

Scaling Inference-Time Search with Vision Value Model for Improved Visual Comprehension

Xiyao Wang^{1,2,†}, Zhengyuan Yang², Linjie Li², Hongjin Lu¹, Yuancheng Xu¹
Chung-Ching Lin², Kevin Lin², Furong Huang^{1,‡}, Lijuan Wang^{2,‡}
¹University of Maryland, College Park ²Microsoft
[†]xywang@umd.edu [‡]Equal advise

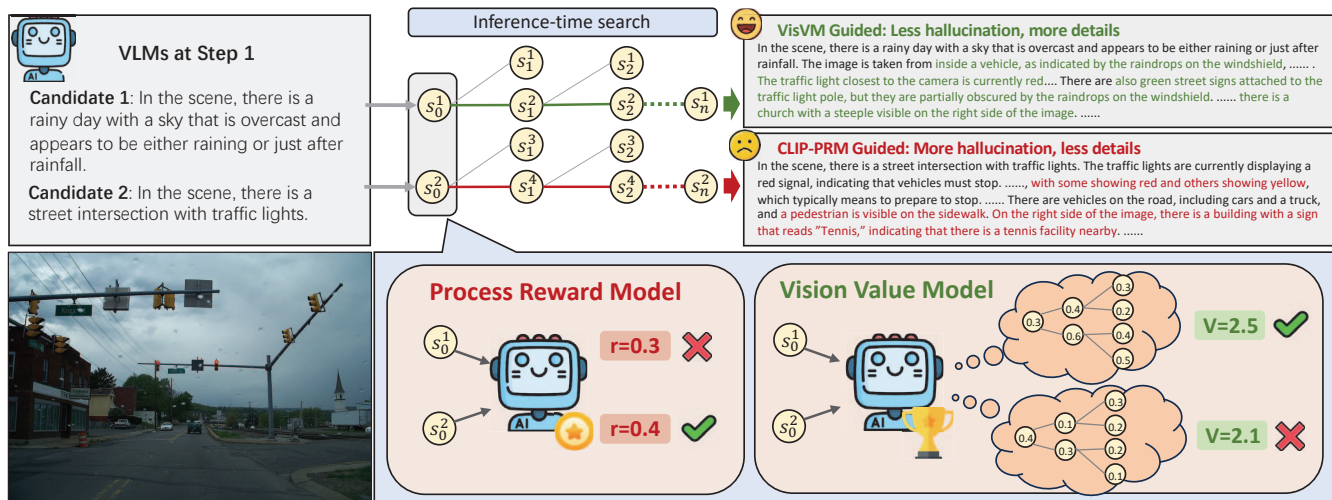


Figure 1. An illustration of how VisVM can better guide vision language model (VLM) during inference-time search. When selecting response candidates at each step, the process reward model (PRM) only considers the immediate reward, whereas VisVM predicts the long-term value by considering potential hallucinations in subsequent generated sentences. This enables VisVM to avoid response candidates with high 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 lack 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 **Vision Value Model (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 improves 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 [3, 12, 33, 39], multiple images [21, 28], and videos [25, 55, 65]. These capabilities stem from large-scale, high-quality

training data, often sourced from web-crawled image-text pairs [20, 37] with effective filtering [7, 19, 70], or enriched through techniques like distillation from stronger VLMs [8], human annotations [4], or added textual descriptions [23]. Despite this progress, VLMs still suffer from visual hallucinations [16, 31, 61] 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) [1, 30, 41, 42, 66] highlight inference-time search as a promising approach for improving response quality, complementary to training time effort. By leveraging a pretrained process reward model [47, 72], 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 **Vision Value Model (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 grounded in two key insights that distinguish it from traditional process reward models in LLM literature [13, 18, 29, 49, 56]: **(1) Forward-looking coherence:** Unlike approaches that rely solely on the local reward of the current sentence, VisVM predicts future consequences to maintain global consistency. It is trained using Temporal Difference (TD) learning [44], enabling it to assess long-term effects rather than just evaluating immediate responses. This forward-looking signal helps mitigate hallucinations by preventing sentences that may lead to inconsistencies in subsequent steps. **(2) Comprehensive visual grounding:** To reduce hallucinations, the reward signal must encapsulate rich visual semantics. We achieve this by leveraging CLIP’s text-image similarity metric, which effectively captures visual concepts and enforces alignment.

