Online Self-Calibration Against Hallucination in Vision-Language Models

Minghui Chen[1,2*] , **Chenxu Yang**[1,2*] , **Hengjie Zhu**[1,2] , **Dayan Wu**[1†] , **Qingyi Si**[3] , **Zheng Lin**[1]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]JD.COM
{chenminghui, yangchenxu, zhuhengjie, wudayan, linzheng}@iie.ac.cn

Abstract

Large Vision-Language Models (LVLMs) often suffer from hallucinations, generating descriptions that include visual details absent from the input image. Recent preference alignment methods typically rely on supervision distilled from stronger models such as GPT. However, this offline paradigm introduces a Supervision-Perception Mismatch: the student model is forced to align with fine-grained details beyond its perceptual capacity, learning to guess rather than to see. To obtain reliable self-supervision for online learning, we identify a Generative-Discriminative Gap within LVLMs, where models exhibit higher accuracy on discriminative verification than open-ended generation. Leveraging this capability, we propose **O**nline **S**elf-**CA**lib**R**ation (OSCAR), a framework that integrates Monte Carlo Tree Search with a Dual-Granularity Reward Mechanism to construct preference data and iteratively refines the model via Direct Preference Optimization. Extensive experiments demonstrate that OSCAR achieves state-of-the-art performance on hallucination benchmarks while preserving general multimodal capabilities.

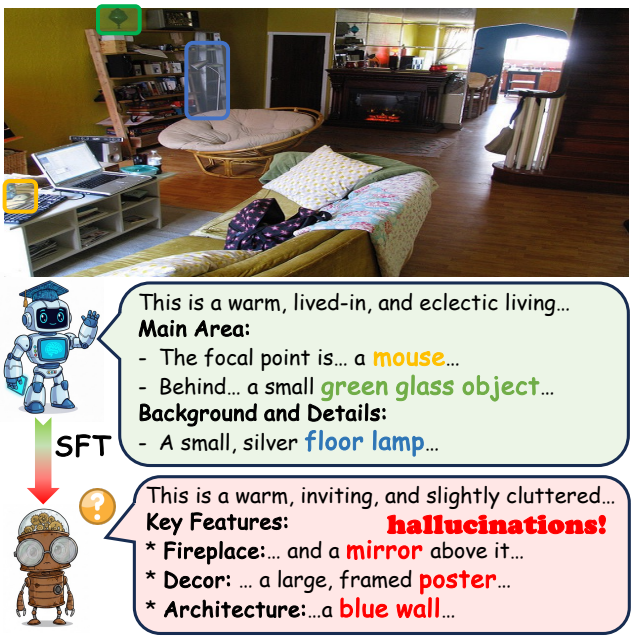

Figure 1: Supervision-Perception Mismatch. Offline supervision from a stronger teacher model forces the student to describe visual details it cannot reliably perceive, resulting in hallucinations.

1 Introduction

Large Vision-Language Models (LVLMs) [Liu *et al.*, 2023c; Zhu *et al.*, 2023; Bai *et al.*, 2023; Dai *et al.*, 2023; Lu *et al.*, 2024] integrate visual encoders with pre-trained Large Language Models (LLMs), achieving strong performance across a wide range of multimodal tasks, from image captioning to visual reasoning. However, these models frequently suffer from hallucinations [Rawte *et al.*, 2023; Bai *et al.*, 2024; Liu *et al.*, 2024], generating content that is inconsistent with or absent from the visual input, such as fabricating nonexistent objects, misinterpreting spatial relationships, or incorrectly describing object attributes. This limitation poses a significant barrier to deploying LVLMs in safety-critical domains like autonomous driving, medical imaging, and robotics, where factual grounding is essential.

Recent efforts to mitigate hallucinations have largely relied on preference alignment techniques, including Reinforcement Learning from Human Feedback (RLHF) [Sun *et al.*, 2024] and DPO [Rafailov *et al.*, 2023]. These methods typically construct preference datasets using human annotations [Sun *et al.*, 2024; Gunjal *et al.*, 2024] or responses distilled from stronger, proprietary models such as GPT [Achiam *et al.*, 2023; Liu *et al.*, 2023a; Li *et al.*, 2023a; Zhou *et al.*, 2024]. While effective to some extent, we argue that this reliance on offline supervision introduces a fundamental Supervision-Perception Mismatch. As illustrated in Fig. 1, teacher models with superior visual capabilities tend to produce highly detailed descriptions that capture subtle visual elements, such as small objects or fine-grained attributes that are difficult to discern. When a weaker student model is supervised by such data, it is compelled to reproduce these fine-grained details that lie beyond its perceptual capacity. Unable to ground these descriptions in actual visual features, the student instead resorts to exploiting language priors and statistical shortcuts, generating different but equally ungrounded

content. In essence, the model learns to guess rather than to see. This observation is further validated by our pilot experiments in Section 3.1. We find that fine-tuning LLaVA-1.5-7B on teacher-distilled data paradoxically increases hallucination rates, with performance degrading further as more training data is added. These findings motivate a shift toward online learning, where training data respects the model's intrinsic perceptual boundaries.

A natural question arises: can we obtain high-quality, truthful supervision from a model that is itself prone to hallucination? To address this, we identify a notable Generative-Discriminative Gap within LVLMs. Our analysis shows that while models often yield to language inertia during autoregressive generation, they exhibit considerably higher accuracy on discriminative tasks, such as verifying whether a specific object exists in an image. This gap arises because discriminative verification, by explicitly conditioning on a specific query, reduces the influence of unconstrained language priors that dominate open-ended generation. This finding suggests that LVLMs possess a latent capacity for self-verification that remains underutilized during generation.

