Mitigating Hallucinations in Large Vision-Language Models by Self-Injecting Hallucinations

Yifan Lu[1,2,3], Ziqi Zhang[1,2], Chunfeng Yuan[1,2,3]*, Jun Gao[4], Congxuan Zhang[5], Xiaojuan Qi[6], Bing Li[1,2,3], Weiming Hu[1,2,3,7],

[1]Beijing Key Laboratory of Super Intelligent Security of Multi-Modal Information, CASIA
[2]State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences
[4]Hello Group [5]Nanchang Hangkong University [6]The University of Hong Kong
[7]School of Information Science and Technology, ShanghaiTech University
luyifan2021@ia.ac.cn,cfyuan@nlpr.ia.ac.cn

Abstract

Large Vision-Language Models (LVLMs) suffer from serious hallucination problems, where the model-generated responses are inconsistent with the visual inputs. Existing hallucination mitigation methods are mainly based on preference alignment and require external human annotations or auxiliary models for preference data collection, which increase costs and limit sustainable improvement. To tackle these challenges, we propose **A**utonomous **P**reference **A**lignment via **S**elf-**I**njection (**APASI**), a novel and generalizable method that mitigates hallucinations without external dependencies. APASI leverages the target LVLM to self-inject hallucinations into a generated response, creating a pair of responses with varying preference levels. During the self-injection process, the dispreferred response is generated based on three key observations of hallucinations, ensuring it simulates real hallucination patterns. This fidelity offers an accurate learning signal for hallucination mitigation. Moreover, APASI incorporates an iterative alignment training strategy combined with curriculum learning to periodically update the preference data with increasing challenge, enabling stable and continuous enhancement of the LVLM. Extensive experiments across six benchmarks show that APASI not only effectively mitigates hallucinations for three baseline models but also achieves comparable or even superior performance to alignment-based methods with external dependency, thereby demonstrating its effectiveness and generalization capability. The code is available at https://github.com/davidluciolu/APASI.

1 Introduction

The study of Large Vision-Language Models (LVLMs) (Li et al., 2023a; Liu et al., 2024b,c; Wang et al., 2024a) has made remarkable progress in recent years. LVLMs substantially enhance cross-modal understanding in vision-language tasks and achieve notable performance across various applications. Despite their demonstrated efficacy, LVLMs frequently encounter hallucination problems which refer to the inconsistency between the factual content of visual input and the corresponding generated textual response (Liu et al., 2024a; Yan et al., 2024). This problem undermines the reliability of LVLMs, making the mitigation of hallucinations a critical area of research.

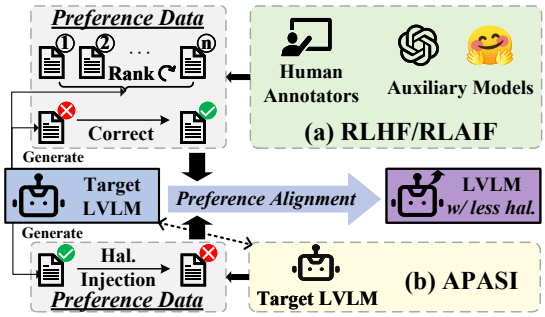


Figure 1: Comparison of different alignment-based hallucination mitigation methods. While all methods employ the framework of preference alignment to optimize models using preference data, they vary in their approaches of collecting the data. Best viewed in color.

Various methods have been proposed to mitigate hallucinations in LVLMs through preference alignment techniques, such as Reinforcement Learning from Human Feedback (RLHF) (Sun et al., 2023; Yu et al., 2024a) and Reinforcement Learning from AI Feedback (RLAIF) (Zhao et al., 2023; Yu et al., 2024b). As shown in Fig.1(a), the target LVLM is trained using ranking or corrective preference data to align with the preference for *no hallucinations* or *reduced hallucinations*. Notably, RLAIF-based methods alleviate the cost and subjectivity issues in RLHF preference data collection (Li et al., 2023b) by replacing human annotators with auxiliary feedback models such as GPT-4 (Achiam et al., 2023). Methods (Zhou et al., 2024c; Ouali et al., 2025) further promote these improvements by replacing

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 13861–13877
November 4-9, 2025 ©2025 Association for Computational Linguistics

proprietary models with open-source models. However, as alignment progresses, the capabilities of the target LVLM become comparable to the fixed and finite capabilities of the feedback models (Yu et al., 2024b), particularly in handling hallucinations. At this point, the feedback models may struggle to detect minor hallucinations in the LVLM's responses or provide effective corrections. Therefore, methods that rely on external models face limitations in achieving sustainable improvement of the target LVLM.

To overcome these limitations, we propose **A**utonomous **P**reference **A**lignment via **S**elf-**I**njection (**APASI**), a novel hallucination mitigation method that constructs preference data using the target model itself. Though following the common framework of using the Direct Preference Optimization (DPO) (Rafailov et al., 2024) algorithm to iteratively align the target LVLM with the preference for *reduced hallucinations*, APASI distinctively utilizes self-generated data, as shown in Fig.1(b). Instead of collecting ranking or corrective feedback as typically done in RLHF/RLAIF methods, APASI generates its data by injecting hallucinations into model-generated responses, creating effective dis-preferred responses that simulate real hallucination patterns. With the original model-generated responses as the preferred, this process forms valid preference pairs for DPO. Meanwhile, the straightforward injection of hallucinations is autonomously executed independently of external auxiliary models, facilitating sustainable model improvement. Moreover, APASI is easily scalable as it requires no annotations.

However, directly injecting hallucinations into responses through bad prompts or visual corruption, as done in (Zhou et al., 2024a; Deng et al., 2024), fails to accurately simulate real hallucination patterns. The self-injection process in APASI is based on three key observations of hallucinations: 1) LVLMs are prone to hallucinate objects that frequently **co-occur** with the existent objects in the image (Li et al., 2023c); 2) LVLMs are prone to generate hallucinated content with an over-reliance on **language priors** (Favero et al., 2024); 3) Hallucinations in LVLMs typically cluster towards the **latter part** of the response (Zhou et al., 2024b). Accordingly, the target LVLM is guided to fabricate sentences about non-existent **co-occurring** objects with **language-only** inputs and then integrates this hallucinated content into the **latter parts** of the preferred response, forming the dis-preferred response. In this way, the preference pair facilitates an accurate learning signal for the mitigation of real hallucination patterns. Moreover, as learning progresses, APASI incorporates a curriculum (Bengio et al., 2009) to gradually reduce the injection of hallucinations, making it more challenging to distinguish subtler differences in the preference pairs, thereby helping to refine the LVLM's ability to identify hallucinations smoothly over iterations.

Our contributions are summarized as follows:

- We propose APASI, a novel hallucination mitigation method for LVLMs. APASI designs a scalable and effective pipeline to autonomously collect preference data by self-injecting hallucinations into model-generated responses, thereby minimizing reliance on external data sources and enabling sustainable improvement.
- APASI leverages key insights into hallucination patterns to accurately construct preference pairs, providing a precise learning signal for hallucination mitigation. APASI further incorporates an iterative alignment strategy with curriculum learning for stable improvement.
- Extensive experiments on various benchmarks validate that APASI effectively mitigates the hallucination problem and enhances performance for baselines including LLaVA-v1.5, LLaVA-v1.6, and Qwen2-VL, showcasing its efficacy and generalization capability.

2 Related Works

2.1 Hallucination in LVLMs

Despite the success, current LVLMs suffer from hallucination problems, where the generated response is inconsistent with the visual inputs. Specifically, this inconsistency is multi-facet, including errors in object, attribute, and relationship (Liu et al., 2024a). Recent research finds key observations of hallucinations including: **Object co-occurrence** (Li et al., 2023c; Zhou et al., 2024b; Leng et al., 2024), **Language prior** (Leng et al., 2024; Favero et al., 2024), and **Positional factor** (Zhou et al., 2024b; Favero et al., 2024). Our proposed APASI is designed to focus on these key observations to construct valid preference pairs.

