

917 A Prompts used in experiments

918 A.1 Prompt for Training Data Generation

919 We provide the prompt used for generating ViCrit task training data in Table 5

920 A.2 Prompt for ViCrit-Bench Evaluation

921 We provide the prompt used for ViCrit-Bench evaluation in Table 3

Table 3: Prompt template used for ViCrit-Bench evaluation.

Prompt Template:

You are provided with an image and the description corresponding to this image. There is one hallucination in this description. Find out the hallucination phase and answer with the hallucination phase directly in a list. Your output should only be a list that contains the hallucination phase you find.

Description:

922 B Task Distribution

923 In Figure 6 we present the distribution of different hallucination types across image categories within ViCrit-Bench.

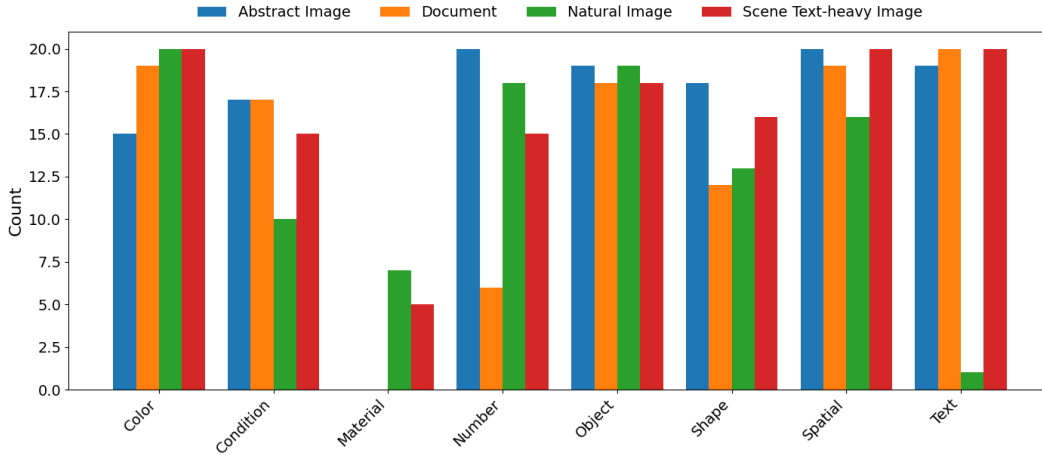


Figure 6: The distribution of the 8 hallucination tasks across 4 different image categories in ViCrit-Bench.

925 C Limitations.

926 We only experiment ViCrit training on standard image-to-text VLMs instead of VLMs that can use vision tools or generate multimodal thoughts, such as O3. However, we note that effective RL proxy task is an equally important and orthogonal directions as MM-CoT and tool using, which can be used to train these models with RL for better perception and vision tool using.

930 D Societal impacts.

931 ViCrit can positively impact society by improving vision-language models' reliability for applications like accessibility and information veracity. However, negative societal risks include the potential

misuse of its error-detection insights for sophisticated disinformation and over-reliance on improved yet imperfect systems, which reduces human scrutiny. Addressing these multifaceted impacts is crucial for responsible AI advancement.

E Comparison with SFT

In this section, we perform SFT on Qwen-2.5-VL-7B and 72B using 900k captioning samples from PixMo-Cap, and compare the results with ViCrit-RL models trained using the same amount of data through ViCrit task RFT. As shown in Table 4, we find that although SFT significantly reduced hallucination in VLMs, it do not lead to notable performance improvements on general benchmarks—in fact, the 7B model even shows a performance drop. This highlights the effectiveness of ViCrit task RFT, which not only reduces hallucinations but also generalizes well to enhance VLM performance on general reasoning tasks.

Table 4: Comparison between ViCrit-RL and ViCrit-RL with using same captioning data for SFT. We find that although hallucinations in the VLM are significantly reduced after SFT, the performance improvement is difficult to generalize to general tasks.

Model	Hallucination Benchmark			General benchamrk								
	CHAIRs ↓	CHAIRi ↓	MMHal ↑	MathVista testmini ↑	MathVision mini ↑	MathVerse mini ↑	MMMU ↑	MMStar ↑	MM-Vet ↑	Blind ↑	Charxiv reasoning ↑	Avg.
Qwen2.5-VL-7B-Instruct	28.0	5.1	3.74	67.8	23.6	44.5	50.6	61.7	66.0	49.3	41.4	50.61
Qwen2.5-VL-7B-CapSFT	25.5	4.4	3.78	67.4	20.1	44.3	52.1	53.4	64.7	47.3	38.0	48.41
ViCrit-RL-7B	25.2	4.5	3.77	70.7	25.7	46.3	52.0	61.9	67.1	52.6	47.8	53.01
Δ (Ours - Qwen2.5-7B)	-2.8	-0.6	+0.03	+2.9	+2.1	+1.8	+1.4	+0.2	+1.1	+3.3	+6.4	+2.40
Qwen2.5-VL-72B-Instruct	26.4	4.8	3.82	74.8	35.2	53.3	63.4	68.4	76.3	61.3	45.5	59.78
Qwen2.5-VL-72B-CapSFT	21.6	3.6	3.89	76.1	34.8	57.9	65.3	68.9	76.5	63.0	44.7	60.78
ViCrit-RL-72B	21.0	3.9	3.91	77.3	40.1	59.8	66.0	69.8	77.1	65.8	49.4	63.16
Δ (Ours - Qwen2.5-72B)	-5.4	-0.9	+0.09	+2.5	+4.9	+6.5	+2.6	+1.4	+0.8	+4.5	+3.9	+3.38

Table 5: Prompt used for training data generation.

