SCALING INFERENCE-TIME SEARCH WITH VISION VALUE MODEL FOR IMPROVED VISUAL COMPREHENSION

Anonymous authors
Paper under double-blind review

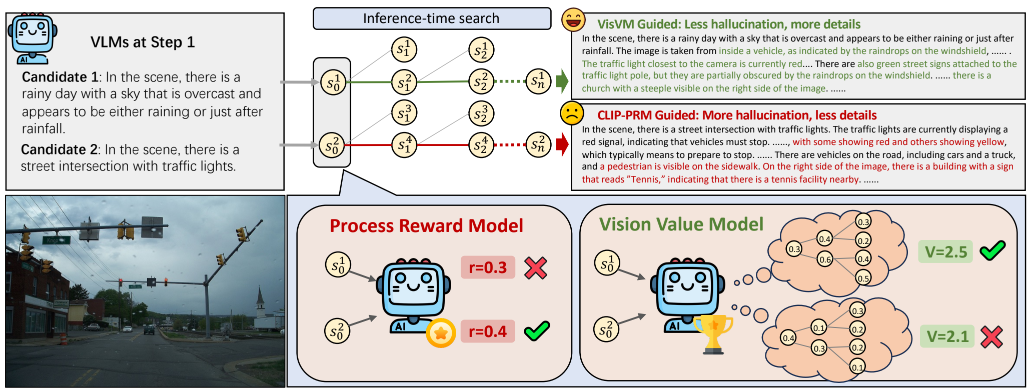


Figure 1: An illustration of how VisVM can better guide vision language model (VLM) during inference-time search. When selecting response candidates at each step, the process reward model (PRM) only considers the immediate reward, whereas VisVM predicts the long-term value by considering potential hallucinations in subsequent generated sentences. This enables VisVM to avoid response candidates with higher hallucination risks and generate image descriptions that are less prone to hallucination and more detailed.

ABSTRACT

Despite significant advancements in vision-language models (VLMs), there lacks effective approaches to enhance response quality by scaling inference-time computation. This capability is known to be a core step towards the self-improving models in recent large language model studies. In this paper, we present **Vis**ion **V**alue **M**odel (VisVM) that can guide VLM inference-time search to generate responses with better visual comprehension. Specifically, VisVM not only evaluates the generated sentence quality in the current search step, but also anticipates the quality of subsequent sentences that may result from the current step, thus providing a long-term value. In this way, VisVM steers VLMs away from generating sentences prone to hallucinations or insufficient detail, thereby producing higher quality responses. Experimental results demonstrate that VisVM-guided search significantly enhances VLMs' ability to generate descriptive captions with richer visual details and fewer hallucinations, compared with greedy decoding and search methods with other visual reward signals. Furthermore, we find that self-training the model with the VisVM-guided captions improve VLM's performance across a wide range of multimodal benchmarks, indicating the potential for developing self-improving VLMs.

1 INTRODUCTION

Vision language models (VLMs) have advanced rapidly, excelling in multimodal tasks involving single images Liu et al. (2023c); Bai et al. (2023); Chen et al. (2024b); Shi et al. (2024), multiple images Jiang et al. (2024); Li et al. (2024d), and videos Li et al. (2024a); Xue et al. (2024); Wang et al. (2024c). These capabilities stem from large-scale, high-quality training data, often sourced from web-crawled image-text pairs Radford et al. (2021); Jia et al. (2021) with effective filtering Changpinyo et al. (2021); Yuan et al. (2021); Hu et al. (2022), or enriched through techniques like distillation

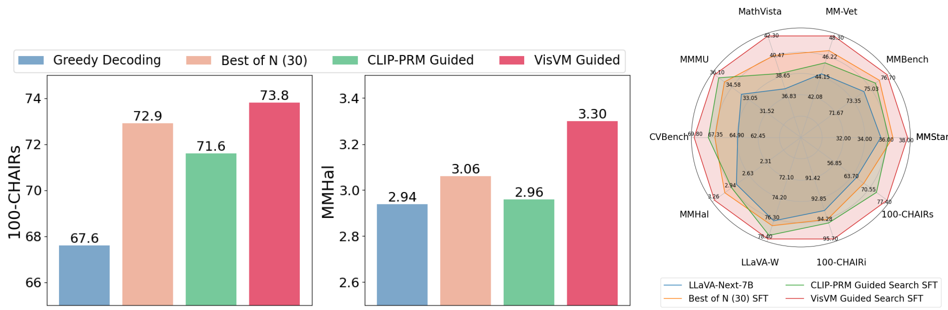


Figure 2: **Left figure**: CHAIRs and MMHal score of descriptive captions generated by LLaVA-Next-7B during inference-time using different search methods. VisVM-guided search clearly outperforms other methods, indicating reduced visual hallucinations. Notably, even with a smaller search budget (search size 6 vs. search size 30), our approach still surpasses the Best-of-N method. **Right figure**: Comparisons of LLaVA-Next-7B after fine-tuning with descriptive captions from different search methods, with VisVM-guided search achieving favorable results across all 9 benchmarks.

from stronger VLMs Chen et al. (2023a), human annotations Betker et al. (2023), or added textual descriptions Lai et al. (2025). Despite this progress, VLMs still suffer from visual hallucinations Liu et al. (2023a); Guan et al. (2024); Wang et al. (2024f); Xia et al. (2024b) and often neglect less salient image regions, limiting their real-world utility. While increasing the scale and quality of training data could help, this approach incurs significant annotation and API costs, making it less scalable. This raises a key question: *Can we enhance VLMs' response quality at inference time, and leverage these improved responses to further advance VLMs' visual comprehension?*

Recent studies on large language models (LLMs) o1b (2024); Lightman et al. (2023b); Snell et al. (2024b); Yang et al. (2024); Snell et al. (2024a) highlight inference-time search as a promising approach for improving response quality, complementary to training time effort. By leveraging a pretrained process reward model Zhang et al. (2024); Tian et al. (2024), LLMs can perform search iterations to produce high-quality outputs, with these refined responses showing potential as synthetic training data to enhance reasoning capabilities. However, extending this approach to VLMs for improved visual comprehension poses unique challenges, particularly in defining a reward signal. While process and outcome rewards are relatively straightforward for LLM tasks like coding and math, VLM tasks—such as descriptive captioning—lack clear outcome measures and require cohesive paragraph image descriptions that consist of multiple global and regional caption sentences. In these cases, each sentence must not only be accurate locally but also contribute to a coherent overall response.

To this end, we propose the **Vis**ion **V**alue **M**odel (VisVM), a value network to guide VLM inference-time search by generating descriptive captions in a step-by-step manner, with each step producing one sentence. As shown in Figure 1, VisVM takes the image and generated sentence at each step as inputs, predicting a **long-term value** to ensure both visual-text alignment and coherence. VisVM is built on two main insights, which differentiates our method from traditional process reward models in LLM literature Uesato et al. (2022); Lightman et al. (2023a); Cobbe et al. (2021); Hosseini et al. (2024); Wang et al. (2024b): *First*, rather than relying only on the local reward of the current sentence, it predicts future consequences to maintain coherence. VisVM is trained using Temporal Difference (TD) learning Sutton (1988), allowing it to go beyond evaluating responses at the current search step. This forward-looking signal helps avoid sentences that could lead to hallucinations in the future, improving global response quality in a single search round. *Second*, the reward signal must capture comprehensive visual concepts to reduce hallucinations, for which CLIP's text-image similarity metric serves effectively.

