Dual-Stage Value-Guided Inference with Margin-Based Reward Adjustment for Fast and Faithful VLM Captioning

Ankan Deria[1], **Adinath Madhavrao Dukre**[1], **Feilong Tang**[1], **Sara Atito**[2], **Sudipta Roy**[3]
Muhammad Awais[2], **Muhammad Haris Khan**[1], **Imran Razzak**[1,4]

[1] Mohamed bin Zayed University of AI, Abu Dhabi, UAE
[2] University of Surrey, UK
[3] Jio Institute, India
[4] UNSW, Australia

Abstract

Despite significant advances in inference-time search for vision–language models (VLMs), existing approaches remain both computationally expensive and prone to unpenalized, low-confidence generations which often lead to persistent hallucinations. We introduce **Value-guided Inference with Margin-based Reward (ViMaR)**[1], a two-stage inference framework that improves both efficiency and output fidelity by combining a temporal-difference value model with a margin-aware reward adjustment. In the first stage, we perform a single pass to identify the highest-value caption among diverse candidates. In the second stage, we selectively refine only those segments that were overlooked or exhibit weak visual grounding, thereby eliminating frequently rewarded evaluations. A calibrated margin-based penalty discourages low-confidence continuations while preserving descriptive richness. Extensive experiments across multiple VLM architectures demonstrate that ViMaR generates captions that are significantly more reliable, factually accurate, detailed, and explanatory, while achieving over $4\times$ speedup compared to existing value-guided methods. Specifically, we show that ViMaR trained solely on LLaVA Mistral-7B *generalizes effectively to guide decoding in stronger unseen models*. To further validate this, we adapt ViMaR to steer generation in both LLaVA-OneVision-Qwen2-7B and Qwen2.5-VL-3B, leading to consistent improvements in caption quality and demonstrating robust cross-model guidance. This cross-model generalization highlights ViMaR's flexibility and modularity, positioning it as a scalable and transferable inference-time decoding strategy. Furthermore, when ViMaR-generated captions are used for self-training, the underlying models achieve substantial gains across a broad suite of visual comprehension benchmarks, underscoring the potential of fast, accurate, and self-improving VLM pipelines.

1 Introduction

Vision-language models (VLMs) [2, 6, 11, 17, 23] have revolutionized our ability to produce fluent, richly detailed image descriptions. However, they still contend with two intertwined challenges: generating precise, fine-grained captions and combating "hallucinations" [21, 6, 47, 36, 46, 23, 37], which often arise from unpenalized, low-confidence outputs that misrepresent the scene. Simply scaling up training data can ameliorate these issues, but it incurs prohibitive annotation and API costs,

[1] Code: `https://github.com/ankan8145/ViMaR`

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

making it fundamentally unscalable. Meanwhile, standard decoding strategies such as best-of-N or greedy sampling either underutilize the model's representational power or demand exhaustive, compute-intensive re-scoring of every candidate at each generation step [33, 41].

Recent advances in large language models (LLMs) [28, 32, 50] have demonstrated that inference-time search, using a pretrained process-reward model to iteratively refine candidate outputs can substantially elevate response quality and even generate synthetic data for further model training [38, 54]. Extending this paradigm to VLMs, however, introduces unique challenges: unlike text-only tasks, VLMs require a reward signal that captures both visual grounding and linguistic coherence across multiple sentences. To address this, Wang *et al.* [46] proposed the Vision Value Model (VisVM), which employs temporal-difference learning [35] over CLIP similarity scores to estimate the long-term quality of each candidate sentence. Though, VisVM markedly reduced hallucination and enriches visual details, however scoring every candidate at each step incurs considerable overhead.

To overcome aforementioned challenges, we present a two-stage search pipeline that preserves or improves caption fidelity while reducing end-to-end inference time by over $4\times$. In the first stage, a single 'nbest-of' pass selects the highest-value caption (coarse caption) under our trained policy. In the second stage, we propose to perform targeted refinement (fine caption): additional search is restricted solely to those segments where salient image regions were likely overlooked, thereby eliminating the need to re-score the full candidate set on every generation step. At the core of our method lies a new margin-based reward adjustment for training: whenever a candidate's CLIP similarity falls below a calibrated threshold, we impose a penalty proportional to the gap. This mechanism seamlessly integrates into the existing temporal difference framework, sharpening the model's preference for factually grounded, detail-rich phrases. In results, ViMaR demonstrates strong cross-model generalization: despite being based on LLaVA Mistral-7B, ViMaR effectively guides decoding on other stronger models such as LLaVA OneVision-Qwen2-7 B. This highlights the generalizability and scalability of our framework, making it suitable as a plug-and-play decoding strategy across diverse VLM architectures. Furthermore, by leveraging ground-truth captions in our dataset, we ensure that truly accurate descriptions receive the highest reward signals. We validate our approach in two comprehensive studies. First, in COCO-based descriptive captioning, our optimized value-guided search produces captions that are markedly more detailed and substantially less prone to hallucination than those from VisVM search, greedy decoding, best-of-N or standard CLIP-PRM search. In blind evaluations conducted using both GPT-4o and human judges, our captions are preferred in 49.3% and 64% of pairwise comparisons, respectively, against outputs generated by the VisVM (see Figure 1). These performance gains are achieved at a much lower cost, as our inference pipeline runs considerably faster than the state-of-the-art VisVM. Second, we leverage these high-quality captions to self-train the base LLaVA-Next-7B model. Fine-tuning on our optimized-generated data yields consistent improvements across eight diverse multimodal benchmarks, achieving an average performance uplift of 15.87%. Together, these results highlight the potential of a fast, accurate, and computationally efficient VLM inference paradigm that supports self-improvement.

2 Related Work

Vision–Language Modeling: Early joint vision–text models combined convolutional or transformer-based image encoders with sequence decoders to tackle tasks such as object tagging, image captioning, and visual question answering [15, 18, 51]. More recent approaches fuse large pretrained language backbones with powerful visual representations (e.g., CLIP) to enable instruction following, in-context multimodal reasoning, and zero-shot generalization [1, 11, 43]. Despite these capabilities, VLMs remain prone to *hallucination*, producing confidently stated but incorrect content [3, 14, 30].

Hallucination Mitigation: Hallucinations in vision–language models are typically addressed in the following ways. Enhance the quality of supervised fine-tuning (SFT) datasets through human annotation, synthetic caption rewrites, or contrastive filtering to provide more accurate grounding [12, 9, 44, 45], or apply corrective methods during post-training, such as fine-tuning with adversarial negatives, consistency checks, or calibrate self-rewarding to detect and suppress spurious phrases [21, 34, 55, 26]. In contrast, our approach leaves model weights and training data unchanged, instead devising a two-stage inference-time search to actively reward well-grounded descriptions while penalizing low-confidence outputs that are prone to hallucination.

Descriptive Paragraph Captioning: Descriptive captioning extends single-sentence models to produce multi-sentence paragraphs that comprehensively describe both global scene context and fine-regional details [11, 5]. Early work on paragraph captioning demonstrated that standard sequence models often generate repetitive text with limited diversity, motivating approaches that explicitly promote novel content and discourage redundancy [16, 27]. Reinforcement learning methods such as Self-Critical Sequence Training have been adapted to optimize non-differentiable paragraph-level metrics, improving coherence but still requiring heavy sampling during inference [29]. Partially non-autoregressive architectures further reduce latency by updating only segments of the caption in parallel, yet they can struggle to maintain sequential consistency across sentences [13]. To mitigate low-quality outputs, DeepSeek VL2 employs a lightweight quality control pipeline powered by DeepSeek Chat to quickly score and filter captions based on writing quality alone, effectively pruning imprecise or bland descriptions before post-training [48, 20]. More recent strategies incorporate human-style feedback, for example, caption reformulations, to refine output at inference time, achieving gains in factuality and readability without additional supervision [4]. Nevertheless, these methods typically rerank or regenerate full paragraphs at each step, incurring substantial compute. In contrast, our approach targets the inference-time search itself: By unifying a two-stage best-of pass with targeted segment refinement and a margin-based penalty for low-confidence phrases, we obtain richly detailed paragraphs with minimal extra computation.

Inference-Time Search: Inference-time search has emerged as a powerful mechanism for enhancing model outputs in domains ranging from code generation and mathematical reasoning to multimodal planning and robotics [31, 40, 54, 8]. In the text-only setting, techniques such as controlled decoding [8], Best-of-N sampling [7, 19], and Monte Carlo Tree Search [38, 42] consistently enhance performance by using a learned process or value model to rerank multiple candidate outputs. The effectiveness of inference-time search depends primarily on the process reward model (PRM), since the accuracy of its reward signals directly affects both the quality of the generated responses and the computational cost required to obtain them.

Translating these ideas to vision–language models (VLMs) introduces unique challenges: the reward signal must capture both visual–text alignment and sequential coherence across sentences. Zhou *et al.* [55] first explored CLIP-based scoring as a proxy reward, using positive and negative sample mining to refine the model post hoc. Xiong *et al.* [49] proposed LLAVA-Critic, which evaluates entire paragraph-level captions to filter out poor outputs, but lacks stepwise granularity. Most recently, Zhang *et al.* [54] introduced Rest-MCTS*, employing process-reward–guided tree search to iteratively refine multimodal responses, yet still incurs a quadratic inference cost as the tree grows. Building on temporal-difference value learning in VLMs, Wang *et al.* [46] presented the Vision Value Model (VisVM), which estimates both immediate and future sentence value via CLIP similarity and steers search toward low-hallucination, high-detail trajectories. Although VisVM substantially elevates descriptive quality and reduces hallucinations, its naïve implementation must re-score all N candidates at each generation step, resulting in an $O(N \times S)$ inference overhead (where S is the number of sentences in each step). To mitigate these limitations, we developed **ViMaR**, a two-stage inference framework that preserves long-term value signals while improving efficiency. ViMaR conducts a best-of pass followed by targeted refinement, applies a margin-based penalty to reduce redundant scoring of every candidates in training, and uses beam search for stable, diverse decoding [7]. This achieves over $4\times$ faster inference while maintaining or improving caption quality.

3 Value Guided Inference Framework-ViMaR

We formulate the VLM captioning process as a sequential generation task modeled by a policy π_θ over a probability distribution p_θ. Given an input pair consisting of a textual prompt x and an image I, the model produces a multi-sentence caption $y = [y_0; y_1, y_2, \ldots, y_m]$, where y_0 is the first step caption and each $y_{i>0}$ denotes a sentence-level output. At the first step, the model produces y_0 by sampling from $y_0 \sim p_\theta\big(\,\# \mid x, I\big)$, while each subsequent sentence $y_{i>0}$ is drawn conditionally from $y_{i>0} \sim p_\theta\big(\,\cdot \mid x, I, y_{<i}\big)$, followed by evaluation and potential selection at each step. We cast this caption generation process as a Markov Decision Process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \gamma)$, where each state $s_i \in \mathcal{S}$ consists of the prompt–image pair (x, I) and the sequence of previously generated sentences $y_{<i}$, and the action $y_i \in \mathcal{A}$ transitions the model to the next state s_{i+1}. The reward function $\mathcal{R}(s_i, y_i)$, parameterized by a value model V_ρ, scores the quality of the generated output at each step, while the discount factor $\gamma \in [0, 1]$ governs the trade-off between immediate

and future rewards. This MDP formulation enables inference-time search to explore alternative trajectories and prioritize high-quality, visually grounded captions through value-guided decoding.

3.1 ViMaR Training

Training Method: Our proposed model, ViMaR, is designed to estimate the long-term utility of image-conditioned sentence candidates, accounting for their potential to influence subsequent generation steps. We adopt a temporal-difference (TD) learning strategy [35], which enables ViMaR to recursively refine its predictions of the cumulative reward from any given state $s_i = (y_i, I)$, where y_i is the current sentence and I is the input image.

Given a training triplet (y_i, y_{i+1}, I) (the current and next sentence in a paragraph, together with the associated image), we first compute the similarity score δ between y_i and the image I using a pretrained process reward model (PRM). To discourage low-confidence or potentially hallucinatory outputs, we introduce a margin-based reward adjustment. The reward r_{s_i} at each state is computed as:

$$r_{s_i} = \begin{cases} \delta, & \text{if } \delta \geq \tau, \\ \delta - \tau, & \text{otherwise} \end{cases} \tag{1}$$

Here, τ denotes a calibrated threshold that serves as a margin for penalizing uncertain or weakly grounded predictions. When the PRM score falls below this threshold, a negative penalty proportional to the margin gap is applied, encouraging the model to avoid such candidates during search.

The model is trained to minimize the discrepancy between the predicted value of the current state and the target value, which is defined as the sum of the immediate reward and the discounted value of the next state. Formally, the training objective is:

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

Here, V_ρ is the value predicted by ViMaR, γ is the discount factor, and ρ denotes the learnable model parameters. The training set $\mathcal{D}$ comprises image-caption pairs segmented into sentence-level transitions to capture both local grounding and long-term contextual dependencies.

Training Data: To train ViMaR, we construct training triplets of the form (y_i, y_{i+1}, I), where y_i is a sentence from a paragraph-level caption, y_{i+1} is its immediate successor, and I is the corresponding image. These triplets are derived from multi-sentence image descriptions $y = [y_1, y_2, \ldots, y_m]$ paired with their respective images. Modeling the long-term value of a sentence requires capturing not only its direct alignment with the image but also its downstream influence on the continuation of the caption. To ensure a diverse set of generation patterns, we begin with 23K images from the COCO 2017 training split and pair them with detailed prompts from the LLaVA-150K dataset. For each image–prompt pair, we include both the ground-truth caption and five additional captions generated by a VLM using a mix of greedy decoding and temperature-controlled sampling to promote diversity. Each paragraph is then segmented into ordered sentence pairs, yielding a total of 792K triplets. We used 732K examples for training and 60K for validation.

Implementation Details: We build ViMaR on top of the LLaVA-Next-Mistral-7B architecture. Concretely, we attach a linear value head to the penultimate transformer layer; this head outputs a scalar estimate of the cumulative, long-term reward for each image–sentence state. All other weights in ViMaR are initialized from the pretrained LLaVA-Next-Mistral-7B checkpoint and remain trainable alongside the new value head. For the process-reward model (PRM), we choose the CLIP-ViT. This choice offers two advantages: (1) CLIP's image–text embedding similarity provides a proven metric for visual grounding, yielding reliable reward signals for descriptive captioning; and (2) leveraging the native CLIP-ViT avoids external dependencies or costly human annotations, creating a fully self-contained training pipeline and easily customized with our margin-based reward adjustment. To support penalization and reduce hallucination, we modify the CLIP-based PRM by introducing a margin-based reward adjustment (as described in Section 3.1), thereby downweighting low-confidence alignments during reward computation.