2.2 Hallucination Mitigation via Alignment

Preference alignment (Ji et al., 2023) has become a prominent strategy for mitigating hallucinations

in models, aiming to align model behavior with the human preference for *no hallucinations* or *reduced hallucinations* (Zhao et al., 2023; Yu et al., 2024a). Early methods (Yu et al., 2024a) use RLHF to collect preference data from human annotators and utilize optimization algorithms such as Proximal Policy Optimization (Schulman et al., 2017) and Direct Preference Optimization (Rafailov et al., 2024) to fine-tune the model. The preference data are typically gathered either by direct collection of rankings for sampled responses (Sun et al., 2023) or by correcting hallucinations in model-generated responses to establish a preferred version (Gunjal et al., 2024). Some studies (Li et al., 2023b; Zhao et al., 2023; Zhou et al., 2024a) use RLAIF to replace human annotators with powerful proprietary models such as GPT-4 (Achiam et al., 2023), reducing costs and enhancing annotation quality. Methods (Zhou et al., 2024c; Yu et al., 2024b; Ouali et al., 2025) further reduce the cost by using open-source models such as CLIP (Radford et al., 2021) and LLaVA-v1.6 (Liu et al., 2024c).

Recent studies have explored self-improvement mechanisms (Huang et al., 2023; Chen et al., 2024), where preference data are derived from the optimized model itself. SIMA (Wang et al., 2024b) uses critic prompts to self-rank sampled responses, but relies on ground-truth references, limiting scalability. STIC (Deng et al., 2024) generates dis-preferred responses through misleading questions or corrupted images, eliminating the need for ground truth. In contrast, APASI directly injects hallucinations based on key hallucination patterns, requiring no external models, annotations, or complex prompt design.

3 Methodology

We first provide preliminaries of the DPO algorithm in Section3.1. Then in Section3.2, we dive into the preference data construction process based on pertinent self-injection of hallucination to provide an accurate learning signal for DPO training. Finally, we introduce the training scheme for the APASI framework using iterative alignment strategy with curriculum learning to achieve continuous and stable optimization in Section3.3.

3.1 Direct Preference Optimization

The LVLM with parameters θ denoted as M_θ, defines a conditional distribution $p_\theta(y|v,x)$, where y denotes the output response for the input image v and the text prompt x. The proposed APASI leverages DPO (Rafailov et al., 2024) to tune the parameters θ and align the LVLM toward the preference of reduced hallucinations. DPO directly learns from the preference data defined as $\mathcal{D} = \{(v_i, x_i, y_i^+, y_i^-)\}_{i=1}^N$, where the preference pair (y_i^+, y_i^-) consists of a preferred and a dis-preferred response for the input v_i and x_i. The optimization target is defined as:

$$\max_\theta \mathbb{E}_{(v_i,x_i,y_i^+,y_i^-)\sim\mathcal{D}} \Big[\log\sigma\Big(\beta\big(\log\frac{p_\theta(y_i^+|v_i,x_i)}{p_{ref}(y_i^+|v_i,x_i)} - \log\frac{p_\theta(y_i^-|v_i,x_i)}{p_{ref}(y_i^-|v_i,x_i)}\big)\Big)\Big], \tag{1}$$

where p_{ref} is defined by a reference model (usually fixed at the initial training checkpoint of M_θ) and β is a hyper-parameter to control the KL-divergence between M_θ and the reference model.

3.2 Hallucination Self-Injection

The proposed APASI designs an autonomous preference data construction pipeline that uses only the target LVLM, without any need for additional annotation or auxiliary model. The data construction pipeline is predicated on the concept that a response y_i^+ which may already contain some hallucinations, and a response y_i^- explicitly corrupted with additional hallucinations relative to y_i^+, form valid preference pair with varying preference levels. Given an unannotated dataset $\mathcal{D}_{un} = \{(v_i, x_i)\}_{i=1}^N$ without ground-truth responses, APASI constructs preference pairs as:

- The **preferred** y_i^+: an original response generated by the target model M_{θ_0} with initial parameters θ_0, i.e., $y_i^+ \sim p_{\theta_0}(\cdot|v_i, x_i)$.
- The **dis-preferred** y_i^-: a response that is created by deliberately injecting additional hallucinations into y_i^+.

Specifically, to better simulate the real hallucination patterns in the dis-preferred response, APASI injects hallucination targeted to key observations of hallucinations including: 1) **Object co-occurrence**, 2) **Language prior**, and 3) **Positional factor**. We provide empirical evidence to support the importance of these observations in Appendix B.1. Accordingly, APASI guides the target model M_{θ_0} to generate sentences about non-existent **co-occurring** objects with **language-only** input and

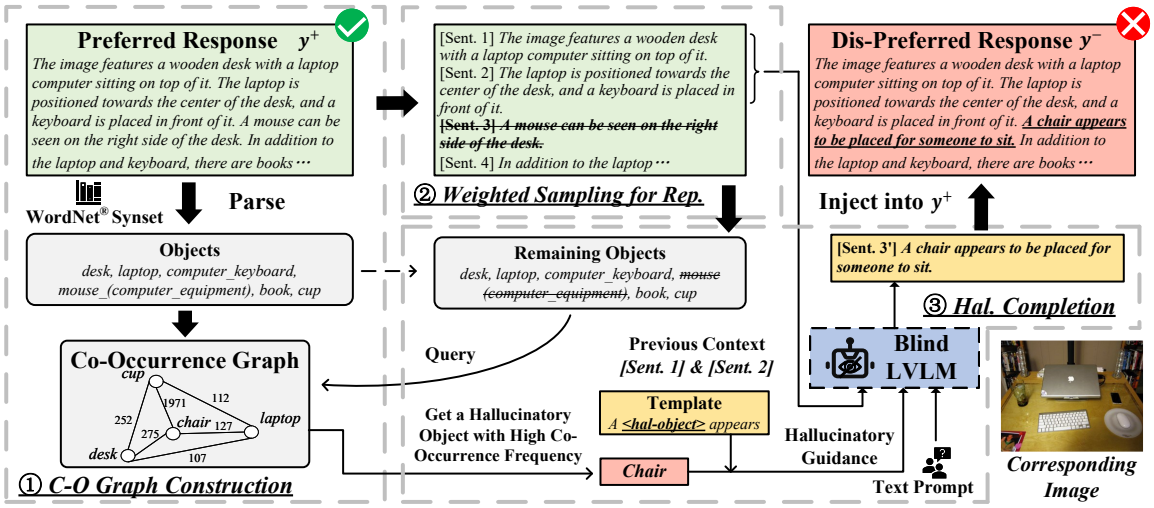


Figure 2: Illustration of the self-injection pipeline: APASI first preprocesses the preferred responses to construct a graph documenting the co-occurrence relationships among objects. APASI then selects sentences to be replaced through weighted sampling, and employs the "blind" target LVLM to make a hallucination completion for replacement with language-only input and hallucinatory guidance. The completed hallucinated sentence is injected into the preferred response to get the dis-preferred one. Best viewed in color.

replaces the **latter part** of y_i^+ with these hallucinated sentences to obtain y_i^-. The detailed process is shown in Fig.2 and is described below.

3.2.1 Co-Occurrence Graph Construction

To simulate co-occurring hallucination, it is necessary to uncover the biased co-occurrence relationships among objects inherent in the M_{θ_0} that exhibits such hallucinations. APASI prepossesses the whole corpus of model-generated preferred responses $\{y_i^+\}_{i=1}^N$ to construct a co-occurrence graph G documenting these relationships. Each preferred response y_i^+ is first parsed into a set object tags o_i, using the WordNet (Miller, 1995) toolbox and synonym sets S. This process consolidates various synonyms into a single category tag for each object to simplify the analysis. APASI then uses these tags $\{o_i\}_{i=1}^N$ to build the co-occurrence graph G, where each node represents an object tag and the edge weight reflects the frequency of co-occurrence between the connected objects. Given an querying object, it's easy to get its co-occurring objects by traversing the nodes connected to the corresponding node.