While the Generative-Discriminative Gap addresses the question of where to obtain reliable supervision, another key challenge lies in how to construct high-quality training data. Standard decoding strategies like greedy or beam search optimize locally, presenting two limitations. First, some tokens that appear safe at the current step may carry high risk of inducing hallucinations in subsequent generation, a cascading effect that local optimization cannot foresee. Second, greedily selecting branches with the lowest immediate hallucination rate at each step may compromise overall response quality in terms of logical consistency and fluency. To overcome these limitations, we integrate Monte Carlo Tree Search (MCTS) to explore the generation space more strategically [Xie *et al.*, 2024a; Tian *et al.*, 2024; Zhang *et al.*, 2024]. We design a Dual-Granularity Reward Mechanism to guide the search. At the node level, we leverage the model's discriminative capability by prompting it to verify whether each generated sentence mentions objects absent from the image, using the probability of a negative response as the process reward. At the trajectory level, we employ a Gated Outcome Reward that evaluates response quality only if the complete response passes a faithfulness check, and returns zero otherwise.

Through MCTS backpropagation, terminal rewards propagate from leaf nodes to the root, enabling the model to identify generation trajectories that balance visual faithfulness with descriptive richness. Building on these insights, we propose **O**nline **S**elf-**CA**lib**R**ation (**OSCAR**), a unified framework for online preference learning. Specifically, OSCAR extracts preference pairs from the MCTS tree at two granularities: global path comparison selects complete trajectories with the highest and lowest cumulative values, while sibling comparison pairs nodes along the optimal path with their worst-performing siblings. These preference pairs are then used to update the model via DPO [Rafailov *et al.*, 2023]. At each iteration, the updated model generates new preference data through MCTS, ensuring the training distribution evolves alongside the model's capabilities and enabling continuous self-improvement. Our contributions are as follows:

- We empirically demonstrate the necessity of online preference learning that respects the model's intrinsic perceptual boundaries, showing that offline distillation from stronger teachers can unexpectedly exacerbate hallucinations.
- We propose OSCAR, a novel training paradigm that exploits the Generative-Discriminative Gap. By integrating MCTS with a Dual-Granularity Reward Mechanism, OSCAR enables lookahead to suppress early tokens that risk inducing downstream hallucinations.
- Extensive experiments show that OSCAR achieves state-of-the-art performance on hallucination benchmarks while simultaneously preserving general multimodal capabilities.

2 Related Work

2.1 Hallucination in LVLMs

Large Vision-Language Models (LVLMs) frequently suffer from hallucinations, generating content that is inconsistent with the visual input [Rawte *et al.*, 2023; Bai *et al.*, 2024]. Various approaches have been proposed to mitigate this issue, including enhancing dataset quality [Liu *et al.*, 2023b; Gunjal *et al.*, 2024; Li *et al.*, 2023a], manipulating the decoding process [Leng *et al.*, 2024; Huang *et al.*, 2024; Suo *et al.*, 2025], leveraging external models for post-hoc correction [Yin *et al.*, 2024; Zhou *et al.*, 2024], and preference optimization [Xie *et al.*, 2024b; Wang *et al.*, 2025].

2.2 Self-Improvement for LLMs

Self-improvement methods enable models to enhance their capabilities using self-generated feedback, reducing reliance on external annotations [Huang *et al.*, 2023; Yuan *et al.*, 2024; Hu *et al.*, 2024; Sun *et al.*, 2025]. In the vision-language domain, recent works such as STIC [Deng *et al.*, 2024] and SIMA [Wang *et al.*, 2025] have explored self-improvement for hallucination mitigation. However, these methods typically construct preference data via simple sampling or beam search, which fails to account for the cascading nature of hallucinations. We instead integrate MCTS with a Dual-Granularity Reward Mechanism, enabling lookahead to suppress tokens that risk inducing downstream hallucinations.

2.3 Monte Carlo Tree Search

Monte Carlo Tree Search (MCTS) has emerged as a powerful technique for enhancing reasoning in LLMs, inspired by its success in game-playing agents [Silver *et al.*, 2016]. Recent works have adapted MCTS to guide text generation by simulating future trajectories and backpropagating rewards, achieving improvements in mathematical reasoning [Xie *et al.*, 2024a; Tian *et al.*, 2024; Zhang *et al.*, 2024] and task planning [Hao *et al.*, 2023; Li and Ng, 2025; Li *et al.*, 2025]. In the context of hallucination mitigation, we are the first to leverage MCTS for preference data construction, enabling lookahead to suppress locally plausible tokens that risk inducing downstream hallucinations.

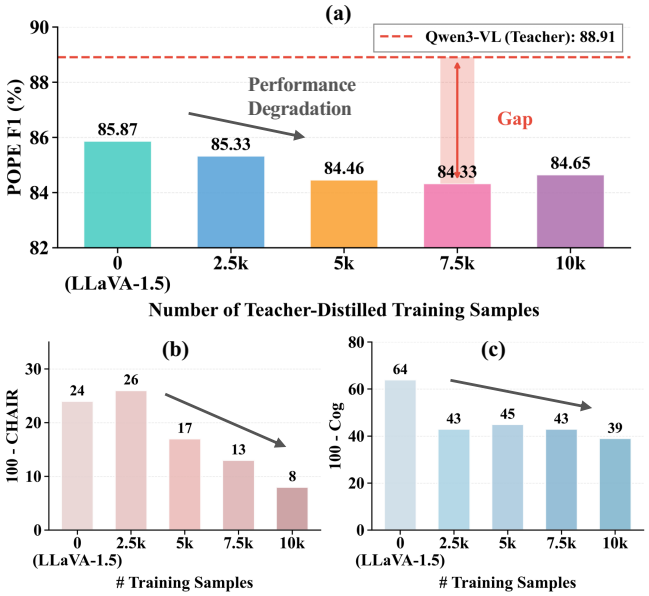

Figure 2: Performance comparison before and after supervised fine-tuning with stronger teacher-distilled data.