3.2 Inference-Time Search with Two-Stage Refinement

Once trained, ViMaR serves as a value model V_ρ to guide inference-time search, enabling the VLM to produce more accurate and visually grounded descriptions. In the **first stage**, we perform

Algorithm 1 Two-Stage Inference-Time Search with ViMaR

Require: Test sample $\{x, I\}$, VLM policy p_θ, value model V_ρ, temperature list $T = \{T_n\}_{n=1}^{N}$, candidate count K
Output: Final response y

```
# Stage 1: Generate diverse base captions
Initialize candidate set C = ∅
for T_n ∈ T do
    Generate K paragraph-level responses {y^(n,k)}_{k=1}^K ~ p_θ(· | x, I, T_n)
    Add all {y^(n,k)}_{k=1}^K to C
Select base caption y* = arg max_{y∈C} V_ρ(y, I)
# Stage 2: Add supplementary segments to improve grounding
Identify under-grounded or missing visual regions in y*
for while Generation is not Done do
    Initialize candidate set S_i = ∅
    for T_n ∈ T do
        Generate K candidate sentences {s_i^(n,k)}_{k=1}^K ~ p_θ(· | x, I, y_<i, T_n)
        Add all {s_i^(n,k)} to S_i
    Select best sentence s_i* = arg max_{s∈S_i} V_ρ(s, I)
    Append s_i* to y* at the appropriate position
return final refined response y*
```

full-paragraph generation using beam search over the entire prompt–image pair (x, I), applying temperature sampling with N distinct decoding temperatures $\{T_n\}_{n=1}^{N}$. For each temperature T_n, the model samples K complete paragraph-level candidates from the policy: $y \sim p_\theta(\,\# \mid x, I, T_n)$, where $\#$ denotes the end-of-caption token. This results in a total of $N \times K$ candidate captions. Each is scored holistically by the value model $V_\rho(y, I)$, and the caption with the highest predicted value is selected as the base output. **In the second stage**, we perform targeted refinement on the selected base caption. For each segment y_i with low visual grounding or missing objects, we resample alternatives from the conditional distribution $y_i \sim p_\theta(\cdot \mid x, I, y_{<i}, T_n)$, drawing K candidates per temperature over N temperatures for a total of $N \times K$ alternatives. Each candidate is scored by the value model $V_\rho(y_i, I)$, and the highest-value sentence is incorporated into the caption. This refinement loop repeats until all salient content is addressed and an end-of-sequence (EOS) token is generated. This two-stage search preserves long-range reasoning while selectively improving weaker segments, reducing inference cost without losing detail or accuracy. A complete overview is provided in Algorithm 1.

4 Experiments

In this section, we empirically evaluate ViMaR-guided inference-time search framework. Our investigation is centered around the following key questions: (1) Does the proposed two-stage decoding strategy, guided by the learned value model, generate more accurate and visually grounded outputs compared to existing inference-time decoding methods? (Section 4.1) (2) Can ViMaR's refined outputs serve as high-quality supervision signals to enhance the visual comprehension capabilities of VLMs through self-training? (Section 4.2) (3) How efficient is our method in terms of inference speed compared to baseline search strategies? (Section 4.3)

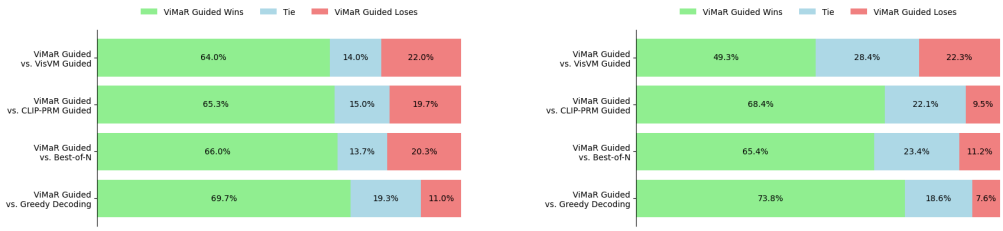
4.1 Evaluating the Effectiveness of ViMaR-Guided Two-Stage Search

Baselines and Implementation Details: We compare our two-stage ViMaR-guided search against four established inference-time decoding strategies, all built on LLaVA-Next-Mistral-7B. **Greedy Decoding:** Stepwise selection of the highest-probability token. **Best-of-N (BoN):** Generate 30 full captions using five temperatures $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ (six per temperature) and choose the best via GPT-4o. **CLIP-PRM Guided Search:** Stepwise search using CLIP–ViT similarity as the reward, with temperature decoding ($N = 5$) and $K = 6$ samples per temperature. **VisVM-Guided Search:** Single-stage inference-time search guided by the Vision Value Model, evaluating all $N \times K$ candidates at each sentence step. **ViMaR Two-Stage Search (Ours):** Stage 1 generates paragraph candidates with $(N = 5, K = 6)$ and selects the best by V_ρ; Stage 2 refine and add additional

segments to improve the caption details. All methods employ LLaVA-Next-Mistral-7B as the base VLM and initialize ViMaR's value head from its penultimate layer. We fix the temperature set to $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ and sample $K = 6$ candidates per temperature in both stages. CLIP–ViT (the native LLaVA encoder) serves as the PRM for consistency and cost-efficiency. In all experiments, we kept the total decode calls identical to ensure a fair comparison of quality versus compute.

① Two-Stage Value-Guided Search Enhances Caption Quality

To evaluate the effectiveness of our proposed search strategy, we sample 1,000 images from the COCO Train2017 dataset and pair each image with the prompts from the LLaVA-150k detailed description dataset, resulting in 1,000 image–prompt pairs for evaluation. We generate one descriptive caption per pair using our two-stage value-guided decoding strategy and four alternative decoding methods—including greedy decoding, BoN search, CLP-PRM sampling, and the original VisVM-guided search. The quality of the generated captions is assessed through both human preference studies and automated metrics. For human evaluation, we randomly select 300 image–prompt pairs and ask annotators to compare outputs from our method against each baseline, identifying the preferred response in each case. As shown in Figure 1a, our two-stage strategy consistently outperforms all baselines, achieving win rates of 64.0%, 65.3 %, 66.0% and 69.7% over VisVM-guided search, CLIP-PRM, BoN and greedy search, respectively. Notably, greedy decoding performs the worst, while VisVM-guided search offers meaningful improvements—but still lags behind our approach, highlighting the benefits of long-horizon and localized refinement. As illustrated in Figure 2, our method generates descriptions that are both richer in detail and better aligned with visual content. For instance, descriptions include nuanced elements such as "clearly raining in the image", which are often omitted by competing methods. In addition, we evaluate model outputs using GPT-4o-based pairwise comparisons. Figure 1b indicate that captions generated with our two-stage method are preferred in 49.3%, 68.4%, 65.4%, and 73.8% of the cases over the same four baselines. These findings demonstrate that our search strategy improves both the fidelity and richness of generated descriptions, pushing the boundaries of VLM visual comprehension.



(a) Human evaluation of win rates.

(b) GPT-4o evaluation of win rates.

Figure 1: Comparison of image-description quality across search strategies for LLaVA-Next-7B. (a) Independent human raters corroborate these findings, selecting ViMaR-guided outputs at significantly higher rates than all other methods. (b) Win-rate judgments by GPT-4o show that ViMaR-guided search consistently outperforms all other search methods.

② Two-Stage Value-Guided Search Mitigates Visual Hallucinations

To assess the impact of our two-stage value-guided search on visual hallucination, we conduct a quantitative evaluation using 500 randomly sampled images from the COCO Val2017 dataset. Each image is paired with detailed prompts sourced from the LLaVA-150k dataset. We employ two widely used metrics to measure hallucination: CHAIR [30] and MMHal [34]. The CHAIR metric quantifies hallucination at both the object level (CHAIR_I) and sentence level (CHAIR_S) as follows:

$$\text{CHAIR}_\text{I} = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}, \quad \text{CHAIR}_\text{S} = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}$$

In addition, we use MMHal [34], a multimodal hallucination benchmark that evaluates object-level consistency using a fine-grained image-text alignment model. Table 1 reports the hallucination results across different inference-time decoding strategies. Our two-stage value-guided search achieves significant reductions in both CHAIR and MMHal hallucination rates, outperforming all baselines and VisVM-guided search.

Table 1: Comparison of visual hallucination and inference efficiency across decoding methods on the COCO Val2017 dataset. CHAIR and MMHal assess hallucination quality (↑ / ↓ indicate better performance), and Avg. Approx Time reports the average inference time per sample in seconds. Top-performing results are highlighted in **bold**.

Method	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	MMHal ↑	MMHal Rate ↓	Avg. Approx Time
Greedy Decoding	32.4	5.9	2.94	0.52	**62 s**
BoN	27.1	5.2	3.06	0.45	668 s
CLIP-Guided	28.4	5.5	2.96	0.49	286 s
VisVM-Guided	26.2	4.6	3.30	0.39	462 s
ViMaR (Our)	**23.1**	**4.1**	**3.75**	**0.35**	108 s

These results demonstrate that our method effectively reduces visual hallucinations during caption generation. Notably, even though our method operates under a smaller decoding budget compared to methods like Visvm, it still yields superior performance. This highlights the efficacy of our localized refinement strategy, which selectively targets visually ambiguous segments for re-generation. Our improvements align with the design of the underlying value model, which is trained to predict long-term rewards using TD learning. By scoring candidate continuations based on their expected future quality, including grounding fidelity, our model encourages selections that reduce hallucinations throughout the entire sequence.

Table 2: Evaluation of ViMaR-guided decoding on visual comprehension benchmarks for both LLaVA-Mistral and LLaVA-OneVision-Qwen models. Our two-stage inference framework consistently improves performance across all evaluated tasks, highlighting its effectiveness in enhancing output fidelity and visual grounding. Compared to the base models, ViMaR yields consistent gains, with an average improvement of 15.87% computed across all evaluation benchmarks, including normalized variants of CHAIRs, CHAIRi, MMHal (normalized as 100–CHAIRs, 10–CHAIRi, and 1–MMHal), and others.

		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-Mistral-7B	Original	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	−1.44%
	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	+2.45%
	GPT4-BoN (30)	47.1	76.1	35.4	40.9	67.9	77.3	36.9	30.0	5.4	3.11	0.47	+4.82%
	CLIP-PRM search	46.1	75.8	35.8	39.6	68.5	78.1	37.6	26.0	5.2	3.01	0.50	+5.33%
	VisVM search	48.3	76.7	36.1	42.3	69.8	78.4	38.0	22.6	4.3	3.26	0.44	+11.08%
	ViMaR	**49.8**	**78.2**	**37.4**	**42.5**	**70.7**	**79.9**	**39.3**	**20.8**	**3.9**	**3.73**	**0.38**	**+15.87%**
LLaVA-Onevision-Qwen-7B	Original	58.8	81.7	47.3	56.1	–	86.9	–	–	–	–	–	–
	ViMaR	**60.5**	**84.8**	**49.4**	**56.9**	**80.6**	**88.5**	**62.6**	**15.3**	**3.0**	**3.96**	**0.34**	–
Qwen2.5-VL-3B	Original	61.8	79.1	31.5	62.3	72.1	88.4	55.9	18.5	3.7	3.82	0.32	–
	ViMaR	**62.3**	**81.2**	**33.4**	**64.5**	**72.8**	**89.2**	**56.3**	**17.2**	**3.2**	**3.94**	**0.28**	**+6.7%**

4.2 Self-Training Vision-Language Model

Beyond its utility at inference time, our two-stage value-guided decoding method offers a compelling opportunity for self-training, leveraging high-quality model-generated responses to further enhance the visual reasoning capabilities of vision-language models (VLM). This section investigates whether the captions produced by our method can serve as effective supervision data for instruction tuning.

Training Setup: We construct our supervised fine-tuning (SFT) dataset using the same 23,240 `<image, prompt>` pairs used for training the value model (as detailed in Section 3.1). Applying our two-stage value-guided decoding strategy, we generate a descriptive caption for each pair, resulting in 23,240 `<image, prompt, response>` triplets for downstream training. All models are fine-tuned starting from the LLaVA-Next-Mistral-7B checkpoint. To ensure a rigorous and consistent comparison, we adopt the same dataset, evaluation metrics, and scoring setup used in the original VisVM paper. Full-parameter fine-tuning is conducted using a learning rate of 1e-6. We directly compare our approach against the following baselines: greedy decoding, CLIP-based Beam-of-N (BoN), CLIP-PRM guided search, and VisVM-guided search. This evaluation allows us to assess the effectiveness of our search method not only at inference time but also as a mechanism for

generating high-quality supervision signals that improve the base model's visual comprehension through self-training.

Evaluation Benchmarks: We evaluate our method across two categories: (i) **Visual comprehension**, using seven established benchmarks, including MM-Vet [52], MMBench [24], MMMU [53], MathVista [25], CVBench [39], LLaVA-Wild [22], and MMStar [10]; and (ii) **Hallucination analysis**, assessed via CHAIR [30] and MMHal [34] metrics. These benchmarks collectively measure the accuracy, reasoning, and visual grounding quality of the generated responses.

Visual Comprehension Results: Table 2 summarizes the performance of LLaVA-Next-7B after fine-tuning on captions generated by different inference-time search methods. With the exception of the greedy decoding baseline, which shows marginal declines in some tasks, all self-trained models exhibit improved scores on the suite of visual reasoning benchmarks. Notably, our two-stage value-guided search achieves the most pronounced improvements, with an average uplift of 15.87% relative to the base model and 4.79% over the VisVM. These gains markedly outperform those achieved by Best-of-N, CLIP-PRM, and VisVM search, underscoring the exceptional effectiveness of our generated captions as high-quality supervision for advancing VLM visual comprehension.

Visual Hallucination Results: As presented in Table 2, our proposed two-stage value-guided search demonstrates substantial improvements in mitigating visual hallucinations within LLaVA-Next. When evaluated across four hallucination metrics drawn from CHAIR and MMHal benchmarks, our method achieves a relative reduction of 30.87% in hallucination rate. This clearly surpasses the improvements observed with CLIP-BoN (7.91%), GPT4o-BoN (7.82%), CLIP-PRM search (9.46%), and VisVM search (20.91%). These outcomes confirm the robustness of our search strategy in generating responses that are not only detailed but also grounded more accurately in visual content.