We validate the effectiveness of VisVM through two main experiments: **(1)** Using VisVM as a guidance signal for VLM inference-time search to generate descriptive image captions, we observe a substantial reduction in hallucinations and more detailed image descriptions. In both GPT and human evaluations, VisVM-guided captions consistently outperform those generated by greedy decoding, best-of-N decoding, and CLIP-PRM guided search, with VisVM-guided captions preferred 74% of the time over greedy decoding results. **(2)** To leverage VisVM's inference-time enhancement of VLM responses, we use VisVM-guided captions as the Supervised Fine-Tuning (SFT) data to self-train the original VLM (LLaVA-Next-7B). Across nine standard benchmarks, VisVM-guided self-training improves the performance by an average of 10.8% , as shown in the right figure in Figure 2.

Our contribution can be summarized as follows:

- We introduce VisVM, a stepwise value model designed to provide long-term vision value signals to guide VLM inference-time search. To the best of our knowledge, VisVM is the first exploration into enhancing VLM visual comprehension through inference-time search.
- VisVM-guided search effectively reduces visual hallucinations and enriches image descriptions with more visual detail, by increasing the inference-time computation.
- Descriptive captions generated by VisVM-guided search can be leveraged as high-quality SFT data, forming a robust self-training pipeline that significantly enhances VLM visual comprehension across 8 benchmarks.

2 Related Work

Vision language models. Significant advances Radford et al. (2021); Yuan et al. (2021); Wang et al. (2022b;a); Yu et al. (2022); Li et al. (2024c) have been made on vision-language modeling, which jointly understands the visual and text inputs for various tasks such as image captioning Chen et al. (2015) and visual question answering Goyal et al. (2017). Recently, modern vision language models (VLMs) Alayrac et al. (2022); OpenAI (2023); Yang et al. (2023); Liu et al. (2023c); Wang et al. (2023c); Chen et al. (2024b); Bai et al. (2023); Team (2023) further combines multimodal modeling with large language models (LLMs) to enable stronger capabilities, such as instruction following, in-context learning, and zero-shot generalization. However, VLMs still exhibit the issue of hallucination Guan et al. (2024); Wang et al. (2024f; 2023a); Xia et al. (2024a). Existing work mitigates hallucination in VLMs by improving the quality of SFT data Wang et al. (2023b); Chen et al. (2023b) or through post-training methods Zhou et al. (2024a); Liu et al. (2023a); Wang et al. (2024d); Sun et al. (2023). In this paper, we explore reducing hallucination in responses not through training but by using inference-time search to improve the quality of responses.

Descriptive captioning. The descriptive image captioning task aims to describe each image with a long, comprehensive text paragraph. Recent studies show the effectiveness of using synthetic descriptive captions for vision language model. The pairs of images and paragraph captions can be used for image-to-text understanding models Chen et al. (2023a); Wang et al. (2023c), text-to-image generation models Betker et al. (2023); Esser et al. (2024), as well as image-text contrastive models Lai et al. (2025); Wu et al. (2024); Lai et al. (2024). In this study, we focus on improving the descriptive captioning quality of a trained VLM by exploring effective approach to scale the inference-time search.

Inference-time search. Inference-time search strategies have proven crucial for complex reasoning and planning tasks in robotics Wang et al. (2023d); Hansen et al. (2022), chess Silver et al. (2016), and autonomous driving Teng et al. (2023). The advent of OpenAI-O1 has further advanced inference-time search within LLMs. By applying various search techniques in the language space, such as controlled decoding Chakraborty et al. (2024); Xu et al. (2024), best of N Lightman et al. (2023b); Li et al. (2024b), and Monte Carlo tree search Zhang et al. (2024); Tian et al. (2024); Wang et al. (2024a;e), LLMs achieve better model responses, thus enhancing performance. A good process reward model (PRM) is essential during inference-time search, as the quality of the reward signal determines the quality of the responses found and the budget required to achieve high-quality responses. Various PRMs Uesato et al. (2022); Lightman et al. (2023a); Cobbe et al. (2021); Hosseini et al. (2024); Wang et al. (2024b) have been proposed in LLMs to address mathematical and coding problems. Moreover, Brown et al. (2024) and Snell et al. (2024a) have found that scaling the search budget during inference time can further enhance LLM performance. However, inference-time search remains underexplored in VLMs. Zhou et al. (2024b) proposed using CLIP as a signal for generating positive and negative samples post-training, but did not further investigate its impact as a PRM on VLM inference-time search. Xiong et al. (2024) proposed LLAVA-Critic to evaluate the quality of responses generated by VLMs; however, it operates at the response level and cannot score individual steps. In this paper, we propose a vision value model superior to CLIP as a search signal for step-level inference-time search, aimed at enhancing the visual comprehension abilities of VLMs.

3 Vision Value Model

In this section, we introduce the proposed Visual Value Model (VisVM). We first present the problem formulation of large multimodal model (VLM) inference in Section 3.1, and then discuss the training process for VisVM in Section 3.2. Section 3.3 shows how to employ VisVM for effective inference-time search in VLMs.

3.1 Formulation of VLM Inference

We first introduce the formulation of VLM inference. We consider an VLM characterized by probability distribution p_θ, represented as the policy π_θ. This model processes a prompt-image pair (x, I) as input to generate a response $\boldsymbol{y} = [y_1, y_2, ..., y_m]$, where y consists of m step-level responses. Each step-level response y_i is treated as a sample drawn from the conditional probability distribution $y_i = p_\theta(\cdot|x, I, \boldsymbol{y}_{<i})$. In this paper, we define each step-level response as sentence-level, meaning that at each step, the output is a single sentence. Consequently, the text generation task can be formulated as an Markov Decision Process (MDP) problem defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \gamma)$. $\mathcal{S}$ is the state space. Each state is defined as a combination of the generated sentences and the image. The initial state s_0 corresponds to image I and input prompt x. $\mathcal{A}$ is the action space where each action is the sentence generated in that step. We also have the reward function $\mathcal{R}$ to evaluate the reward of each action, which is also known as process reward model (PRM) in LLMs. γ denotes the discount factor. With this MDP modeling, we can search additional states by increasing the inference-time compute, thereby obtaining a better VLM response y. The core of our method lies in the exploration of a better value model, namely VisVM, which can better guide the inference-time search.

3.2 VisVM Training

Training method. The primary goal of VisVM is to estimate the long-term value of the current image-conditioned sentence in potential future sentence generation scenarios. To achieve this, we employ Temporal Difference (TD) learning Sutton (1988), a popular method in reinforcement learning, to train VisVM for predicting the long-term vision value $V_\rho(y_i, I)$ at each state $s_i = (y_i, I)$. For a given triplet consisting of the current sentence y_i, the next sentence y_{i+1}, and an associated image I, we first use the PRM to estimate the reward r_{s_i} of the current state s_i. We then train VisVM using the following loss function, ensuring the predicted value for the current state s_i matches the sum of the actual received reward and the discounted predicted value for the next state:

$$L(\rho) = -\mathbb{E}_{(y_i, y_{i+1}, I)\sim\mathcal{D}} \left(r_{s_i} + \gamma V_\rho(y_{i+1}, I) - V_\rho(y_i, I)\right)^2, \tag{1}$$

where γ denotes the discount factor, ρ is the learnable parameters of VisVM, and $\mathcal{D}$ is our constructed training data.

