7. Appendix

7.1. Additional Related Works

Hallucination in LVLMs Large Vision-Language Models (LVLMs) have achieved remarkable progress in multimodal understanding and reasoning tasks[4, 20, 31, 52]. However, hallucination remains a critical challenge, where models generate content inconsistent with visual inputs[72]. This issue stems from multiple sources: the inherent properties of LLMs[14, 33], misalignment between visual and textual modalities[24, 49], and quality limitations in instruction tuning data[39, 93]. To evaluate hallucination severity, various benchmarks have been proposed. POPE[43] assesses hallucinations related to object existence through discriminative questions, while CHAIR[61] quantifies the proportion of hallucinated objects in image descriptions at both instance-level (CHAIR_i) and sentence-level (CHAIR_s). More comprehensive evaluation metrics have emerged to cover broader hallucination types, including attributes, spatial relations, and emotions[84, 100]. From a visual semantics perspective, hallucinations can be categorized into three primary types: object hallucinations (generating non-existent entities), attribute hallucinations (incorrect descriptions of visual properties), and relation hallucinations (inaccurate assertions about spatial or semantic relationships)[77].

Hallucination Mitigation Existing approaches for mitigating hallucinations in LVLMs fall into two categories: training-based and training-free methods[24, 33, 91]. Training-based methods include supervised fine-tuning (SFT) with hallucination-free data[93, 97], such as LRV[49] and InstructBLIP[20], and preference optimization techniques that leverage paired data to align model outputs with human preferences[29, 42, 71, 94]. However, SFT approaches require carefully curated datasets and substantial computational resources, while preference learning methods often rely on external models like GPT-4V for annotation[100]. Training-free approaches primarily employ contrastive decoding strategies[23, 39, 55, 73, 79, 81], which contrast output distributions from original and modified inputs (e.g., masked images or enhanced visual features) to reduce reliance on language priors[39]. While these methods avoid additional training costs, they introduce inference-time computational overhead and do not fundamentally improve the model's intrinsic generation capabilities. In contrast to these paradigms, our proposed HAIT adopts a reinforcement learning approach built upon the standard SFT+RL paradigm. This method integrates hallucination-aware mechanisms through a novel reward design that is applied directly to SFT-finetuned models.

Self-Improvement through Iterative Training Iterative training enables models to progressively refine their capabilities through self-generated feedback, eliminating dependence on external supervision. Recent work has explored diverse strategies for self-iteration: Pang et al.[57] construct preference pairs via rule-based reward functions, Chen et al.[13] leverage adversarial learning grounded in human-annotated responses, Dong et al.[21] perform iterative updates with fixed reward models, and Wu et al.[83] frame the training process as zero-sum games with learned preference models. Additionally, Rosset et al.[62] introduce Direct Nash Optimization (DNO), employing expert models to estimate win rates through batch policy iteration. These methods differ fundamentally from offline approaches[82, 95] by operating in online frameworks where policies adapt dynamically. In the context of multimodal models, self-play has also been explored for vision-language alignment[69, 70], demonstrating the potential of iterative refinement across modalities. HAIT distinguishes itself through a hallucination-aware reward mechanism that integrates three complementary signals. HAIT employs a staged weighting strategy that prioritizes establishing fundamental visual grounding in early iterations before transitioning to fine-grained error correction, a curriculum-inspired approach particularly effective for multimodal caption generation where outputs exhibit complex mixtures of grounded and hallucinated content.

7.2. Experiments

7.2.1. Metric Calculation

CHAIR CHAIR (Caption Hallucination Assessment with Image Relevance) quantifies object hallucination for image captioning systems. It measures the extent to which generated captions mention objects that are not supported by the image, both at the level of individual object mentions and at the level of entire sentences.

Following the original work, our calculation relies on the predefined vocabulary $\mathcal{V}$ of 80 MS-COCO object classes. This vocabulary incorporates synonym sets, ensuring that different lexical expressions referring to the same object type (e.g., "couch" and "sofa") are mapped to a single canonical class within $\mathcal{V}$. This practice confines the evaluation to a consistent and manageable set of object types.

For an image and its caption reference r with ground-truth object set $\mathcal{O}_{\text{gt}}(r)$ and a generated caption y, let $\mathcal{O}_{\text{gen}}(y)$ denote the set of objects mentioned in y. The hallucinated objects can be represented as:

$$\mathcal{H}(r) = (\mathcal{O}\text{gen}(y) \cap \mathcal{V}) \setminus (\mathcal{O}\text{gt}(r) \cap \mathcal{V}) \tag{20}$$

and we write $\mathcal{M}(r) = (\mathcal{O}\text{gen}(y) \cap \mathcal{V})$ for the set of all mentioned objects.

Instance-level hallucination (CHAIR_i). On a dataset $\mathcal{D}$,

the instance-level variant CHAIR_i is defined as

$$\text{CHAIR}_i = \frac{\sum_{r\in\mathcal{D}} |\mathcal{H}(r)|}{\sum_{r\in\mathcal{D}} |\mathcal{M}(r)|}. \tag{21}$$

Lower values indicate fewer hallucinated object mentions relative to the total number of object mentions.

Sentence-level hallucination (CHAIR_s). The sentence-level variant CHAIR_s measures how frequently a caption contains at least one hallucinated object. For each image v and its generated caption y, define

$$\mathbf{1}_{\text{halluc}}(r) = \begin{cases} 1, & \text{if } \mathcal{H}(r) \neq \emptyset, \\ 0, & \text{otherwise,} \end{cases} \tag{22}$$

where $\mathcal{H}(r)$ is as defined above. CHAIR_s is then given by

$$\text{CHAIR}_s = \frac{1}{|\mathcal{D}|} \sum_{r\in\mathcal{D}} \mathbf{1}_{\text{halluc}}(r). \tag{23}$$

Smaller CHAIR_s values indicate that hallucination occurs in fewer captions.

BLEU BLEU is an n-gram overlap metric originally proposed for machine translation and now widely used in text generation tasks, including image captioning. It measures modified n-gram precision between a candidate caption and a set of reference captions, combined with a brevity penalty to discourage overly short outputs.

