Mitigating Object Hallucinations via Sentence-Level Early Intervention

Shangpin Peng[1*] **Senqiao Yang**[2*] **Li Jiang**[3] **Zhuotao Tian**[1✉]
[1]Harbin Institute of Technology, Shenzhen
[2]The Chinese University of Hong Kong [3]The Chinese University of Hong Kong, Shenzhen

Abstract

*Multimodal large language models (MLLMs) have revolutionized cross-modal understanding but continue to struggle with hallucinations - fabricated content contradicting visual inputs. Existing hallucination mitigation methods either incur prohibitive computational costs or introduce distribution mismatches between training data and model outputs. We identify a critical insight: hallucinations predominantly emerge at the early stages of text generation and propagate through subsequent outputs. To address this, we propose **SENTINEL** (**S**entence-level **E**arly i**N**tervention **T**hrough **IN**-domain pr**E**ference **L**earning), a framework that eliminates dependency on human annotations. Specifically, we first bootstrap high-quality in-domain preference pairs by iteratively sampling model outputs, validating object existence through cross-checking with two open-vocabulary detectors, and classifying sentences into hallucinated/non-hallucinated categories. Subsequently, we use context-coherent positive samples and hallucinated negative samples to build context-aware preference data iteratively. Finally, we train models using a context-aware preference loss (C-DPO) that emphasizes discriminative learning at the sentence level where hallucinations initially manifest. Experimental results show that SENTINEL can reduce hallucinations by over 90% compared to the original model and outperforms the previous state-of-the-art method on both hallucination benchmarks and general capabilities benchmarks, demonstrating its superiority and generalization ability. The models, datasets, and code are available at* https://github.com/pspdada/SENTINEL.

1. Introduction

Recent advancements in multimodal large language models (MLLMs) have demonstrated significant progress in aligning visual and textual representations through cross-modal feature integration, marking a pivotal step toward the development of general-purpose AI systems [2, 9, 33–35, 44, 65, 87]. However, a critical challenge persists in multimodal settings: the phenomenon of hallucinations [4, 36, 53], wherein models generate factually inconsistent or fabricated information that deviates from the image content provided by users. This issue not only degrades user trust and experience but also poses substantial risks in real-world applications of MLLMs, thereby impeding the realization of trustworthy general AI systems [5, 18, 68].

To address this challenge, recent work has explored enhanced decoding strategies [8, 19, 27] as a means to mitigate hallucinations. While these approaches show promise, they often introduce trade-offs, including increased computational overhead during inference, higher latency, and re-

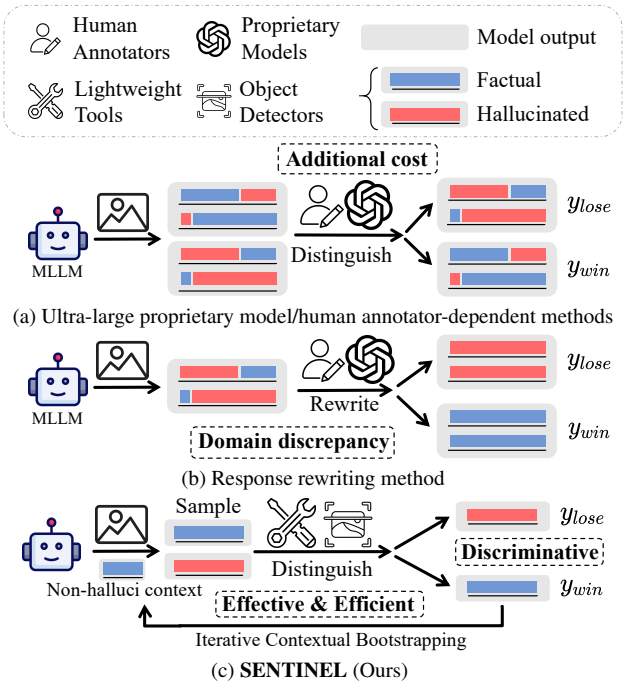


Figure 1. **Comparative analysis of data construction strategies for hallucination mitigation in MLLMs.** Our proposed approach demonstrates superior efficiency and effectiveness in generating high-quality, domain-specific preference learning datasets, offering a robust solution for reducing hallucination in MLLMs.

* Equal contribution.
✉ Corresponding author (tianzhuotao@hit.edu.cn).

liance on specific dependencies, which may limit their scalability and practicality in resource-constrained scenarios.

