the beautiful veil.\n\nIn the background, two mountains are visible, adding to the elegant and classical atmosphere of the portrait. The work of art highlights the details of the subject and her surroundings, making it an outstanding and stunning visual experience.

[End of Assistant 2]

GPT4-V Output:

Accuracy:
Scores of the two answers:
Assistant 1: 9
Assistant 2: 4
Reason: Assistant 1 correctly identifies the painting as the Mona Lisa and describes the enigmatic smile and attire accurately. However, the veil is not black but rather a dark translucent veil that doesn't cover her face; hence the slight deduction. Assistant 2 incorrectly describes the veil as brown and squared with an intricate design, which is inaccurate as the Mona Lisa's veil is quite simple and not brown. Furthermore, Assistant 2 mentions that the woman's profile is captured, which is incorrect as the Mona Lisa is facing the viewer directly, not in profile. Also, the description of the attire as "richly dressed" is misleading, since her clothing in the painting is relatively modest.

Detailedness:
Scores of the two answers:
Assistant 1: 7
Assistant 2: 5
Reason: Assistant 1 provides a detailed description of the background and attire, contributing to a rich narrative about the painting. However, it could have included more detail about the specific elements of the landscape and the technique used. Assistant 2, while less accurate, attempts to describe the elegance and details of the painting and the background. However, the description lacks the necessary detail and specific elements that are present in the painting, such as the absence of intricate design on the veil and the incorrect mention of the subject's profile.

Figure 9. Case illustrating the evaluation of GPT-4V in open-ended generation task. “Assistant 1” and “Assistant 2” correspond to “visual contrastive decoding” and “regular decoding”.