You are a helpful assistant designed to manipulate text with precision. Your task is as follows:

1. Identify all noun phrases in a given paragraph. A noun phrase consists of a noun and its modifiers (e.g., "the wooden bridge," "a flock of birds"). Noun phrase is two to five words long. Do not output a list of multiple noun phrases.
2. Randomly select one noun phrase from the list, it can be small background objects, scene text, foreground objects. Try to select scene text and small background objects more often when possible.
3. Replace the chosen noun phrase with another phrase that is visually similar, such as changing the object attributes, replacing the object with a visually similar noun, or adding and removing characters within the scene text. The replacement should be visually similar but not identical to the original phrase. Be creative and don't always focus on the most obvious or common replacements such as color.
4. However, the replacement should introduce clear change, such that it is impossible to be ambiguous. The change should be directly related to image and be a visual description. Do not only change words to its synonyms or make ambiguous changes. Do not merely change words to its synonyms. Do not merely change words to its synonyms.
5. Ensure the edited paragraph is still be a plausible image description, and the change is not too obvious.
6. Group the original phrase in <Before>original</Before>, and changed phrase in <After>changed</After>. <Caption> is used to give input caption and should not be generated. Perform this transformation accurately and naturally.

Here are some examples:

1. <Caption>This image appears to be a screenshot taken from an iPhone displaying the interface of a food delivery app, likely DoorDash, around the Chicago and Gary, Indiana area. The top of the screen indicates the time as 2:30 PM, with the phone connected to an LTE network. The battery icon suggests a low battery level of 15-20%. Central to the image is a map highlighting various regions with color codes: red areas represent high traffic or demand, likely meaning those areas are "busy" for delivery drivers, as indicated by a red text banner. Lighter red and green sections represent varying levels of demand.

At the top of the image, a black banner labeled "Promos" is displayed, accompanied by a blue notification bell icon with the number two beside it, indicating two notifications.

The bottom of the screen shows a black navigation bar. It contains options for "Dash," "Schedule," "Account," "Ratings," and "Earnings." The "Dash" option is highlighted in red, suggesting it is currently selected. Centrally located in this bar is a red "Dash Now" button, implying that the user can begin delivering immediately. An additional black banner, located just above the navigation bar, reads "In... Hammond."

Overall, this detailed caption gives a comprehensive idea of the app's functionality, likely indicating areas of high demand where food delivery services are needed the most.</Caption>

<Before>a low battery level of 15-20%</Before>

<After>a high battery level of 75-80%</After>

2. <Caption>The image depicts a screenshot from a strategy video game with a third-person, aerial view. The central character, named Anselm, navigates through a complex, industrial-style building that evokes the aesthetic of games like Metal Gear. The environment is dark and futuristic, with certain areas illuminated, revealing various paths and stairs. The top left corner displays the yellow text "Instructor Eastwood," alongside a graph-like design. The upper center features game-related instructions in white text stating "Defensive Measures – Use Range to Your Advantage," with the notations "5" and "5.1" accompanying the instructions. Additionally, a map of the area is situated in the lower left corner, while the bottom right corner features interactive elements or a key potentially indicating available weapons. The overall scene suggests a mission-focused gameplay scenario requiring strategic maneuvering and tactical decision-making.</Caption>

<Before>text "Instructor Eastwood,"</Before>

<After>text "Instructor Westwood,"</After>

3. <Caption>The image depicts a three-dimensional panoramic view of a conference room where a business meeting is taking place. The setting is a typical meeting room with white walls, fluorescent lighting, and windows equipped with blinds on some, including wooden slats on one. At the center of the room is a round, yellow table that appears slightly distorted due to the panoramic effect. On this table, there are various items, including a box of tissues, a white mug, a teacup, and pamphlets.

Surrounding the table, seated in a circle, are eight individuals. They appear to be a mixed group of men and women, predominantly of Asian descent, and are dressed in a variety of attire ranging from business to casual. All attendees are wearing name tags on the left side of their chests, indicating their participation in the meeting. Their seating arrangement includes black chairs with green backs, and each person either has their hands folded in their laps or is holding something, possibly a drink.

From left to right, the attendees include a woman in a peach-colored t-shirt with writing, a man in a blue shirt, a woman in a gray sweater, a woman in a green shirt, another woman in a blazer, an empty chair, a man in a yellowish shirt, a man wearing a ball cap, and a man in a blue and green jacket. One notable aspect is that the meeting environment, though professional, is quite understated with minimalistic decor and standard conference room furnishings.</Caption>

<Before>eight individuals</Before>

<After>seven individuals</After>

Here is the input caption: <Caption>{CAPTION}</Caption>