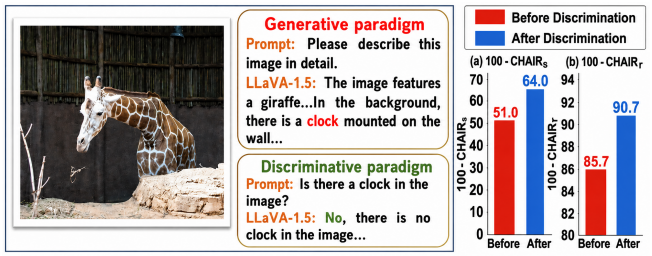

Figure 3: Illustration of the Generative-Discriminative Gap.

3 Observations and Motivations

3.1 Supervision-Perception Mismatch

Recent preference alignment methods predominantly rely on supervision signals distilled from stronger, proprietary models such as GPT [Achiam *et al.*, 2023]. We hypothesize that this reliance on offline supervision introduces a fundamental *Supervision-Perception Mismatch*: when a target model with limited visual perception capacity is trained on data generated by a teacher model with superior perceptual abilities, the student is compelled to align with fine-grained visual details that exceed its intrinsic perceptual capabilities. Consequently, the model may learn to minimize training loss by exploiting language priors and statistical shortcuts rather than grounding its outputs in visual features. As illustrated in Fig. 1, teacher models with superior visual capabilities tend to capture subtle visual elements that weaker models cannot reliably perceive. For instance, the teacher model identifies a "mouse" near the laptop, a "green glass object" in the corner, and a "floor lamp" in the background. When supervised by such data, the student model fails to ground these details in actual visual features, instead generating different but equally ungrounded content, such as hallucinating a "mirror" above the fireplace, a "poster", and a "blue wall" that does not exist.

To empirically validate this hypothesis, we employ Qwen3-VL-8B-Instruct [Yang *et al.*, 2025] to generate detailed image descriptions on randomly sampled images from the LLaVA-150k dataset, and use these teacher-generated descriptions to fine-tune LLaVA-1.5-7B with varying data scales (2.5k, 5k, 7.5k, and 10k samples). As shown in Fig. 2 (a), despite Qwen3-VL achieving 88.91% on POPE F1, the fine-tuned LLaVA models consistently underperform the original baseline (85.87%), with performance degrading further as more training data is added. Similar trends are observed on the AMBER benchmark, where both CHAIR and Cog scores deteriorate after fine-tuning (Fig. 2b-c). This counterintuitive finding confirms that *offline* supervision can paradoxically exacerbate hallucinations, motivating our shift toward *online* learning.

3.2 Generative-Discriminative Gap

To explore whether a hallucination-prone model can still provide useful self-supervision, we investigate its behavior across different inference paradigms. Our analysis reveals a notable *Generative-Discriminative Gap*: while models frequently yield to language inertia during autoregressive generation, where plausible linguistic patterns overshadow visual grounding, they exhibit improved accuracy when tasked with discriminative verification against visual evidence. Fig. 3 illustrates this gap with a concrete example. When prompted to describe an image in detail, LLaVA-1.5 hallucinates a "clock mounted on the wall" that does not exist. However, when the same model is explicitly asked "Is there a clock in the image?", it correctly answers "No". This discrepancy suggests that discriminative verification, by conditioning on a specific query, reduces the influence of unconstrained language priors that dominate open-ended generation.

To quantify this gap, we conduct the following analysis. We first generate image captions using LLaVA-1.5-7B on 500 randomly sampled images from the COCO dataset and compute the CHAIR metrics. For each hallucinated object x detected in the generated captions, we construct a discriminative query: "Is there a/an x in the image?" and prompt the same model to respond with "Yes" or "No". If the model correctly answers "No", we remove this hallucinated object from the caption and recompute the CHAIR metrics. As shown in Fig. 3, this simple self-verification procedure reduces CHAIR_S from 49.0% to 36.0% and CHAIR_I from 14.3% to 9.3%, confirming that LVLMs possess a latent capacity for self-verification that remains underutilized during standard generation. This finding motivates our approach: by leveraging the model's discriminative ability to curate training data, we can align the granularity of generated descriptions with the model's intrinsic perceptual capabilities while improving factual accuracy.

4 Methodology

Building upon the insights from Section 3, we present **O**nline **S**elf-**CA**lib**R**ation (OSCAR) , a framework illustrated in Fig. 4 that leverages the model's capability to con-