We validate the effectiveness of VisVM through two

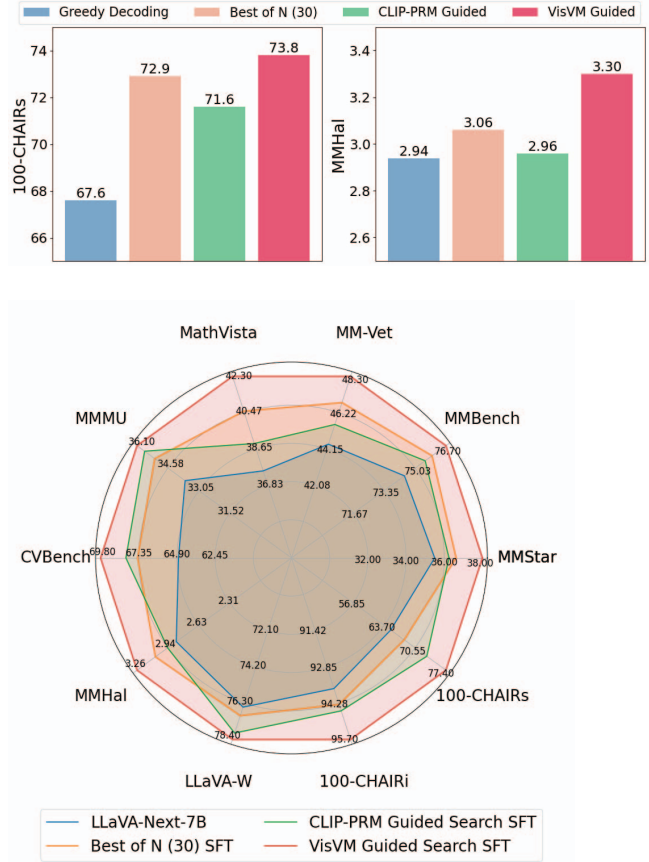


Figure 2. Upper: CHAIRs and MMHal score of descriptive captions generated by LLaVA-Next-7B during inference-time using different search methods. VisVM-guided search clearly outperforms other methods, indicating reduced visual hallucinations. Notably, even with a smaller search budget (search size 6 vs. search size 30), our approach still surpasses the Best-of-N method. **Lower:** 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.

main experiments: inference-time VisVM-guided search and self-improvement training. **(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, captions generated with VisVM consistently outperform those produced by greedy decoding, best-of-N decoding, and CLIP-PRM-guided search. Notably, VisVM-guided captions are preferred 74% of the time over those from greedy decoding. **(2)** To better 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 and Qwen2-VL-7B). Across nine standard benchmarks, VisVM-guided self-training improves the performance of the original VLMs by an average of 10.8% and 7.3%, respectively.

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 9 benchmarks.

2. Related Work

Vision language models. Significant advances [26, 37, 51, 62, 68, 70] have been made on vision-language modeling, which jointly understands the visual and text inputs for various tasks such as image captioning [10] and visual question answering [15]. Recently, modern vision language models [2, 3, 12, 33, 36, 45, 57, 67] further combine multimodal modeling with large language models to enable stronger capabilities, such as instruction following, in-context learning, and zero-shot generalization. However, VLMs still exhibit the issue of hallucination [16, 52, 61]. Existing work mitigates hallucination in VLMs by improving the quality of SFT data [11, 54] or through post-training methods [31, 43, 59, 74]. 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. Descriptive captioning 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 [8, 57], text-to-image generation models [4, 14], as well as image-text contrastive models [22, 23, 63]. In this study, we focus on improving the descriptive caption quality of a trained VLM by exploring effective approaches to scale the inference-time search.