3.2.2 Weighted Sampling for Injection

APASI injects hallucinations into y_i^+ by replacing a proportion ρ (injection rate) of the L original sentences with hallucinated counterparts. The indices of the replaced sentences are sampled from a multinomial distribution defined by parameters $w_{1:L}$. Recognizing that hallucinations commonly occur in the latter part of the responses, we empirically set the last sentence twice as likely to be sampled as the first one. The weight of k-th sentence is defined as $w_k = 1 + \frac{k-1}{L}$ by linear interpolation.

3.2.3 Hallucination Completion and Injection

Suppose APASI samples the k-th sentence in y_i^+ to be replaced with a hallucinated sentence $y_{i,k}^-$, resulting in the dis-preferred response:

$$y_i^- = (y_{i,1}^+, \ldots, y_{i,k-1}^+, y_{i,k}^-, y_{i,k+1}^+, \ldots, y_{i,L}^+). \tag{2}$$

Considering the hallucination patterns of co-occurrence and language prior, the hallucinated $y_{i,k}^-$ is generated by a language-only "blind" LVLM under the guidance of a hallucinated co-occurring object $o_{i,k}^{hal}$, forming a description of $o_{i,k}^{hal}$. Specifically, $o_{i,k}^{hal}$ is obtained by querying the co-occurrence graph G for an object that frequently co-occurs with objects in the remaining sentences, while ensuring $o_{i,k}^{hal}$ not in o_i. $o_{i,k}^{hal}$ is then put into a pre-defined guiding template, *e.g., "A ($o_{i,k}^{hal}$) appears"*, which serves as hallucinatory guidance for the "blind" M_{θ_0} to make a hallucination completion $\tilde{y}_{i,k}$ as:

$$\tilde{y}_{i,k} \sim p_{\theta_0}\big(\cdot\,|x_i, y_{i,<k}^+, temp(o_{i,k}^{hal})\big), \tag{3}$$

where the text-only input includes the original prompt x_i, the previous context $y_{i,<k}^+$ of k-th sentence, and the filled template $temp(o_{i,k}^{hal})$. Note

that if sentences after the k-th one are sampled to be replaced, the previous context includes $y^-_{i,k}$ instead of $y^+_{i,k}$. $y^-_{i,k}$ is finally obtained by composing the template and the completion as $y^-_{i,k} = (temp(o^{hal}_{i,k}), \tilde{y}_{i,k})$.

3.3 Iterative Alignment with Curriculum Learning

To alleviate the distribution shift problem in preference alignment (Gao et al., 2023; Yu et al., 2024b), APASI employs an iterative alignment strategy. At iteration t, the latest optimized model $M_{\theta_{t-1}}$ is used for preference data construction including preferred responses generation and hallucination injection. $M_{\theta_{t-1}}$ is then optimized with DPO target in Equation (1) and preference data to get M_{θ_t} for data construction in the next iteration.

Furthermore, APASI incorporates a curriculum that progressively increases the difficulty of the alignment task, thereby facilitating a smoother and more effective learning trajectory over iterations (Bengio et al., 2009). The curriculum specifically reduces the injection rate ρ with each iteration t, according to a monotonically decreasing curriculum function $f_c(t)$. As ρ decreases, the gap between the preferred response y^+ and the dis-preferred response y^- narrows, intensifying the challenge of distinguishing subtle differences within the preference pairs (Xu et al., 2023). This progressive increase in task difficulty is crucial for refining the LVLM's ability to accurately detect and reduce hallucinations. The detailed algorithm for Iterative Alignment with Curriculum Learning outlined in Appendix A.3.

4 Experiments

4.1 Experimental Setups

Preference Data. The construction of preference data in APASI leverages self-injection based on the detailed description task, which requires LVLMs to accurately perceive and describe visual elements, thus directly reflecting hallucination issues. We construct the **SI-23k** dataset derived from images and descriptive responses in the *detail-23k* subset of the LLaVA's instruction tuning dataset, excluding ground-truth responses. We further construct the scaled-up **SI-130k** by adding unannotated images from the VisualGenome (VG) dataset (Krishna et al., 2017). The descriptive prompts in SI-130k are from SI-23k. For object parsing, we employ LVIS object synonym sets (Gupta et al., 2019), which are based on WordNet. The default injection rate ρ is set to 0.2.

Implementation Details. In this study, we choose the supervised fine-tuned LLaVA-v1.5-7B (Liu et al., 2024b), a widely used baseline, as the target model for hallucination mitigation. Unless specified otherwise, all experiments are conducted with LLaVA-v1.5-7B and SI-23k. We use SI-23k and SI-130k to train the target LVLM in a single iteration to obtain **APASI-Base** and **APASI-Scaled**, respectively. For iterative alignment with curriculum learning (IACL), we train the LVLM for $T = 3$ iterations with a curriculum function $f_c(t) = 0.8 - 0.2t$ and get **APASI-IACL**. A single iteration takes about 520 minutes running with 8 V100 GPUs and data construction accounts for about 31.2% of time.

Evaluation. For quantitative analysis, we conduct evaluation on both hallucination benchmarks (Object-Hal (Rohrbach et al., 2018), AMBER (Wang et al., 2023b), and POPE (Li et al., 2023c)) and comprehensive benchmarks (MMBench (Liu et al., 2025), MMVet (Yu et al., 2023), and LLaVABench (Liu et al., 2024d)). Object-Hal provides CHAIR-i (C-i) and CHAIR-s (C-s) metrics measuring the ratio of hallucinated objects and responses respectively in generative tasks. POPE uses F1 scores in discriminative tasks. AMBER reports both C-i and F1. We report overall scores for the comprehensive benchmarks.

4.2 Main Results

4.2.1 Performance Comparison with the Baseline Model

Results show that APASI effectively mitigates the hallucination problem in the LLaVA-v1.5-7B baseline model, as shown in Tab.1. Specifically, APASI-Base reduces the ratio of hallucinated objects by 4.5/1.8 in generative tasks on Object-Hal and AMBER, respectively, and reduces the ratio of hallucinated responses by 12.9 on Object-Hal. APASI-Base also improves the performance on all three comprehensive benchmarks for the baseline. Further incorporating IACL makes improvement on most of the benchmarks, suggesting the effectiveness of IACL. Notably, both APASI-Base and APASI-IACL are trained with SI-23k, where the images and the textual prompts are already used in the supervised fine-tuning of the baseline. The enhancement brought about by seen data indicates

Model	Hallucination Benchmarks					Comprehensive Benchmarks		
	Object-Hal		AMBER		POPE	MMBench TEST-v1.1	MMVet	LLaVA BENCH
	CHAIR-s ↓	CHAIR-i ↓	CHAIR-i ↓	F1	F1			
LLaVA-v1.5-7B	51.0	13.7	7.8	74.7	85.9 (86.9)	62.3	30.5	63.4
+ POVID †	33.6	9.0	5.2	86.5	86.9	64.9	31.8	68.7
+ HA-DPO †	-	-	3.7	82.9	84.3	-	-	67.2
+ RLAIF-V ‡	**20.8**	6.0	**2.8**	84.5	78.9	63.6	30.1	64.9
+ CLIP-DPO ‡	-	-	7.2	80.5	85.8	-	-	-
+ CSR ‡	28.0	7.5	4.4	86.5	**87.0**	65.4	33.9	71.1
+ STIC ¶	-	-	-	-	-	65.3	32.6	68.9
+ SIMA ¶	41.6	13.0	6.6	**86.9**	85.8	64.9	31.6	66.1
+ OPERA §	47.8	14.6	-	-	85.4	64.4	-	60.3
+ VCD §	48.6	14.9	-	74.9	84.5	-	-	65.8
+ Less is more §	36.8	11.3	-	75.8	86.0	-	-	60.9
+ APASI-Base	38.1	9.2	6.0	86.1	85.6 (87.0)	66.7	33.5	67.3
+ APASI-IACL	31.7	7.2	5.7	85.7	85.0 (87.0)	65.6	**34.4**	**71.2**
+ APASI-Scaled	23.2	**5.1**	3.5	86.7	85.0 (87.4)	**67.2**	32.2	70.1

Table 1: Performance of APASI with the LLaVA-v1.5-7B baseline compared with other hallucination mitigation methods across various benchmarks. †, ‡, ¶, and § respectively indicate RLAIF methods with proprietary models, RLAIF methods with open-source models, self-improvement methods, and non-alignment-based methods.

the effectiveness of the preference-alignment training paradigm in deepening the LVLM's utilization of existing data. Scaling up preference data also takes effect, especially in reducing hallucinations. APASI-Scaled with SI-130k further reduces hallucination ratios of the baseline by 27.8/8.6/4.3 on Object-Hal and AMBER.