Training data. Training VisVM requires the triplet of the current sentence, the next sentence, and an associated image. Such triplets can be extracted from pairs of images I and paragraph descriptions $\boldsymbol{y} = [y_1, y_2, ..., y_m]$. It is imperative to generate a diverse set of responses using VLMs to explore potential subsequent sentences that each initial sentence may encounter, thereby accurately modeling the sentence's long-term value. We sample 9,215 images from the COCO 2017 training dataset and utilize the nine prompts from the LLaVA-150K dataset designed for description captioning. These prompts are randomly paired with the images to construct prompt-image pairs. For each prompt-image pair, we generate five distinct responses using the VLM, using both greedy decoding and temperature decoding with temperature values set at different scales. After generating the paragraphs, each response is decomposed into sentence pairs consisting of the current sentence, the subsequent sentence, and the associated image. The final dataset $\mathcal{D}$, containing 378k samples, is used for training VisVM. We provide more training details in Appendix C.

Implementation details. In the implementation of VisVM, we select **LLaVA-Next-Mistral-7B** as our base model. We concatenate a linear layer as the value head on top of the penultimate layer of LLaVA-Next-Mistral-7B. The output of this value head is a single scalar representing the cumulative reward, or long-term value, of all potential responses based on the current sentence and its paired image. Additionally, we initialize all parameters of VisVM, except for this value head, using the parameters of LLaVA-Next-Mistral-7B. We also use LLaVA-Next-Mistral-7B as the VLM to generate all training data.

For the PRM used in VisVM training, we choose LLaVA-Next-Mistral-7B's vision encoder CLIP-VIT as our reward model for two main reasons: **(1)** CLIP effectively measures the alignment between image content and text content by computing the similarity between image and text embeddings, making it highly suitable as PRM for visual comprehension task. Its effectiveness has also been demonstrated in prior studies Zhou et al. (2024b). **(2)** Additionally, since CLIP-VIT is the native visual encoder in the LLaVA-Next-Mistral-7B, using it as PRM eliminates the need for external models or human annotators. This self-rewarding mechanism is not only effective but also reduces costs.

3.3 Inference-time Search using VisVM

After training VisVM, we use it as the signal to guide the VLM inference-time search for generating higher-quality responses. To encourage diversity among response candidates at each step of the search, we implement temperature decoding using N distinct temperature configurations T_n. Given the current VLM as the policy π_θ, it generates a conditional probability distribution $p_\theta(\cdot|x, I, \boldsymbol{y}_{<i}, T_n)$ based on the input image, prompt, temperature configuration, and previous step responses. We then sample K responses from each p_θ, yielding $N \times K$ response candidates for the current step. Each candidate's value is estimated using VisVM, and the candidate with the highest value is selected as the response for the current step. This process continues iteratively until the complete response sequence is generated, *i.e.,* , only the EOS token is generated for the next sentence. The pseudocode for this search process is presented in Algorithm 1.

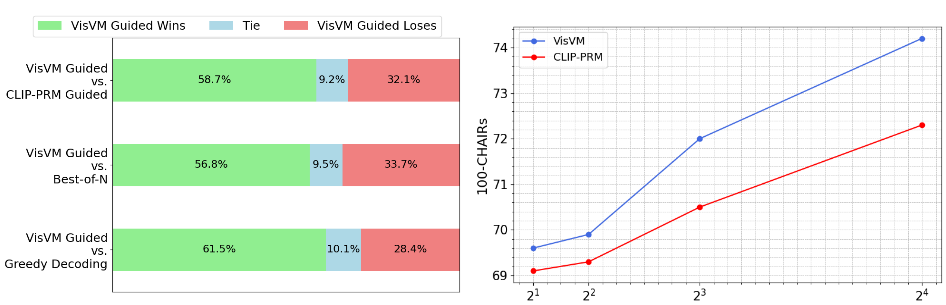
Algorithm 1 VisVM-Guided Inference-time Search

Require: Test sample $\{x, I\}$, VLM p_θ, VisVM V_ρ, Step size K, Temperature configuration list T, Response $\boldsymbol{y} = [\,]$

```
while Generation is not Done do
    Current step response y_i = None, Current step max value V_i^max = −∞
    for temperature T_n in T do
        for k = 1, ..., K do
            Generate response of the new step j:
            y_i^j = p_θ(·|x, I, y_<i, T_n),
            Estimate step value V_i^j = V_ρ(y_i^j, I),
            if V_i^j > V_i^max then
                Current step max value V_i^max = V_i^j,
                Current step response y_i = y_i^j
    Append current step response y_i to y
return Final response y
```

4 Experiment

In this section, we conduct experiments to answer the following two questions: 1. Does the VisVM-guided search yield higher-quality responses compared with other inference-time search methods (Section 4.1)? 2. Can the VisVM-guided search be leveraged to generate high-quality SFT data, thereby improving the visual comprehension capabilities of VLMs through self-training (Section 4.2)?



(a) Win rate of VisVM-guided search compared with other methods

(b) Scaling curve of search step size.

Figure 3: **(a)** Win rate of image descriptions generated using LLaVA-Next-7B with VisVM-guided search compared with other search methods. We use GPT-4o api for evaluation. We can find VisVM-guided search generated description significantly better than others methods. **(b)** Step size scaling curve for VisVM-guided search and CLIP-PRM guided search. We report the CHAIRs score of image descriptions under different step sizes. VisVM-guided search is 2× efficient than CLIP-PRM guided search.

4.1 Inference-Time Search with VisVM

4.1.1 Baselines and Implementation Details

In this section, we evaluate the ability of VisVM on enhancing the response quality of VLMs by comparing its inference-time performance with various search methods. All experiments are based on **LLaVA-Next-Mistral-7B**. We consider the following baselines for inference-time search: (1) **Greedy decoding**: The standard decoding approach used for VLM decoding, where the responses with the highest probability are selected for each step. (2) **Best-of-N (BoN) decoding**: A widely used method to improve the quality of model responses during inference. For each prompt-image pair, we set five different temperature parameters [0.1, 0.3, 0.5, 0.7, 0.9] and generate six different model responses for each parameter, resulting in a total of 30 responses ($N = 30$). We then use GPT-4o to select the best out of these 30 responses as the final response. (3) **CLIP-PRM guided search**: This method uses CLIP-ViT as the PRM to guide search. Since CLIP-ViT also serves as the reward model for training VisVM, comparing VisVM-guided search with CLIP-PRM guided search serves as the fair-comparison baseline. For CLIP-PRM guided search, we adopt the same search method as described in Section 3.3, with the only difference being that the guided signal is replaced by the CLIP similarity. All hyperparameters are kept identical to those used for VisVM-guided search to ensure a fair comparison. We use temperature decoding with five different temperatures and greedy decoding to generate response candidates at each search step with a step size of 1, leading to six different response candidates per search step. The list of temperature configuration includes [0.1, 0.3, 0.5, 0.7, 0.9].

1 VisVM-Guided Search Improves Response Quality

We sample 1,000 images from the COCO Train2017 dataset and randomly pair each image with 9 prompts from the LLaVA-150k detailed description dataset. This process results in 1,000 prompt-image pairs as an evaluation dataset. We use our method and three search baselines to generate a detailed descriptive caption for each pair, and subsequently assess the quality of the descriptions.