Inference-time search. Inference-time search strategies have proven crucial for complex reasoning and planning tasks in robotics [17, 58], chess [40], and autonomous driving [46]. 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 [6, 64], best of N [27, 30], and Monte Carlo tree search [47, 50, 60, 72], 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 deter-

mines the quality of the responses found and the budget required to achieve high-quality responses. Various PRMs [13, 18, 29, 49, 56] have been proposed in LLMs to address mathematical and coding problems. Moreover, Brown et al. [5] and Snell et al. [41] 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. [75] 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. In this paper, we propose a vision value model superior to CLIP as a search signal for 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 $\mathbf{y} = [y_1, y_2, \dots, y_m]$, 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 $y_i = p_\theta(\cdot | x, I, \mathbf{y}_{<i})$. 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 [44], 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:

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

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 $\mathbf{y} = [y_1, y_2, \dots, y_m]$. 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 B.

Implementation details. For the implementation of VisVM, we take **LLaVA-Next-Mistral-7B** as our base model for an example. The implementation of VisVM in the experiment section for both LLaVA-OV-7B and Qwen2-VL-7B follows the same procedures. 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. For the training data, we use the base model corresponding to VisVM to generate descriptions for all images and decompose them into training data.

For the PRM used in VisVM training, we choose each VLM’s vision encoder. For LLaVA-Next-Mistral-7B and Qwen2-VL-7B, we use CLIP-ViT, while for LLaVA-OV-7B, we use SigLIP. There are two main reasons for this: **(1)** CLIP-like neural networks effectively measure 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 [75]. **(2)** Additionally, since CLIP-ViT and SigLIP are the native visual encoders in base VLMs, using them as PRM eliminates the need for external models or human annotators. This self-rewarding mechanism is not only effective but also reduces costs.

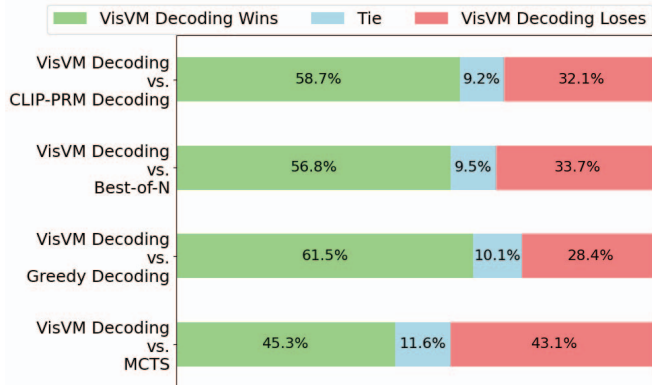
3.3. Inference-time Search using VisVM

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 $p_\theta(\cdot|x, I, \mathbf{y}_{<i}, T_n)$ 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 pseudo code for this search process is in Algorithm 1.

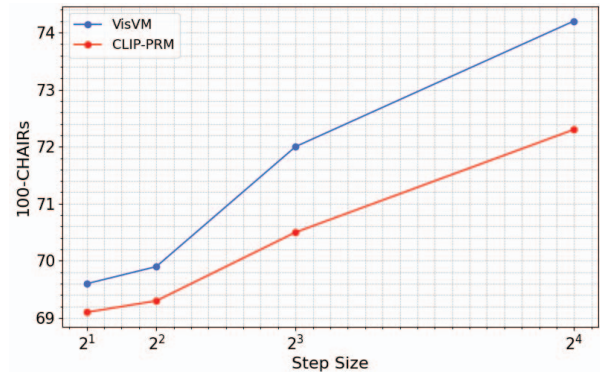
Algorithm 1 VisVM-Guided Inference-time Search

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

- 1: **while** Generation is not Done **do**
- 2: Current step response $y_i = \text{None}$, Current step max value $V_i^{max} = -\infty$
- 3: **for** temperature T_n in T **do**
- 4: **for** $k = 1, \dots, K$ **do**
- 5: Generate response of the new step j :
 $y_i^j = p_\theta(\cdot|x, I, \mathbf{y}_{<i}, T_n)$,
- 6: Estimate step value $V_i^j = V_\rho(y_i^j, I)$,
- 7: **if** $V_i^j > V_i^{max}$ **then**
- 8: Current step max value $V_i^{max} = V_i^j$,
- 9: Current step response $y_i = y_i^j$
- 10: Append current step response y_i to \mathbf{y}
- 11: **return** Final response \mathbf{y}



(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. 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)?