In contrast to generative performance, APASI shows inconsistent discriminative results, particularly under-performing the baseline on POPE. This discrepancy stems from APASI's alignment goal, which does not directly optimize discriminative capabilities. To better ground discriminative ability to the hallucination-reduced descriptions, we prompt the LVLMs with *Describe the image and answer the question.* during POPE testing. The modified F1 scores, displayed next to the original scores in Table 1, show improvements: the baseline increases by 1.16%, while APASI-Base/IACL/Scaled improve by $1.64\%/2.35\%/2.82\%$, These results indicates that the LVLMs with better generative performance exhibit greater improvement.

4.2.2 Performance Comparison with the SOTA Methods

We also compare APASI with the SOTA hallucination mitigation methods in Tab.1. The methods are categorized as follows: 1) RLAIF methods with proprietary models: POVID (Zhou et al., 2024a) and HA-DPO (Zhao et al., 2023); 2) RLAIF methods with open-source models: RLAIF-V (Yu et al., 2024b), CLIP-DPO (Ouali et al., 2025), and CSR (Zhou et al., 2024c); 3) self-improvement methods: STIC (Deng et al., 2024) and SIMA (Wang et al., 2024b); 4) methods without preference alignment: OPERA (Huang et al., 2024), VCD (Leng et al., 2024) and Less is more (Yue et al., 2024). Methods in categories 1) and 2) rely on external resources, whereas those in 3) and 4) do not.

Without the need for any external support, APASI achieves comparable or even better performance across all three hallucination benchmarks compared to methods dependent on external resources. APASI also outperforms all methods without external dependencies on four out of five hallucination metrics. For comprehensive abilities, APASI performs the best across all three benchmarks, highlighting its effectiveness in both hallucination mitigation and overall ability enhancement.

4.3 Ablation Studies and Analysis

4.3.1 Ablation Studies on Injection Settings

As shown in Tab.2, we design ablative experiments to verify the effectiveness of the settings in getting the dis-preferred responses trough self-injection. We specifically compare APASI-Base with five variants when injecting hallucinations. Qualitative analyses are in Appendix C.1.

Results show that it will bring performance drop if using random non-existent objects instead of co-occurring ones for hallucinatory guidance. Fur-

Setting	Object-Hal		AMBER
	C-s ↓	C-i ↓	C-i ↓
LLaVA-v1.5-7B	51.0	13.7	7.8
APASI-Base	38.1	9.2	6.0
random guide	44.4	12.2	7.0
w/o guide	66.7	21.0	11.3
replace object	49.6	13.5	8.1
w/o w.s.	39.7	9.7	6.0
GT as preferred	38.1	9.7	5.9

Table 2: Ablation studies on injection settings.

ther removing the guiding template for completion even causes the model to underperform the baseline. This is because the language-only model may not reliably generate hallucinated sentences only based on textual prompt and the previous context, making the preference pairs invalid. These results verify the effectiveness of the hallucinatory guidance with co-occurring objects.

It will also bring performance drop if injecting hallucinations by replacing with the co-occurring object word, instead of a sentence describing the object. Only replacing words may result in the injected target words being markedly inconsistent with the context. Such dis-preferred responses with ridiculous mistakes fail to provide effective learning signal, verifying the importance of employing the language-only model for injection to maintain reasonableness in the dis-preferred responses.

The decline observed from removing weighted sampling when deciding sentences to be replaced shows the necessity of considering positional factor in self-injection. Using the ground-truth responses instead of the model-generated ones as the preferred responses to be injected does not result in performance changes. However, this will limit APASI to scale up to unannotated data.

4.3.2 Analysis on IACL and Sustainability of Improvement.

We first select 1000 samples from SI-23k to plot the histograms of the preferred and dis-preferred log probabilities with different injection rates in Fig.3. Results show that as the injection rate decreases, the dis-preferred responses are more likely to be generated and less easily distinguishable by the target LVLM, thus making the alignment task more challenging. This observation supports our strategy of employing a curriculum with decreasing injection rate.

To evaluate the effectiveness of iterative alignment with curriculum learning, we plot performance on MMVet across iterations in Fig.4(a). Besides APASI-IACL with a decreasing injection rate ρ (0.6 to 0.2), we evaluate three variants: 1) a reverse curriculum where ρ increases from 0.2 to 0.6; 2) ρ fixed at 0.6; 3) ρ fixed at 0.2. In single-round alignment, lower ρ yields better performance, suggesting harder preference pairs promote learning. Detailed results for different ρ are in the Appendix B.5. Over iterations, APASI with fixed ρ and the reverse curriculum show fluctuating improvements, while APASI-IACL with decreasing ρ demonstrates a smooth trajectory and achieves the best performance at iter3 with a 3.6% gain on MMVet compared to iter1. These results verify that a well-designed curriculum fosters smooth learning to benefit sustainable improvement over iterations.

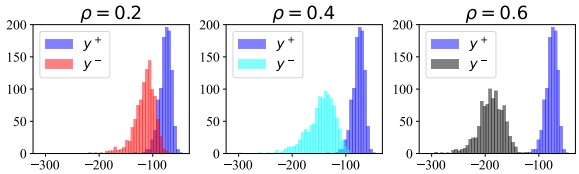


Figure 3: Comparative histograms of preferred vs dis-preferred log probabilities with different injection rates. Best viewed in color.

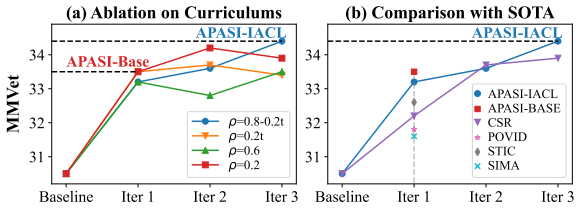


Figure 4: MMVet scores of: **(a)** APASI at each iterations with different curriculums for injection rate; **(b)** different models at each iterations. Best viewed in color.

We further compare APASI-IACL with SOTA methods in 4(b). Most methods only perform a single-round alignment and are outperformed by APASI-IACL. CSR (Zhou et al., 2024c) supports iterative alignment. However, CSR's reliance on an external model for preference data collection limits its effectiveness, thus it yields inferior results to the autonomous APASI-IACL. These results validate APASI's advantage in facilitating sustainable model improvement.