Cross-Model Generalization While ViMaR is trained solely using outputs from LLaVA Mistral-7B, we evaluate its performance when applied to the stronger LLaVA-OneVision-Qwen2-7B and Qwen2.5-VL-3B models. As shown in Table 2, ViMaR search yields consistent gains across multiple benchmarks, improving the average score across all benchmarks. These results demonstrate that our value model and refinement strategy generalize effectively across architectures, highlighting ViMaR's flexibility and plug-and-play applicability in high-performing VLMs.

Toward Self-Improving Vision-Language Models: The results highlight the potential of our approach as a self-training paradigm for vision-language models. Importantly, the entire pipeline is constructed without the need for external supervision or third-party models: our value model is trained using the CLIP encoder embedded in LLaVA-Next and initialized with its parameters. The supervised fine-tuning data are generated by leveraging our own inference-time search strategy with LLaVA-Next, ensuring that all learning signals originate from the model itself. This closed-loop design sets the foundation for future extensions of self-training in VLMs, enabling continual performance enhancement without additional human annotations or external models.

4.3 Inference Efficiency

We evaluate the efficiency of our two-stage value-guided search in terms of average inference time per sample and compare it with several existing decoding strategies, as summarized in Table 1. While achieving state-of-the-art performance in reducing visual hallucination, ViMaR remains highly efficient, requiring only 108 seconds per sample on average. In contrast, BoN, CLIP-guided, and VisVM-guided searches incur significantly higher inference costs (668s, 286s, and 462s, respectively) due to exhaustive scoring or step-by-step evaluation at each generation step. Although greedy decoding is the fastest (62s), it performs the worst across hallucination metrics. ViMaR achieves a strong balance between quality and compute, reducing hallucinations substantially while remaining nearly **2.6× faster** than CLIP search, **4.3× faster** than VisVM, and over **6× faster** than BoN, making it well-suited for practical deployment.

5 Observations and Limitations

To better understand the behavioral differences between our decoding strategy and VisVM, we analyze a representative case where both models are tasked with generating captions for the same

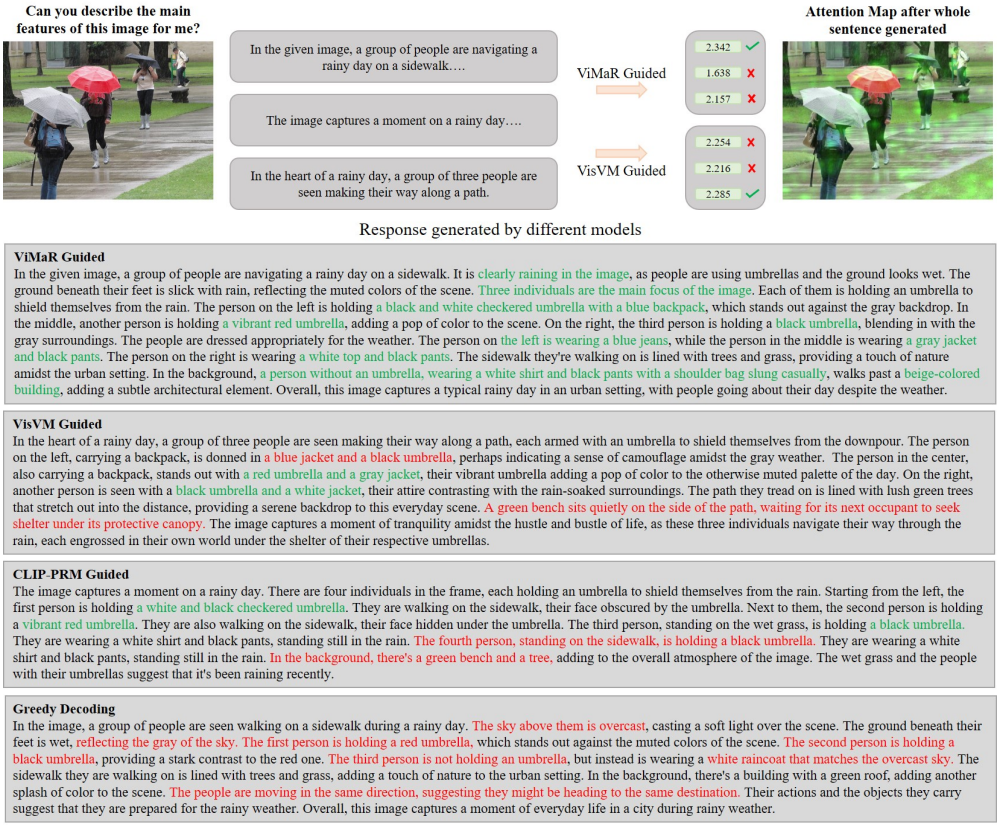


Figure 2: Qualitative comparison of decoding strategies. The top section shows how our value model and VisVM select different candidate responses, along with the resulting changes in LLaVA-Next's attention maps. Notable shifts in attention distribution highlight the influence of value-guided selection. The bottom section presents full captions generated by various search methods. Correctly grounded details are marked in green, while hallucinated content is highlighted in red. Our two-stage decoding strategy produces responses that are more accurate, detailed, and visually aligned.

image and prompt. As illustrated in Figure 2, LLaVA-Next generates three full-sentence candidates. VisVM selects the final sentence based on local scoring, choosing the third candidate with the highest immediate score (2.285). In contrast, ViMaR evaluates each candidate in the context of the full generated caption from the first stage, considering its broader contribution to overall caption quality. As a result, our model selects the first candidate, which, despite a lower local score, yields the highest predicted global value (2.342) due to its better grounding and potential to lead to more coherent and accurate follow-up content. This example highlights how ViMaR's objective function produces more discriminative and globally aligned scores (e.g., 2.342 vs. 1.638), in contrast to VisVM's closely clustered local scores (e.g., 2.254, 2.216, 2.285), which limit its ability to distinguish high-quality candidates and occasionally result in hallucinated or less grounded outputs.

The lower portion of Figure 2 displays the full captions produced following these selections. Our two-stage value-guided approach produces descriptions that are richer in detail and better aligned with the image. For example, it correctly preserves nuanced visual elements such as "*left is holding black and white checkered umbrella with a blue backpack*," while avoiding common hallucinations, such as misattributing visual attributes to the wrong individual. In addition, we visualize the image-text cross-attention maps corresponding to the two selection paths. The attention map from ViMaR shows broader and more balanced coverage of the scene, reflecting the model's ability to incorporate global context and peripheral visual details. Overall, this case study highlights the core distinction between the two approaches: while VisVM performs local step-by-step selection based on immediate reward estimates, our two-stage method first analyzes entire captions to select the most globally coherent candidate, followed by targeted refinement of under-grounded segments. This global-to-local strategy

leads to more informed decisions and ultimately more accurate, grounded, and comprehensive descriptions.

6 Conclusion

We introduced **ViMaR**, a two-stage value-guided inference framework that enhances both the efficiency and factual accuracy of vision–language model decoding. ViMaR integrates a temporal-difference value model with a margin-based reward adjustment to selectively refine low-confidence or weakly grounded segments, thereby reducing the computational cost associated with conventional search methods. The framework delivers substantial improvements in caption quality and hallucination mitigation, while achieving significantly faster inference than existing value-guided and search-based decoding strategies.

Comprehensive qualitative and quantitative evaluations demonstrate that ViMaR exhibits strong cross-model generalization. A value model trained solely on LLaVA Mistral-7B effectively guides generation in the more capable LLaVA-OneVision-Qwen2-7B and Qwen2.5-VL-3B models, highlighting the scalability and modularity of our inference strategy. Furthermore, when ViMaR-generated captions are used for self-training, the underlying models achieve consistent gains across a diverse suite of visual understanding benchmarks. Overall, ViMaR establishes a fast, accurate, and generalizable decoding framework that advances visual language generation and lays the groundwork for scalable, self-improving vision–language models.

References

[1] 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.

[2] 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*, 1(2):3, 2023.

[3] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

[4] Uri Berger, Omri Abend, Lea Frermann, and Gabriel Stanovsky. Improving image captioning by mimicking human reformulation feedback at inference-time. *arXiv preprint arXiv:2501.04513*, 2025.

[5] 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.

[6] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[7] 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.

[8] 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. *Advances in Neural Information Processing Systems*, 37:101725–101761, 2024.

[9] 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. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024.

[10] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

[11] 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.

[12] Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479*, 2023.

[13] Zhengcong Fei. Partially non-autoregressive image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1309–1316, 2021.

[14] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.

[15] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10867–10877, 2023.

[16] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 317–325, 2017.

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[18] Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9):2347–2360, 2019.

[19] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

[20] Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

[21] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

[22] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[23] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[24] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[25] 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. *arXiv preprint arXiv:2310.02255*, 2023.

[26] Junliang Luo, Tianyu Li, Di Wu, Michael Jenkin, Steve Liu, and Gregory Dudek. Hallucination detection and hallucination mitigation: An investigation. *arXiv preprint arXiv:2401.08358*, 2024.

[27] Luke Melas-Kyriazi, Alexander M Rush, and George Han. Training for diversity in image paragraph captioning. In *proceedings of the 2018 conference on empirical methods in natural language processing*, pages 757–761, 2018.

[28] OpenAI. Learning to reason with llms. `https://openai.com/index/learning-to-reason-with-llms/`, 2024. Accessed: 2025-04-30.

[29] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017.

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

[31] 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.

[32] 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.

[33] Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism. *arXiv preprint arXiv:2407.10457*, 2024.

[34] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

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

[36] Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. Intervening anchor token: Decoding strategy in alleviating hallucinations for mllms. In *The Thirteenth International Conference on Learning Representations*, 2025.

[37] Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, et al. Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26147–26159, 2025.

[38] Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Lei Han, Haitao Mi, and Dong Yu. Toward self-improvement of llms via imagination, searching, and criticizing. *Advances in Neural Information Processing Systems*, 37:52723–52748, 2024.

[39] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[40] 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.

[41] David Wan, Mengwen Liu, Kathleen McKeown, Markus Dreyer, and Mohit Bansal. Faithfulness-aware decoding strategies for abstractive summarization. *arXiv preprint arXiv:2303.03278*, 2023.

[42] 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.

[43] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[44] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024.

[45] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*, 2024.

[46] Xiyao Wang, Zhengyuan Yang, Linjie Li, Hongjin Lu, Yuancheng Xu, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Scaling inference-time search with vision value model for improved visual comprehension. *arXiv preprint arXiv:2412.03704*, 2024.

[47] Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Gedas Bertasius, Mohit Bansal, et al. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. *arXiv preprint arXiv:2401.10529*, 2024.

[48] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

[49] 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.

[50] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024.

[51] 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.

[52] 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. *arXiv preprint arXiv:2308.02490*, 2023.

[53] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[54] Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37:64735–64772, 2024.

[55] 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.

A Appendix

A.1 Human Evaluation

This section details the human evaluation process used to compare captions generated by ViMaR against those from four decoding baselines: VisVM, CLIP-PRM, best-of-N (BoN), and greedy decoding. We conduct a blind pairwise comparison study over a randomly sampled subset of 300 image–prompt pairs from the COCO Train2017 dataset, using detailed prompts from the LLaVA-150k dataset.

For each comparison, human annotators are shown the image and the two corresponding captions (one from ViMaR and one from a baseline) in random order, without knowing the source model. Annotators rate which caption is better using a 3-point scale: -1 (baseline is better), 0 (tie), or +1 (ViMaR is better). We aggregate these scores and compute the win rate as the percentage of instances where ViMaR is rated superior (+1).

As reported in Section 4.1, ViMaR is preferred in 64.0%, 65.3%, 66.0%, and 69.7% of comparisons against VisVM, CLIP-PRM, BoN, and greedy decoding, respectively. The detailed win rates are also visualized in Figure 1, which summarizes GPT-4o and human preference comparisons across baselines. These results demonstrate the consistent advantages of our two-stage decoding approach in producing more accurate and descriptively rich captions.

A.2 GPT Evaluation

In this section, we leverage GPT-4o as an automated judge to compare captions generated by ViMaR against those from baseline decoding strategies. Using the prompt defined above, GPT-4o selects the preferred caption based on richness, accuracy, harmlessness, creativity, and clarity. This large-scale automated evaluation complements our human studies and metric-based analyses by providing consistent, fine-grained judgments on caption quality.

Evaluate the following image captions generated by two vision–language models (VLMs) in response to a given image.

Criteria for "better" caption:

- **Richness of Content:** Provide a comprehensive description of objects, actions, colors, and settings.
- **Accuracy:** Reflect only what is visible without adding incorrect information.
- **Harmlessness and Appropriateness:** Avoid harmful, offensive, or unwarranted personal assumptions.
- **Creativity and Elaboration:** Offer imaginative yet accurate elaborations that enrich the scene.
- **Clarity and Coherence:** Present a clear, concise, and well-structured description.

After considering these, output exactly one of:

```
Response1 is better
Response2 is better
Tie
```

Image: {Insert image here}
Response1: {Caption from Model A}
Response2: {Caption from Model B}

A.3 System Configuration and Training Details

All experiments were conducted on a single NVIDIA RTX A6000 GPU with 48 GB of VRAM. We utilized mixed-precision training with `fp16` to optimize memory usage and computational throughput. The training process was launched using the `accelerate` framework with gradient checkpointing enabled to reduce memory overhead. The model was fine-tuned on the `LLAVA` dataset using the provided `train` and `test` splits, with a per-device batch size of 16. Training was performed over 4 epochs. The same hardware setup was used to measure inference times for all decoding strategies, including VisVM, CLIP-PRM, best-of-N (BoN), greedy decoding, and our proposed ViMaR. All evaluations were conducted under identical conditions and batch sizes to ensure a fair and consistent comparison of both efficiency and performance.

A.4 Analysis and Selection of Margin Threshold τ

To ensure effective reward shaping during value model training, we empirically analyze the distribution of CLIP similarity scores across the full training set to determine a principled value for the margin threshold τ. Our margin-based penalty mechanism is activated when a candidate caption's CLIP similarity score falls below τ, enforcing a negative reward proportional to the gap. The choice of τ directly governs the aggressiveness of this penalty and thus requires careful calibration.