4.1. Inference-Time Search with VisVM

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]. (4) **Monte Carlo Tree Search (MCTS)**: MCTS is a widely used inference-time search method for enhancing the performance of LLM. Thus, we adopt MCTS for VLM as another key baseline. To ensure a fair comparison, we continue to use CLIP-ViT as the PRM. At each step, during the expansion of child nodes, we generate six child nodes using five different temperature values along with greedy decoding. The number of MCTS iterations is set to 10. We also provide comparison with other finetuning and decoding methods in Appendix C due to space limitation.

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. Then, we pair captions generated by VisVM-guided search with those from other decoding methods for the same image and subsequently assess the quality of the descriptions.

GPT evaluation. We use GPT-4o to compare VisVM-guided search against other baselines, as shown in Figure 3a. The prompt used for evaluation is in Appendix A. We observe a notable superiority in the win rate of the VisVM-guided search compared with CLIP-PRM, BoN, and Greedy, with the win rate of 58.7%, 56.8%, and 61.5%. Under GPT-based evaluation, while the advantage of VisVM-guided search over MCTS is less pronounced than against other baselines, it still achieves a higher win rate, outperforming MCTS with 45.3% compared to 43.1%.

Table 1. 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	62.4%	6.7%	30.9%
vs. MCTS	44.9%	15.2%	39.9%
vs. BoN	60.2%	9.6%	30.2%
vs. Greedy	75.8%	5.4%	18.8%

Human evaluation. We randomly select 200 prompt-image pairs and corresponding captions for human evaluation. We recruit 10 human evaluators to perform blind selections between these pairs to calculate the win rates of each method. We average their evaluations to obtain the final result in Table 1. We find that descriptions generated by VisVM-guided search are significantly preferred over those from CLIP-PRM, BoN, and Greedy decoding, with win rates of 62.4%, 60.2%, and 75.8%, respectively. Compared to GPT-based evaluation, VisVM exhibits a clearer advantage over MCTS under human evaluation. Human evaluators report more instances where captions from VisVM and MCTS are of comparable quality, leading to a higher tie rate. Notably, under the more reliable human evaluation, VisVM achieves a 44.9% win rate and a 15.2% tie rate against MCTS, reinforcing its effectiveness.

Results from both GPT and human evaluations consistently demonstrate that VisVM-guided search substantially enhances the response quality of VLMs in captioning.

Computational cost. We further compare the computational cost of various test-time compute methods to demonstrate the superiority of our approach. Specifically, we measure GPU hours required by each method to generate 1,000 image captions, utilizing an 8×80GB A100 GPUs setup. The results, summarized in Figure 4, reveal that all test-time compute methods significantly increased computational cost compared to greedy decoding. Among these, CLIP-guided search and VisVM-guided search incur the most minor increases. Furthermore, under identical step size and temperature settings, while MCTS achieves performance comparable to VisVM, it demands approximately seven times more computational resources. Additionally, MCTS must relearn the value function for each new prompt-image pair, highlighting its limited generalization capability. These findings further underscore the efficiency and effectiveness of VisVM, demonstrating its superiority in both performance and scalability.

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. Following the setting in previous works [74, 75], we randomly sample 500 images from the COCO Val2014 dataset and use prompts

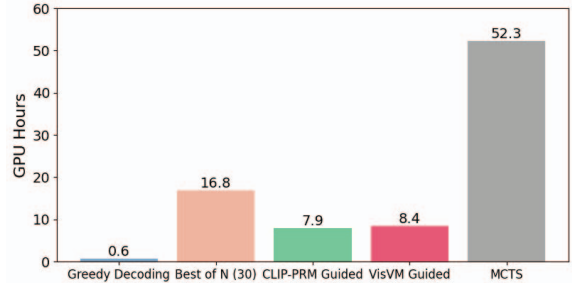


Figure 4. Comparison of GPU hours required to generate 1000 image captions by different test-time compute methods. The GPU hour consumed by VisVM-guided search is significantly lower than Best-of-N and MCTS.