4.3.3 Analysis on generalization capability

To assess APASI's applicability across different LVLMs, we apply APASI to two additional base-

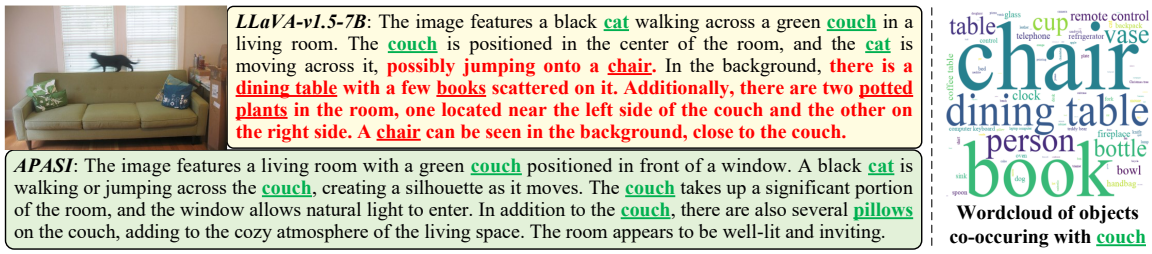


Figure 5: **Left**: Comparison of baseline and APASI responses. Correct and incorrect objects are underlined in red and green respectively. Other hallucinated contents are in red. **Right**: The wordcloud of objects co-occurring with couch obtained from baseline's generations.

Model	Object-Hal		MMVet	LLaVA BENCH
	C-s ↓	C-i ↓		
LLaVA-v1.6-7B	38.1	8.8	42.5	75.6
+ APASI-Base	**28.8**	**7.2**	**44.2**	**80.4**
Qwen2-VL-7B	44.4	9.5	62.0	92.3
+ APASI-Base	**23.8**	**6.2**	**63.7**	**93.5**

Table 3: Performance of APASI with LLaVA-v1.6-7B and Qwen2-VL-7B baselines.

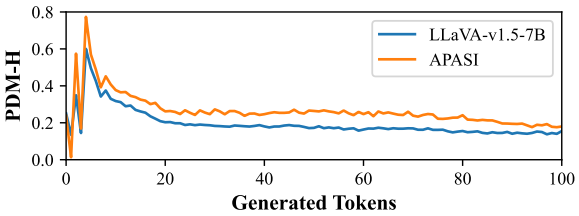


Figure 6: Averaged PDM-H for responses generated by the baseline and APASI. Best viewed in color.

line models: LLaVA-v1.6-7B(vicuna) (Liu et al., 2024c), an improved version of LLaVA-v1.5-7B, and Qwen2-VL-7B (Wang et al., 2024a), which has an different architecture from the LLaVA series and is among the most advanced open-source LVLMs. Note that LLaVA-v1.6-7B and Qwen2-VL-7B autonomously generate their versions of the SI-23k dataset for preference alignment. Results in Tab.3 show that even these advanced LVLMs are not entirely unaffected by hallucinations. APASI effectively mitigates this problem and improves overall performance on comprehensive benchmarks for both LLaVA-v1.6-7B and Qwen2-VL-7B. These results demonstrate the APASI's robust generalization capabilities, particularly its compatibility with difference model architectures.

4.3.4 Analysis on Reliance of Language Priors

To assess the reliance on **language priors**, a key hallucination pattern, we calculate the Prompt Dependency Measure based on Hellinger distance (PDM-H) (Favero et al., 2024). PDM-H quantifies the divergence in the LVLM's probability distribution when generating a token under image-language input versus language-only input. A lower PDM-H indicates greater reliance on the textual input. We randomly sample 1000 images from COCO-test2017 (Lin et al., 2014) and generate detailed descriptions using the LLaVA-v1.5-7B baseline and APASI respectively. The averaged PDM-H curves, depicted in Fig.6, show that APASI exhibits higher PDM-H than the baseline. This result indicates APASI's effectiveness in mitigating the over-reliance on **language priors**. More details about PDM-H are given in A.5.

4.4 Qualitative Analysis

To intuitively show the effectiveness of APASI in mitigating hallucinations, we compare responses from LLaVA-v1.5-7B baseline and APASI in Fig.5. Both models are given with the same image from COCO-test2017 with a detailed description prompt. The response of the baseline exhibited significant hallucinations with non-existent objects in the latter part. Notably, incorrect objects like dining table, chair, and book. usually appears in indoor scenes together with couch. This pattern of co-occurrence is further verified by the wordcloud of **co-occurring** objects in Fig.5, which derived from statistical analysis of the baseline's outputs. Remarkably, APASI effectively eliminates the **co-occurring** hallucination, while capturing all major objects in the image.

5 Conclusions

We propose APASI, a novel method for mitigating hallucinations in LVLMs via preference alignment without external dependencies. This self-injecting process, based on key observations, has provided

an accurate learning signal for effective preference alignment. APASI has further employed iterative alignment with curriculum learning for improving the training process. Extensive experiments have demonstrated our effectiveness and superiority across three baselines and various benchmarks.

Limitations

Our experiments were conducted under computational resource constraints, which restricted the application of APASI to relatively small-scale models (e.g., 7B parameters). In contrast, real-world deployments increasingly rely on larger models, and assessing APASI's effectiveness in such settings remains an important direction for future work. Additionally, similar to many prior studies, APASI primarily addresses inconsistencies between visual inputs and generated textual responses, i.e., fidelity-related hallucinations. While our results suggest that APASI also yields improvements on knowledge-intensive tasks, hallucinations stemming from factual inaccuracies about real-world knowledge are not explicitly modeled in the current framework. Addressing this limitation will be the focus of future research.

Acknowledgments

This work is supported by Beijing Natural Science Foundation (L243015, L223003), the Natural Science Foundation of China (No. 62036011, 62192782), the Project of Beijing Science and Technology Committee (No. Z231100005923046).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. In *Forty-first International Conference on Machine Learning*.

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, James Zou, Kai-Wei Chang, and Wei Wang. 2024. Enhancing large vision language models with self-training on image comprehension. *arXiv preprint arXiv:2405.19716*.

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

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.

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

Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. 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.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, and 1 others. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing.

2024. 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.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

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

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024d. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2025. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. 2025. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In *European Conference on Computer Vision*, pages 395–413. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

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

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. 2023a. V3det: Vast vocabulary visual detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19844–19854.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023b. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Furong Huang, and 1 others. 2024b. Enhancing visual-language modality alignment in large vision language models via self-improvement. *arXiv preprint arXiv:2405.15973*.

Canwen Xu, Corby Rosset, Luciano Del Corro, Shweti Mahajan, Julian McAuley, Jennifer Neville, Ahmed Hassan Awadallah, and Nikhil Rao. 2023. Contrastive post-training large language models on data curriculum. *arXiv preprint arXiv:2310.02263*.

Bei Yan, Jie Zhang, Zheng Yuan, Shiguang Shan, and Xilin Chen. 2024. Evaluating the quality of hallucination benchmarks for large vision-language models. *arXiv preprint arXiv:2406.17115*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024a. 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.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and 1 others. 2024b. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Zihao Yue, Liang Zhang, and Qin Jin. 2024. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *arXiv preprint arXiv:2402.14545*.

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

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024a. Aligning modalities in vision large language models via preference fine-tuning. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024b. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*.

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

A Details about the Method and Implementation

A.1 Details of Hallucination Self-Injection

For a preferred response with L sentences and an injection rate of ρ, the number of the sentence to be replaced is rounded ρL. If the number is zero, the sample will be discarded. Note that the first sentence in the preferred response is excluded from weighted sampling, to avoid insufficient previous text for hallucination completion.

The average sentence number is 5.5 for responses generated by LLaVA-v1.5-7B on the source dataset of LLaVA-detail-23k (Liu et al., 2024d). The corresponding SI-23k dataset with ρ set to 0.2 contains 23,196 samples after discarding invalid samples.

- "A <hal-object>appears"
- "There is a <hal-object>"
- "There are <hal-object>"
- "<hal-object>can also be seen"
- "<hal-object>can be seen"
- "You can see a <hal-object>"
- "There are multiple <hal-object>"
- "Several <hal-object>can be observed"
- "Some <hal-object>are present"
- "Among the items, there is a <hal-object>"
- "In the image, there is a <hal-object>"
- "On the right, there is a <hal-object>"
- "On the left, a <hal-object>is present"
- "In the center, you see a <hal-object>"
- "At the top, there is a <hal-object>"
- "At the bottom, a <hal-object>is visible"
- "In the background, a <hal-object>can be seen"
- "In the foreground, there is a <hal-object>"
- "To the side, a <hal-object>is located"
- "Near the edge, a <hal-object>appears"
- "Close to the center, a <hal-object>is seen"

Table 4: The list of hallucinatory guiding templates.