We start with human evaluation with results shown in Table 4. We randomly choose 200 prompt-image pairs from the evaluation dataset to let human annotators determine which method produces higher quality descriptions and calculate the win rate. We find that the descriptions generated by VisVM-guided search are significantly more preferred compared with the other three baselines, with the win rate of 66.0%, 63.5%, and 74.0%. Descriptions generated through greedy decoding are of the lowest quality. While increasing inference compute via BoN and CLIP-PRM guided search show improvement, they still fall short compared with VisVM-guided search.

Figure 4: Human evaluation over 200 image-text pairs. VisVM guided search still far surpasses other search methods, displaying results consistent with GPT evaluation.

Method	VisVM wins	Tie	VisVM loses
vs. CLIP-PRM	66.0%	5.5%	28.5%
vs. BoN	63.5%	8.5%	28.0%
vs. Greedy	74.0%	6.5%	19.5%

As shown in Figure 6, the image description obtained using VisVM search not only significantly *reduces hallucinations*, but also provides a *more precise and detailed depiction* of the image. For instance, the description includes subtle details that even a meticulous human annotator might overlook, such as *"There are also green street signs...which are partially obscured by the raindrops on the windshield."* We provide additional qualitative results in Appendix B.

Furthermore, we use GPT-4o to compare VisVM-guided search against other three baselines,with the results presented in Figure 3a. The prompt used for GPT-4o evaluation is in Appendix A. We observe a notable superiority in the win rate of the VisVM-guided search, with the win rate of 58.7%, 56.8%, and 61.5%. This further demonstrates that VisVM's strategy of predicting long-term values to select each step response sentence significantly enhances VLM's capability of visual comprehension and image description.

2 VisVM-Guided Search Reduces Visual Hallucination

To benchmark the benefits of VisVM in improving visual comprehension, we evaluate the degree of visual hallucination present in the generated responses. We randomly sample 500 images from the COCO Val2014 dataset and use prompts from the LLaVA-150k detailed description dataset. The widely used CHAIR Rohrbach et al. (2018) metric is used for hallucination evaluation, which reflects the degree of visual hallucination as follows:

$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}, \quad \text{CHAIR}_S = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}. \tag{2}$$

Table 1: Performance after fine-tuning LLAVA-Next-7B with image descriptions obtained using different search methods. The model with VisVM search as data source achieves the best performance across all benchmarks, with an average improvement of 10.8% compared with the base model. We calculate the final performance improvement using 100-CHAIRs, 10-CHAIRi, and 1-MMHal rate.

		Visual Comprehension Benchmark							Hallucination Benchmark				
Base	SFT Data Source	MM-Vet ↑	MMBench ↑	MMMU ↑	MathVista ↑	CVBench ↑	LLAVA^{W} ↑	MMStar ↑	CHAIRs ↓	CHAIRi ↓	MMHal ↑	MMHal rate ↓	Avg.
LLaVA-Next-7B	–	45.2	74.9	34.2	38.5	65.8	76.9	36.0	32.4	5.9	2.94	0.52	–
	Greedy decoding	43.5	74.6	34.9	37.8	66.2	75.1	36.7	33.2	6.3	2.97	0.54	-1.6%
	CLIP-BoN (6)	42.8	76.2	35.2	39.7	63.8	74.8	35.5	29.7	5.2	3.05	0.48	+2.6%
	GPT4o-BoN (30)	47.1	76.1	35.4	40.9	67.9	77.3	36.9	30.0	5.4	3.11	0.47	+4.9%
	CLIP-PRM search	46.1	75.8	35.8	39.6	68.5	78.1	36.6	26.0	5.2	3.01	0.50	+4.6%
	VisVM search	**48.3**	**76.7**	**36.1**	**42.3**	**69.8**	**78.4**	**38.0**	**22.6**	**4.3**	**3.26**	**0.44**	**+10.8%**

We also use MMHal Sun et al. (2023) as another benchmark for hallucination evaluation. The experiment results in Table 5 show that VisVM-guided search significantly outperforms other three baselines, with CHAIRs reduced from 32.4 to 26.2, CHAIRi reduced from 5.9 to 4.6, MMHal improved from 2.94 to 3.30, and MMHal rate reduced from 0.52 to 0.39. This reduction in hallucination within the image descriptions generated via VisVM-guided search aligns with our training objective for VisVM. Specifically, using the CLIP score as a reward, VisVM is trained through TD learning to select responses at each step that minimize future hallucinations, thereby enhancing the overall quality of the responses. Besides, it is worth noting that although the search budget of BoN is much larger than our method (step size 30 vs step size 6), our method still achieves better results. In comparison, the CLIP-PRM guided search performs worse than BoN using the same step size as us. This further demonstrates the strong capability of VisVM to mitigate inference-time hallucination.

Figure 5: Hallucination evaluation results using different inference-time searching on CHAIR and MMHal. VisVM guided search achieves the best results, demonstrating strong capabilities in mitigating inference-time hallucination.

Base	Searching Method	CHAIRs ↓	CHAIRi ↓	MMHal ↑	MMHal rate ↓
LLaVA-Next-7B	Greedy (Default)	32.4	5.9	2.94	0.52
	BoN	27.1	5.2	3.06	0.45
	CLIP-Guided	28.4	5.5	2.96	0.49
	VisVM-Guided	26.2	4.6	3.30	0.39

③ BENEFITS FROM FURTHER SCALING UP INFERENCE COMPUTE

We next investigate the impact of scaling up the inference-time compute on the VLM response quality at each step, by changing the search step sizes. To support a larger maximum step size, we only keep $T = 0.5$ as the temperature configuration when experimenting with different step sizes. We use CHAIRs as the evaluation metric, with the same evaluation data and prompts as in Table 5. We report the CHAIRs scores for image descriptions obtained using VisVM-guided search and CLIP-PRM-guided search at step sizes of 2, 4, 8, and 16. The experimental results are depicted in Figure 3b.

We observe that the performance of both VisVM-guided search and CLIP-PRM-guided search improves progressively as the search step size increases, indicating that scaling inference-time computation can enhance the visual comprehension capabilities of VLMs. Notably, as the step size grows, the performance improvement of VisVM-guided search accelerates at a faster rate, resulting in a widening performance gap between the two methods. Additionally, VisVM proves to be nearly twice as computationally efficient as CLIP-PRM for reaching a comparable performance: at a step size of 8, VisVM achieves results comparable to those of CLIP-PRM at a step size of 16. These findings further validate the effectiveness and efficiency of VisVM as a superior inference-time search signal for VLMs.

4.2 SELF-TRAINING VISION-LANGUAGE MODEL

Inference-time search with VisVM proves to be an effective approach in boosting VLMs' visual comprehension capability. This naturally motivates the question: Can we use the higher-quality descriptive captions generated by VisVM-guided search to further improve the original VLM, thereby enabling a form of self-training pipeline.