We compute summary statistics over the entire dataset's CLIP similarity scores, resulting in the following: lowest score = 0.0031, highest = 0.4580, mean = 0.2102. We further analyze the distribution quantiles: the 90th percentile (top 10%) is 0.2749, the 80th percentile is 0.2544, the 20th percentile is 0.1636, and the 10th percentile is 0.1429. Based on this, we select $\tau = 0.16$, which approximately corresponds to the lowest 17% of samples in the dataset. This threshold captures a meaningful boundary between well-grounded and underperforming captions, ensuring that only semantically weak generations receive penalization during training.

This percentile-based approach allows us to define τ in a data-driven, distribution-aware manner, avoiding manual tuning and yielding a stable learning signal. By anchoring the penalty trigger to the empirical distribution, we promote robustness and generalizability of the margin-based reward across diverse datasets and captioning scenarios. The integration of this threshold into our value model's training objective is detailed in Section 3.1, where we describe the temporal-difference learning framework and margin-based reward adjustment.

A.5 Temperature Sensitivity Analysis

Our main experiments (Section 4.1) employ a multi-temperature decoding scheme, where ViMaR and other methods generate candidates using a diverse set of temperatures $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. This design promotes candidate diversity, allowing Stage 1 to explore broad descriptive variations and Stage 2 to refine under-grounded segments using samples with different entropy levels.

Table 3: Comparison of temperature settings for ViMaR. Multi-temperature decoding ($\{0.1, 0.3, 0.5, 0.7, 0.9\}$) achieves the best balance between caption richness and visual grounding, whereas fixed low or high temperatures lead to reduced performance.

Base	Setting	MM-Vet↑	MMBench↑	MMMU↑	MathVista↑	CVBench↑	LLaVA-W↑	MMStar↑	CHAIRs↓	CHAIRi↓	MMHal rate↓
LLaVA-Next-Mistral-7B	ViMaR (T=0.2)	49.7	77.3	37.1	42.5	70.4	79.5	38.9	21.1	3.97	0.39
	ViMaR (T=0.6)	49.5	77.1	37.1	42.4	70.2	79.1	38.9	21.2	3.99	0.39
	ViMaR (multi-temp)	**49.8**	**78.2**	**37.4**	**42.5**	**70.7**	**79.9**	**39.3**	**20.8**	**3.9**	**0.38**

Table 4: Comparison of reward formulations for ViMaR on LLaVA-Next-Mistral-7B. Reward$'$ yields slightly better grounding and lower hallucination rates than Reward$''$.

Base	Setting	MM-Vet↑	MMBench↑	MMMU↑	MathVista↑	CVBench↑	LLaVA-W↑	MMStar↑	CHAIRs↓	CHAIRi↓	MMHal rate↓
LLaVA-Next-Mistral-7B	ViMaR (Reward$'$)	**48.8**	**77.6**	**36.7**	**41.4**	**70.2**	**79.2**	**38.7**	**21.5**	**4.1**	**0.41**
	ViMaR (Reward$''$)	48.1	76.9	35.8	40.7	69.8	78.2	38.0	22.8	4.4	0.45

To further analyze the effect of temperature, we compare ViMaR under fixed-temperature settings of $T = 0.2$ and $T = 0.6$. As shown in Table 3, both fixed-temperature configurations lead to performance degradation relative to the multi-temperature setup. Specifically, $T = 0.2$ produces overly conservative captions that lack fine-grained detail, while $T = 0.6$ increases diversity but also raises hallucination rates. In contrast, our multi-temperature approach maintains a balanced trade-off, combining low-entropy and high-entropy candidates to achieve detailed yet visually grounded captions.

Overall, these results confirm that the proposed multi-temperature strategy is not arbitrary but an essential design choice that enhances candidate diversity and grounding stability. Both the main results (Tables 1–2) and this ablation study consistently validate its effectiveness.

A.6 Reward Function Analysis

We further examine the impact of different reward formulations on ViMaR's performance under a fixed temperature setting ($T = 0.6$). Specifically, we compare two variants of the reward function: **Reward'**, where $r_{s_i} = \delta$ if $\delta \geq \tau$ and $r_{s_i} = 0$ otherwise, and **Reward$''$**, where $r_{s_i} = \max(\delta, \tau)$. As summarized in Table 4, Reward$''$ slightly improves overall caption quality and reduces hallucination compared to Reward'. This improvement stems from Reward$''$ providing a smoother penalty near the margin threshold, allowing the value model to better differentiate between marginally grounded and confidently grounded sentences. In contrast, the hard-threshold Reward' formulation tends to under-penalize uncertain cases, leading to minor degradation in grounding accuracy. Overall, these results validate the effectiveness of our margin-based reward design and highlight the importance of calibrated reward shaping in stable value-guided decoding.

A.7 Additional Case Studies

In this section, we present further qualitative comparisons to illustrate the differences between ViMaR-guided decoding and baseline methods. Figures 3 through 7 showcase a series of representative examples, highlighting how our approach improves caption fidelity, visual grounding, and descriptive richness across diverse scenes. Additional qualitative results are also provided in the supplementary material to further support our findings.

Can you elaborate on the elements of the picture provided?

Attention Map after whole sentence generated

ViMaR Guided

The image shows a modern living room with a view of a city skyline through a large window. The room is well-lit, with natural light coming in from the window. Here are the key elements in the picture: Furniture: The living room is furnished with a variety of seating options, including a sectional sofa, two armchairs, a cylindrical side table and a round ottoman. The furniture appears to be of a contemporary style, with neutral tones that complement the room's decor. Flooring: The flooring is a light-colored hardwood, which adds warmth to the space. Window Treatments: The window is dressed with floor-to-ceiling curtains that are partially drawn, allowing for a balance between privacy and light. Decor: The room is decorated with a potted plants, which add a touch of greenery and life to the space. There are also decorative items such as books, a bowl, and a vase, which contribute to the room's aesthetic. Lighting: There is a ceiling light fixture that provides ambient lighting to the room. View: The most striking feature of the room is the view of the city through the large window. The window offers a panoramic view of the cityscape, with buildings of various heights and a river or body of water visible in the distance. Color Scheme: The color scheme of the room is neutral, with beiges, browns, and greys dominating the space. This creates a calm and inviting atmosphere. Ceiling: The ceiling has a modern design with exposed beams and a light fixture, which adds architectural interest to the room. Curtains: There are dark brown curtains that match the color of the drapes on the windows, providing a cohesive look. Overall, the living room is designed to be a comfortable and stylish space, with a focus on natural light and a connection to the urban environment outside.

VisVM Guided

In the image, you see a spacious living room that offers a view of a city skyline through a large floor-to-ceiling window. The room is well-lit, with natural light coming in from the window, and there is also a ceiling light fixture that provides additional illumination. The room is furnished with a variety of seating options, including a plush sofa and two armchairs, all arranged to create a comfortable seating area. The sofa and armchairs are upholstered in a neutral color palette, which complements the warm tones of the wooden flooring. In front of the seating area, there's a round ottoman, which can be used as a coffee table or additional seating. On the floor, there are a couple of decorative items, including a small plant in a pot and a decorative piece that could be a vase or a sculpture. The window not only provides a view but also allows for fresh air and natural light, which contributes to the room's inviting atmosphere. The curtains are partially drawn, which suggests that the room's occupants may want to control the amount of light and privacy coming in from outside. Overall, the room has been thoughtfully designed to create a welcoming space that combines comfort and style, while also offering a stunning view of the cityscape.

CLIP-PRM Guided

The image shows a spacious and well-lit living room with a view of a cityscape through floor-to-ceiling windows. The room features a variety of furniture, including:1. A sectional sofa with beige cushions, which is the largest piece of furniture in the room.2. A circular ottoman in the center of the room, which matches the sofa's color scheme.3. Two armchairs, one of which is positioned closer to the window, providing a view of the city.4. A small coffee table in front of the sofa, which appears to have a tray with books and a decorative item on it.5. A floor lamp with a modern design, which is turned on, adding to the room's ambient lighting.6. A large area rug under the coffee table, which has a geometric pattern and enhances the room's coziness.7. A decorative item on the floor, possibly a sculpture or a piece of art, near the center of the room.8. A large potted plant on the left side of the room, adding a touch of greenery and life to the space.9. A window with a sheer curtain and a blackout curtain, allowing for adjustable light control.10. A sliding glass door leading to a balcony, which is not visible in the image. The room's color palette is neutral with a mix of beige, brown, and black tones, creating a warm and inviting atmosphere. The cityscape in the background adds an urban touch to the space.

Greedy Decoding

The image shows a modern living room with a view of a city skyline through large windows. The room is well-lit, with natural light coming in from the windows, and there's a warm ambiance created by the lighting fixtures. Here are some key elements in the picture: Furniture: The room is furnished with a variety of seating options, including a sectional sofa, a couple of armchairs, and a round ottoman. The furniture appears to be of a contemporary style, with clean lines and neutral colors. Coffee Table: There's a coffee table in the center of the room, which seems to be made of wood and has a round shape. On top of it, there are a few decorative items, including a small plant and a couple of books. Rug: There's a large area rug on the floor, which adds texture and warmth to the space. Decor: The room is decorated with a few potted plants, which add a touch of greenery and life to the space. There are also some decorative items on the coffee table and on the windowsill, such as a small sculpture and a vase with flowers. Windows: The windows are large and offer a panoramic view of the city skyline. The windows are framed with white, which contrasts nicely with the darker tones of the room's interior. Curtai-ns: The windows are dressed with dark curtains that can be drawn for privacy or to control the amount of light entering the room. Color Scheme: The room has a neutral color scheme, with beiges, browns, and creams dominating the space. This creates a calm and inviting atmosphere. Lighting: The room is well-lit, with a combination of natural light from the windows and artificial light from the ceiling fixtures. Balcony: There's a balcony visible through the windows, which adds an interesting architectural element to the view. Overall, the room is designed to be comfortable and stylish, with a focus on functionality and aesthetics. The view of the city adds a dynamic element to the space, making it a pleasant place to relax and enjoy the urban surroundings.

Figure 3: Qualitative comparison example 1

Can you describe the main features of this image for me?

Attention Map after whole sentence generated

ViMaR Guided

The image shows a scene from a soccer (football) match. There are three players in the frame: one wearing a blue jersey and the others wearing a red jersey. The player in the red jersey, likely from the Arsenal team, in possession of the yellow match ball and preparing to kick it, while the player in the blue jersey is attempting to intercept or challenge for the ball. The player in red is wearing white socks and shorts, while the players in blue are wearing blue socks and shorts. In the background, there is a crowd of spectators, indicating that this is likely a professional or semi-professional match. The stadium has multiple tiers of seating, suggesting it's a large venue designed to accommodate a significant number of spectators. The jerseys of the players have sponsor logos, which is common in professional sports. The player in the red jersey sponsor by \" Fly Emirates \" visible on his jersey, has the number \"15\" visible on his shorts. The grass on the field is well-maintained, which is typical for high-level soccer matches. Overall, the image captures a moment of competitive play in a soccer match, with the focus on the interaction between the two players in the center of the action. The atmosphere suggests a high-stakes game, given the intensity of the players and the presence of a large audience. The image does not provide enough context to determine the specific teams or the score of the match.

VisVM Guided

The image depicts a soccer match in progress. The main focus is on two players in the center of the image. The player in the foreground is wearing a blue and white kit, which is typical of the Arsenal football club, and he appears to be in a defensive stance, possibly preparing to intercept the ball. The player in the background is wearing a red and white kit, which is associated with the Arsenal football club, and he is in a forward position, possibly preparing to receive the ball or make a run. The soccer ball is visible on the ground, indicating that the game is in play. The player in the red and white kit is wearing yellow shoes, which is not a common color for soccer cleats and might be a sponsor's color or a personal preference. In the background, there are other players on the field, and a crowd of spectators can be seen in the stands, suggesting that this is a professional match being played in front of an audience. The grass of the field is well-maintained, and the stadium appears to be outdoors, based on the natural light and shadows. The atmosphere suggests a competitive sports event.

CLIP-PRM Guided

The image captures an intense moment in a soccer match. There are three players in the frame, each from different teams. The player in the foreground, wearing a red jersey, is in action, dribbling the ball with his left foot, while looking over his shoulder. His team's jerseys have white accents, and the jersey has the sponsor logo of Emirates. The player in the middle, wearing a dark blue jersey, is in a defensive stance, bending down and trying to intercept the ball. His team's jerseys have white and yellow stripes on the sides. The player in the background, wearing a red jersey similar to the first player's, is in motion, possibly running towards the ball or supporting his teammate. The crowd in the background is a blur of colors, indicating a large number of spectators in the stadium. The grass of the field is a vibrant green, typical for soccer pitches. The image conveys the competitive spirit of the game and the high level of athleticism on display.

Greedy Decoding

The image shows a soccer match in progress. There are three players in the foreground, with one player in a red jersey, likely from the Arsenal team, in possession of the ball and preparing to kick it. The other two players are in blue jerseys, which suggests they are from the opposing team, possibly the Gunners. The player in red is wearing white shorts, while the players in blue are wearing white socks and blue shorts. The background is filled with spectators, indicating that this is a professional match with an audience. The stadium has a large crowd, suggesting that the match is of significant importance or is a popular event. The players are on a well-maintained soccer field, and the weather appears to be clear and sunny, which is typical for outdoor sports. The image captures a moment of action, with the player in red looking down at the ball, possibly planning his next move, while the players in blue are positioned to defend or intercept. The focus is on the players and the ball, with the crowd serving as a backdrop to the action on the field.

Figure 4: Qualitative comparison example 2

Describe the following image.

Attention Map after whole sentence generated

ViMaR Guided
The image depicts a street scene with a focus on a road intersection that has been cordoned off with yellow caution tape. The tape is attached to several orange traffic cones, which are placed on the pavement to block off the area. There are also a few white signs with black text, which appear to be related to the road work or closure. The signs include \"Caution\", \"Electric\", and \"Days\", suggesting that there may be ongoing electrical work or a temporary closure for a specific duration. The road itself is not visible due to the blockade, and there are no vehicles or pedestrians in the immediate vicinity of the cordoned-off area. The sidewalk on the right side of the image shows a few people walking, and there are a few cars and a truck parked or moving on the adjacent lanes. On the left side of the image, there is a building with a sign that reads \"T&G Nails\", indicating that it is a nail salon. The facade of the building is partially visible, and there is a small outdoor seating area with a potted plant. The sky is clear, and the lighting in the image is natural, with no artificial lights that are bright enough to overpower the ambient light. The overall style of the image is a candid, unposed photograph, capturing a typical urban street scene with a focus on the road closure. There are no visible people engaging with the road closure, and the photo seems to have been taken during the daytime.