Given a candidate caption c and a reference set $\mathcal{R}$, let p_n denote the modified precision for n-grams of order $n = 1, \ldots, N$, computed with clipped counts. Let $|c|$ denote the length of c and let $|r^*|$ be the length of the reference whose length is closest to $|c|$. The brevity penalty is

$$\text{BP} = \begin{cases} 1, & |c| > |r^*|, \\ \exp\left(1 - \dfrac{|r^*|}{|c|}\right), & |c| \leq |r^*|. \end{cases} \tag{24}$$

The BLEU score of order N is defined as

$$\text{BLEU}_N = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \tag{25}$$

where $w_n \geq 0$ are weights such that $\sum_n w_n = 1$ (a common choice is $N = 4$ and $w_n = 1/4$). Higher BLEU scores indicate stronger n-gram agreement with the references.

CIDEr CIDEr (Consensus-based Image Description Evaluation) is tailored to image captioning. It represents captions as TF–IDF weighted n-gram vectors and measures the degree of consensus between a candidate and multiple references via cosine similarity.

For n-grams of order n, let $\mathbf{g}_n(x) \in \mathbb{R}^{d_n}$ denote the TF–IDF feature vector of caption x. Given a candidate c and references $\mathcal{R}$, the CIDEr score for order n is

$$\text{CIDEr}_n(c, \mathcal{R}) = \frac{1}{|\mathcal{R}|} \sum_{r\in\mathcal{R}} \frac{\mathbf{g}_n(c)^\top \mathbf{g}_n(r)}{\|\mathbf{g}_n(c)\|_2 \, \|\mathbf{g}_n(r)\|_2}. \tag{26}$$

The final CIDEr score averages over $n = 1, \ldots, N$:

$$\text{CIDEr}(c, \mathcal{R}) = \frac{1}{N} \sum_{n=1}^{N} \text{CIDEr}_n(c, \mathcal{R}). \tag{27}$$

In practice, N is typically set to 4. Larger CIDEr values indicate that the candidate more closely matches the consensus of human references.

METEOR METEOR is a sentence-level metric that combines unigram precision and recall, and allows matches based on stemming and synonymy. It further incorporates a penalty for fragmented word order, making it more sensitive to fluency and adequacy than pure n-gram precision.

For a candidate c and a reference r, an alignment is constructed to maximize the number of matched unigrams m. Let $|c|$ and $|r|$ denote the lengths of c and r, respectively. Unigram precision and recall are defined as

$$P = \frac{m}{|c|}, \qquad R = \frac{m}{|r|}. \tag{28}$$

They are combined using a weighted harmonic mean

$$F_{\text{mean}} = \frac{10\,P\,R}{R + 9P}, \tag{29}$$

which places more weight on recall. Let ch be the number of contiguous matched chunks in the alignment. The fragmentation penalty is

$$\text{Pen} = \gamma \left(\frac{ch}{m}\right)^{\theta}, \tag{30}$$

with typical parameter values $\gamma = 0.5$ and $\theta = 3$. The METEOR score is then

$$\text{METEOR} = (1 - \text{Pen})\, F_{\text{mean}}. \tag{31}$$

When multiple references are available, the score is computed against each reference and the maximum is reported.

ROUGE-L ROUGE-L is based on the length of the longest common subsequence (LCS) between a candidate and a reference sentence. Unlike exact n-gram overlap, the LCS can capture similarity while allowing gaps, and is therefore less sensitive to small local reorderings.

Given a candidate c and a reference r, let $\mathrm{LCS}(c, r)$ denote the length of their longest common subsequence. We define

$$P_{\mathrm{LCS}} = \frac{\mathrm{LCS}(c, r)}{|c|}, \qquad R_{\mathrm{LCS}} = \frac{\mathrm{LCS}(c, r)}{|r|}. \tag{32}$$

The ROUGE-L score is the corresponding F-measure:

$$\text{ROUGE-L} = \frac{(1 + \beta^2)\, P_{\mathrm{LCS}}\, R_{\mathrm{LCS}}}{R_{\mathrm{LCS}} + \beta^2 P_{\mathrm{LCS}}}, \tag{33}$$

where β controls the relative weight of recall (often $\beta = 1$, yielding the standard F_1 score). With multiple references, ROUGE-L is computed with respect to each reference and then aggregated (e.g., by taking the maximum or the average).

SPICE SPICE (Semantic Propositional Image Caption Evaluation) evaluates captions at the level of semantic content rather than surface word overlap. Each caption is parsed into a scene graph with objects, attributes, and relations, which are represented as tuples. The candidate and reference graphs are then compared in this tuple space.

Let $\mathcal{T}(c)$ be the set of semantic tuples extracted from a candidate caption c, and let $\mathcal{T}(\mathcal{R})$ be the union of tuples from the references $\mathcal{R}$. Precision and recall in the semantic tuple space are

$$P_{\text{sem}} = \frac{|\mathcal{T}(c) \cap \mathcal{T}(\mathcal{R})|}{|\mathcal{T}(c)|}, \qquad R_{\text{sem}} = \frac{|\mathcal{T}(c) \cap \mathcal{T}(\mathcal{R})|}{|\mathcal{T}(\mathcal{R})|}. \tag{34}$$

The SPICE score is the corresponding F_1-measure:

$$\mathrm{SPICE} = \frac{2\, P_{\text{sem}}\, R_{\text{sem}}}{P_{\text{sem}} + R_{\text{sem}}}. \tag{35}$$

Higher SPICE values indicate that the candidate more accurately recovers the objects, attributes, and relations expressed in human-written captions.

7.2.2. Hyperparameters and Implementation Details

We provide a comprehensive summary of our training and evaluation configuration in Table 4. The training process consists of two main stages: Supervised Fine-Tuning (SFT) and HAIT training.

Training Steps Calculation: The number of training steps for HAIT phase is determined based on the dataset size and batch configuration. Specifically, with a batch size of 1024 and approximately 118k training samples, we compute the total training steps as 115 (118k/1024), ensuring adequate exposure to the training data.