A.2 Hallucinatory Guiding Templates

APASI employs guiding templates to guide the target LVLM to fabricate content about non-existent co-occurring objects for hallucination self-injection. We pre-define the templates according to

Algorithm 1 APASI incorporating iterative alignment with curriculum learning

Input: Unannotated dataset: $\mathcal{D}_{un} = \{(v_i, x_i)\}_{i=1}^{N}$. Initial target LVLM: M_{θ_0}. Synonym sets for objects: S. Curriculum function for injection rate: $f_c(\cdot)$. Number of iterations: T.
Output: M_{θ_T} with reduced hallucination

for $t = 1, \ldots, T$ **do**
  $\{y_i^+\}_{i=1}^{N} \leftarrow$ Generation$(M_{\theta_{t-1}}, \mathcal{D}_{un})$
  $\{o_i\}_{i=1}^{N}, G \leftarrow$ Preprocessing$(\{y_i^+\}_{i=1}^{N}, S)$
  $\rho = f_c(t)$
  $\{y_i^-\}_{i=1}^{N} \leftarrow$ HalInjection$(M_{\theta_{t-1}}, \{y_i^+\}_{i=1}^{N}, \{o_i\}_{i=1}^{N}, G, \rho)$
  $\mathcal{D} \leftarrow \{(v_i, x_i, y_i^+, y_i^-) \mid i = 1 \text{ to } N\}$
  $M_{\theta_t} \leftarrow$ DPOTraining$(M_{\theta_{t-1}}, \mathcal{D})$
end for
return M_{θ_T}

the usual structure of descriptive sentence, as listed in Tab.4. During every self-injection, one of the templates is randomly sampled and instantiated by filling in the co-occurring object <hal-object>. The target LVLM then follows the template to make a hallucinatory completion.

A.3 Details of Iterative Alignment with Curriculum Learning

As outlined in Algorithm 1, at each iteration, the preferred responses $\{y_i^+\}_{i=1}^{N}$ are generated using the latest model $M_{\theta_{t-1}}$ and are preprocessed to get the object tags $\{o_i\}_{i=1}^{N}$ and the co-occurrence graph G. Following this, $M_{\theta_{t-1}}$ inject hallucinations into $\{y_i^+\}_{i=1}^{N}$ to get the dis-preferred responses $\{y_i^-\}_{i=1}^{N}$, with a injection rate $\rho = f_c(t)$ determined by the curriculum. Subsequently, $M_{\theta_{t-1}}$ is trained with DPO optimization target with the preference dataset $\mathcal{D}$, resulting in the optimized M_{θ_t} for data construction in the next iteration. In this way, APASI operates smoothly in an iterative manner, ensuring sustainable improvement.

A.4 Descriptive Prompts

In the experiments, we scale-up the preference data to images from the VisualGenome (VG) (Krishna et al., 2017), where the textual prompts are provided for the images. For each image in VG, we simply pair it with a random descriptive prompt in SI-23k, which is sourced from the detail-23k subset of the LLaVA's instruction tuning dataset (Liu et al., 2024d). The prompts are listed in Tab.5

- "What is this photo about?"
- "Explain the visual content of the image in great detail."
- "Describe the following image."
- "What do you see happening in this image?"
- "Analyze the image in a comprehensive and detailed manner."
- "Write a detailed description of the given image."
- "What's happening in the scene?"
- "What do you think is going on in this snapshot?"
- "Can you elaborate on the elements of the picture provided?"
- "What are the key elements in this picture?"
- "Can you describe the main features of this image for me?"

Table 5: The list of descriptive prompts used in preference data construction.

A.5 Prompt Dependency Measure

We calculate the Prompt Dependency Measure based on Hellinger distance (PDM-H) (Favero et al., 2024) measuring the difference of LVLM's generative probability distribution when given image-language input and language-only input. It is calculated at the j-th step as follows:

$$\text{PDM-H}(j) = H\big(p(\cdot|v, x, y_{<j}), p(\cdot|x, y_{<j})\big), \tag{4}$$

The Hellinger distance is defined as:

$$H(p, q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{d} (\sqrt{p_i} - \sqrt{q_i})^2}, \tag{5}$$

Baseline Model	LLaVA-v1.5-7B	LLaVA-v1.6-7B	Qwen2-VL-7B
β	0.1	0.1	0.1
#epoch	3	1	1
learning rate	4e-7	1e-6	1e-6
batchsize	64	64	64
lora_r	128	64	64
lora_alpha	256	128	128

Table 6: Training configurations for different baseline models.

Removed Sentence	None	1st	2nd	3rd	4th	5th
Object-Hal C-s ↓	51.0	49.8	49.6	47.0	39.4	37.0

Table 7: Sentence-level hallucination rate of the response after removing sentences at difference positions.

where $p = (p_1, p_2, \ldots, p_d)$ and $Q = (q_1, q_2, \ldots, q_d)$ are two d-dimension discrete probability distributions.

A.6 Implementation Details of Training

During the alignment training, we employ LoRA (Hu et al., 2021) for efficient tuning. Visual encoders of all three models are frozen. The hyperparameters for different baseline models are listed in Tab.6. All experiments are conducted on 8 V100 32GB GPUs. It takes about 6/8/8 hours for training LLaVA-v1.5-7B, LLaVA-v1.6-7B, and Qwen2-VL-7B respectively.

B Quantitative Results

B.1 Empirical Evidence on Hallucination Observations

The hallucination self-injection process in APASI is based on three key observations: **object co-occurrence**, **language prior** and **positional factor**. These observations are not only consistent with prior studies but are also substantiated by our empirical analyses. For **object co-occurrence**, Object-Hal evaluations of LLaVA-v1.5-7B's responses reveal that 98% of hallucinated objects co-occur with correct ones, with 25.9% appearing as the top-1 co-occurring object and 67.9% within the top 5. For **language prior**, Fig.6 shows that models with fewer hallucinations yield higher PDM-H scores, indicating reduced reliance on language priors. For **positional factor**, as shown in Tab.7, the hallucination rate decreases progressively as later sentences (from 1st to 5th) are removed from the response, indicating that hallucinations are more likely to appear in later parts of the response. Together, these findings confirm the soundness of our design principles in capturing the key factors underlying hallucination generation.

Model	Object-Hal C-s↓	MMVet
LLaVA-1.5-7B	51.0	30.5
LLaVA-1.5-13B	49.2	33.5
LLaVA-1.5-7B+APASI-Base	**38.1**	**33.5**
LLaVA-1.6-7B	38.1	42.5
LLaVA-1.6-13B	30.2	43.9
LLaVA-1.6-7B+APASI-Base	**28.8**	**44.2**

Table 8: Performance comparison with larger models.

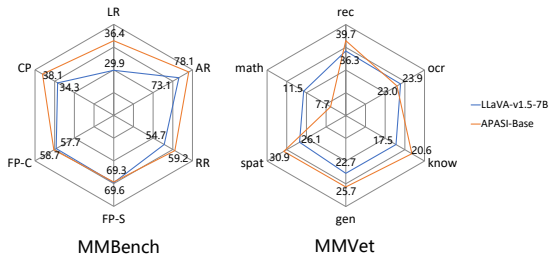


Figure 7: Detailed Results on MMBench (Left) and MMVet (Right). Best viewed in color.