Table 2. 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 ↓	AMBER Cov ↑
LLaVA-Next-7B	Greedy (Default)	32.4	5.9	2.94	0.52	63.9
	MCTS	25.9	4.7	3.24	0.37	67.3
	BoN	27.1	5.2	3.06	0.45	65.3
	CLIP-Guided	28.4	5.5	2.96	0.49	66.1
	VisVM-Guided	26.2	4.6	3.30	0.39	66.8
LLaVA-OV-7B	Greedy (Default)	35.0	5.5	3.12	0.36	74.5
	CLIP-Guided	29.4	4.8	3.27	0.34	75.5
	VisVM-Guided	27.0	3.7	3.34	0.31	76.2
Qwen2-VL-7B	Greedy (Default)	30.8	5.2	3.27	0.37	69.4
	CLIP-Guided	27.3	3.9	3.33	0.32	70.2
	VisVM-Guided	24.5	3.3	3.39	0.29	73.5

from the LLaVA-150k detailed description dataset. The widely used CHAIR [38] metric is used for hallucination evaluation and we also use MMHal [43] as another benchmark for hallucination evaluation. Besides, we adopt the coverage metric from AMBER [53] to evaluate the object coverage of generated captions, thus preventing artificially low hallucination scores caused by overly short captions

The experiment results based on LLaVA-Next-7B in Table 2 show that VisVM-guided search significantly outperforms greedy decoding, BoN, and CLIP-guided search, reducing CHAIRs from 32.4 to 26.2, CHAIRi from 5.9 to 4.6, MMHal rate from 0.52 to 0.39, and improving MMHal from 2.94 to 3.30. Meanwhile, object coverage improves from 63.9 to 66.8, indicating that the reduced hallucination brought by VisVM is not through generating short captions. Compared to MCTS, VisVM achieves comparable or superior performance while requiring significantly lower computation cost, highlighting its efficiency and effectiveness.

The 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 response quality.

To validate the robustness of VisVM, we retrain the corresponding VisVM based on LLaVA-OV-7B and Qwen2-VL-7B-Instruct, following the procedure in Section 3.2. Table 2’s results indicate that VisVM can effectively mitigate

Table 3. Ablation study of different PRMs for VisVM training. We observe that stronger PRM lead to better VisVM performance.

Searching Method	CHAIRs ↓	CHAIRi ↓	MMHal ↑	MMHal rate ↓	AMBER Cov ↑
Greedy (Default)	32.4	5.9	2.94	0.52	63.9
CLIP-VisVM-Guided	26.2	4.6	3.30	0.39	66.8
SigLIP-VisVM-Guided	25.6	4.4	3.31	0.36	67.5

hallucinations even when applied to stronger VLMs.

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 2. 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 performance 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 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 Stronger PRM can Further Enhance VisVM

In the previous and next sections, motivated by self-improvement, we consistently select the visual encoder corresponding to the base VLMs as PRM for VisVM training. In this subsection, we conduct an ablation study to demonstrate the generality of the VisVM training pipeline. Specifically, we utilize a more powerful model, SigLIP, as the PRM to train VisVM, while maintaining LLaVA-Next-7B as the base model. The remaining training procedures are identical to those used when CLIP served as the PRM. We evaluate the performance of VisVM trained with different PRMs using the CHAIR, MMHal, and AMBER Cov metrics; the results are presented in Table 3. Notably, using SigLIP as the PRM results in significantly reduced hallucinations in captions generated through VisVM guided search, with clear improvements observed particularly in CHAIR and AMBER Cov scores. This finding indicates that leveraging a stronger PRM further enhances VisVM capabilities, underscoring the generalizability and strong potential of the

VisVM training framework.