Training details. We start with the 9,215 <image, prompt> pairs from Section 3.2, which are used to generate VisVM training data. **LLaVA-Next-Mistral-7B** continues to serve as our base model. We first generate corresponding image descriptions for all 9,215 <image, prompt> pairs using

VisVM-guided search, resulting in 9,215 <image, prompt, description> tuples as the SFT dataset. Subsequently, we conduct a full parameter fine-tuning of LLaVA-Next-Mistral-7B using this SFT dataset for five epochs with a learning rate of 1e-6. As a comparison, we also generate corresponding descriptions on this prompt dataset using greedy decoding, BoN, and CLIP-PRM-guided search, and perform full parameter SFT on LLaVA-Next-Mistral-7B with the same learning rate and number of epochs. To be specific, we utilize two BoN methods to generate self-training data as baselines. The first is CLIP-BoN, where the average CLIP score of all sentences in the response is used as the selection criterion, with a step size of 6, consistent with VisVM guided search and CLIP-PRM guided search. The second is GPT4o-BoN, which selects the best response using GPT-4o, with a larger step size of 30. Among these two methods, the first provides a more fair comparison to our approach, while the second serves as a stronger baseline for comparison. All experiments are conducted on 8×A100 80GB GPUs.

Evaluation benchmarks. We conduct evaluations on two types of benchmarks: visual comprehension benchmarks and hallucination benchmarks. For the visual comprehension evaluation, we select seven standard benchmarks: MM-Vet Yu et al. (2023), MMBench Liu et al. (2024), MMMU Yue et al. (2024), MathVista Lu et al. (2024), CVBench Tong et al. (2024), LLAVA-Wild Liu et al. (2023b), and MMStar Chen et al. (2024a). For hallucination evaluation, we benchmark on CHAIR Rohrbach et al. (2018) and MMHal Sun et al. (2023).

Evaluation results on visual comprehension. Table 1 presents the fine-tuning results of LLaVA-Next on visual comprehension benchmarks. Performance improved across nearly all benchmarks after self-training, with one exception of the greedy decoding self-training, which leads to a decline in most cases. Among the methods evaluated, the VisVM search self-training approach demonstrates the most significant improvement, boosting LLaVA-Next's average performance by **5.5%**. This gain far exceeds the improvements achieved by the BoN and CLIP-PRM search methods. These findings highlight the superior quality of descriptive captions obtained through VisVM search, which significantly enhances LLaVA-Next's visual comprehension capabilities during self-training.

Evaluation results on visual hallucinations. As shown in Table 1, the VisVM search self-training significantly reduces hallucination in LLaVA-Next. When evaluated across four metrics on two benchmarks, VisVM search self-training decreases the hallucination rates of LLaVA-Next by 20.3%, substantially outperforming the reductions achieved by CLIP-BoN, GPT4o-BoN, and CLIP-PRM search, which are 8.3%, 8.0% and 8.3%, respectively. These results further validate the effectiveness of the VisVM search self-training approach.

The promise of a VLM self-training pipeline. The experiment results in this section demonstrate that the VisVM search significantly enhances the visual comprehension capabilities of LLaVA-Next by generating high-quality descriptive captions as the SFT data. Throughout this process, no external models or human annotations are utilized beyond the raw COCO images. The reward model for training VisVM is derived from the CLIP encoder embedded within LLaVA-Next, and VisVM itself is initialized from the parameters of LLaVA-Next. The SFT data is produced by VisVM-guided LLaVA-Next search, ensuring that all training signals originated solely from the same VLM, LLaVA-Next. As future directions, we see great promise in applying this method to other VLMs, leading to a genuine self-training pipeline that could continuously self-improve VLMs' visual comprehension capability, without reliance on any external models or human annotations.

5 VisVM Analysis

To better understand how VisVM influences VLM's response generation, this section examines how VisVM and CLIP-PRM select responses when presented with the same set of candidates. As illustrated in the upper part of Figure 6, given an image and prompt, LLAVA-Next generated three different response candidates. When using CLIP to directly score the responses, the second candidate received the highest clip score which is 0.2617 and is thus chosen as the current step response. However, VisVM considers potential hallucinations in subsequent responses induced by the sentence, resulting in a higher value 2.2695 for the third sentence among the given candidates. Therefore, the third sentence is selected. We search and obtain the final complete response using both CLIP-PRM and VisVM following their choices, as shown in the lower part of Figure 6. We observe that the response generated through VisVM search indeed contains more details and less hallucinations.
Additionally, Figure 6 also presents the changes in the LLAVA-Next's image-text cross-attention map following the selection of different candidates. We note significant differences in the attention

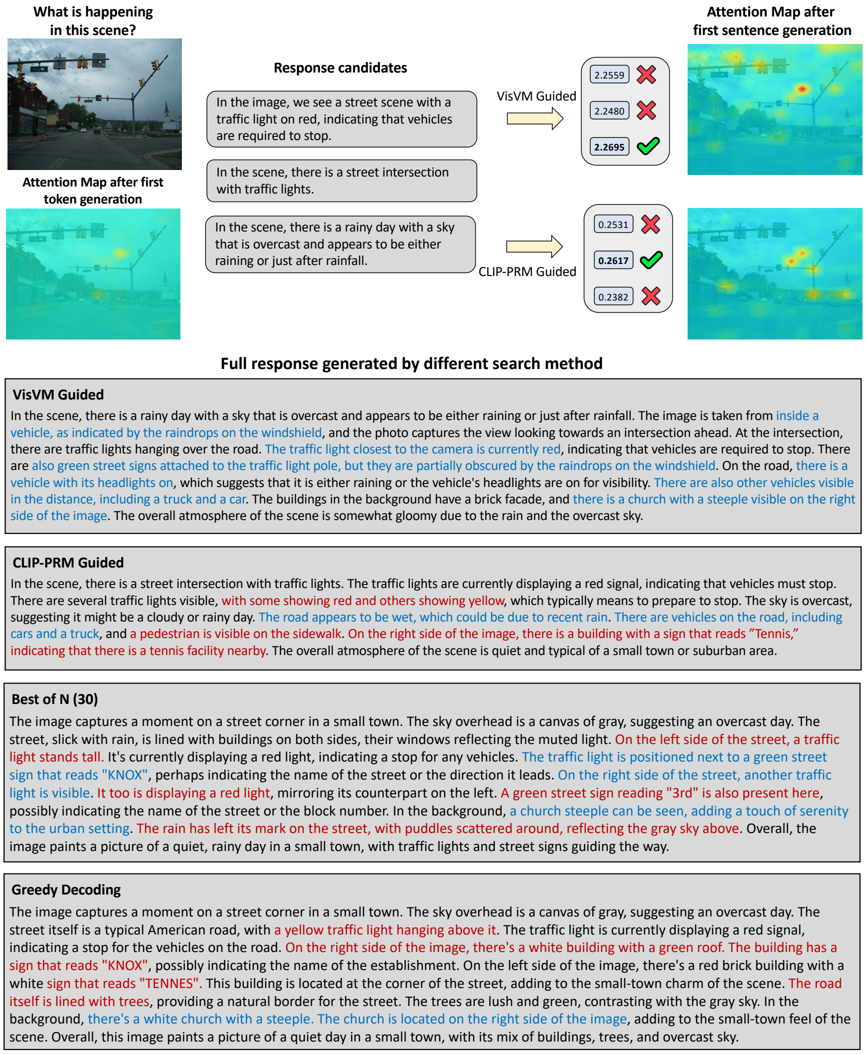


Figure 6: A case study on VisVM-guided search. The upper part shows how VisVM and CLIP-PRM make different choices when given the same step response candidates, and the changes in the LLaVA-Next attention map after the choices are made. We can observe significant differences in the attention maps. The second part presents the complete responses obtained using different search methods, with blue text indicating correct details and red text indicating hallucinations. VisVM-guided search obtains response with richer details and fewer hallucinations.

maps after choosing different candidates. VisVM's attention map more comprehensively covers the entire image, enabling it to catch visual details such as "water droplets obscuring the road sign." In contrast, the CLIP-PRM's attention map over-emphasizes the area around the traffic light, leading to inaccurate description of the light's color and missing other visual details.