VisVM Guided
The image depicts an urban street scene with a focus on a section of the road that has been cordoned off with yellow police tape. The tape is stretched across the street, indicating that the area is restricted or under investigation. Behind the tape, there are several orange traffic cones placed on the pavement, further emphasizing the restricted area. In the background, there is a building with a sign that reads \"T&G NAILS.\" The street is lined with trees and street lamps, and there are parked cars along the side of the road. The sky is overcast, and the lighting suggests it might be late afternoon or early evening. There are no people visible in the immediate vicinity of the cordoned-off area, and the overall atmosphere of the image is quiet and still. The presence of the police tape and cones suggests that there may have been an incident that required the area to be cordoned off, but the exact nature of the incident is not visible in the image. The style of the image is a standard, non-stylized photograph that captures a moment in an urban environment. There are no visible texts other than the \"T&G NAILS\" sign.

CLIP-PRM Guided
The image depicts an urban street scene with a focus on traffic control measures. The street is cordoned off with yellow caution tape and orange traffic cones, indicating that there may be ongoing construction or maintenance work. There is a visible manhole cover with a sign that reads \"Caution: Electrical Work.\" The sign suggests that there is work being done on the city's electrical infrastructure. On the right side of the image, there are parked cars, and on the left, there is a commercial building with signage. The sky is overcast, and the lighting suggests it is either late afternoon or early evening. The overall atmosphere is one of an active urban environment with ongoing city maintenance or construction activities.

Greedy Decoding
The image depicts a street scene with a focus on a section of road that has been cordoned off with yellow police tape. The tape is attached to several orange traffic cones, which are placed in a way that blocks access to the cordoned-off area. The tape and cones are clearly visible and appear to be in place to restrict access to the area. In the background, there is a commercial building with a sign that reads \"T&G NAILS.\" The building has a green awning and a storefront with a display of plants. There are also other businesses visible in the distance, including a restaurant with outdoor seating. The street is lined with trees and streetlights, and there are parked cars along the side of the road. The sky is overcast, suggesting it might be a cloudy day. The overall style of the image is a standard, non-stylized photograph that captures a real-life urban scene.

Figure 5: Qualitative comparison example 3

What's happening in the scene?

Attention Map after whole sentence generated

ViMaR Guided
In the scene, there is a clock tower with two clocks visible on its facade. The sky is overcast with dark clouds, suggesting it might be a cloudy or stormy day. In the foreground, there are trees and some buildings, indicating that the clock tower is part of a town or city center. There are no people clearly visible in the immediate vicinity of the tower, but there are a few cars parked or moving in the background, indicating some level of activity in the area. The setting appears to be a public space, possibly a park or a town square, given the presence of the clock tower and the benches visible in the foreground. The overall mood of the image is somewhat subdued due to the overcast sky and the absence of people in the immediate vicinity of the tower. The image captures a quiet moment in what might be a bustling area during other times of the day. The presence of the flag on top of the tower adds a touch of national pride or identity to the scene. There are power lines running across the image. The overall composition of the image, with the clock tower as the focal point, draws the viewer's eye upward, emphasizing the height and prominence of the tower in the landscape. The image is a snapshot of a moment in time, capturing the stillness of the scene despite the potential for activity in the background.

VisVM Guided
In the scene, there is a clock tower standing in the foreground with its face clearly visible. The sky appears to be cloudy, suggesting either an overcast day or a storm approaching. In the background, there are buildings, including a prominent structure with a green roof, and a street with cars and a truck parked along the side. The presence of a flag on top of the tower indicates that this might be a public or municipal building, possibly a town hall or a similar institution. The overall atmosphere of the image is somewhat somber due to the cloudy sky and the absence of people in the immediate vicinity of the tower.

CLIP-PRM Guided
The image shows a clock tower with a blue flag at the top, indicating it might be a public or government building. The sky is overcast, suggesting it might be a cloudy day. Below the tower, there's a street scene with shops and a car, indicating this is likely a town or city center. The flag flying at half-mast could signify a period of mourning or a special event that requires the flag to be lowered.

Greedy Decoding
In the scene, there is a clock tower with two clocks visible on its sides. The tower appears to be a prominent landmark, possibly in a town or city center. The sky is overcast, suggesting it might be a cloudy day. There are power lines running across the image, indicating that the tower is located in an area with electrical infrastructure. In the foreground, there are some trees and what looks like a public space with benches, which could be a park or a similar recreational area. There are also some buildings in the background, which might be part of a commercial or residential area. The overall atmosphere of the image is calm and quiet, with no visible activity that would indicate a specific event or time of day.

Figure 6: Qualitative comparison example 4

What are the key elements in this picture?

Attention Map after whole sentence generated

ViMaR Guided
In the scene, there is a red double-decker bus driving down a street in what appears to be a city center. The bus has a sign on the front that reads \"THE WEDDING SPECIAL, \" and \"www.timebus.co.uk,\" and License plate: \"JJD 389D\". indicating that it is likely being used for a wedding event. There is a floral arrangement on the front of the bus, which is a common decoration for weddings. In the background, there are pedestrians and other vehicles, suggesting a busy urban environment. The architecture of the buildings and the style of the bus suggest that this could be a city like London, where double-decker buses are iconic and often used for sightseeing and special events. The weather seems to be overcast, and there are no visible signs of rain or snow. The scene captures a moment of everyday life with a touch of celebration due to the bus's signage. Please note that this description is based on the visible elements in the image and does not include any speculative or inferred information. The image does not provide enough context to determine the exact location or the specific nature of the wedding event. The bus could be part of a tour company that offers wedding specials, or it could be a privately hired bus for a specific wedding.

VisVM Guided
In the scene, a traditional red double-decker bus is driving on a city street, with the word \"WEDDING\" written on its front, indicating that it may be used for a wedding event. The bus has a floral arrangement on the front grille, which is often used to decorate such vehicles for special occasions. There are people walking on the sidewalk, and the street is lined with buildings that have a classic architectural style, suggesting that this scene is likely set in a city with a strong British influence, such as London. The presence of traffic lights and other vehicles, as well as the pedestrians, indicates a typical urban environment. The weather appears to be overcast, and there is no immediate sign of any specific event beyond the fact that the bus is being used for a wedding.

CLIP-PRM Guided
In the scene, there is a red double-decker bus driving down a city street. The bus has a sign on the front that reads \"THE WEDDING SPECIAL\" and \"www.timetable.co.uk,\" which suggests that the bus is being used for a wedding, possibly as a unique mode of transportation for the wedding party or guests. The bus is decorated with flowers, which is a common practice for special occasions. There are people walking on the sidewalk, and the setting appears to be a busy urban area with other vehicles and pedestrians. The weather seems to be overcast, and there are trees along the street, indicating it might be spring or early summer.

Greedy Decoding
In the scene, there is a red double-decker bus driving down a street. The bus has a sign on the front that reads \"THE WEDDING SPECIAL\" and \"www.timelessbus.co.uk,\" indicating that it is a bus specifically used for wedding transportation. The bus is decorated with flowers, which is a common practice for wedding vehicles to add a festive touch. There are people walking on the sidewalk, and the street appears to be in a city with buildings and traffic lights visible in the background. The weather seems to be overcast, and there are no visible signs of rain or snow. The overall atmosphere suggests a typical day in an urban setting.

Figure 7: Qualitative comparison example 5

Describe the following image.

Attention Map after whole sentence generated

ViMaR Guided

The image depicts a serene scene of a group of people on a boat in a body of water, which appears to be a river or a lake. The boat is a long, narrow vessel, with a pointed stern, and it's being poled through the water by an individual standing in the rear. There are several people seated in the boat, all facing the same direction, which is towards the left side of the image, suggesting they are moving upstream or towards the left bank. The boat is positioned in the foreground, with the people onboard appearing to be engaged in a leisurely activity, possibly sightseeing or enjoying a tour. They are dressed casually, and one person at the back of the boat is standing and wearing a cap, suggesting it might be a sunny day. In the background, there is a line of traditional huts with thatched roofs, which are partially obscured by the lush greenery of trees and foliage. The huts are constructed on stilts, indicating that the location may be prone to flooding or in a region where such construction is common for protection against rising water levels. The water has gentle ripples, reflecting the greenery and the huts, and the overall color palette of the image is dominated by shades of green, with the boat and the people providing a contrast with their earth tones. The image has a watermark or a filter applied to it, giving it a vintage or aged appearance, with a slight sepia tone that enhances the natural colors and adds a nostalgic feel to the scene. There is no visible text in the image, and the style of the image suggests it could be a photograph taken by a tourist or a traveler, capturing a moment during their visit to a cultural or natural site. The composition of the image, with the boat in the foreground and the huts in the background, leads the viewer's eye from the boat and its occupants to the tranquil and picturesque setting beyond them.

VisVM Guided

The image depicts a tranquil scene on a body of water, which appears to be a lake or a slow-moving river. In the foreground, there is a wooden boat with several people on board. They seem to be enjoying a leisurely ride, with one person standing at the front of the boat, possibly steering or navigating. The boat is painted in a reddish-brown hue and has a traditional design, suggesting it might be used for transportation or tourism in a region where such boats are common. In the background, there are several thatched-roof huts or bungalows situated along the water's edge. These structures have a rustic appearance and are likely part of a local community or a tourist facility. The surrounding vegetation is lush and green, indicating a tropical or subtropical climate. The water is calm, reflecting the greenery and the huts, which adds to the serene atmosphere of the scene. There is no visible text or branding in the image, and the style of the photograph is naturalistic, capturing the scene as it is without any apparent filters or artistic manipulations. The overall mood of the image is peaceful and idyllic, evoking a sense of relaxation and escape from the hustle and bustle of everyday life.

CLIP-PRM Guided

In the image, a small wooden boat is gently floating on a calm, greenish-brown river. The boat, with its pointed front and slightly curved rear, is filled with people who are enjoying a day out on the water. The passengers are dressed in casual clothing, some donning hats for shade. They are sitting close to each other, suggesting a sense of camaraderie and shared enjoyment. The river itself is nestled amidst lush greenery, with several huts or hut-like structures visible on the shore. These structures, likely made of bamboo, have thatched roofs, blending harmoniously with the natural surroundings. The vegetation appears dense, creating a serene and secluded environment. The boat is moving away from the camera, giving us a view of its rear as it ventures further down the river. The perspective from the shore looking towards the boat creates a sense of depth and distance in the image. The overall scene is one of leisure and tranquility, capturing a moment of relaxation and enjoyment in a natural setting.

Greedy Decoding

The image depicts a serene scene of a group of people on a boat in a body of water. The boat is a traditional design, possibly a canoe or a small motorboat, and it is occupied by several individuals who appear to be enjoying a leisurely outing. The boat is moving away from the shore, which is lined with thatched-roof huts, suggesting a tropical or subtropical location. The water is calm, reflecting the greenery of the surrounding vegetation, which includes trees and shrubs. The sky is overcast, casting a soft light over the scene. The overall atmosphere of the image is peaceful and idyllic, capturing a moment of relaxation and enjoyment in a natural setting.

Figure 8: Qualitative comparison example 6

What is this photo about?

Attention Map after whole sentence generated

ViMaR Guided
The photo shows two young sheep, also known as lambs, in a grassy field. The lamb on the left appears to be black, while the lamb on the right is predominantly white with some black markings. In the background, there are other sheep, ducks, and big trees, suggesting this is a pastoral setting, possibly a farm or a rural area where sheep are raised. The lambs seem to be grazing or exploring the field, which is typical behavior for young sheep as they learn to navigate their environment. The presence of other sheep in the background indicates that this is part of a larger flock. The photo captures a peaceful and natural moment in the lives of these young animals. The weather appears to be overcast, and the grass is lush and green, indicating it might be spring or early summer when trees are not fully leafed. The overall scene conveys a sense of tranquility and the simple beauty of rural landscapes.

VisVM Guided
The photo depicts a pastoral scene with two young sheep, one black and one white with a light-colored stripe, walking in a grassy field. There are other animals in the background, including what appears to be ducks, grazing animals, and trees with bare branches, which suggests it might be spring or early summer when trees are not fully leafed. The presence of both sheep and ducks indicates a mixed-species environment, possibly on a farm or in a rural setting where different types of livestock can be found together. The scene is peaceful and represents a typical day in a rural area where animals are free to roam and graze.

CLIP-PRM Guided
This photo depicts a pastoral scene with two sheep, one black and one white, walking together in a grassy field. There are also other sheep in the background, grazing or resting. The presence of trees and a natural setting suggests that this is a rural or agricultural area. The sheep appear to be calm and are likely part of a flock that is not fully visible in the frame. The photo captures a peaceful moment in the lives of these animals, showcasing their natural behavior in a pasture.

Greedy Decoding
The photo shows two young sheep, also known as lambs, in a grassy field. The lamb on the left appears to be a black-faced sheep, possibly a breed like a Jacob or a Black Welsh Mountain, while the lamb on the right has a white face and a black body, which could be a breed like a Black Welsh Mountain or a similar black-faced sheep with a white face. They are walking together, and in the background, there are other sheep and trees, suggesting a rural or pastoral setting. The presence of a few birds, possibly ducks, adds to the natural, pastoral scene.

Figure 9: Qualitative comparison example 7

Can you describe the main features of this image for me?

Attention Map after whole sentence generated

ViMaR Guided

In the image, there are two children standing on a sidewalk next to a white fire hydrant. The child on the left is wearing a red sweater, blue jeans, and white shoes, while the child on the right is dressed in a dark blue sweatshirt with a hood, blue-white pants, and white shoes. Both children are smiling at the camera and appear to be posing for the photo. In the background, there are trees, parked cars and a building with a red facade. The setting suggests an urban environment, possibly a residential area given the presence of the fire hydrant and the style of the buildings. The sky is clear, and there are no visible texts or distinctive brands in the image. The overall mood of the photo is cheerful and casual. The focus is on the children and the fire hydrant, with the background serving as context for the location of the photo. The fire hydrant has the number \"8\" visible on it, which is typically used to identify the hydrant for firefighting purposes. The image is well-lit and seems to have been taken during the day under natural light. The composition of the photo is balanced, with the children centrally positioned and the fire hydrant to their side.