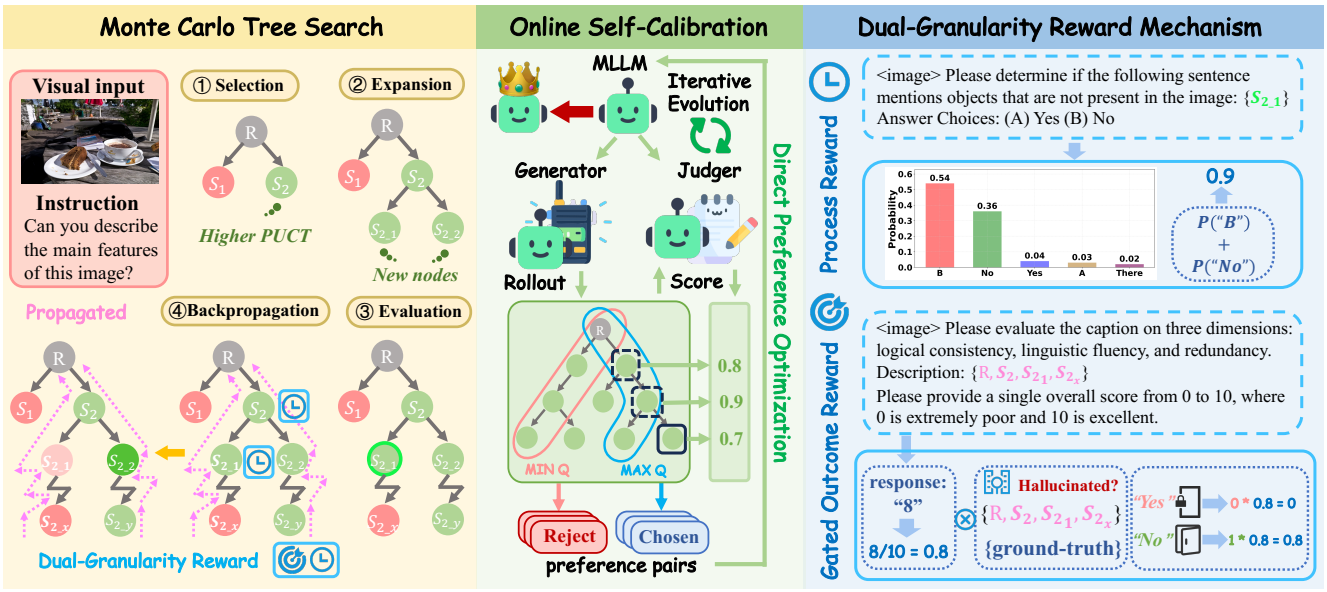

Figure 4: Overview of Online Self-CAlibRation (OSCAR). Left: Monte Carlo Tree Search explores the generation space through selection, expansion, evaluation, and backpropagation. Middle: Preference pairs are extracted from the MCTS tree to refine the model via Direct Preference Optimization. Right: The Dual-Granularity Reward Mechanism combines a process reward that verifies each sentence, and a gated outcome reward that evaluates response quality only when the trajectory is hallucination-free.

struct high-quality online preference data and iteratively refines the model through DPO. In this section, we first detail our MCTS-guided generation with a dual-granularity reward mechanism (§4.2), and then describe the preference learning procedure (§4.3).

4.1 LVLM Inference

A Large Vision-Language Model (LVLM) $\mathcal{M}_\theta$ with parameters θ takes as input a visual image $\mathbf{v}$ and a textual instruction $\mathbf{q}$, and generates a textual response $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ in an autoregressive manner. Specifically, at each generation step t, the model computes the probability distribution over the vocabulary conditioned on the visual input, the textual instruction, and the previously generated tokens:

$$p(y_t|\mathbf{v}, \mathbf{q}, y_{<t}; \theta) = \text{Softmax}(\ell_t), \tag{1}$$

where ℓ_t denotes the logit vector for the next token y_t, and $y_{<t} = (y_1, \ldots, y_{t-1})$ represents the sequence of tokens generated before step t. The complete response is generated by sequentially sampling tokens from this distribution until the end-of-sequence token is produced.

4.2 MCTS-Guided Generation

A key challenge in constructing faithful training data is that standard decoding strategies such as greedy or beam search optimize locally, which suffers from two limitations. First, tokens that seem acceptable at the current step may induce hallucinations in later generation, a long-term risk that local optimization cannot anticipate. Second, prioritizing branches with minimal immediate hallucination at each step may degrade overall response quality, sacrificing logical consistency and fluency. To overcome these limitations, we integrate Monte Carlo Tree Search (MCTS) into the generation process. By simulating future trajectories and backpropagating terminal rewards, MCTS enables the model to evaluate the long-term value of each token, identifying generation paths that balance visual faithfulness with descriptive richness.

MCTS Procedure. We decompose the generation process into sentence-level steps, where each node in the search tree represents a partial response. Formally, let $s_t = (\mathbf{v}, \mathbf{q}, a_1, a_2, \ldots, a_{t-1})$ denote the state at step t, where a_i represents the i-th generated sentence. An action a_t corresponds to generating a complete sentence, delimited by terminal punctuation marks. Each MCTS iteration consists of four phases:

- **Selection.** Starting from the root node, we traverse the tree by selecting child nodes according to the PUCT (Predictor + Upper Confidence bounds applied to Trees) criterion [Rosin, 2011]:

$$a^* = \arg\max_a \left[Q(s,a) + c_{\text{puct}} \cdot p(a|s) \cdot \frac{\sqrt{N(s)}}{1 + N(s,a)} \right], \tag{2}$$

  where $Q(s,a)$ denotes the action value, $p(a|s) = \pi_\theta(a|s)/|a|^\lambda$ is the length-normalized policy probability with penalty λ, $N(s)$ and $N(s,a)$ are visit counts, and c_{puct} controls the exploration-exploitation trade-off.

- **Expansion.** At a leaf node, we sample K candidate sentences from the policy π_θ using temperature sampling. To ensure diversity, we filter candidates with embedding similarity exceeding a threshold τ_{sim}.

- **Evaluation.** Each expanded node receives a value score

through our *Dual-Granularity Reward Mechanism*, detailed below.

- **Backpropagation.** After evaluation, statistics are propagated from the leaf node back to the root. The Q-value and state value are updated as:

$$Q(s_t, a) = r(s_t, a) + \gamma \cdot V(s_{t+1}), \tag{3}$$

$$V(s_t) = \frac{\sum_a N(s_{t+1}) \cdot Q(s_t, a)}{\sum_a N(s_{t+1})}, \tag{4}$$

where $r(s_t, a) = \text{value}(s_{t+1}) - \text{value}(s_t)$ represents the immediate reward and γ is the discount factor.