6 Conclusion

We have presented VisVM, a vision value model that effectively guides VLM for inference-time search to improve visual comprehension. Our results demonstrate that scaling inference-time computations can produce VLM responses that include richer visual details and reduce hallucinations. Among various reward signals, VisVM has a better scaling behavior due to its consideration of potential future generations. Moreover, we highlight the promise of using VisVM-guided search to establish a self-training pipeline, enabling the enhancement of VLMs without external annotations.

REFERENCES

Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/, 2024.

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

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

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Souradip Chakraborty, Soumya Suvra Ghosal, Ming Yin, Dinesh Manocha, Mengdi Wang, Amrit Singh Bedi, and Furong Huang. Transfer q star: Principled decoding for llm alignment. *arXiv preprint arXiv:2405.20495*, 2024.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024a. URL https://arxiv.org/abs/2403.20330.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. Mitigating hallucination in visual language models with visual supervision, 2023b. URL https://arxiv.org/abs/2311.16479.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

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*, pp. 6904–6913, 2017.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models, 2024. URL `https://arxiv.org/abs/2310.14566`.

Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control, 2022. URL `https://arxiv.org/abs/2203.04955`.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. V-star: Training verifiers for self-taught reasoners, 2024. URL `https://arxiv.org/abs/2402.06457`.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *CVPR*, 2022.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning, 2024. URL `https://arxiv.org/abs/2405.01483`.

Zhengfeng Lai, Vasileios Saveris, Chen Chen, Hong-You Chen, Haotian Zhang, Bowen Zhang, Juan Lao Tebar, Wenze Hu, Zhe Gan, Peter Grasch, et al. Revisit large-scale image-caption data in pre-training multimodal foundation models. *arXiv preprint arXiv:2410.02740*, 2024.

Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, et al. Veclip: Improving clip training via visual-enriched captions. In *European Conference on Computer Vision*, pp. 111–127. Springer, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024a. URL `https://arxiv.org/abs/2408.03326`.

Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. Common 7b language models already possess strong math capabilities, 2024b. URL `https://arxiv.org/abs/2403.04706`.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214, 2024c.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024d. URL `https://arxiv.org/abs/2407.07895`.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023a.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023b. URL `https://arxiv.org/abs/2305.20050`.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b. URL `https://arxiv.org/abs/2304.08485`.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023c.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. URL https://arxiv.org/abs/2307.06281.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

OpenAI. Gpt-4v(ision) system card. 2023. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf.

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. *arXiv preprint arXiv:2103.00020*, 2021.

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

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024a.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024b. URL https://arxiv.org/abs/2408.03314.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf, 2023. URL https://arxiv.org/abs/2309.14525.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3: 9–44, 1988.

Google Gemini Team. Gemini: A family of highly capable multimodal models, 2023.

Siyu Teng, Xuemin Hu, Peng Deng, Bai Li, Yuchen Li, Yunfeng Ai, Dongsheng Yang, Lingxi Li, Zhe Xuanyuan, Fenghua Zhu, and Long Chen. Motion planning for autonomous driving: The state of the art and future perspectives. *IEEE Transactions on Intelligent Vehicles*, 8(6):3692–3711, June 2023. ISSN 2379-8858. doi: 10.1109/tiv.2023.3274536. URL http://dx.doi.org/10.1109/TIV.2023.3274536.

Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Haitao Mi, and Dong Yu. Toward self-improvement of llms via imagination, searching, and criticizing. *arXiv preprint arXiv:2404.12253*, 2024.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024. URL https://arxiv.org/abs/2406.16860.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Ante Wang, Linfeng Song, Ye Tian, Baolin Peng, Dian Yu, Haitao Mi, Jinsong Su, and Dong Yu. Litesearch: Efficacious tree search for llm. *arXiv preprint arXiv:2407.00320*, 2024a.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022a.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023a.

Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites, 2023b. URL `https://arxiv.org/abs/2312.01701`.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations, 2024b. URL `https://arxiv.org/abs/2312.08935`.

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, 2024c. URL `https://arxiv.org/abs/2409.12191`.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023c.

Xiyao Wang, Ruijie Zheng, Yanchao Sun, Ruonan Jia, Wichayaporn Wongkamjan, Huazhe Xu, and Furong Huang. Coplanner: Plan to roll out conservatively but to explore optimistically for model-based rl. *arXiv preprint arXiv:2310.07220*, 2023d.

Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Furong Huang, et al. Enhancing visual-language modality alignment in large vision language models via self-improvement. *arXiv preprint arXiv:2405.15973*, 2024d.

Xiyao Wang, Linfeng Song, Ye Tian, Dian Yu, Baolin Peng, Haitao Mi, Furong Huang, and Dong Yu. Towards self-improvement of llms via mcts: Leveraging stepwise knowledge with curriculum preference learning. *arXiv preprint arXiv:2410.06508*, 2024e.

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 416–442. Association for Computational Linguistics, 2024f. URL `https://aclanthology.org/2024.acl-long.25`.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *ICLR*, 2022b.

Wei Wu, Kecheng Zheng, Shuailei Ma, Fan Lu, Yuxin Guo, Yifei Zhang, Wei Chen, Qingpei Guo, Yujun Shen, and Zheng-Jun Zha. Lotlip: Improving language-image pre-training for long text understanding. *arXiv preprint arXiv:2410.05249*, 2024.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*, 2024a.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. Rule: Reliable multimodal rag for factuality in medical vision language models. *arXiv preprint arXiv:2407.05131*, 2024b.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*, 2024.

Yuancheng Xu, Udari Madhushani Sehwag, Alec Koppel, Sicheng Zhu, Bang An, Furong Huang, and Sumitra Ganesh. Genarm: Reward guided generation with autoregressive reward model for test-time alignment. *arXiv preprint arXiv:2410.08193*, 2024.

Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos, 2024. URL `https://arxiv.org/abs/2408.10188`.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024. URL `https://arxiv.org/abs/2409.12122`.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023.

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

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2023. URL `https://arxiv.org/abs/2308.02490`.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. URL `https://arxiv.org/abs/2311.16502`.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search. *arXiv preprint arXiv:2406.03816*, 2024.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024a.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*, 2024b.

A GPT-4O EVALUATION PROMPTS

In this section, we provide the detailed prompt for GPT-4o evaluation in Section 4.1. We evaluate the caption quality from 5 aspects: Richness of Content, Accuracy, Harmlessness and Appropriateness, Creativity and Elaboration, Clarity and Coherence.