Reward Weight Configuration: We employ distinct reward weight schemes across the two training epochs. In Epoch-1, we utilize a balanced weighting with Base=0.05, HA=0.4, HA-λ_1=0.4, HA-λ_2=0.6, AHA=0.4, and ME=0.15. For Epoch-2, we transition to Base=0.0, HA=0.1, AHA=0.7, and ME=0.2, emphasizing different components during the iterative training process.

Adversarial Model Updates: The adversarial model is updated iteratively every 10 training steps, with each update referred to as an iteration. This frequency strikes a balance between training stability and adversarial effectiveness, ensuring consistent adversarial guidance while maintaining training efficiency.

Table 4. Training and Evaluation Configuration

Component	Parameters
Supervised Fine-Tuning (SFT)	
	Batch size: 16
	Epochs: 2
	Learning rate: 1×10^{-5}
	Warmup steps ratio: 0.1
	Time: 2 hours/epoch
HAIT Training	
	Batch size: 1024
	Training steps: 115 (118k samples / 1024)
	Rollout: 8
	Learning rate: 1×10^{-6}
	Temperature: 1.0
	Top-p: 1.0
	Top-k: -1(vLLM no limit)
	Adversarial model update frequency: 10 steps
	Time: 30 hours/epoch
Reward Weights	
	Epoch-1: Base=0.05, HA=0.4
	λ_1=0.4, λ_2=0.6
	AHA=0.4, ME=0.15
	Epoch-2: Base=0.0, HA=0.1
	AHA=0.7, ME=0.2
Evaluation on Flickr30k/HA-DPO	
	Temperature: 1.0
	Top-p: 1.0
	Top-k: -1(vLLM no limit)
	Package: spaCy, NLTK, COCO-caption
Hardware Configuration	
	GPU: 8 $\times$ NVIDIA A800
	RAM: 512 GB
	CPU: 96 cores

7.2.3. Model Evaluator Reward

Prompt Text Display

You are an AI evaluator tasked with assessing the quality of image captions. Your role is to rate a **candidate caption** based on its alignment with a given list of **reference captions**. Provide a score from **1 to 5**, with **1** being the lowest and **5** the highest, based on the caption's accuracy, relevance, and detail.
Rating Criteria: - **5:** The caption perfectly describes the image, is highly consistent with the references, detailed, fluent, and error-free. - **4:** The caption is accurate and mostly consistent with the references but may have minor inaccuracies, omissions, or slightly reduced fluency. - **3:** The caption is partially consistent with the references but may lack important details, contain errors, or use unnatural language. - **2:** The caption has significant errors, low consistency with the references, and is partially relevant or inaccurate. - **1:** The caption is completely irrelevant, meaningless, or inconsistent with the references.
Instructions: For each input, you will be given: 1. A **candidate caption**. 2. A list of **reference captions**. Evaluate the candidate caption against the reference list and provide your response in the following strict format: `"Rating: X \n Reason: ..."` (where **X** is the score). Ensure your reasoning clearly explains the score based on the criteria above.

We first evaluate three candidate models for human preference alignment: Qwen3-4B-Instruct-2507, Qwen2.5-3B-Instruct, and Llama-3.2-3B-Instruct. We construct a test set of 50 samples randomly selected from the COCO-2017 validation dataset to assess model alignment with human preferences. For each of the 50 image samples, five human annotators independently provide captions and rate them on a 1–5 scale. This results in 250 annotated descriptions for the complete test set. Each description is then evaluated by the candidate models, and model ratings are compared against human judgments. As shown in Table 5, we report the average score of the model for each level of caption. We select the model with the smallest deviation from human ratings, which is **Qwen3-4B-Instruct-2507**.

Table 5. Model rating performance comparison (average scores)

Model	1	2	3	4	5
Qwen3-4B	1.02	1.96	3.02	3.92	4.90
Qwen2.5-3B	1.00	2.10	3.12	3.88	4.84
Llama-3.2-3B	1.00	2.06	3.10	3.80	4.88

Notably, while model rating consistency may be influenced by human annotation variability, the close performance across models suggests that the rating mechanism exhibits reasonable robustness. When using the model as a model evaluator to assess human preference alignment, we consistently used the same prompt template, as shown in Box in 7.2.3.

7.2.4. Additional Experiment Result

Chair Changes We separately visualize the changes in the CHAIR metric during the training process under different reward settings. As shown in Fig 5 and Fig 6, HAIT demonstrates a relatively stable convergence speed, whereas HA achieves the fastest convergence. In contrast, both AHA and ME converge at relatively slower rates.

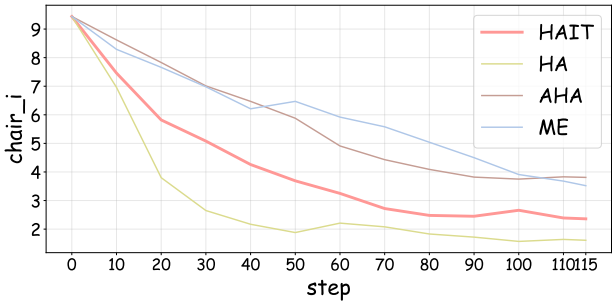

Figure 5. The changes of $CHAIR_i$ during epoch-1, showing how it evolves under different reward settings.

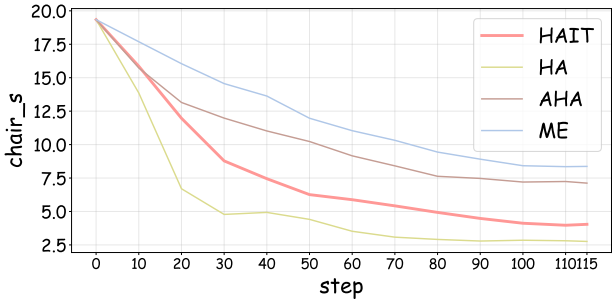

Figure 6. The changes of $CHAIR_s$ during epoch-1, showing how it evolves under different reward settings.