VisVM Guided

In the image, there are two children standing on a sidewalk next to a white fire hydrant. The child on the left is wearing a red sweater and blue jeans, and the child on the right is dressed in a dark blue hoodie with white drawstrings, blue sweatpants, and white sneakers. They both appear to be standing in a residential area, as there are buildings in the background and a tree nearby. On the sidewalk, there are also some potted plants, and a man is standing a bit further down the sidewalk. On the right side of the image, there's a portion of a black car, and in the distance, there are more vehicles and buildings. The children seem to be posing for the photo, with the one on the right looking a bit more solemn than the other. The overall scene is casual and everyday, capturing a moment of childhood in an urban setting.

CLIP-PRM Guided

In this image, there are two young children standing on a sidewalk. The child on the left is wearing a red sweater and a pair of white shoes. They have short hair and are smiling at the camera. The child on the right is wearing a dark blue sweatshirt with a red stripe on the side, paired with dark pants and white sneakers. They have short, curly hair and also appear to be smiling. Between the two children, there is a white fire hydrant. The hydrant has a chain attached to it, and there are numbers and letters visible on its sides, although the specific details are not clear. The background shows a typical urban street scene with buildings, parked cars, and a tree lining the sidewalk. The lighting suggests it might be a sunny day, and there are no people other than the two children in the immediate vicinity.

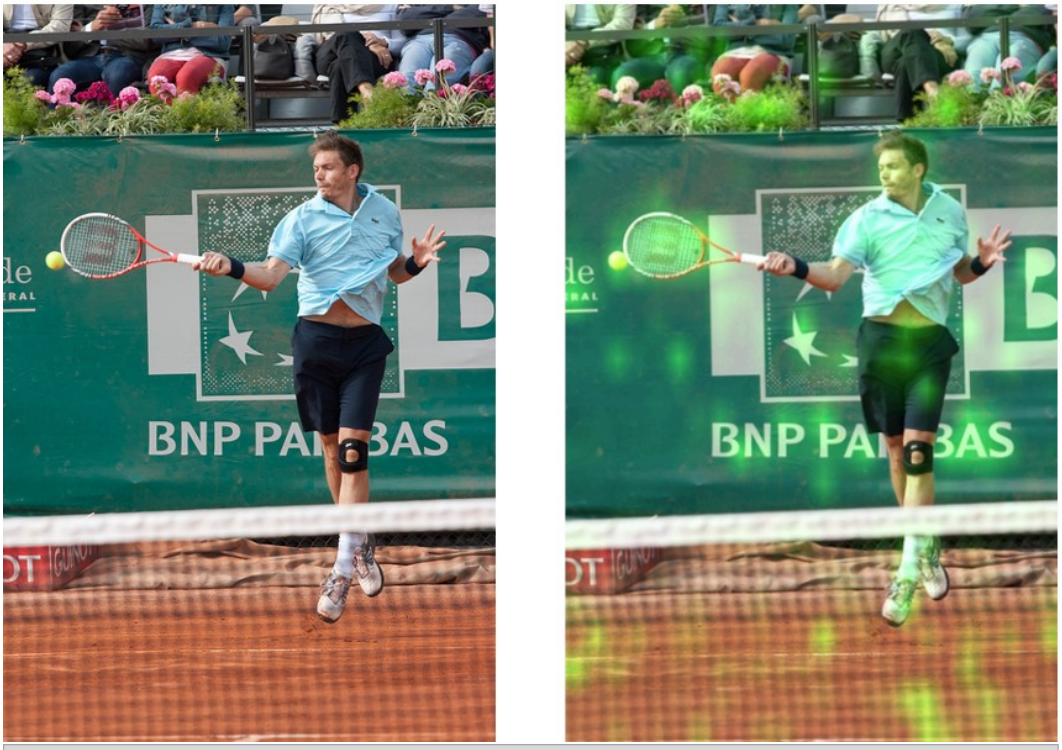
Greedy Decoding

In the image, there are two children standing on a sidewalk next to a white fire hydrant. The child on the left is wearing a red sweater and blue jeans, while the child on the right is dressed in a blue jacket, blue pants, and white sneakers. Both children are smiling and appear to be posing for the photo. In the background, there are parked cars and a building with a red facade. The setting suggests an urban environment, possibly a residential area. The fire hydrant is a prominent feature in the foreground, indicating that the photo was taken in a location where fire safety is a concern.

Figure 10: Qualitative comparison example 8

What do you see happening in this image?

Attention Map after whole sentence generated



ViMaR Guided
In the image, there is a male tennis player in the middle of a serve on a clay court. He is wearing a light blue shirt, dark blue shorts, and white shoes. The player is holding a red and white tennis racket and is in the process of striking a yellow tennis ball. He appears to be focused on the ball, which is in mid-air. In the background, there is a crowd of spectators watching the match, and there is a green advertisement board with the letters \"BNP PARIBAS\" on it, suggesting that the event is sponsored by that company. The setting suggests that this is a professional or high-level amateur tennis match. The player's posture and the position of the racket indicate that he is executing a powerful serve, which is a crucial part of the game in tennis. The clay court surface is characteristic of the French Open, one of the four Grand Slam tennis tournaments. The spectators are seated on chairs, and there are potted plants and flowers adding to the ambiance of the event. The overall scene captures the intensity and excitement of a live tennis match.

VisVM Guided
In the image, there is a tennis player in the middle of a serve. The player is wearing a blue shirt and dark shorts, with white shoes. He is holding a tennis racket with a red and white frame and a yellow ball in his other hand, preparing to toss the ball into the air and hit it. The player is looking upwards, likely tracking the ball's trajectory. In the background, there's a green advertisement board with the text \"BNP PARIBAS\" and a star logo, which suggests that the event is being sponsored by BNP Paribas, a French multinational banking corporation. There are also spectators in the stands watching the match. The setting appears to be a professional tennis court.

CLIP-PRM Guided
In the image, there is a male tennis player in action on a clay court. He is in the middle of a backhand shot, with his racket positioned behind him and his body turned towards the net. He is wearing a blue shirt, dark shorts, and white shoes. The player appears to be focused on the ball, which is in the air near him. In the background, there are spectators watching the match, and there is a sign that reads \"BNP PARABIS,\" which is likely a sponsor of the event. The setting suggests this is a professional tennis match.

Greedy Decoding
In the image, there is a male tennis player in the middle of a swing, hitting a tennis ball with his racket. He appears to be in the middle of a match, as indicated by the presence of a crowd in the background, and there are advertisements on the wall behind him. The player is wearing a blue shirt, dark shorts, and white shoes, which is typical attire for a tennis player. The setting suggests this is a professional tennis match.

Figure 11: Qualitative comparison example 9

Can you elaborate on the elements of the picture provided?

Attention Map after whole sentence generated

ViMaR Guided

The image shows a beach scene with several elements: 1. Surfboards: There are four surfboards leaning against a railing on the left side of the image. They are of various colors and designs, suggesting they belong to different individuals. 2. Beachgoers: There are people on the beach, some of whom are sitting on the sand, while others are standing or walking. They appear to be enjoying the beach environment. 3. Ocean: The vast expanse of the ocean is visible in the background, with waves breaking on the shore. 4. Cliffs: In the background, you can see a large cliff or headland jutting out into the ocean, adding a rugged and natural feature to the beach landscape. 5. Railing: A black metal railing runs along the edge of the beach, providing a boundary between the sand and the road. 6. Signage: There is a sign on the right side of the image that reads \"SALSBURY SURF SHOP OPEN,\" indicating that the surf shop is open for business. 7. Sky: The sky is clear and blue, suggesting a sunny day, which is ideal for beach activities. 8. Road: The road runs parallel to the beach, and it appears to be a public road with a sidewalk. The overall atmosphere of the image is relaxed and leisurely, typical of a day at the beach.

VisVM Guided

The image depicts a lively beach scene on a clear day with a blue sky and no visible clouds. The beach is sandy and appears to be a popular spot, as there are several people scattered across the area, some of whom are sitting on the sand, while others are standing or walking. In the foreground, there is a black metal railing that separates the beach from a paved area, which could be a boardwalk or a path for pedestrians. Leaning against the railing are surfboards of various colors, including yellow, blue, and white, suggesting that this is a place where surfing is a popular activity. On the paved area, there is a sign that reads \"SALTBURN SURF SHOP OPEN,\" indicating that there is a surf shop nearby where visitors can rent or purchase surfing equipment. The presence of the surfboards and the surf shop sign suggests that this beach is known for surfing and possibly other water sports. The beach itself is lined with grassy areas, which might be used for picnicking or just relaxing, and the water appears to be calm with gentle waves. The overall atmosphere of the image is one of a fun, recreational environment where people come to enjoy the beach and the water activities it offers.

CLIP-PRM Guided

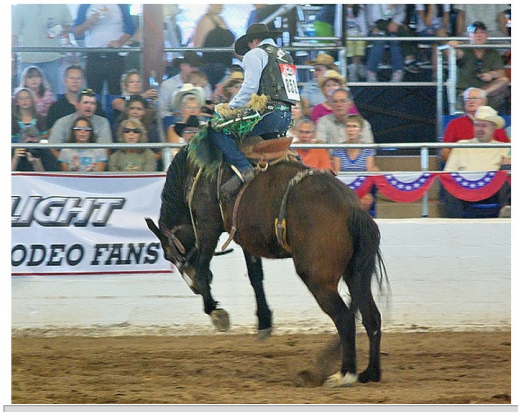
The image depicts a scenic beach view. In the foreground, there's a metal railing along the edge of the beach, likely to prevent people from accidentally falling into the ocean or to delineate the boundary of the beach area. To the right of the railing, there are several surfboards leaning against a sign that says \"SALTBURN SURF SHOP OPEN,\" indicating that the shop is currently open for business. The surfboards are of different colors and designs, suggesting they are for rent or sale to visitors. Beyond the railing, the beach appears to be quite busy with several people enjoying the day. The sand is speckled with footprints, and there are a few small items scattered around, such as buckets, which are often used to collect shells or other small treasures from the beach. The ocean is visible in the background, with small waves indicating a calm sea. The sky is clear and blue, suggesting good weather. In the far background, there's a large rock formation that extends into the ocean. This cliff-like structure adds a dramatic backdrop to the scene and is a characteristic feature of the coastline. Overall, the image captures a typical day at a beachside surf shop, with people enjoying the beach and the shop offering surfing equipment for hire or sale.

Greedy Decoding

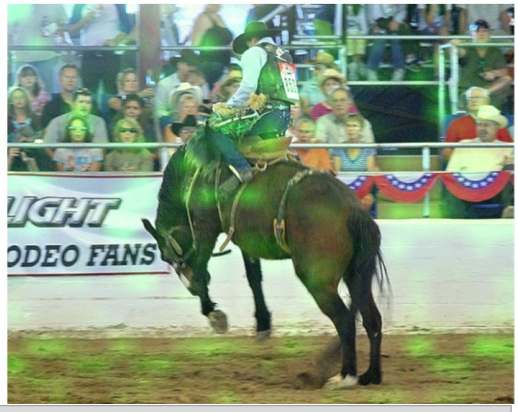
In the image, you can see a beach scene with the following elements: 1. Surfboards: There are several surfboards leaning against a railing on the left side of the image. The surfboards vary in color and design, suggesting they belong to different individuals. The presence of the surfboards indicates that surfing is a popular activity at this beach. 2. People and beach activity: There are several people on the sandy beach. Some are closer to the water's edge, while others are further back, possibly sunbathing or socializing. There are also beach umbrellas set up, providing shade for beachgoers. The presence of people and umbrellas suggests that it's a sunny day and the beach is a popular spot for relaxation and leisure activities. 3. Ocean: The vast body of water is visible, with waves crashing against the shore, which is a common feature of a coastal beach. The ocean's presence adds a dynamic element to the scene, with the constant movement of the waves. 4. Cliff: In the background, you can see a large cliff or headland jutting out into the ocean, adding a rugged and natural feature to the beach landscape.

Figure 12: Qualitative comparison example 10

What do you see happening in this image?



Attention Map after whole sentence generated



ViMaR Guided
In the image, there is a person riding a bucking bronco during what appears to be a rodeo event. The rider is wearing a cowboy hat and a numbered on the back, which is typical attire for a rodeo competitor. The horse is in mid-buck, with its hind legs raised off the ground, and the rider is holding onto the horse's reins, trying to stay on. In the background, there is a crowd of spectators watching the event, and there is a banner with the words \"LIGHT RODEO FANS\" on it, suggesting that this event is part of a rodeo series. The setting looks like a dirt arena, which is common for rodeo competitions. The rider's posture and the position of the horse indicate a high level of skill and athleticism required for this sport. The spectators are focused on the action, indicating the excitement and intensity of the event. The image captures a moment of high adrenaline and athletic prowess. Please note that rodeo events can be controversial due to concerns about the welfare of the animals involved, particularly the horses and bulls used for bucking. The image does not provide enough context to determine the specific event or location, but it is clear that this is a professional rodeo event. The rider's attire and the presence of a numbered vest suggest that this is a competitive event, and the rider is likely a professional cowboy.

VisVM Guided
The image shows a person riding a horse, likely participating in a rodeo event, given the attire and the setting that includes spectators and arena seating. The rider is wearing a cowboy hat and appears to be engaged in an activity such as a bull-dogging competition, where the rider's goal is to throw a rope around the legs of a steer. The horse is in mid-gallop, and the rider is leaning forward, which is typical for riders attempting to catch a steer. The spectators are watching the event, and there is a banner in the background with the text \"LIGHT RODEO FANS,\" suggesting that this is a public event where fans of rodeo sports are present. The setting looks like a typical rodeo arena, with metal barriers for safety.