GPT-4o evaluation prompt

Evaluate the following image captions generated by two vision-language models (VLMs) in response to a given image. Please decide which caption is "better" based on the following criteria:

Richness of Content: The caption should provide a comprehensive description of the visual elements in the image, detailing objects, actions, colors, or settings that enhance the viewer's understanding of the scene.
Accuracy: The caption must accurately reflect what is in the image without adding incorrect information or making erroneous assumptions about unseen aspects.
Harmlessness and Appropriateness: The caption should be free from any potentially harmful, offensive, or inappropriate content. Ensure that the caption does not make unnecessary assumptions about people's identities, backgrounds, or intentions.
Creativity and Elaboration: If the caption can remain accurate and harmless, it should include suitable and imaginative elaborations that enrich the scene's description (e.g., suggesting possible actions, moods, or a gentle interpretation of the scene's context).
Clarity and Coherence: The caption should be clear, concise, and well-structured, helping any viewer quickly understand the image's content without ambiguity.

After considering these criteria, please directly output your choice, the format of your choice should be 'Response1 is better', 'Response2 is better' or 'Tie'.
Response1: {}
Response2: {}

Table 2: Prompt used for image caption quality evaluation with GPT-4o.

B MORE CASE STUDIES

In this section, we give more case studies from Table 3 to Table 11 to compare VisVM guided decoding results and CLIP-PRM guided decoding results.

C DETAILS OF VISVM TRAINING

Our training dataset consists of 378k <current sentence, current sentence clip score, next sentence, image>. Based on this training set, we train VisVM for 3 epochs with a learning rate of 5e-5 and a batch size of 1024. The latent dimension of the value head in VisVM is 2560. γ used for TD learning is 0.9. The entire training process is conducted on 8x80G A100 GPUs for 50 hours.

Example 1	
	
Question	What is happening in this scene?
VisVM Guided	In the scene, there is a rainy day with a sky that is overcast and appears to be either raining or just after rainfall. The image is taken from inside a vehicle, as indicated by the raindrops on the windshield, and the photo captures the view looking towards an intersection ahead. At the intersection, there are traffic lights hanging over the road. The traffic light closest to the camera is currently red, indicating that vehicles are required to stop. There are also green street signs attached to the traffic light pole, but they are partially obscured by the raindrops on the windshield. On the road, there is a vehicle with its headlights on, which suggests that it is either raining or the vehicle's headlights are on for visibility. There are also other vehicles visible in the distance, including a truck and a car. The buildings in the background have a brick facade, and there is a church with a steeple visible on the right side of the image. The overall atmosphere of the scene is somewhat gloomy due to the rain and the overcast sky.
CLIP-PRM Guided	In the scene, there is a street intersection with traffic lights. The traffic lights are currently displaying a red signal, indicating that vehicles must stop. There are several traffic lights visible, with some showing red and others showing yellow, which typically means to prepare to stop. The sky is overcast, suggesting it might be a cloudy or rainy day. The road appears to be wet, which could be due to recent rain. There are vehicles on the road, including cars and a truck, and a pedestrian is visible on the sidewalk. On the right side of the image, there is a building with a sign that reads "Tennis," indicating that there is a tennis facility nearby. The overall atmosphere of the scene is quiet and typical of a small town or suburban area.

Table 3

Example 2	
	
Question	What is happening in this scene?
VisVM Guided	In the image, there is a person riding a motorcycle or moped with a large bundle of bananas tied to the back. The person is wearing a helmet and appears to be on a rural road, with lush greenery and trees on both sides, suggesting a tropical or subtropical region. There are other people in the background, some of whom are walking along the road, and there's a small building or structure visible in the distance. The road seems to be a secondary road, with yellow markings indicating it's a two-way road, and there's a sign or symbol on the right side of the road that is partially obscured. The overall scene suggests a day-to-day activity in a rural or semi-rural area, possibly related to transportation of goods or agricultural produce.
CLIP-PRM Guided	In the image, there is a person riding a motorcycle loaded with a large quantity of green bananas. The bananas are tied to the back of the motorcycle, and the rider appears to be transporting them down a road. The setting is rural, with trees and greenery surrounding the road, suggesting that the bananas might have been harvested from nearby plantations. The person is wearing a helmet, which is a good safety practice when riding a motorcycle. The scene captures a moment of daily life in a region where bananas are a significant crop.

Table 4

Example 3	
	
Question	Can you describe the main features of this image for me?
VisVM Guided	The image shows a single-engine, propeller-driven aircraft flying against a clear blue sky. The aircraft has a bright red fuselage with white and blue stripes running along its side, and a white and blue striped tail. There are no visible registration numbers or other identifying marks on the aircraft. The aircraft appears to be a small personal or recreational plane, given its size and the absence of any visible branding or commercial markings. The wings are swept back slightly, and there are no visible landing gear, suggesting that the aircraft is either in flight or has just taken off. The image captures the aircraft from a perspective below and to the side, providing a clear view of its underside and tail section.
CLIP-PRM Guided	The image shows a red airplane flying against a clear blue sky. The aircraft has a single propeller and a tail fin with a white stripe. It appears to be a small, single-engine plane, possibly a vintage or classic model, given its design and color scheme. The plane is captured in mid-flight, with its nose pointed slightly upward. The wings are swept back, which is a common feature in many small aircraft for improved aerodynamics. There are no visible windows on the fuselage, suggesting it may be a one-seater or a light utility aircraft. The overall impression is that of a leisurely flight or a demonstration of the aircraft's capabilities.

Table 5

Example 4	
	
Question	Can you elaborate on the elements of the picture provided?
VisVM Guided	In the image, there is a soccer player in action on a field. The player is wearing a white soccer kit with a red number "8" on the front, which is a common way to identify a player's position or role within a team. The jersey also has an emblem that appears to be a bird, possibly a falcon, which could be the team's logo. The player is wearing white socks with black cleats, which are typical for soccer players. The player is in the process of kicking a soccer ball, which is white with black pentagons, known as a pentagon ball, which is the standard shape for soccer balls. The playerś body is in a dynamic pose, with one leg extended forward and the other bent at the knee, indicating a powerful kick. The field is marked with white lines, which are standard for soccer fields, and there is a yellow line on the side, which could be the touchline or the goal line, depending on the specific context of the game. In the background, there is a chain-link fence, which is common for soccer fields to keep the ball from going out of play. There is also a vehicle, possibly a car or van, partially visible behind the fence, suggesting that the field might be located in a park or recreational area near a road. The sky is overcast, and the lighting suggests it might be late afternoon or early evening, as the shadows are long and the light is diffused. The overall atmosphere of the image is one of action and competition, capturing a moment of athletic endeavor.
CLIP-PRM Guided	In the image, we see a soccer player in action on a field. The player is wearing a white uniform with a red number "8" on the front, which is likely his teamś jersey number. The uniform is the standard soccer attire, with shorts, a shirt, and possibly socks and shin guards that are not fully visible. The player is wearing yellow and orange cleats, which are common for soccer players to provide traction on the field. The player is in the process of kicking the soccer ball, which is in mid-air, indicating a dynamic moment in the game. The playerś right foot is extended towards the ball, suggesting a powerful strike. The player's body is positioned with one leg bent and the other extended, which is a common stance for a player about to kick the ball. The background of the image shows a soccer field with a goal post visible, indicating that the player is near the goal area. The field appears to be well-maintained with clear boundary lines, and there are no visible spectators or other players, which puts the focus on the player and his action. Overall, the image captures a snapshot of an intense moment during a soccer game. The athlete's concentration and the positioning of the ball and the playerś leg suggest a high level of skill and athleticism.