B.2 Performance Comparison with Larger Models

To evaluate the practical competitiveness of APASI, we conduct comparisons against larger-scale models, specifically LLaVA-1.5-13B and LLaVA-1.6-13B. As shown in Tab.8, despite the substantial parameter gap, our 7B models trained with APASI attain performance that is comparable to, or even surpasses that of the corresponding 13B models in the same series. These results further underscore the effectiveness and efficiency of our approach.

B.3 Detailed Results on Comprehensive Benchmarks

We evaluate comprehensive abilities of LVLMs on MMBench (Liu et al., 2025), MMVet (Yu et al., 2023), and LLaVABench (Liu et al., 2024d). MMBench evaluates the reasoning and perception capabilities and subdivides them into six Level-2 capability dimensions including: logic reasoning (LR), attribute reasoning (AR), relation reasoning (RR), fine-grained perception-single instance (FP-S), fine-grained perception-cross instance (FP-C), and coarse perception (CP). MMVet evaluates six core capabilities of the LVLMs including recognition (rec), knowledge (know), OCR, Spatial aware-

Model	complex	conv	detail	overall
LLaVA-v1.5-7B	79.5	51.8	56.6	65.5
APASI-Base	80.5	53.1	60.7	67.3
Δ	1.26%	2.51%	**7.24%**	2.74%

Table 9: Detailed Results on LLaVA-BENCH.

Synonym Sets	Object-Hal		AMBER
	C-s ↓	C-i ↓	C-i ↓
WordNet(LVIS)	38.1	9.2	33.5
V3Det	**31.7**	**8.7**	**33.8**

Table 10: Ablation studies on injection rates.

ness (Spat), Language generation (gen), and math. LLaVA-BENCH evaluates the model's on three tasks including conversation (conv), detailed description (detail), and complex reasoning (complex).

Fig.7 shows detailed results on MMBench and MMVet. APASI-Base outperforms the baseline on all capabilities except math and OCR. This can be explained by the alignment target of hallucination mitigation. APASI focuses on improving the ability of perception for objects, but fails to cover perception for characters or mathematical symbols. Tab.9 shows detailed results on LLaVA-BENCH. APASI-Base outperforms the baseline on all three tasks. The improvement on the detailed description task is the highest among all. Notably, APASI achieves improvement on the reasoning capabilities on all three benchmarks, though the preference alignment in APASI doesn't optimize these capabilities directly. This suggests that reasoning abilities can benefit from the improvement of perception abilities.

B.4 Analysis on Synonym Sets

The synonym set is applied during preprocessing to group object mentions when constructing the co-occurrence graph. This step offers lightweight, plug-and-play lexical normalization (e.g., merging "bicycle" and "bike") to enhance the coverage of object relationships. To assess their impact, we replace the originally used WordNet (LVIS) (Miller, 1995; Gupta et al., 2019) with V3Det (Wang et al., 2023a), a more modern synonym set featuring broader and finer-grained categories. As shown in Tab.10, this substitution yields improvements

ρ	Object-Hal		AMBER
	C-s ↓	C-i ↓	C-i ↓
0.1	49.2	15.3	8.7
0.2	**38.1**	**9.2**	**6.0**
0.3	**38.1**	10.5	6.1
0.4	42.4	11.2	6.2
0.5	41.3	10.9	6.2
0.6	46.4	12.3	6.4

Table 11: Ablation studies on injection rates.

Method	Data Source	Size
POVID	LLaVA-Instruct	17k
HA-DPO	VG	6k
RLAIF-V	COCO, MovieNet,... (7 total)	83k
CLIP-DPO	COCO, SAM,... (12 total)	750k
CSR	LLaVA-Instruct	13k
STIC	COCO, LLaVA-Instruct	11k
SIMA	LLaVA-Instruct	17k
Ours SI-23k	LLaVA-Instruct	23k
Ours SI-130k	LLaVA-Instruct, VG	130k

Table 12: Comparison of preference data sources and sizes.

of +6.4/+0.5/+0.3 on ObjectHal-C-s/ObjectHal-C-i/MMVet, respectively. These results highlight the flexibility and extensibility of our pipeline with respect to synonym sources.

B.5 Ablation Studies on Injection Rate

We design ablative experiments on the injection rate ρ. As shown in Tab.11, APASI achieves the best performance when setting ρ to 0.2. When ρ is greater than 0.2, APASI performs poorer as ρ increases. This is consistent with the observation that larger gap within the preference pair makes the alignment task more easy, which is less effective for model improvement. Setting ρ to 0.1 results in worse performance than 0.2. An injection of 0.1 leads to a situation where number of replaced sentences is zero in about half of the preferred sentences, thus making the corresponding pair invalid.

B.6 Analysis on Data Source

To enable a fair comparison with existing methods, in Tab. 12 we provide a detailed analysis of data sources and dataset sizes across all methods compared. Our SI-23k and SI-130k datasets are constructed from the simplest sources, relying solely

Preference Data	Object-Hal		MMVet
	C-s ↓	C-i ↓	
LLaVA-v1.5-7B	51.0	13.7	30.5
SI-23k (APASI-Base)	38.1	9.2	33.5
SI-6k	46.0	12.9	31.7
SI-23k + POVID-17k	**34.9**	**8.3**	**34.1**

Table 13: Performance of APASI trained on different preference data.

on the supervised fine-tuning (SFT) dataset of the baseline model (LLaVA-v1.5). Notably, APASI-Base and APASI-IACL, trained on the moderate-scale SI-23k dataset, already achieve competitive performance.

To further evaluate performance under limited data conditions, we trained an APASI variant using only 6k samples randomly drawn from SI-23k. As shown in Tab. 13, APASI with 6k data still outperforms the baseline by +5.0/+0.8/+1.2 on ObjectHal-C-s/ObjectHal-C-s/MMVet, respectively. The performance gap between the 6k and 23k settings underscores the critical role of data scale in enhancing model effectiveness.

We also extend the data sources by incorporating GPT-4V-labeled preference pairs, combining SI-23k with POVID-17k (Zhou et al., 2024a). Tab. 13 reports the results of APASI trained on SI-23k alone versus the combined dataset. Incorporating the external dataset yields modest improvements of +3.2/+0.9/+0.6 on ObjectHal-C-s/ObjectHal-C-i/MMVet, respectively. Although APASI is designed to function without reliance on external resources, these results indicate that it can also be effectively scaled when such resources are available.

B.7 Computational Cost of APASI

We analyze the computational cost of APASI using the SI-23k dataset and LLaVA-v1.5-7B as the target LVLM.

Preference data construction includes the following stages: **1) Preferred generation.** The target LVLM generates preferred responses with input images and textual questions taking about 80 mins on 8 V100 32GB GPUs. **2) Co-occurrence graph construction.** Objects in preferred responses are parsed via WordNet toolbox and synonym sets. We traverse these objects to build the co-occurrence graph stored as a dictionary, where each key-value pair represents an object and its co-occurring objects with frequencies, requiring about 6MB storage. This takes about 3 mins with one Intel E5-2698 CPU process. **3) Hallucination injection.** The visually disabled target LVLM generates hallucinated sentences guided by co-occurring objects, replacing the sampled sentences in the preferred responses. This takes about 80 mins on 8 V100 32GB GPUs, with the resulting dataset occupying about 8MB.

As training takes about 360 mins, data construction accounts for only about 31.2% ($\frac{80+3+80}{80+3+80+360}$) in a full iteration's time. Storage and hardware demands also remain acceptable, making the data construction practically feasible.

C Qualitative Results

C.1 Analysis on Injection Settings

We compare dis-preferred responses under different hallucination injection settings, as shown in Fig.8. For hallucination injection in APASI, the target model is guided to generate sentences about non-existent table to obtain the dis-preferred responses. In this way, the injected sentence is ensured is guaranteed to contain hallucinations, and the injected hallucination remains linguistically reasonable. If removing the guidance of the co-occurring object, the model generates an actually correct sentence describing the kite for injection. The dis-preferred response without hallucinations makes the preference pair invalid. If injecting hallucinations only by replacing with the word car and chair, the dis-preferred response includes absurd mistakes in the like "man holding cars". These absurd mistakes are inconsistent with typical hallucinations with linguistic reasonability, thereby the preference pair less effective.