Language Prior Evaluation First, we describe our method for evaluating language priors, which serves as a supplement to Section 3. We selected four scenarios: kitchen, street, bathroom, and office, from the validation set of COCO2017, with 100 images for each scenario. Notably, during the selection process, we intentionally included some "adversarial" images with missing elements. These images lack certain scene-related contents, making them similar to the Adversarial subset in the POPE[43]

benchmark. This experiment is conducted based on Qwen3-VL-8B-Instruct.

We compare the performance of the original model with that of the model after undergoing the HAIT-iteration-3 process. As shown in Fig 7, the experimental results indicate that the frequency of hallucinations caused by co-occurring words decreases significantly. This demonstrates that our method effectively alleviates object hallucinations induced by language priors. Furthermore, when analyzing hallucination errors, we find that, given the rapid development of current visual encoders, it is challenging for the model to fabricate visual content without any evidence. As a result, object hallucination can mainly be categorized into two types. The first type involves small objects, such as a kitchen knife. Due to limitations in image resolution and other factors, it is often impossible for humans to determine with certainty whether such an object exists in the image. The model, however, tends to identify objects that might plausibly exist but cannot be confirmed, relying on language priors. For large objects, such as a kitchen refrigerator, hallucination cases usually occur when only a small part of the possible object is visible, which is insufficient to support a definitive judgment. In such instances, the model tends to "imagine" the complete object based on partial features. In summary, these hallucination errors arise primarily from insufficient visual information, which allows language priors to dominate the model's predictions.

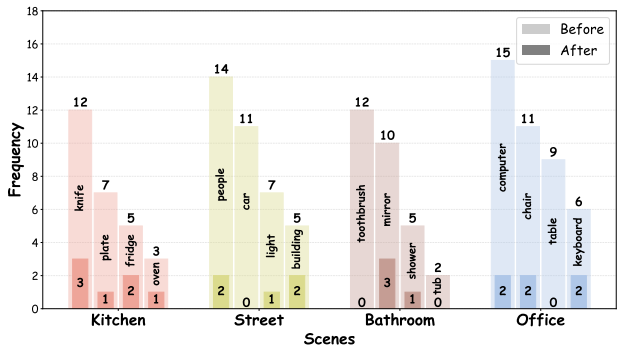

Figure 7. Reduction in hallucination frequency after HAIT-iteration-3.

7.2.5. General Capability Preservation

To ensure that HAIT's adversarial iterative training does not degrade the model's general capabilities, we evaluate Qwen3-VL-8B-Instruct on five widely-used benchmarks from the HuggingFace Open LLM Leaderboard [7]. These benchmarks assess diverse abilities including mathematical reasoning, commonsense inference, multitask language understanding, and factual accuracy. Specifically, we evaluate on GSM8K (5-shot) [19] for multi-step mathematical reasoning, HellaSwag (10-shot) [96] for commonsense inference, WinoGrande (5-shot) [63] for pronoun disambiguation and commonsense reasoning, MMLU (5-shot) [32] for multitask accuracy across 57 subjects, and TruthfulQA (0-shot) [47] for measuring truthfulness in answering questions. We compare the base model with both standard supervised fine-tuning and HAIT-trained variants across multiple iterations.

Table 6 presents comprehensive performance comparison across training approaches. Throughout this section, we refer to "SFT-epoch-1 + HAIT-epoch-1" as "HAIT-iteration-0". Results demonstrate that HAIT training maintains stable general capabilities throughout the iterative process. The base model achieves an average score of 71.64, while SFT-epoch-1 shows modest improvement to 71.86. However, SFT-epoch-2 drops to 71.60, confirming overfitting issues with continued supervised training. In contrast, HAIT maintains consistent performance with iteration-0 achieving 72.00, iteration-2 reaching 72.12. The average score fluctuation across HAIT iterations remains within 0.6 percentage points of the base model. These results confirm that HAIT's adversarial training for hallucination reduction does not compromise the model's core reasoning and knowledge capabilities.

7.2.6. Results on POPE

We also report the performance of HAIT on the POPE[43] benchmark. The Polling-based Object Probing Evaluation (POPE) is a streamlined approach to assess object hallucination in LVLMs. LVLMs are required to respond to formatted questions in the form: "Is there a <object> in the image?" with "Yes" or "No". The label distribution is balanced across the dataset, with 50% "Yes" and 50% "No" answers. The complete POPE test is divided into three splits: **random**, **popular**, and **adversarial**, in which the queried objects are randomly selected, selected from the most frequent categories in the dataset, and selected to be highly correlated with those present in the image, respectively. The dataset consists of 500 randomly selected images from the MSCOCO[48] validation set.

To facilitate testing, we add the prompt "Please only use yes or no to answer this question." to restrict LVLM responses to "Yes" or "No". Additionally, we use greedy during inference to ensure deterministic output generation, with all other experimental setups identical to those described in Section 5.3. Four key evaluation metrics are generated: Accuracy, Recall, Precision and F1 score. We report the results across the three splits, and the outcomes are presented in Table 7.

We observe that HAIT significantly improves Accuracy and is superior to other methods, while the Recall metric does not show a notable increase. This occurs because HAIT mitigates the model's tendency to generate responses based on language priors rather than visual evidence, particularly in scene-dependent scenarios like the Adversarial

Table 6. Performance comparison of HAIT across general benchmarks

Models	Iteration Methods	GSM8K	Hellaswag	WinoGrande	MMLU	TruthfulQA	Leaderboard Ave
Qwen3 -VL-8B -Instruct	Original	85.2	74.6	71.9	72.5	54.0	71.64
	SFT-epoch-1	85.5	74.8	72.1	72.7	54.2	71.86
	SFT-epoch-2	85.3	74.5	71.8	72.6	53.8	71.60
	SFT-epoch-1 + HAIT-epoch-1	85.6	74.9	72.3	72.8	54.4	72.00
	HAIT-iteration-1	85.9	74.9	72.2	72.0	54.3	71.86
	HAIT-iteration-2	85.1	75.2	72.9	72.9	54.5	72.12
	HAIT-iteration-3	85.7	75.0	72.1	72.5	54.5	71.96