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. To demonstrate the robustness of our method, we conduct experiments using two different VLMs, **LLaVA-Next-Mistral-7B** and **Qwen2-VL-7B-Instruct**, as the base models. 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 on base VLMs using this SFT dataset for three 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 base models with the same learning rate and number of epochs. All experiments are conducted on $8 \times 80GB$ A100 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 [69], MMBench [34], MMMU [71], Math-Vista [35], CVBench [48], LLaVA-Wild [32], and MM-Star [9]. For hallucination evaluation, we benchmark on CHAIR [38] and MMHal [43].

Evaluation results on visual comprehension. Table 4 presents the fine-tuning results of LLaVA-Next and Qwen2-VL 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 and Qwen2-VL average performance by **5.5%** and **1.8%**, respectively. 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 VLM’s visual comprehension capabilities during self-training.

Evaluation results on visual hallucinations. As shown in Table 4, the VisVM search self-training can also significantly reduce hallucination in VLM. When evaluated

Table 4. Performance after fine-tuning LLaVA-Next-Mistral-7B and Qwen2-VL-7B-Instruct 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% and 7.3% compared with the base model, respectively. We calculate the final performance improvement using 100-CHAIRs, 10-CHAIRi, and 1-MMHal rate respectively.

Base	SFT Data Source	Visual Comprehension Benchmark							Hallucination Benchmark				Avg.
		MM-Vet ↑	MMBench ↑	MMMU ↑	MathVista ↑	CVBench ↑	LLAVA ^w ↑	MMS _{tar} ↑	CHAIRs ↓	CHAIRi ↓	MMHal ↑	MMHal rate ↓	
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%
	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%
Qwen2-VL-7B	–	58.4	83.0	49.3	58.2	74.5	87.1	56.3	30.8	5.2	3.27	0.37	–
	Greedy decoding	58.3	83.1	49.4	58.7	74.1	86.3	56.5	29.7	5.1	3.13	0.42	-0.8%
	GPT4o-BoN (30)	58.8	83.7	49.3	60.2	74.6	87.2	56.7	25.4	4.0	3.31	0.35	+3.9%
	CLIP-PRM search	58.5	83.5	49.5	59.2	74.9	87.9	56.5	23.6	3.7	3.31	0.32	+5.0%
	VisVM search	58.9	84.1	49.7	61.1	76.2	88.2	57.0	21.4	3.4	3.34	0.28	+7.3%

Table 5. Hallucination comparison of VisVM and CLIP selection starting from same sentence candidates.

Selection Model	CHAIRs ↓	CHAIRi ↓
CLIP	31.6	5.7
VisVM	30.9	5.3

across four metrics on two benchmarks, VisVM search self-training decreases the hallucination rates of LLaVA-Next and Qwen2-VL by 20.3% and 16.9%, substantially outperforming the reductions achieved by BoN and CLIP-PRM search. 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 and Qwen2-VL 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 visual encoder embedded within LLaVA-Next and Qwen2-VL, and VisVM itself is initialized from the parameters of LLaVA-Next and Qwen2-VL. The SFT data is produced by VisVM-guided search using base VLMs, ensuring that all training signals originated solely from the same VLM. 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.

4.3. VisVM Analysis

To further understand how VisVM enhances response quality by predicting future values, we design a quantitative experiment in this section to compare the effects on image captioning when selecting step candidates using

VisVM versus CLIP. We follow the experimental settings described in Section 4.1, randomly sampling 500 images from the COCO Val2014 dataset and using prompts from the LLaVA-150k detailed description dataset. For each <image, prompt> pair, we employ the LLaVA-Next-7B model to generate six candidate sentences, including greedy decoding and five different temperature settings. Subsequently, we select one candidate sentence from these six candidates using the VisVM and CLIP models independently. We then utilize the LLaVA-Next-7B model to continue generating a complete image description via greedy decoding based on the selected candidate sentence. Finally, we evaluate hallucinations within generated descriptions using the CHAIR metric, with results shown in Table 5.

Despite the selection being made from the same set of sentence candidates, differences arise in the selected candidates due to VisVM’s ability to predict long-term value, resulting in fewer hallucinations in captions generated by greedy decoding. In VisVM-guided search, this predictive selection by VisVM is applied at each step, significantly minimizing the occurrence of hallucinations in the final response. We provide a more detailed case study in Appendix D to further illustrate this.