Dual-Granularity Reward Mechanism. Central to our approach is a reward mechanism that combines node-level process supervision with trajectory-level outcome evaluation. This design enables the search to identify early tokens that, while locally plausible, carry high risk of inducing downstream hallucinations.

Process Reward (Node-Level). For each generated sentence a_t, we employ the model's discriminative capability to assess whether the sentence mentions objects absent from the image. Specifically, we construct a verification prompt:

> `<image>` Please determine if the following sentence mentions objects that are not present in the image: {`sentence`}
> Answer Choices: (A) Yes (B) No

The process reward r_{proc} is computed as the probability that the model judges the sentence as hallucination-free:

$$r_{\text{proc}}(a_t) = p_\theta\big(\text{“No”} \mid \mathbf{v}, \mathcal{P}_{\text{proc}}(a_t)\big), \tag{5}$$

where $\mathcal{P}_{\text{proc}}(a_t)$ denotes the verification prompt instantiated with the candidate sentence a_t.

Gated Outcome Reward (Trajectory-Level). To evaluate the quality of a complete trajectory, we perform a greedy rollout from the current state to a terminal state, yielding a complete response $\mathbf{y}_{\text{rollout}}$. The outcome reward incorporates a *gating mechanism* that enforces strict faithfulness requirements. First, we assess whether the complete response contains any hallucinated content through a faithfulness check. Specifically, we extract all objects mentioned in the generated description and compare them against the ground-truth objects provided by the dataset. Specifically, we extract all object nouns from the generated description and map them to canonical COCO category names using a predefined synonym dictionary. If any mapped object does not appear in the ground-truth object set, the response is considered to contain hallucinations. The gating function is defined as:

$$g(\mathbf{y}_{\text{rollout}}) = \mathbb{1}[\mathcal{O}(\mathbf{y}_{\text{rollout}}) \subseteq \mathcal{O}_{\text{gt}}], \tag{6}$$

where $\mathcal{O}(\mathbf{y}_{\text{rollout}})$ denotes the set of canonical object names extracted from the generated response, and $\mathcal{O}_{\text{gt}}$ denotes the ground-truth object set. Second, for trajectories that pass the gate, we assess the response quality along dimensions of logical consistency, linguistic fluency, and redundancy:

> `<image>` Please evaluate the following caption on three dimensions: logical consistency, linguistic fluency, and redundancy.
> Caption: {$\mathbf{y}_{\texttt{rollout}}$}
> Please provide a single overall score from 0 to 10, where 0 is extremely poor and 10 is excellent.

Let $\text{score}_{\text{quality}} \in [0, 10]$ denote the extracted quality score. The gated outcome reward is defined as:

$$r_{\text{out}}(\mathbf{y}_{\text{rollout}}) = \begin{cases} \text{score}_{\text{quality}}/10 & \text{if } g(\mathbf{y}_{\text{rollout}}) = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The final value assigned combines both granularities:

$$\text{value}(s_t, a_t) = r_{\text{proc}}(s_t, a_t) + r_{\text{out}}(\mathbf{y}_{\text{rollout}}). \tag{8}$$

Through MCTS backpropagation, this trajectory-level reward signal propagates from leaf nodes to the root, elevating the estimated value of early tokens that lead to faithful and high-quality completions.

4.3 Iterative Preference Learning

Preference Pair Extraction. We extract preference pairs from the MCTS tree at two levels of granularity. For *global path comparison*, we identify the complete path with the highest cumulative Q-value as the chosen response $\mathbf{y}^+$ and the one with the lowest Q-value as the rejected response $\mathbf{y}^-$:

$$\mathbf{y}^+ = \arg\max_{\mathbf{y} \in \mathcal{T}} Q(\mathbf{y}), \quad \mathbf{y}^- = \arg\min_{\mathbf{y} \in \mathcal{T}} Q(\mathbf{y}), \tag{9}$$

where $\mathcal{T}$ denotes the set of all complete trajectories in the tree. For *sibling comparison*, we traverse each depth along the optimal path and pair the selected node with its worst-performing sibling, if their Q-value difference exceeds a threshold δ_Q:

$$(\mathbf{y}_d^+, \mathbf{y}_d^-) = (s_{<d} \oplus a_d^*,\ s_{<d} \oplus a_d^{\text{worst}}),\ \text{if } Q(a_d^*) - Q(a_d^{\text{worst}}) \geq \delta_Q, \tag{10}$$

where $s_{<d}$ denotes the partial response up to depth d, and $\oplus$ denotes concatenation. This step-wise comparison enables the extraction of multiple preference pairs from a single MCTS tree, maximizing the utilization of information accumulated during the search process.