C.2 Examples of Model Response

Fig.9 and Fig.10 show more comparative examples of APASI and the LLaVA-v1.5-7B baseline. Both models are given with descriptive prompts and images from COCO-test2017 (Lin et al., 2014) or MMHal (Sun et al., 2023). Results show that APASI effectively mitigates the hallucination problem for LLaVA-v1.5-7B while capturing all major objects in the image.

Image

Preferred Response	**Dis-Preferred Guided Gen.**
The image features a man standing in a courtyard, holding a kite with long streamers. He appears to be enjoying his time outdoors, possibly preparing to fly the kite. There are several other people in the scene, some of whom are walking around or standing near the man with the kite	*The image features a man standing in a courtyard, holding a kite with long streamers. **There is a table nearby, possibly for holding the kite or other items.** There are several other people in the scene, some of whom are walking around or standing near the man with the kite .....*
Dis-Preferred No Guidance	**Dis-Preferred Object Rep.**
*The image features a man standing in a courtyard, holding a kite with long streamers. **The kite is flying high in the sky, capturing the attention of those around him.** There are several other people in the scene, some of whom are walking around or standing near the man with the kite .....*	*The image features a man standing in a courtyard, holding a **car** with long streamers. He appears to be enjoying his time outdoors, possibly preparing to fly the kite. There are several other **chair** in the scene, some of whom are walking around or standing near the man with the kite .....*

Figure 8: Comparison of the dis-preferred responses under different hallucination injection settings. The preferred response is in the light green box. The dis-preferred response obtained by guided model generation, which is actual used in APASI, is in the light red box. The two ineffective dis-preferred responses are in the gray boxes. The injected hallucination parts are underlined. Best viewed in color.

LLaVA-v1.5-7B: The image depicts a rainy day in a city, with a wet street reflecting the surrounding environment. The street is lined with trees, creating a serene atmosphere despite the rain. There are several cars parked along the street, and **a bus is visible in the background**. A few people can be seen walking on the sidewalk, with one person closer to the left side of the image and another person further to the right. **A traffic light is also present in the scene, indicating the presence of an intersection.** The overall scene captures the essence of a typical rainy day in a city, with people going about their daily routines despite the wet weather.

APASI: The image depicts a rainy day in a city, with a wet city street reflecting the surrounding environment. The street is lined with trees, creating a scenic atmosphere. There are several cars parked along the street, and some are driving through the rain. Numerous people can be seen walking on the sidewalk, with some of them carrying umbrellas to shield themselves from the rain. The pedestrians are scattered throughout the scene, with some closer to the foreground and others further back. The overall atmosphere of the image is that of a typical rainy day in a city, with people going about their daily routines despite the wet weather.

LLaVA-v1.5-7B: The image features a cozy bedroom with two bunk beds, each with a yellow pillow. The beds are positioned close to each other, creating a sense of warmth and comfort. The beds are adorned with colorful pillows, adding a vibrant touch to the room. **In addition to the beds, there are two chairs in the room, one located near the left side of the beds and the other near the right side. A potted plant can be seen in the room, adding a touch of greenery to the space.** The overall atmosphere of the room is inviting and welcoming.

APASI: The image features a cozy bedroom with two bunk beds, each with a colorful comforter. The beds are positioned next to each other, creating a comfortable sleeping arrangement. There are several pillows on the beds, with some placed on the top bunk and others on the bottom bunk. The pillows come in various sizes and colors, adding a vibrant touch to the room. The overall atmosphere of the bedroom is inviting and warm, making it an ideal space for rest and relaxation.

LLaVA-v1.5-7B: The image features a zoo setting with a giraffe standing next to a woman who is feeding it. The giraffe is positioned close to the woman, and they are both standing near a fence. Another person is visible in the scene, standing further back and observing the interaction between the woman and the giraffe. In the background, **there are a few more people, likely visitors to the zoo. A bottle can be seen on the ground, possibly belonging to one of the visitors.** The scene captures a moment of interaction between the woman and the giraffe, showcasing the unique experience of being up close with these majestic animals.

APASI: The image features two women standing next to a giraffe at a zoo. The giraffe is inside an enclosure, and the women are on the other side of a fence, observing the animal. The women appear to be enjoying their time at the zoo and interacting with the giraffe. The scene also includes a few trees in the background, adding to the zoo setting. The women and the giraffe are the main focus of the image, creating a memorable moment for them.

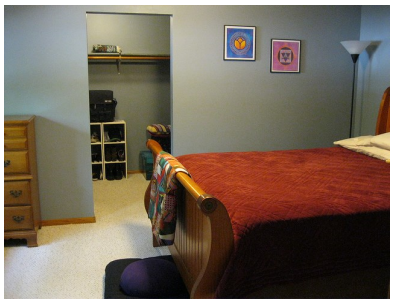
Figure 9: Comparison of the responses of the baseline and APASI. Correct and wrong objects are underlined in red and green respectively. Other hallucinated contents are in red. Best viewed in color.

LLaVA-v1.5-7B: The image depicts a spacious living room with a large couch and a chair placed in the center of the room. The couch is positioned against the wall, while the chair is located closer to the center. A coffee table is situated in the middle of the room, surrounded by the seating area. The living room is connected to a kitchen, which features a sink and a refrigerator. A dining table is also present in the room, with chairs placed around it. **A person can be seen in the background, possibly enjoying the space or preparing to use the kitchen. Additionally, there are two cups placed on surfaces in the room, one near the couch and the other closer to the dining table.**

APASI: The image showcases a spacious and clean living room with a modern design. The living room is furnished with two black leather couches, a glass coffee table, and several chairs. The room also features a kitchen area with a counter and a bar stool. The living room is well-lit and has a large window, allowing natural light to fill the space. The furniture is arranged in a way that creates a comfortable and inviting atmosphere. The overall appearance of the living room is modern and stylish.

LLaVA-v1.5-7B: The image depicts a group of three people sitting around a dining table in a living room, enjoying each other's company. They are engaged in conversation and drinking wine, with several wine glasses placed on the table. There are also cups and a bottle visible on the table. The living room is furnished with a couch and a chair, providing comfortable seating for the group. A laptop is placed on the table, and a potted plant can be seen in the background, adding a touch of greenery to the room. A vase is also present, adding to the room's decor. **In the background, a clock is mounted on the wall**, and a book can be spotted on a surface, indicating that the room is used for various activities.

APASI: The image features a group of three people, two men and a woman, sitting together in a living room. They are gathered around a coffee table, which is filled with various items such as books, a laptop, and a bunch of presents. The group appears to be enjoying each other's company and having a good time. The living room is furnished with chairs and a couch, and there are also several cups, wine glasses, and a potted plant present in the room. The atmosphere seems relaxed and social as the friends engage in conversation and share the presents.



LLaVA-v1.5-7B: The image depicts a cozy bedroom with a large bed situated in the center of the room. The bed is covered with a red blanket, and there are two pillows on it. The room also features a dresser and **a chair**, both placed near the bed. In addition to the bedroom furniture, **there are several books scattered around the room, indicating that the occupant enjoys reading. A TV is mounted on the wall, providing entertainment options for the room's occupant. A clock is also visible on the wall, allowing the person to keep track of time.**

APASI: The image features a cozy bedroom with a large bed situated in the center of the room. The bed has a red comforter and is positioned against a wall. The bedroom also contains a dresser and a closet. There are several pictures hanging on the wall above the bed, and the room has a clean and organized appearance. The bedroom appears to be a part of a small apartment, with the bed being the main focal point of the space.

Figure 10: Comparison of the responses of the baseline and APASI. Correct and wrong objects are underlined in red and green respectively. Other hallucinated contents are in red. Best viewed in color.