Table 7. Performance comparison between HAIT and other methods with Qwen3-VL-8B-Instruct on POPE

Subset	Methods	Accuracy	Recall	Precision	F1-Score
Random	Original	86.20	86.46	86.00	86.23
	VCD	87.67	88.80	86.83	87.80
	OPERA	88.33	87.86	88.69	88.27
	Nullu	88.43	87.46	89.19	88.32
	ICT	88.17	89.00	87.54	88.26
	HAIT	**89.57**	86.93	91.77	89.28
Popular	Original	83.63	86.46	81.83	84.08
	VCD	85.86	89.13	83.67	86.31
	OPERA	85.93	87.86	84.59	86.20
	Nullu	86.00	87.66	84.39	86.22
	ICT	86.03	89.00	84.02	86.43
	HAIT	**88.36**	86.93	89.50	88.19
Adversarial	Original	82.30	86.53	79.78	83.01
	VCD	84.70	89.26	81.80	85.36
	OPERA	85.03	87.60	83.32	85.40
	Nullu	85.36	87.80	83.72	85.71
	ICT	84.93	89.06	82.27	85.53
	HAIT	**86.66**	86.93	86.47	86.70

subset.

7.2.7. Generation Examples

To qualitatively demonstrate HAIT's effectiveness, we present representative generation examples in Table 8. We show three cases comparing the base Qwen3-VL-8B-Instruct model with HAIT iteration 3: two short captions from the COCO validation set and one long description from the HA-DPO test set.

The examples reveal consistent hallucination reduction patterns. In short captions, the base model generates non-existent objects or misidentifies visual elements, while HAIT iteration 3 produces accurate descriptions faithful to the visible content. The long description example demonstrates that HAIT maintains this improvement even for extended captions, eliminating spurious object mentions while preserving descriptive richness. These qualitative observations align with our quantitative results, confirming that HAIT effectively grounds generation in visual content without compromising informativeness.

7.2.8. Potential Discussion

Comparison with Other Methods We run SFT for two epochs to obtain both SFT-epoch-1 and SFT-epoch-2 baselines. HAIT is initialized from SFT-epoch-1, followed by one HAIT epoch and three HAIT iterations. In contrast, other training-free methods compared in our experiments are based on the original model weights with only decoding strategies modified. This distinction is important as HAIT involves parameter updates while the competing methods maintain the original model parameters.

Comparison with RL-based Approaches We primarily compare with training-free methods rather than reinforcement learning (RL) based approaches for object hallucination mitigation. This choice is motivated by the current research landscape where the majority of object hallucination mitigation methods focus on training-free techniques. While RL-based methods for this problem are relatively scarce, we note that the ablation variants in our experiments can be viewed as approximations of GRPO variants. Our comprehensive evaluation demonstrates that HAIT outperforms these variants across multiple metrics.

Table 8. Representative generation examples comparing original Qwen3-VL-8B-Instruct and HAIT iteration 3

Image	Generated Captions
	**References:** [A buffet of casserole dishes on a kitchen counter. A kitchen area with a table full of various foods. Different platters of food are set in the kitchen. Many containers of food are covered up on the table. Covered food is sitting on a kitchen counter.] **Original:** A cozy kitchen scene with meal prep underway. Covered dishes, a set of tableware, jars of sauce, and a bottle of wine sit on the counter, while a TV in the background shows a talk show, blending home cooking with casual entertainment. **HAIT:** Covered dishes, jars of sauce, and a bottle of wine are arranged on a kitchen counter, with a television visible in the background. Various food containers and platters fill the kitchen space during meal preparation.
	**References:** [A bathroom that has a person's hand holding an object. A person is holding up a hair dryer in the bathroom. A woman holds a dryer in a bathroom. Someone steaming the wrinkles out of a garment. This is a blow dryer in a bathroom.] **Original:** A hand holding a hair dryer in a bathroom, with towels hanging nearby and toiletries on the toilet lid. **HAIT:** A person's hand holding a hair dryer in a bathroom setting, with towels visible in the background.
	**Reference:** A table is depicted in the image, showcasing a miniature shopping cart brimming with bananas. The shopping cart contains precisely five ripe bananas. Moreover, positioned on the right side of the table is an elegant glass of water. No bottles are present on the table; however, there are narrative cards and a family photograph. Surrounding the table are chairs, with one noticeable chair on the left side. The intriguing arrangement of bananas within the shopping cart and the assortment of items on the table contribute to an engaging scene. **Original:** This is a detailed, eye-level photograph of a dining table setting, likely in a restaurant or café, with a focus on a small, wire-frame shopping cart filled with bananas.Central Focus: The Banana Cart- A small, silver-colored metal shopping cart, resembling a miniature grocery cart, sits prominently on the table.- Inside the cart are several ripe yellow bananas with some brown spots, and at least one green banana, suggesting a mix of ripeness.- The bananas are arranged in a casual, slightly overflowing manner, with some leaning against each other and others resting on the cart's wire mesh floor.Surrounding Elements:- Signage: A white rectangular sign is mounted above the cart, featuring the logo and text "FOOD for LIFE" in a blue circular emblem, followed by "Club rooms 15" in blue and black lettering. This suggests the setting is a "Club Rooms" dining area, possibly within a hotel or upscale establishment, with table number 15.- Table Setting: The cart is placed on a white tablecloth. To the right, a clear glass of water and a folded white napkin on a silver napkin ring are visible. A dark-colored bottle, possibly a condiment or [truncated at max_new_tokens=256] **HAIT:** The image depicts a dining table setting with a miniature wire-frame shopping cart filled with bananas. The shopping cart contains precisely several ripe bananas, including yellow ones and at least one green banana. Positioned above the cart is a white rectangular sign displaying "FOOD for LIFE" and "Club rooms 15" in blue and black lettering. On the right side of the table sits an elegant glass of water alongside a folded white napkin. No bottles are present on the table; however, there are narrative cards and a family photograph arranged nearby. Chairs surround the table, with one noticeable chair on the left side. The intriguing arrangement of bananas within the shopping cart and the assortment of items on the table contribute to an engaging scene, showcasing a thoughtful table display reminiscent of educational or awareness events focused on food themes. Overall, the layout appears deliberate and uncluttered, centering attention on the cart and signage. The sign sits above the cart, while nearby cards add context without overwhelming the focal point.