DPO Training. Given the preference dataset $\mathcal{D} = \{(\mathbf{v}_i, \mathbf{q}_i, \mathbf{y}_i^+, \mathbf{y}_i^-)\}_{i=1}^N$ constructed via MCTS, we update the model using Direct Preference Optimization (DPO) [Rafailov *et al.*, 2023]. The DPO objective directly optimizes the policy to prefer chosen responses over rejected ones:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(\mathbf{v}, \mathbf{q}, \mathbf{y}^+, \mathbf{y}^-) \sim \mathcal{D}} \left[\log \sigma\left(\beta \cdot h_\theta(\mathbf{y}^+, \mathbf{y}^-)\right)\right], \tag{11}$$

where $\sigma(\cdot)$ is the sigmoid function, β is a temperature parameter, and:

$$h_\theta(\mathbf{y}^+, \mathbf{y}^-) = \log \frac{\pi_\theta(\mathbf{y}^+|\mathbf{v}, \mathbf{q})}{\pi_{\text{ref}}(\mathbf{y}^+|\mathbf{v}, \mathbf{q})} - \log \frac{\pi_\theta(\mathbf{y}^-|\mathbf{v}, \mathbf{q})}{\pi_{\text{ref}}(\mathbf{y}^-|\mathbf{v}, \mathbf{q})}. \tag{12}$$

Here, π_{ref} denotes the reference policy, initialized as the model checkpoint before the current iteration. The training proceeds iteratively: at each iteration m, we use the current

Method	Generative Task							Discriminative Task		
	Object HalBench		AMBER-Gen.				MM-VET	AMBER-Dis.		POPE
	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	CHAIR↓	Cover↑	Hal↓	Cog↓	Overall↑	Acc↑	F1↑	F1↑
InstructBLIP[Dai *et al.*, 2023]	42.3	**7.9**	9.4	51.8	40.5	4.8	25.7	74.3	79.9	78.56
MiniGPT-4[Zhu *et al.*, 2023]	31.4	11.1	13.6	**63.0**	65.3	11.3	22.1	63.6	64.7	61.50
mPLUG-Owl2[Ye *et al.*, 2024]	54.4	12.0	10.6	52.2	39.9	4.5	**37.6**	75.6	78.5	86.20
LLaVA-1.5-7B [Liu *et al.*, 2023c]	49.0	14.3	7.6	49.6	31.2	3.6	32.5	72.2	75.5	85.87
+STIC [Deng *et al.*, 2024]	-	-	7.6	52.1	35.8	4.4	31.8	71.6	74.2	-
+POVID [Zhou *et al.*, 2024]	33.6	9.0	5.0	50.1	28.6	3.0	31.7	71.9	74.7	**86.90**
+SIMA [Wang *et al.*, 2025]	40.9	10.4	6.4	47.4	26.1	3.2	31.6	73.4	76.4	-
+Self-Rewarding	38.4	11.2	6.8	48.2	27.5	3.0	32.8	73.1	76.8	85.93
+OSCAR (Iter1)	32.0	9.7	5.6	46.0	22.1	2.1	32.9	74.4	78.3	86.04
+OSCAR (Iter2)	28.6	9.0	5.1	45.4	19.4	1.9	33.8	75.3	79.4	86.07
+OSCAR (Iter3)	**27.6**	8.2	**4.5**	45.7	**17.2**	**1.6**	34.6	**75.8**	**80.2**	86.22
LLaVA-1.5-13B [Liu *et al.*, 2023c]	44.8	11.8	6.4	**50.9**	30.3	3.1	**37.6**	67.9	69.1	86.67
+Self-Rewarding	35.2	9.6	5.4	48.3	24.6	2.4	36.8	69.5	71.2	86.78
+OSCAR (Iter1)	16.4	5.8	3.3	45.8	13.3	0.9	37.4	71.9	74.1	86.89
+OSCAR (Iter2)	7.8	2.8	2.8	48.7	9.9	0.6	35.6	70.3	71.9	87.20
+OSCAR (Iter3)	**5.4**	**2.6**	**2.6**	47.6	**8.0**	**0.5**	36.5	**72.4**	**74.6**	**87.26**

Table 1: Comparison with state-of-the-art methods on hallucination benchmarks. We evaluate on both generative tasks and discriminative tasks. The best results are shown in **bold** and the second best results are underlined. ↓ indicates lower is better, ↑ indicates higher is better.

policy $\pi_\theta^{(m)}$ to construct new preference data via MCTS, then update the model to obtain $\pi_\theta^{(m+1)}$. This online paradigm ensures that the training distribution evolves alongside the model's improving capabilities, progressively tightening the alignment between generated content and the model's perceptual boundaries.

5 Experiments

5.1 Experimental Setup

Evaluation Benchmarks. We conduct evaluations on both generative and discriminative hallucination tasks. For the generative task, we evaluate on Object-HalBench [Rohrbach *et al.*, 2018], AMBER [Wang *et al.*, 2023], and MM-VET [Yu *et al.*, 2023]. For the discriminative task, we report results on AMBER [Wang *et al.*, 2023] and POPE [Li *et al.*, 2023b].

Baselines. We compare OSCAR with three categories of methods: (1) other open-source LVLMs, including InstructBLIP [Dai *et al.*, 2023], MiniGPT-4 [Zhu *et al.*, 2023], and mPLUG-Owl2 [Ye *et al.*, 2024]; (2) SoTA data-driven preference learning methods for hallucination mitigation, including STIC [Deng *et al.*, 2024], POVID [Zhou *et al.*, 2024], and SIMA [Wang *et al.*, 2025]; (3) a Self-Rewarding baseline that employs the same hallucination detection reward but constructs preference data via beam search instead of MCTS.

Implementation Details. We adopt LLaVA-1.5-7B and LLaVA-1.5-13B as base models. For preference data construction, we sample images and prompts from LLaVA-150k [Liu *et al.*, 2023c], yielding 120k preference pairs per iteration. The MCTS search is configured with $c_{\text{puct}} = 1.0$, length penalty $\lambda = 1.25$, and Q-value difference threshold $\delta_Q = 0.05$. For DPO training, we employ LoRA [Hu *et al.*, 2022] with rank 128 and $\alpha = 256$, using the Adam optimizer with a learning rate of 1×10^{-5} and temperature $\beta = 0.1$. The model is iteratively trained for 3 iterations.