5. 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.

Acknowledgment

Wang, Xu, and Huang are supported by DARPA Transfer from Imprecise and Abstract Models to Autonomous Technologies (TIAMAT) 80321, DARPA HR001124S0029-AIQ-FP-019, DOD-AFOSR-Air Force Office of Scientific Research under award number FA9550-23-1-0048, National Science Foundation NSF-IIS-2147276 FAI, National Science Foundation NAIRR240045, National Science Foundation TRAILS Institute (2229885). Private support was provided by Peraton.

References

- [1] Learning to reason with llms. <https://openai.com/index/learning-to-reason-with-llms/>, 2024. 2
- [2] 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. 3
- [3] 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. 1, 3
- [4] 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. 2, 3
- [5] 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. 3
- [6] 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. 3
- [7] 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*, pages 3558–3568, 2021. 2
- [8] 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*, 2023. 2, 3
- [9] 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?, 2024. 7
- [10] 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. 3
- [11] Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. Mitigating hallucination in visual language models with visual supervision, 2023. 3
- [12] 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*, pages 24185–24198, 2024. 1, 3
- [13] 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. 2, 3
- [14] 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. 3
- [15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 3
- [16] 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. 2, 3
- [17] Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control, 2022. 3
- [18] Arian Hosseini, Kingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordani, and Rishabh Agarwal. V-star: Training verifiers for self-taught reasoners, 2024. 2, 3
- [19] 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. 2
- [20] 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. 2
- [21] Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhui Chen. Mantis: Interleaved multi-image instruction tuning, 2024. 1
- [22] 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. 3

- [23] 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*, pages 111–127. Springer, 2025. 2, 3
- [24] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding, 2023. 1
- [25] 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, 2024. 1
- [26] 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, 2024. 3
- [27] 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, 2024. 3
- [28] 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, 2024. 1
- [29] 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*, 2023. 2, 3
- [30] 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, 2023. 2, 3
- [31] 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*, 2023. 2, 3
- [32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 7
- [33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 3
- [34] 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. 7
- [35] 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. 7
- [36] OpenAI. Gpt-4v(ision) system card. 2023. 3
- [37] 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. 2, 3
- [38] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 6, 7
- [39] 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. 1
- [40] 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. 3
- [41] 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*, 2024. 2, 3
- [42] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. 2
- [43] 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. 3, 6, 7
- [44] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3:9–44, 1988. 2, 4
- [45] Google Gemini Team. Gemini: A family of highly capable multimodal models, 2023. 3
- [46] 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, 2023. 3
- [47] 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. 2, 3
- [48] 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. 7
- [49] 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. 2, 3
- [50] 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*, 2024. 3

- [51] 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*, 2022. 3
- [52] 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*, 2023. 3
- [53] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation, 2024. 6
- [54] 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, 2023. 3
- [55] 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, 2024. 1
- [56] 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, 2024. 2, 3
- [57] Weihang 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*, 2023. 3
- [58] 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*, 2023. 3
- [59] 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*, 2024. 3
- [60] 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*, 2024. 3
- [61] 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)*, pages 416–442. Association for Computational Linguistics, 2024. 2, 3
- [62] 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*, 2022. 3
- [63] 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. 3
- [64] Yuancheng Xu, Udari Madhushani Sehwag, Alec Koppel, Sicheng Zhu, Bang An, Furong Huang, and Sumittra Ganesh. Genarm: Reward guided generation with autoregressive reward model for test-time alignment. *arXiv preprint arXiv:2410.08193*, 2024. 3
- [65] 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. 1
- [66] 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. 2
- [67] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of llms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 2023. 3
- [68] 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. 3
- [69] 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. 7
- [70] 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. 2, 3
- [71] 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 Wenhua Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. 7
- [72] 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. 2, 3
- [73] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing llms through hallucination-aware direct preference optimization, 2023. 1
- [74] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large lan-

guage models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024. [3](#), [6](#), [1](#)

- [75] 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*, 2024. [3](#), [4](#), [6](#), [1](#)