Analysis of CIDEr Scores Regarding the relatively high CIDEr scores observed in our ablation studies, we attribute this phenomenon to the inherent characteristics of the Flickr30K dataset, where reference captions tend to be semantically similar. This similarity may inflate CIDEr metrics compared to other datasets with more diverse captions.

Output Length Analysis We also analyzed the length of generated content before and after training. Our analysis shows that the output texts consistently reach or approach the set maximum new token limit, indicating that the reduction in hallucinations is not attributable to shorter output lengths. This finding confirms that HAIT's effectiveness stems from genuine improvement in content quality.

7.3. Proof of Theorems in Section 4

7.3.1. Proof of Theorem 1

Theorem 1 (Convergence of HAIT). Under standard assumptions on the loss function ℓ (monotonicity and convexity), HAIT converges to a stationary point where the masked caption distributions align:

$$\pi_{\theta^*}(\tilde{\mathbf{y}}'|\mathbf{v},\mathbf{x}) = p_{\text{data}}(\tilde{\mathbf{y}}|\mathbf{v},\mathbf{x},\mathbf{y}' \sim \pi_{\theta^*}) \tag{36}$$

Proof. We establish convergence by examining the fixed-point properties of the integrated reward objective. At iteration $t+1$, the total reward function is:

$$\begin{aligned} R_{\text{HAIT}}(\mathbf{y}',\mathbf{v},\mathbf{x}) = r_{\text{AHA}}(\theta,\theta_t) + \alpha \cdot r_{\text{HA}}(\mathbf{y}',\mathbf{v}) \\ +\beta \cdot r_{\text{LLM}}(\mathbf{y}',\mathbf{y}) \end{aligned} \tag{37}$$

where the Anti-Hallucination Adversarial (AHA) loss is defined as:

$$\begin{aligned} r_{\text{AHA}}(\theta,\theta_t) = -\mathbb{E}_{v,x,y,y'}\Bigg[\ell\Bigg(\beta \log \frac{\pi_\theta(\tilde{y}|v,x)}{\pi_{\theta_t}(\tilde{y}|v,x)} \\ -\beta \log \frac{\pi_\theta(\tilde{y}'|v,x)}{\pi_{\theta_t}(\tilde{y}'|v,x)}\Bigg)\Bigg] \end{aligned} \tag{38}$$

At the fixed point $\theta^* = \theta_t$, convergence requires the gradient to vanish: $\nabla_\theta R_{\text{total}}(\theta^*,\theta^*) = 0$. We first analyze the AHA component. Taking the gradient at the fixed point yields:

$$\begin{aligned} \nabla_\theta r_{\text{AHA}}(\theta,\theta^*)|_{\theta=\theta^*} = \mathbb{E}_{\mathbf{v},\mathbf{x},\mathbf{y},\mathbf{y}'}\left[\ell'(0)\cdot\beta\left(\nabla_\theta \log \pi_\theta(\tilde{\mathbf{y}}|\mathbf{v},\mathbf{x})\right.\right. \\ \left.\left. -\nabla_\theta \log \pi_\theta(\tilde{\mathbf{y}}'|\mathbf{v},\mathbf{x})\right)|_{\theta=\theta^*}\right] \end{aligned} \tag{39}$$

Given that $\mathbf{y}' \sim \pi_{\theta^*}(\cdot|\mathbf{v},\mathbf{x})$ and $\tilde{\mathbf{y}}$, $\tilde{\mathbf{y}}'$ represent sequences with LCS removed, the gradient expectation equals zero when:

$$\mathbb{E}_{\mathbf{y}'\sim\pi_{\theta^*}}\left[\nabla_\theta \log \pi_\theta(\tilde{\mathbf{y}}'|\mathbf{v},\mathbf{x})\right] = \mathbb{E}_{\mathbf{y}\sim p_{\text{data}}}\left[\nabla_\theta \log \pi_\theta(\tilde{\mathbf{y}}|\mathbf{v},\mathbf{x})\right] \tag{40}$$

This equality holds if and only if the masked caption distributions coincide:

$$\pi_{\theta^*}(\tilde{\mathbf{y}}'|\mathbf{v},\mathbf{x}) = p_{\text{data}}(\tilde{\mathbf{y}}|\mathbf{v},\mathbf{x}) \tag{41}$$

For the Hallucination-Aware (HA) component, at convergence the reward saturates when the model generates captions without object hallucinations, consistent with the aligned masked distributions. Similarly, the LLM evaluator component reaches its maximum when generated captions match the semantic quality of references, which occurs when masked distributions align. Therefore, HAIT converges to a stationary point where the policy cannot differentiate between masked reference captions and its own masked generations in hallucination-prone regions, achieving alignment on error-containing subsequences while preserving correctly predicted content.