5.2 Main Results

Overall Performance. Tab. 1 presents comprehensive comparisons between OSCAR and existing methods on hallucination benchmarks. Our method achieves state-of-the-art performance on both generative and discriminative tasks, demonstrating its effectiveness in mitigating hallucinations while preserving general multimodal capabilities.

Generative Task. On Object-HalBench, OSCAR substantially reduces hallucination metrics for LLaVA-1.5-7B, decreasing $CHAIR_S$ from 49.0 to 27.6 and $CHAIR_I$ from 14.3 to 8.2. These improvements significantly surpass prior methods such as POVID (33.6/9.0) and SIMA (40.9/10.4). On AMBER generative metrics, OSCAR achieves the lowest Hal score (17.2) and Cog score (1.6), indicating fewer hallucinated responses and reduced reliance on human-like cognitive shortcuts. Notably, on MM-VET, which evaluates general multimodal understanding, OSCAR improves the overall score from 32.5 to 34.6, demonstrating that our hallucination mitigation does not compromise response quality or descriptive richness. For the larger LLaVA-1.5-13B model, OSCAR yields even more substantial gains, achieving $CHAIR_S$ of 5.4 and $CHAIR_I$ of 2.6, which represent reductions of 87.9% and 78.0% respectively compared to the baseline.

Discriminative Task. On discriminative benchmarks, OSCAR also demonstrates consistent improvements. For AMBER discrimination, OSCAR improves accuracy from 72.2% to 75.8% and F1 score from 75.5% to 80.2% on LLaVA-1.5-7B. On POPE, OSCAR achieves an F1 score of 86.22%, comparable to POVID (86.90%), while significantly outperforming it on generative tasks. These results confirm that OSCAR enhances the model's visual grounding capability across both generative and discriminative paradigms.

Iterative Improvement. A key advantage of our approach is its ability to enable continuous self-improvement through iterative training. As shown in Tab. 1, performance improves progressively from Iter1 to Iter3 across the major-

Index	PR	GOR	MCTS	$CHAIR_s$	$CHAIR_i$	POPE
1	-	-	-	49.0	14.3	85.87
2	✓	-	-	46.7	13.8	86.01
3	✓	-	✓	45.6	13.5	86.00
4	-	✓	✓	44.0	12.6	86.03
5	✓	✓	✓	32.0	9.7	86.04

Table 2: Ablation study on key components of OSCAR. PR: Process Reward; GOR: Gated Outcome Reward.

Method	**CHAIR↓**	**Hal↓**	**Cog↓**
LLaVA-1.5-7B	7.6	31.2	3.6
+SFT (Qwen3-VL)	9.2	62.7	6.1
+SFT (LLaVA)	7.5	30.6	3.4
+SFT (OSCAR)	**4.5**	**15.4**	**1.4**

Table 3: Comparison of different training data sources on AMBER benchmark. All methods use 10k training samples.

ity of metrics. For LLaVA-1.5-7B, $CHAIR_S$ decreases from 32.0 (Iter1) to 28.6 (Iter2) and further to 27.6 (Iter3), while the AMBER Hal score drops from 22.1 to 19.4 to 17.2 over the same iterations. Similar trends are observed for LLaVA-1.5-13B, where $CHAIR_S$ decreases from 16.4 (Iter1) to 7.8 (Iter2) and finally to 5.4 (Iter3). This progressive enhancement validates our online learning paradigm: as the model improves, the quality of MCTS-generated preference data also increases, creating a virtuous cycle that enables sustained capability growth.

5.3 Analysis

Ablation Studies To analyze the contribution of each component in OSCAR, we conduct ablation studies on LLaVA-1.5-7B using a single iteration. Results are shown in Tab. 2. **Effect of Process Reward.** Comparing Index 4 and 5, adding the process reward substantially reduces $CHAIR_S$ from 44.0 to 32.0, demonstrating that node-level hallucination feedback is essential for fine-grained guidance during tree search. **Effect of Gated Outcome Reward.** Comparing Index 3 and 5, incorporating the gated outcome reward reduces $CHAIR_S$ from 45.6 to 32.0, ensuring trajectory-level faithfulness that complements the local process reward. **Effect of MCTS.** Comparing Index 2 and 5, integrating MCTS dramatically reduces $CHAIR_S$ from 46.7 to 32.0, confirming that lookahead search is crucial for identifying tokens that may induce downstream hallucinations. The full model significantly outperforms all partial configurations, indicating that the three components work synergistically.

Analysis of online Learning To validate the effectiveness of on-policy learning, we compare three training strategies using 10k samples: (1) SFT with Qwen3-VL-8B-Instruct distilled data, (2) SFT with LLaVA's own generated data, and (3) SFT with chosen samples from our OSCAR-constructed preference data. Results are shown in Tab. 3. SFT with Qwen3-VL distilled data increases CHAIR from 7.6 to 9.2 and Hal from 31.2 to 62.7, confirming the Supervision-Perception Mismatch discussed in Section 3.1. SFT with LLaVA's own generated data maintains similar performance but yields no improvement. In contrast, SFT with our OSCAR-constructed data substantially reduces CHAIR to 4.5, Hal to 15.4, and Cog to 1.4, demonstrating that our MCTS-guided data construction effectively leverages the model's discriminative capability to generate high-quality online training data.