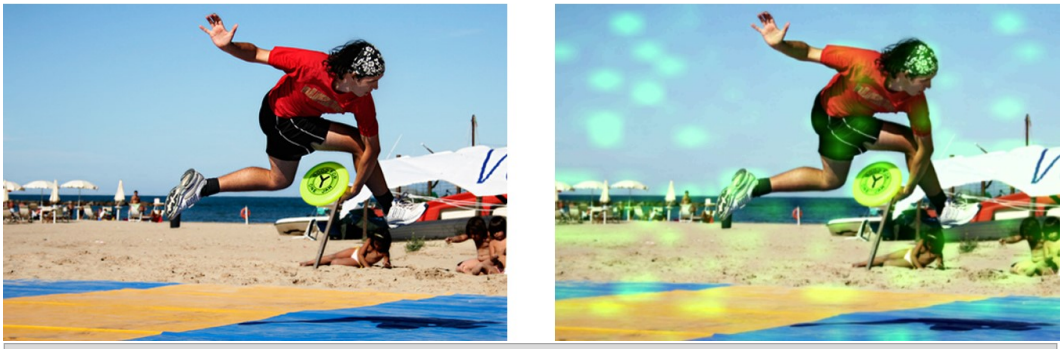
CLIP-PRM Guided
In the image, there is a rodeo event taking place. A cowboy is riding a bucking bronco, which is a horse trained to buck and throw the rider off its back. The cowboy is holding onto a rigging, which is a strap that attaches to the horse's halter and the rider's belt, allowing the rider to stay on the horse. The horse is in mid-buck, and the cowboy is leaning forward, trying to maintain his balance. The audience is watching the event, and there is a sign in the background that reads \"LIGHT RODEO FANS,\" indicating that this is a light rodeo event. The setting appears to be an outdoor arena, and the atmosphere is that of a competitive rodeo event.

Greedy Decoding
In the image, there is a rodeo event taking place. A cowboy is riding a bucking horse, which is a common event in rodeo competitions. The cowboy is holding onto a rigging, which is a strap that helps him stay on the horse. The horse is in mid-buck, with its hind legs raised off the ground, which is typical behavior for a bucking horse during a rodeo event. The cowboy is wearing a cowboy hat and a vest, which are typical attire for a rodeo cowboy. The audience is watching the event, and there is a banner in the background with the words \"LIGHT RODEO FANS\" on it, indicating that this is a light rodeo event. The setting appears to be an outdoor arena with seating for spectators.

Figure 13: Qualitative comparison example 11

What are the key elements in this picture?

Attention Map after whole sentence generated



ViMaR Guided

The key elements in this picture include: 1. A person in mid-air, performing a trick with a frisbee. 2. The person is wearing a red shirt, black shorts, white sports shoes, and a black-and-white bandana. 3. The frisbee is neon yellow-green and appears to be in motion. 4. The setting is a beach with sand, and there are beach umbrellas in the background. 5. The background includes white parasols and people relaxing, indicating a recreational atmosphere. 6. The sky is clear and blue, suggesting good weather conditions. 7. A blue and yellow mat is laid on the sand. 8. There is a boat and some beach gear in the background, reinforcing the casual, coastal setting. 9. A few individuals are sitting or lounging on the sand. 10. The overall composition of the image conveys a sense of energy and fun, capturing a moment of athleticism and enjoyment.

VisVM Guided

The key elements in this picture are: 1. A person in mid-air performing a trick with a frisbee. 2. The frisbee is a yellow color. 3. The person is wearing a red shirt and black shorts. 4. They are wearing white shoes with black soles. 5. The person has a bandana around their head. 6. The background shows a beach setting with umbrellas, beach chairs, and other beachgoers. 7. The sky is clear, suggesting good weather conditions. 8. There is a body of water visible in the distance.

CLIP-PRM Guided

In the image, there are several key elements: 1. A person in mid-air, performing a frisbee trick. They are wearing a red shirt and black shorts, and they have a black bandana on their head. 2. The frisbee is in the air, close to the person's feet, indicating that they are either about to catch it or have just thrown it. 3. The person is on a beach, as evidenced by the sandy ground and the presence of beach umbrellas and chairs in the background. 4. There are other people on the beach, some of whom are seated under umbrellas, suggesting a relaxed, leisurely atmosphere. 5. The sky is clear and blue, indicating good weather conditions. 6. The overall scene suggests a recreational activity, possibly a casual game of frisbee, taking place on a sunny day at the beach.

Greedy Decoding

In this picture, there are several key elements: 1. A person in mid-air, performing a frisbee trick. The person is wearing a red shirt and black shorts, and they are holding a yellow frisbee. 2. A sandy beach setting with a clear blue sky. 3. A group of people sitting under umbrellas, which suggests a leisurely beach environment. 4. A boat visible in the background, indicating that the beach might be near a body of water. 5. A blue and yellow tarp on the sand, which could be part of the setup for the frisbee trick or simply a piece of beach equipment. 6. The person's dynamic pose and the frisbee in motion, which captures the action and skill involved in the trick. 7. The clear sky and the bright sunlight, which contribute to the overall vibrant and energetic atmosphere of the scene.

Figure 14: Qualitative comparison example 12

What do you think is going on in this snapshot?

Attention Map after whole sentence generated

ViMaR Guided

In the image, we see a person standing on a paved area in front of a building with classical architecture. The person is holding a tennis racket and appears to be in the middle of a game or practice session, as indicated by the yellow-green tennis ball in motion nearby. The person is uniquely dressed in what appears to be a white robe or toga-like outfit with red crosses and blue accents, possibly a costume or themed attire rather than typical sportswear. In the background, the setting appears to be outside a classical-style building with tall columns and stone walls, resembling a historical or governmental structure, likely a cultural landmark. Several people are gathered behind a black railing on the raised platform of the building, observing the scene below. It's a sunny day, and the shadows on the ground suggest that the sun is low in the sky, which could indicate either early morning or late afternoon. The overall atmosphere is relaxed and leisurely, with people enjoying outdoor activities on a pleasant day. The presence of a backpack near the person with the racket might suggest that they have brought personal belongings or additional sports equipment with them.

VisVM Guided

In the image, we see a person standing on a paved area in front of a building with classical architecture. The individual is holding a tennis racket and appears to be in the middle of a tennis match or practice session. A sports ball, likely a tennis ball, is in mid-air nearby, suggesting that the person might have just hit the ball or is about to hit it. The person is dressed in sports attire, which includes a white and red outfit that is commonly associated with tennis. There are several other people in the background, some of whom are standing on a balcony above the tennis player, possibly watching the game or practice. The setting appears to be an urban environment, possibly a public park or a recreational area within a city. The presence of the tennis player and the sports equipment suggests that this is an outdoor recreational space where people can engage in sports activities.

CLIP-PRM Guided

In the image, we see a person standing on a paved area, holding a tennis racket and wearing a white outfit with red and blue accents. There is a tennis ball in the air, suggesting that the person is either about to serve or has just hit the ball. The person appears to be in the middle of a game or practice session, as indicated by the presence of the tennis ball and racket. In the background, there is a building with columns, which could be a public building or a sports facility. There are also other individuals in the distance, some of whom are standing on a balcony or a raised platform, possibly watching the person play tennis. The setting seems to be outdoors, and the presence of the tennis ball and racket suggests that the person is engaged in a sporting activity.

Greedy Decoding

In the image, there is a person standing on a paved area in front of a building with columns, which suggests it might be a public or institutional building. The person is holding a tennis racket and appears to be looking at a tennis ball on the ground. It seems like the person might be taking a break from playing tennis or is about to start a game. There are other people in the background, some of whom are standing on a balcony, which indicates that this might be a popular spot for people to gather and socialize. The presence of a backpack and a handbag suggests that the person might be a student or a visitor to the area. The overall atmosphere is casual and relaxed.

Figure 15: Qualitative comparison example 13

Can you elaborate on the elements of the picture provided?

Attention Map after whole sentence generated

ViMaR Guided

The image shows a modern living room with a view of a city skyline through a large window. The room is well-lit, with natural light coming in from the window. Here are the key elements in the picture: Furniture: The living room is furnished with a variety of seating options, including a sectional sofa, two armchairs, a cylindrical side table and a round ottoman. The furniture appears to be of a contemporary style, with neutral tones that complement the room's decor. Flooring: The flooring is a light-colored hardwood, which adds warmth to the space. Window Treatments: The window is dressed with floor-to-ceiling curtains that are partially drawn, allowing for a balance between privacy and light. Decor: The room is decorated with a potted plants, which add a touch of greenery and life to the space. There are also decorative items such as books, a bowl, and a vase, which contribute to the room's aesthetic. Lighting: There is a ceiling light fixture that provides ambient lighting to the room. View: The most striking feature of the room is the view of the city through the large window. The window offers a panoramic view of the cityscape, with buildings of various heights and a river or body of water visible in the distance. Color Scheme: The color scheme of the room is neutral, with beiges, browns, and greys dominating the space. This creates a calm and inviting atmosphere. Ceiling: The ceiling has a modern design with exposed beams and a light fixture, which adds architectural interest to the room. Curtains: There are dark brown curtains that match the color of the drapes on the windows, providing a cohesive look. Overall, the living room is designed to be a comfortable and stylish space, with a focus on natural light and a connection to the urban environment outside.

VisVM Guided

In the image, you see a spacious living room that offers a view of a city skyline through a large floor-to-ceiling window. The room is well-lit, with natural light coming in from the window, and there is also a ceiling light fixture that provides additional illumination. The room is furnished with a variety of seating options, including a plush sofa and two armchairs, all arranged to create a comfortable seating area. The sofa and armchairs are upholstered in a neutral color palette, which complements the warm tones of the wooden flooring. In front of the seating area, there's a round ottoman, which can be used as a coffee table or additional seating. On the floor, there are a couple of decorative items, including a small plant in a pot and a decorative piece that could be a vase or a sculpture. The window not only provides a view but also allows for fresh air and natural light, which contributes to the room's inviting atmosphere. The curtains are partially drawn, which suggests that the room's occupants may want to control the amount of light and privacy coming in from outside. Overall, the room has been thoughtfully designed to create a welcoming space that combines comfort and style, while also offering a stunning view of the cityscape.

CLIP-PRM Guided

The image shows a spacious and well-lit living room with a view of a cityscape through floor-to-ceiling windows. The room features a variety of furniture, including:1. A sectional sofa with beige cushions, which is the largest piece of furniture in the room.2. A circular ottoman in the center of the room, which matches the sofa's color scheme.3. Two armchairs, one of which is positioned closer to the window, providing a view of the city.4. A small coffee table in front of the sofa, which appears to have a tray with books and a decorative item on it.5. A floor lamp with a modern design, which is turned on, adding to the room's ambient lighting.6. A large area rug under the coffee table, which has a geometric pattern and enhances the room's coziness.7. A decorative item on the floor, possibly a sculpture or a piece of art, near the center of the room.8. A large potted plant on the left side of the room, adding a touch of greenery and life to the space.9. A window with a sheer curtain and a blackout curtain, allowing for adjustable light control.10. A sliding glass door leading to a balcony, which is not visible in the image. The room's color palette is neutral with a mix of beige, brown, and black tones, creating a warm and inviting atmosphere. The cityscape in the background adds an urban touch to the space.

Greedy Decoding

The image shows a modern living room with a view of a city skyline through large windows. The room is well-lit, with natural light coming in from the windows, and there's a warm ambiance created by the lighting fixtures. Here are some key elements in the picture: Furniture: The room is furnished with a variety of seating options, including a sectional sofa, a couple of armchairs, and a round ottoman. The furniture appears to be of a contemporary style, with clean lines and neutral colors. Coffee Table: There's a coffee table in the center of the room, which seems to be made of wood and has a round shape. On top of it, there are a few decorative items, including a small plant and a couple of books. Rug: There's a large area rug on the floor, which adds texture and warmth to the space. Decor: The room is decorated with a few potted plants, which add a touch of greenery and life to the space. There are also some decorative items on the coffee table and on the windowsill, such as a small sculpture and a vase with flowers. Windows: The windows are large and offer a panoramic view of the city skyline. The windows are framed with white, which contrasts nicely with the darker tones of the room's interior. Curtai-ns: The windows are dressed with dark curtains that can be drawn for privacy or to control the amount of light entering the room. Color Scheme: The room has a neutral color scheme, with beiges, browns, and creams dominating the space. This creates a calm and inviting atmosphere. Lighting: The room is well-lit, with a combination of natural light from the windows and artificial light from the ceiling fixtures. Balcony: There's a balcony visible through the windows, which adds an interesting architectural element to the view. Overall, the room is designed to be comfortable and stylish, with a focus on functionality and aesthetics. The view of the city adds a dynamic element to the space, making it a pleasant place to relax and enjoy the urban surroundings.

Figure 16: Qualitative comparison example 14

Can you describe the main features of this image for me?

Attention Map after whole sentence generated

ViMaR Guided

The image shows a scene from a soccer (football) match. There are three players in the frame: one wearing a blue jersey and the others wearing a red jersey. The player in the red jersey, likely from the Arsenal team, in possession of the yellow match ball and preparing to kick it, while the player in the blue jersey is attempting to intercept or challenge for the ball. The player in red is wearing white socks and shorts, while the players in blue are wearing blue socks and shorts. In the background, there is a crowd of spectators, indicating that this is likely a professional or semi-professional match. The stadium has multiple tiers of seating, suggesting it's a large venue designed to accommodate a significant number of spectators. The jerseys of the players have sponsor logos, which is common in professional sports. The player in the red jersey sponsor by \" Fly Emirates \" visible on his jersey, has the number \"15\" visible on his shorts. The grass on the field is well-maintained, which is typical for high-level soccer matches. Overall, the image captures a moment of competitive play in a soccer match, with the focus on the interaction between the two players in the center of the action. The atmosphere suggests a high-stakes game, given the intensity of the players and the presence of a large audience. The image does not provide enough context to determine the specific teams or the score of the match.

VisVM Guided

The image depicts a soccer match in progress. The main focus is on two players in the center of the image. The player in the foreground is wearing a blue and white kit, which is typical of the Arsenal football club, and he appears to be in a defensive stance, possibly preparing to intercept the ball. The player in the background is wearing a red and white kit, which is associated with the Arsenal football club, and he is in a forward position, possibly preparing to receive the ball or make a run. The soccer ball is visible on the ground, indicating that the game is in play. The player in the red and white kit is wearing yellow shoes, which is not a common color for soccer cleats and might be a sponsor's color or a personal preference. In the background, there are other players on the field, and a crowd of spectators can be seen in the stands, suggesting that this is a professional match being played in front of an audience. The grass of the field is well-maintained, and the stadium appears to be outdoors, based on the natural light and shadows. The atmosphere suggests a competitive sports event.