7.3.2. Proof of Theorem 2

Theorem 2 (Efficiency of Anti-Hallucination Adversarial Training) Let $\gamma = \mathbb{E}[|\text{LCS}(\mathbf{y},\mathbf{y}')|/|\mathbf{y}|]$ denote the average fraction of correctly predicted tokens. The gradient variance of HAIT is reduced by a factor of $(1-\gamma)^2$ compared to holistic sequence training:

$$\text{Var}[\nabla_\theta R_{\text{HAIT}}] \le (1-\gamma)^2 \cdot \text{Var}[\nabla_\theta R_{\text{holistic}}] \tag{42}$$

Proof. We denote the gradient of the holistic sequence reward (without masking) as:

$$g_{\text{holistic}} = \nabla_\theta\left[\beta \log \frac{\pi_\theta(\mathbf{y}|\mathbf{v},\mathbf{x})}{\pi_{\theta_t}(\mathbf{y}|\mathbf{v},\mathbf{x})} - \beta \log \frac{\pi_\theta(\mathbf{y}'|\mathbf{v},\mathbf{x})}{\pi_{\theta_t}(\mathbf{y}'|\mathbf{v},\mathbf{x})}\right] \tag{43}$$

and the gradient with hallucination-aware masking as:

$$g_{\text{HAIT}} = \nabla_\theta\left[\beta \log \frac{\pi_\theta(\tilde{\mathbf{y}}|\mathbf{v},\mathbf{x})}{\pi_{\theta_t}(\tilde{\mathbf{y}}|\mathbf{v},\mathbf{x})} - \beta \log \frac{\pi_\theta(\tilde{\mathbf{y}}'|\mathbf{v},\mathbf{x})}{\pi_{\theta_t}(\tilde{\mathbf{y}}'|\mathbf{v},\mathbf{x})}\right] \tag{44}$$

The log-probability can be decomposed token-wise:

$$\log \pi_\theta(\mathbf{y}|\mathbf{v},\mathbf{x}) = \sum_{j=1}^{|\mathbf{y}|} \log \pi_\theta(y_j|\mathbf{v},\mathbf{x},\mathbf{y}_{<j}) \tag{45}$$

For the masked sequence, only non-LCS tokens contribute:

$$\log \pi_\theta(\tilde{\mathbf{y}}|\mathbf{v},\mathbf{x}) = \sum_{j\in\mathcal{I}_{\text{diff}}} \log \pi_\theta(y_j|\mathbf{v},\mathbf{x},\mathbf{y}_{<j}) \tag{46}$$

where $\mathcal{I}_{\text{diff}}$ denotes the index set of tokens not in the LCS. Since $|\mathcal{I}_{\text{diff}}| = (1-\gamma)|\mathbf{y}|$, the gradient magnitude scales as:

$$\|g_{\text{HAIT}}\|^2 \approx (1-\gamma)^2 \|g_{\text{holistic}}\|^2 \tag{47}$$

Applying the law of total variance:

$$\text{Var}[g_{\text{HAIT}}] = \mathbb{E}[\text{Var}[g_{\text{HAIT}}|\gamma]] + \text{Var}[\mathbb{E}[g_{\text{HAIT}}|\gamma]] \tag{48}$$

Because masking proportionally reduces both gradient magnitude and variability:

$$\text{Var}[g_{\text{HAIT}}] \leq (1-\gamma)^2 \cdot \text{Var}[g_{\text{holistic}}] \tag{49}$$

This variance reduction yields more stable gradient updates and accelerated convergence. The effect is particularly pronounced in multimodal caption generation, where γ tends to be large as models typically produce substantial correctly grounded content alongside hallucinated elements, making the focused optimization on error regions significantly more efficient than holistic sequence training.

References

[1] Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29915–29926, 2025. 1

[2] Peter Anderson, Basura T. Murphy, Mark J. Warden, et al. Spice: Semantic propositional image caption evaluation. *European Conference on Computer Vision (ECCV)*, page 382–398, 2016. 6

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1 (2):3, 2023. 1

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 9

[5] 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. 1

[6] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, page 65–72, 2005. 6

[7] Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard. https://huggingface.co/spaces/open-llm-leaderboard, 2023. 13

[8] Rina Buoy, Sovisal Chenda, Nguonly Taing, Marry Kong, Masakazu Iwamura, and Koichi Kise. Addressing the attention drift problem for khmer long textline recognition: R. buoy et al. *International Journal on Document Analysis and Recognition (IJDAR)*, pages 1–26, 2025. 1

[9] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13817–13827, 2024. 2

[10] Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. Perturbollava: Reducing multimodal hallucinations with perturbative visual training. *CoRR*, abs/2503.06486, 2025. 1

[11] Jiuhai Chen, Jianwei Yang, Haiping Wu, Dianqi Li, Jianfeng Gao, Tianyi Zhou, and Bin Xiao. Florence-vl: Enhancing vision-language models with generative vision encoder and depth-breadth fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24928–24938, 2025. 2

[12] Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4209–4221, 2025. 2, 7

[13] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024. 4, 9

[14] 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. 2, 9

[15] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024. 2

[16] Zhiyuan Chen, Yuecong Min, Jie Zhang, Bei Yan, Jiahao Wang, Xiaozhen Wang, and Shiguang Shan. A survey of multimodal hallucination evaluation and detection. *arXiv preprint arXiv:2507.19024*, 2025. 1

[17] Xu Chu, Xinrong Chen, Guanyu Wang, Zhijie Tan, Kui Huang, Wenyu Lv, Tong Mo, and Weiping Li. Qwen look again: Guiding vision-language reasoning models to re-attention visual information. *arXiv preprint arXiv:2505.23558*, 2025. 1

[18] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023. 2

[19] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. 13

[20] 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. 9

[21] Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024. 9

[22] Z. Yue et al. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 2025. MiMo-VL-7B-SFT / RL models. 1, 2, 6

[23] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312, 2024. 9

[24] Jinlan Fu, Shenzhen Huangfu, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, and See-Kiong Ng. Chip: Cross-modal hierarchical direct preference optimization for multimodal llms. *arXiv preprint arXiv:2501.16629*, 2025. 9

[25] Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. Visual description grounding reduces hallucinations and boosts reasoning in lvlms. *arXiv preprint arXiv:2405.15683*, 2024. 1

[26] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024. 1

[27] Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. Damro: Dive into the attention mechanism of lvlm to reduce object hallucination. *arXiv preprint arXiv:2410.04514*, 2024. 2

[28] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, pages 14375–14385, 2024. 1

[29] 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*, pages 18135–18143, 2024. 9

[30] Zhihui Guo, Xin Man, Hui Xu, and Jie Shao. Lisa: A layer-wise integration and suppression approach for hallucination mitigation in multimodal large language models, 2025. 1