Case Study Fig. 5 presents a qualitative comparison between LLaVA-1.5 and our OSCAR method. LLaVA-1.5 generates numerous hallucinated objects (highlighted in red) that

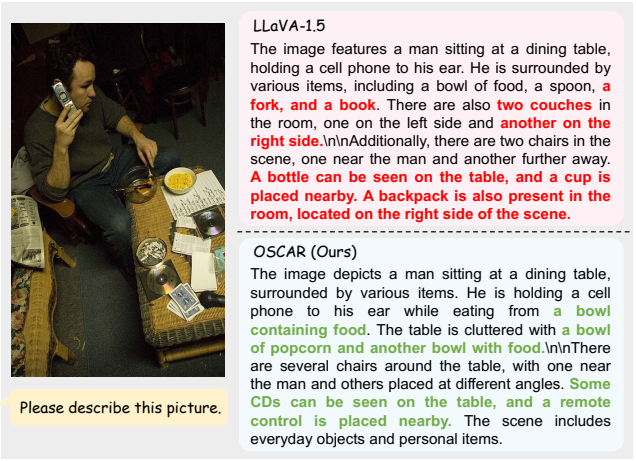

Figure 5: Qualitative comparison between LLaVA-1.5 and OSCAR (Ours). Hallucinated content is highlighted in red, while correct descriptions are shown in green. Our method generates fewer hallucinations.

are entirely absent from the image, while in contrast, our OSCAR method produces significantly fewer hallucinations. Moreover, the response generated by OSCAR exhibits better fluency and reduced redundancy, resulting in a more concise and coherent description. This demonstrates that our dual-granularity reward mechanism effectively balances visual faithfulness with overall response quality.

6 Conclusion

In this paper, we identified two key observations: a Supervision-Perception Mismatch in offline preference learning that unexpectedly worsen hallucinations, and a Generative-Discriminative Gap that provides reliable self-supervision signals. Building on these insights, we proposed Iterative Self-Calibration (OSCAR), which integrates MCTS with a Dual-Granularity Reward Mechanism for online preference learning. Experiments demonstrated that OSCAR achieves state-of-the-art performance on hallucination benchmarks while preserving general multimodal capabilities. Our work highlights the importance of respecting the model's intrinsic perceptual boundaries, offering a new perspective for building more reliable vision-language systems.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62472420.

Contribution Statement

Minghui Chen and Chenxu Yang contributed equally to this work. Dayan Wu is the corresponding author.

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[Bai *et al.*, 2024] 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.

[Dai *et al.*, 2023] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

[Deng *et al.*, 2024] Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Y Zou, Kai-Wei Chang, and Wei Wang. Enhancing large vision language models with self-training on image comprehension. *Advances in Neural Information Processing Systems*, 37:131369–131397, 2024.

[Gunjal *et al.*, 2024] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143, 2024.

[Hao *et al.*, 2023] Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, 2023.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[Hu *et al.*, 2024] Chi Hu, Yimin Hu, Hang Cao, Tong Xiao, and Jingbo Zhu. Teaching language models to self-improve by learning from language feedback. *arXiv preprint arXiv:2406.07168*, 2024.

[Huang *et al.*, 2023] Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 1051–1068, 2023.

[Huang *et al.*, 2024] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024.

[Leng *et al.*, 2024] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

[Li and Ng, 2025] Junyi Li and Hwee Tou Ng. Think&cite: Improving attributed text generation with self-guided tree search and progress reward modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9928–9942, 2025.

[Li *et al.*, 2023a] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023.

[Li *et al.*, 2023b] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[Li *et al.*, 2025] Zhigen Li, Jianxiang Peng, Yanmeng Wang, Yong Cao, Tianhao Shen, Minghui Zhang, Linxi Su, Shang Wu, Yihang Wu, Yuqian Wang, et al. Chatsop: An sop-guided mcts planning framework for controllable llm dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17637–17659, 2025.

[Liu *et al.*, 2023a] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *CoRR*, 2023.

[Liu *et al.*, 2023b] 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.

[Liu *et al.*, 2023c] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[Liu *et al.*, 2024] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou,

Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.

[Lu *et al.*, 2024] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[Rafailov *et al.*, 2023] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[Rawte *et al.*, 2023] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.

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

[Rosin, 2011] Christopher D Rosin. Multi-armed bandits with episode context. *Annals of Mathematics and Artificial Intelligence*, 61(3):203–230, 2011.

[Silver *et al.*, 2016] 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.

[Sun *et al.*, 2024] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, 2024.

[Sun *et al.*, 2025] Yutao Sun, Mingshuai Chen, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, and Jianwei Yin. The self-improvement paradox: Can language models bootstrap reasoning capabilities without external scaffolding? *arXiv preprint arXiv:2502.13441*, 2025.

[Suo *et al.*, 2025] Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. Octopus: Alleviating hallucination via dynamic contrastive decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29904–29914, 2025.

[Tian *et al.*, 2024] 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.

[Wang *et al.*, 2023] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *CoRR*, 2023.

[Wang *et al.*, 2025] Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Taha Kass-Hout, et al. Enhancing visual-language modality alignment in large vision language models via self-improvement. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 268–282, 2025.

[Xie *et al.*, 2024a] Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P Lillicrap, Kenji Kawaguchi, and Michael Shieh. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*, 2024.

[Xie *et al.*, 2024b] Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. *arXiv preprint arXiv:2411.02712*, 2024.

[Yang *et al.*, 2025] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[Ye *et al.*, 2024] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13040–13051, 2024.

[Yin *et al.*, 2024] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

[Yu *et al.*, 2023] 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.

[Yuan *et al.*, 2024] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*, 2024.

[Zhang *et al.*, 2024] Di Zhang, Xiaoshui Huang, Dongzhan Zhou, Yuqiang Li, and Wanli Ouyang. Accessing gpt-4 level mathematical olympiad solutions via monte carlo tree self-refine with llama-3 8b. *arXiv preprint arXiv:2406.07394*, 2024.

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

[Zhu *et al.*, 2023] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.