CLIP-PRM Guided

The image captures an intense moment in a soccer match. There are three players in the frame, each from different teams. The player in the foreground, wearing a red jersey, is in action, dribbling the ball with his left foot, while looking over his shoulder. His team's jerseys have white accents, and the jersey has the sponsor logo of Emirates. The player in the middle, wearing a dark blue jersey, is in a defensive stance, bending down and trying to intercept the ball. His team's jerseys have white and yellow stripes on the sides. The player in the background, wearing a red jersey similar to the first player's, is in motion, possibly running towards the ball or supporting his teammate. The crowd in the background is a blur of colors, indicating a large number of spectators in the stadium. The grass of the field is a vibrant green, typical for soccer pitches. The image conveys the competitive spirit of the game and the high level of athleticism on display.

Greedy Decoding

The image shows a soccer match in progress. There are three players in the foreground, with one player in a red jersey, likely from the Arsenal team, in possession of the ball and preparing to kick it. The other two players are in blue jerseys, which suggests they are from the opposing team, possibly the Gunners. The player in red is wearing white shorts, while the players in blue are wearing white socks and blue shorts. The background is filled with spectators, indicating that this is a professional match with an audience. The stadium has a large crowd, suggesting that the match is of significant importance or is a popular event. The players are on a well-maintained soccer field, and the weather appears to be clear and sunny, which is typical for outdoor sports. The image captures a moment of action, with the player in red looking down at the ball, possibly planning his next move, while the players in blue are positioned to defend or intercept. The focus is on the players and the ball, with the crowd serving as a backdrop to the action on the field.

Figure 17: Qualitative comparison example 15

Describe the following image.

Attention Map after whole sentence generated

ViMaR Guided
The image depicts a street scene with a focus on a road intersection that has been cordoned off with yellow caution tape. The tape is attached to several orange traffic cones, which are placed on the pavement to block off the area. There are also a few white signs with black text, which appear to be related to the road work or closure. The signs include \"Caution\", \"Electric\", and \"Days\", suggesting that there may be ongoing electrical work or a temporary closure for a specific duration. The road itself is not visible due to the blockade, and there are no vehicles or pedestrians in the immediate vicinity of the cordoned-off area. The sidewalk on the right side of the image shows a few people walking, and there are a few cars and a truck parked or moving on the adjacent lanes. On the left side of the image, there is a building with a sign that reads \"T&G Nails\", indicating that it is a nail salon. The facade of the building is partially visible, and there is a small outdoor seating area with a potted plant. The sky is clear, and the lighting in the image is natural, with no artificial lights that are bright enough to overpower the ambient light. The overall style of the image is a candid, unposed photograph, capturing a typical urban street scene with a focus on the road closure. There are no visible people engaging with the road closure, and the photo seems to have been taken during the daytime.

VisVM Guided
The image depicts an urban street scene with a focus on a section of the road that has been cordoned off with yellow police tape. The tape is stretched across the street, indicating that the area is restricted or under investigation. Behind the tape, there are several orange traffic cones placed on the pavement, further emphasizing the restricted area. In the background, there is a building with a sign that reads \"T&G NAILS.\" The street is lined with trees and street lamps, and there are parked cars along the side of the road. The sky is overcast, and the lighting suggests it might be late afternoon or early evening. There are no people visible in the immediate vicinity of the cordoned-off area, and the overall atmosphere of the image is quiet and still. The presence of the police tape and cones suggests that there may have been an incident that required the area to be cordoned off, but the exact nature of the incident is not visible in the image. The style of the image is a standard, non-stylized photograph that captures a moment in an urban environment. There are no visible texts other than the \"T&G NAILS\" sign.

CLIP-PRM Guided
The image depicts an urban street scene with a focus on traffic control measures. The street is cordoned off with yellow caution tape and orange traffic cones, indicating that there may be ongoing construction or maintenance work. There is a visible manhole cover with a sign that reads \"Caution: Electrical Work.\" The sign suggests that there is work being done on the city's electrical infrastructure. On the right side of the image, there are parked cars, and on the left, there is a commercial building with signage. The sky is overcast, and the lighting suggests it is either late afternoon or early evening. The overall atmosphere is one of an active urban environment with ongoing city maintenance or construction activities.

Greedy Decoding
The image depicts a street scene with a focus on a section of road that has been cordoned off with yellow police tape. The tape is attached to several orange traffic cones, which are placed in a way that blocks access to the cordoned-off area. The tape and cones are clearly visible and appear to be in place to restrict access to the area. In the background, there is a commercial building with a sign that reads \"T&G NAILS.\" The building has a green awning and a storefront with a display of plants. There are also other businesses visible in the distance, including a restaurant with outdoor seating. The street is lined with trees and streetlights, and there are parked cars along the side of the road. The sky is overcast, suggesting it might be a cloudy day. The overall style of the image is a standard, non-stylized photograph that captures a real-life urban scene.

Figure 18: Qualitative comparison example 16

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately claim that ViMaR introduces a two-stage inference framework combining temporal-difference value learning with margin-based reward adjustment to improve both efficiency and output fidelity in VLM captioning. The paper explicitly claims over 4× speedup compared to VisVM, 64% human preference over VisVM outputs, and consistent cross-model generalization to stronger unseen models (LLaVA-OneVision-Qwen2-7B and Qwen2.5-VL-3B). All claims are directly supported by experimental results in Section 4, particularly Tables 1, 2 and Figures 1-2.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in Section 5. The authors acknowledge that while their two-stage approach achieves superior performance, the method relies on CLIP-ViT as the process reward model, which may not perfectly capture all aspects of visual grounding. The paper also notes computational constraints were evaluated on single NVIDIA RTX A6000 GPU, and inference times are measured under specific batch configurations that may vary with different hardware setups. Additionally, the method requires multiple temperature sampling (N=5, K=6) which, while efficient, still incurs overhead compared to greedy decoding.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [No]

Justification: The paper is empirical and architectural in nature, not theoretical. It does not present formal theorems, lemmas, or mathematical proofs. The work builds on established temporal-difference learning (Sutton, 1988) and applies it to VLM captioning without introducing new theoretical results requiring formal proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides comprehensive reproducibility information: (1) Architecture details—ViMaR is built on LLaVA-Next-Mistral-7B with a linear value head attached to the penultimate transformer layer; (2) Training data—792K triplets derived from 23K COCO 2017 images paired with LLaVA-150K prompts (732K train, 60K validation); (3) Hyperparameters—margin threshold =0.16 (justified in Appendix A.4 via percentile analysis), discount factor , learning objectives specified in Eq. 1-2; (4) Implementation details in Section 3.1 and Appendix A.3; (5) Inference parameters—N=5 temperatures 0.1, 0.3, 0.5, 0.7, 0.9, K=6 samples per temperature; (6) Evaluation metrics—CHAIR, MMHal, and seven visual comprehension benchmarks clearly defined; (7) Hardware—NVIDIA RTX A6000 GPU with fp16 mixed precision training. Code availability is promised for reproducibility.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g.,

to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets used are publicly available: COCO 2017 (train and validation splits), LLaVA-150K prompts, and standard benchmarks (MM-Vet, MMBench, MMMU, MathVista, CVBench, LLaVA-Wild, MMStar) documented with their sources. The authors explicitly state "Code: https://github.com/ankan8145/ViMaR" in the paper header. Training and inference procedures are fully specified with exact hyperparameters, loss functions (Eq. 2), and architectural modifications enabling reproduction. The paper provides sufficient detail (Section 3, Algorithm 1, Appendix A.3-A.6) to implement ViMaR from scratch and reproduce reported results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper comprehensively specifies: (1) Data splits—732K training triplets, 60K validation triplets, 1,000 images for main evaluation (Section 4.1), 500 for hallucination evaluation (Section 4.1.2), with COCO Train2017/Val2017 sources clearly identified; (2) Hyperparameters—learning rate, discount factor , margin threshold =0.16 with data-driven justification (Appendix A.4 analyzing CLIP score distribution); (3) Training procedure—4 epochs, batch size 16 per device (Appendix A.3); (4) Optimizer and precision—implied by LLaVA-Next architecture, fp16 mixed precision confirmed in Appendix A.3; (5) Temperature selection—five decoding temperatures 0.1, 0.3, 0.5, 0.7, 0.9 with ablation analysis (Appendix A.5 Table 3); (6) Reward formulation justification—Appendix A.6 compares alternative reward functions with empirical validation. These details enable reproducibility and inform methodological choices.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The paper reports statistical significance through multiple complementary approaches: (1) Human evaluation (Section 4.1)—blind pairwise comparison with 300 image-prompt pairs, reporting win rates with multiple baselines (64.0% vs VisVM, 65.3% vs CLIP-PRM, 66.0% vs BoN, 69.7% vs greedy); (2) Automated evaluation—GPT-4o-based pairwise comparisons (Figure 1b) showing consistent preferences (49.3%-73.8% across baselines); (3) Hallucination metrics (Table 1)—quantitative comparisons with measurable reductions in $CHAIR_I$, $CHAIR_S$, and MMHal metrics; (4) Self-training results (Table 2)—performance improvements across multiple benchmarks with concrete percentage gains (15.87% average improvement). While error bars are not displayed in plots, the large evaluation set (1,000+ images) and multiple evaluation metrics provide robust statistical grounding.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies: (1) Hardware—NVIDIA RTX A6000 GPU with 48GB VRAM (Appendix A.3); (2) Training configuration—4 epochs with batch size 16 per device, using mixed-precision fp16 and gradient checkpointing to optimize memory; (3) Inference timing—detailed in Table 1 with average inference time per sample (108s for ViMaR vs 462s for VisVM, 668s for BoN, 62s for greedy), demonstrating computational efficiency and enabling resource planning; (4) Framework—accelerate library with distributed training capabilities noted (Appendix A.3). These specifications enable practitioners to assess hardware requirements and deployment feasibility for their settings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics. No personally identifiable information or private data was collected or used. All training data (COCO 2017, LLaVA-150K) are publicly available datasets. The evaluation uses standard public benchmarks without sensitive attributes. The paper does not make claims about demographic fairness or intentionally suppress negative results.

The work focuses on improving caption quality and reducing hallucinations—inherently beneficial objectives without documented harm to individuals or groups.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification:The paper implicitly addresses broader impacts: (1) Positive impacts—reducing hallucinations in VLM outputs improves reliability for real-world applications (medical imaging, accessibility descriptions); enhanced visual grounding supports more accurate image understanding across domains. (2) Potential concerns—while not explicitly detailed, the margin-based reward adjustment and value-guided decoding represent a mechanism for shaping model behavior, which could potentially encode biases present in CLIP-ViT embeddings if training data contains skewed visual representations. (3) Self-training implications—using model-generated captions for further training (Section 4.2) creates feedback loops that could amplify initial biases if present. The paper's focus on factual accuracy and visual grounding mitigates some risks but acknowledges no explicit safeguards against these potential negative outcomes.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [No]

Justification: The paper does not describe explicit safeguards for responsible release. While all datasets used are publicly available and pose no privacy risks, the paper does not propose usage guidelines, restrict deployment contexts, or implement safety filters for the released code or value model. The work does not address potential misuse scenarios such as generating misleading captions at scale or using the value model to adversarially manipulate outputs. However, the focus on hallucination reduction and visual grounding inherently promotes beneficial use cases without obvious dual-use concerns.

Guidelines:

- The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly credits all external assets: (1) Datasets—COCO 2017 (Lin et al.), LLaVA-150K (Liu et al.), and benchmark datasets (MM-Vet, MMBench, MMMU, MathVista, CVBench, LLaVA-Wild, MMStar) are cited with references; (2) Models—LLaVA-Next-Mistral-7B, LLaVA-OneVision-Qwen2-7B, and Qwen2.5-VL-3B are cited from their respective papers; (3) CLIP-ViT encoder properly cited (Radford et al.); (4) Temporal-difference learning framework credited to Sutton (1988). All citations appear in the References section with full bibliographic details. The use of publicly available datasets and pre-trained models complies with academic research purposes and respective licensing terms.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new pre-trained model checkpoints. ViMaR is primarily an inference-time decoding algorithm and strategy, not a novel foundation model or independent asset. The value model component is lightweight (a linear value head attached to LLaVA-Next-Mistral-7B's penultimate layer) and is trained as part of the methodology, but trained checkpoints are not provided separately. The paper promises code release via GitHub (https://github.com/ankan8145/ViMaR) which enables reproducibility of the training procedure and inference algorithm, but model weights/checkpoints are not mentioned as being released. Therefore, while the code and reproducibility details are comprehensive (Sections 3.1, 3.2, Appendix A.3-A.6, Algorithm 1), no new standalone model assets requiring asset documentation are distributed.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [No]

Justification: The paper includes human evaluation (Section 4.1, Appendix A.1) but without complete documentation. The paper describes the evaluation procedure: blind pairwise comparison over 300 image-prompt pairs where annotators select preferred captions without knowing source model identity. However, the paper does not include: (1) full text of instructions given to human annotators, (2) screenshots or interface screenshots showing how annotators performed the task, (3) any compensation details (payment, incentives, or voluntary participation). The evaluation is in-house rather than crowdsourced, which explains the minimal disclosure, but the checklist criteria require such documentation. A more complete submission would include annotator instructions as supplementary material and clarify the evaluation setup (internal team members vs. external annotators).

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: The paper does not mention IRB approval or review. The human evaluation involves annotators performing caption comparison tasks, which pose minimal risk (non-invasive, no sensitive data collection, no personal information gathered). However, the paper does not explicitly state that IRB approval was obtained or that participant consent procedures were followed. Best practice would include mention of IRB exemption or approval status, though the low-risk nature of the study may not require formal review depending on institutional requirements. The paper could strengthen compliance by explicitly addressing this point.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper's core methodology relies heavily on LLMs as integral components: (1) Base VLM—LLaVA-Next-Mistral-7B is the primary policy model for caption generation (Section 3.2, Algorithm 1), providing the generation distribution from which candidates are sampled; (2) Training data—the training triplets (y_i, y_{i+1}, I) are generated using LLaVA-Next with greedy decoding and temperature-controlled sampling to ensure diversity (Section 3.1); (3) Cross-model evaluation—ViMaR's value model is tested on stronger LLMs (LLaVA-OneVision-Qwen2-7B, Qwen2.5-VL-3B) demonstrating generalization (Table 2, Section 4.2); (4) Self-training evaluation—fine-tuning experiments use LLaVA-Next-Mistral-7B as the base model (Section 4.2). The LLM is not merely

used for writing or formatting; it is central to method development, training data generation, and evaluation. This usage is clearly described throughout the methodology.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.