[31] Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective, 2024. 9

[32] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. 13

[33] 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. 1, 2, 7, 9

[34] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[35] Junho Kim, Hyunjun Kim, Kim Yeonju, and Yong Man Ro. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. *Advances in Neural Information Processing Systems*, 37:133571–133599, 2024. 2

[36] Sihyeon Kim, Boryeong Cho, Sangmin Bae, Sumyeong Ahn, and Se-Young Yun. Vacode: Visual augmented contrastive decoding. *arXiv preprint arXiv:2408.05337*, 2024. 2

[37] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22 (1):79–86, 1951. 5

[38] Wei Lan, Wenyi Chen, Qingfeng Chen, Shirui Pan, Huiyu Zhou, and Yi Pan. A survey of hallucination in large visual language models. *arXiv preprint arXiv:2410.15359*, 2024. 1

[39] 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. 1, 2, 7, 9

[40] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[41] 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. 2

[42] 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. 9

[43] 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. 1, 9, 12, 13

[44] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges. *arXiv preprint arXiv:2501.02189*, 2025. 2

[45] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL Workshop*, page 74–81, 2004. 6

[46] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699, 2024. 2

[47] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022. 13

[48] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 6, 13

[49] 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. 9

[50] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[51] 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 (CVPR)*, pages 26296–26306, 2024. 2

[52] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. 9

[53] 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. 1

[54] Jiacai Liu, Chaojie Wang, Chris Yuhao Liu, Liang Zeng, Rui Yan, Yiwen Sun, Yang Liu, and Yahui Zhou. Improving multi-step reasoning abilities of large language models with direct advantage policy optimization. *arXiv preprint arXiv:2412.18279*, 2024. 2

[55] Avshalom Manevich and Reut Tsarfaty. Mitigating hallucinations in large vision-language models (lvlms) via language-contrastive decoding (lcd). *arXiv preprint arXiv:2408.04664*, 2024. 9

[56] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443, 1997. 4

[57] Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. *Advances in Neural Information Processing Systems*, 37:116617–116637, 2024. 9

[58] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, page 311–318, 2002. 6

[59] Suzanne Petryk, David Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph Gonzalez, and Trevor Darrell. Aloha: A new measure for hallucination in captioning models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 342–357, 2024. 1

[60] 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. 2

[61] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 6, 9

[62] Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*, 2024. 9

[63] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021. 13

[64] Kanato Sato, Haruto Kaneko, and Mei Fujimura. Reducing cultural hallucination in non-english languages via prompt engineering for large language models. *OSF Preprints*, 10, 2024. 2

[65] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2, 3

[66] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. 2, 3

[67] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, page 1279–1297. ACM, 2025. 2

[68] Zhang Shengyu, Dong Linfeng, Li Xiaoya, Zhang Sen, Sun Xiaofei, Wang Shuhe, Li Jiwei, Runyi Hu, Zhang Tianwei, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023. 1

[69] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587): 484–489, 2016. 9

[70] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Laurent Sifre, George Van Den Driessche, Veda Panneershelvam, Marc Lanctot, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Master-

ing the game of go without human knowledge. *Nature*, 550 (7676):354–359, 2017. 9

[71] 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. 9

[72] 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. 9

[73] 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. 9

[74] Qwen Team. Qwen3 technical report, 2025. 2, 6

[75] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4566–4575, 2015. 6

[76] Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. *CoRR*, abs/2410.11779, 2024. 1

[77] Kaishen Wang, Hengrui Gu, Meijun Gao, and Kaixiong Zhou. Damo: Decoding by accumulating activations momentum for mitigating hallucinations in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 9

[78] Qinxin Wang, Hao Tan, Sheng Shen, Michael Mahoney, and Zhewei Yao. Maf: Multimodal alignment framework for weakly-supervised phrase grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2030–2038, 2020. 1

[79] 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. 9

[80] Zecheng Wang, Xinye Li, Zhanyue Qin, Chunshan Li, Zhiying Tu, Dianhui Chu, and Dianbo Sui. Can we debias multimodal large language models via model editing? In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3219–3228, 2024. 1

[81] Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. *arXiv preprint arXiv:2405.17820*, 2024. 9

[82] Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *arXiv preprint arXiv:2407.19594*, 2024. 9

[83] Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. Self-play preference optimization for language model alignment, 2024. 9

[84] Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 25543–25551, 2025. 9

[85] Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. *CoRR*, abs/2412.13817, 2024. 1

[86] Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14635–14645, 2025. 2, 7

[87] Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Mitigating hallucination in large vision-language models via modular attribution and intervention. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[88] Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2

[89] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2

[90] Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14625–14634, 2025. 1

[91] 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. 2, 9

[92] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 6

[93] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953, 2024. 9

[94] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via

behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024. 9

[95] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D Goodman. Star: Bootstrapping reasoning with reasoning, 2022. *URL https://arxiv. org/abs/2203.14465*, 2203, 2022. 9

[96] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence?, 2019. 13

[97] Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. In *European Conference on Computer Vision*, pages 196–213. Springer, 2024. 9

[98] Yaming Zhang, Jianfei Yu, Wenya Wang, Li Yang, Jia Yang, and Rui Xia. Enhancing multimodal object-entity relation extraction via multi-aspect contrastive learning in large multimodal models. *IEEE Transactions on Audio, Speech and Language Processing*, 2025. 1

[99] Z. Zhao, Y. Deng, W. Zhang, and Q. Gu. Enhancing lvlms through hallucination-aware direct preference optimization. In *arXiv preprint arXiv:2311.16839*, 2023. Introduces the HA-DPO dataset for hallucination mitigation. 6

[100] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023. 9

[101] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025. 2

[102] Tsinghua University Zhipu AI KEG. Glm-4.1v-thinking: Towards versatile multimodal reasoning. *arXiv preprint arXiv:2507.01006*, 2025. GLM-4.1V vision-language model (9B parameters). 2, 6

[103] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023. 2

[104] 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. 2

[105] Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1624–1633, 2025. 1

[106] Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, and Yuexian Zou. Vasparse: Towards efficient visual hallucination mitigation via visual-aware token sparsification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4189–4199, 2025. 1