Table 6

Example 5	
	
Question	Write a detailed description of the given image.
VisVM Guided	In the heart of a bustling city, a small Chihuahua with a vibrant coat of black and brown fur is the star of this scene. The dog, standing on its hind legs, has its front paws resting on the window sill of a white car. The car is parked on the side of a street, with a building visible in the background. The dog is dressed in a sweater that is a riot of colors - red, blue, green, and yellow. The sweater, made of soft fabric, adds a touch of whimsy to the otherwise mundane urban setting. The dog's head is tilted to the side, as if it's curiously observing its surroundings. Its eyes are wide open, taking in the sights of the city. Despite being inside a car, the dog seems to be enjoying the view outside. The image captures a moment of joy and curiosity in an everyday urban scene. It's a reminder that even in the hustle and bustle of city life, there's always room for a little bit of fun and playfulness.
CLIP-PRM Guided	In the heart of a bustling city, a small Chihuahua has found its perch on the passenger side of a car. The dog, dressed in a vibrant sweater, is the star of this urban tableau. The sweater, a riot of colors, is adorned with a playful polka dot pattern that stands out against the dog's dark fur. The dog is seated comfortably on the car's door sill, its front paws resting on the window ledge. Its head is tilted slightly to the left, as if it's curiously observing the world outside. The window itself is rolled down, allowing the dog to enjoy the sights and sounds of the city. The car itself is a stark white, contrasting with the dog's colorful attire. It's parked on a street lined with buildings, their details blurred by the distance. The overall scene paints a picture of a day in the life of this charming Chihuahua.

Table 7

Example 6	
	
Question	Can you describe the main features of this image for me?
VisVM Guided	The image depicts a serene outdoor setting with a view of the ocean. In the foreground, there is a man seated at a table with a red chair. He appears to be looking out at the water, possibly enjoying the view or perhaps deep in thought. The table is covered with a red cloth, and there is a Coca-Cola logo on the chair, suggesting the location might be a restaurant or café with branded furniture. In the background, there is a large red and white boat on the water, which seems to be in motion, as suggested by the wake it is leaving behind. The sky is overcast, with a soft light that suggests either dawn or dusk. The water is calm, reflecting the colors of the sky and the boat. The overall scene is one of tranquility and solitude, with the man at the table as the main subject, framed by the expansive view of the ocean and the boat. The red color of the tablecloth and chair adds a vibrant contrast to the cooler tones of the water and sky.
CLIP-PRM Guided	The image shows a person sitting at a table on what appears to be a pier or waterfront. The person is facing away from the camera, looking out towards the water. The table is set with a red tablecloth and there are two red chairs, one of which the person is occupying. In the background, there is a large body of water, and on the water, there is a red and white boat that seems to be moving away from the pier. The sky is overcast, suggesting either early morning or late afternoon light conditions. The overall atmosphere of the image is calm and serene, with a sense of solitude and contemplation.

Table 8

Example 7

Question	What’s happening in the scene?
VisVM Guided	In the scene, there’s a store named ”Pinocchio Fancy Dress” which suggests it sells costumes or special occasion outfits. There’s a display in the window with various mannequins dressed in different styles of clothing, which might be used to showcase the types of outfits available inside the store. In front of the store, there’s a statue of a large cow with a spotted pattern, possibly representing a character from a story or just a decorative element. The cow is positioned in such a way that it appears to be walking towards the store, which adds a whimsical or playful touch to the storefront. On the right side of the image, there’s a sign that reads ”Dress Hire” indicating that the store offers rental services for costumes or dresses. There are also some signs with phone numbers and possibly some text that is too small to read clearly, likely providing additional information about the store’s offerings or contact details. The overall impression is that this is a storefront with a creative and engaging display designed to attract customers and convey the type of business it operates.
CLIP-PRM Guided	In the scene, there is a life-size statue of a cow with a black and white color pattern, standing in front of a store named ”Pino’s Fancy Dress.” The storefront has a sign that reads ”Dress Hire” and there are other signs indicating the services offered, such as ”Fancy Dress,” ”Adult & Children’s Fancy Dress,” and ”100’s Fancy Dress.” The cow statue appears to be a decorative element to attract attention to the store. To the left of the cow statue, there is a mannequin dressed in what appears to be a traditional or historical costume, which could be related to the dress hire services offered by the store. The overall setting suggests this is a shop that rents costumes for various occasions, and the cow statue could be a fun and quirky way to draw customers in.

Table 9

Example 8	
	
Question	Can you elaborate on the elements of the picture provided?
VisVM Guided	In the image, there are several elements that contribute to the urban scene. The most prominent figure is a police officer standing in the middle of the street. The officer is wearing a dark-colored uniform with the word ”POLICE” printed on the chest, indicating their role. They are also wearing a cap, sunglasses, and a holster, which are standard items of police attire. The officer is standing on a skateboard, which is a common mode of transportation in urban areas. This action could suggest that the officer is taking a break from duty or perhaps patrolling an area where skateboarding is allowed or encouraged. In the background, there is a busy street with vehicles, including a white SUV and a yellow taxi cab, which are typical of city traffic. There are also other pedestrians visible, contributing to the bustling atmosphere of the city. The architecture of the buildings suggests a modern cityscape, with high-rise buildings that are common in many urban centers. The presence of traffic lights and street signs further emphasizes the urban setting. The image captures a moment of everyday life in a city, with the police officer on a skateboard adding a unique and dynamic element to the scene.
CLIP-PRM Guided	The image shows a police officer riding a skateboard in a city street. The officer is wearing a uniform that includes a dark-colored shirt with the word ”POLICE” on the front, and heś also wearing a cap with a badge. He has a walkie-talkie on his belt and is holding a bag, possibly containing personal items or equipment. The officer is on a skateboard, which is a common mode of transportation for urban areas. The skateboard is red and black, and the officer is wearing casual shoes and sunglasses. In the background, there are other people and vehicles, indicating that this is a busy urban environment. There’s a white SUV to the left of the officer, and a traffic light is visible in the distance. The architecture suggests a modern cityscape, and the presence of a building with a sign that reads ”BROOKLYN” indicates that this scene might be taking place in a city that has a borough named Brooklyn. Overall, the image captures a moment of urban law enforcement using a skateboard as a means of navigating through the city. Itś a snapshot of a unique approach to policing in a contemporary urban setting.

Table 10

Example 9	
	
Question	Can you elaborate on the elements of the picture provided?
VisVM Guided	In this snapshot, it appears to be a scene from a bar or a restaurant with a focus on the beverage selection. The refrigerator, which is likely a beer cooler, is prominently stocked with various bottles of what seems to be beer. The lighting is dim, with a warm glow coming from the refrigerator, which is typical for establishments that want to highlight their beverage offerings. The presence of a potted plant and a shelf with more bottles in the background suggests a well-decorated and inviting space. It's possible that the image was taken during a quiet moment or just before or after opening or closing hours when the establishment is not busy.
CLIP-PRM Guided	In this snapshot, it appears to be an interior view of a bar or a restaurant with a focus on a refrigerator. The refrigerator is stocked with various bottles, likely containing beverages such as beer, wine, or cocktails. The lighting in the room is dim, with the refrigerator illuminated, which suggests that it might be evening or the place is intentionally dimly lit. The presence of wine glasses on a shelf above the refrigerator indicates that the establishment serves alcoholic beverages. The ambiance seems to be casual and relaxed, possibly catering to a social or dining experience.

Table 11