On the other hand, preference alignment methods [31, 51, 60] avoid additional inference costs but face other challenges. As shown in Fig. 1a, many of them rely on large proprietary models (*e.g.*, GPT [1]) [21, 69, 75, 80, 82, 86] or human annotators [13, 76], incurring high costs. Additionally, Fig. 1b highlights that output rewriting [69, 82, 86] can create distributional discrepancies, while Lai et al. [25] and our experiments in Tab. 2 show that out-of-domain training data harms generalization. *Therefore, the high costs and the distribution disparities inherent in the curated training data may compromise hallucination mitigation efforts.*

Key observations. To address hallucination with greater efficacy and efficiency, we investigate the dynamics of hallucination within the model's output. Our analysis reveals that hallucination intensity escalates with the length of generated text, while mitigating hallucinations at specific sentences significantly reduces their prevalence in subsequent outputs, as detailed in Figs. 2a and 2b. These findings suggest that early intervention—targeting hallucinations at their initial occurrence—is crucial to preventing their propagation in later generations. This raises a key question: *How can we effectively implement an early intervention strategy to address hallucinations of MLLMs as they arise?*

Our solution. In this work, we propose **SENTINEL** (**S**entence-level **E**arly i**N**tervention **T**hrough **IN**-domain pr**E**ference **L**earning), which provides early intervention for the initial occurrence of hallucinations during generation. Unlike existing methods, SENTINEL operates without relying on external large language models for rewriting, ensuring that the learning targets remain strictly within the domain of the model's original outputs. This approach preserves the model's intrinsic distribution and expression patterns while effectively curbing hallucination propagation.

Specifically, SENTINEL first employs an in-domain candidate bootstrapping strategy, which performs multiple sampling rounds on the current model, extracts objects from the outputs, and applies consistency cross-checking to classify objects as *hallucinated*, *uncertain*, or *factual*. This is followed by a context-aware preference data generation process, which constructs preference pairs using non-hallucinated positive samples and hallucinated negative ones, enhanced by iterative contextual bootstrapping. Finally, context-aware preference learning is performed using the modified context-aware DPO loss, maximizing the likelihood of generating context-coherent positive samples while minimizing hallucinated negative ones. By focusing on captions where hallucinations first emerge, SENTINEL effectively halts their propagation in subsequent outputs.

Experimental results across various benchmarks demonstrate that SENTINEL effectively mitigates object hallucination while preserving the generalization capabilities of MLLMs. Specifically, on Object Halbench [55] and AMBER [63], hallucinations are reduced by about 92% and 65%, respectively, with consistent improvements on HallusionBench [12]. Furthermore, SENTINEL preserves its performance on VQAv2 [10] and TextVQA [59], and achieving decent gains on both ScienceQA [41] and MM-Vet [78].

To summarize, our contributions are as follows:

- We demonstrate that early intervention at the first occurrence of hallucination is crucial for preventing its propagation in subsequent model outputs of MLLMs.
- We propose SENTINEL, which effectively and efficiently mitigates hallucinations without requiring extensive external resources or manual effort.
- The model-agnostic SENTINEL achieves state-of-the-art performance on hallucination benchmarks without compromising MLLMs' general capabilities.

2. Background and Motivation

In this section, we briefly introduce the foundational concepts and methods relevant to this study in Sec. 2.1, establishing the necessary background. Following this, in Sec. 2.2, we outline our key insights and elucidate the motivations behind our proposed designs.

2.1. Related Work and Preliminaries

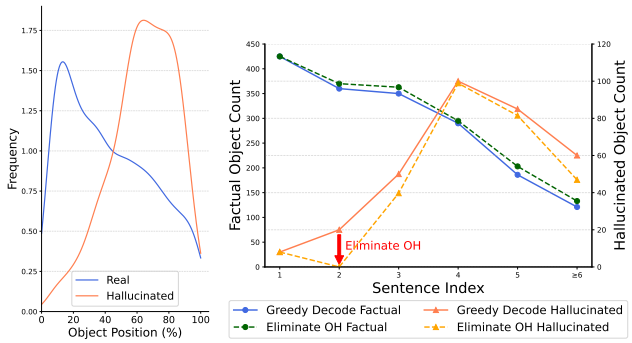
Object Hallucination (OH) in Multimodal Large Language Models (MLLMs) is characterized by the generation of text that is semantically coherent yet inconsistent with the visual content of the provided image [4, 53]. To mitigate this issue, recent advancements have focused on innovative decoding strategies, which aim to reduce the prevalence of OH by refining the generation process of MLLMs [6, 8, 19, 27].

Concurrently, preference learning has emerged as an alternative approach for addressing OH, leveraging its capacity to align MLLMs with human expectations for truthfulness and traceability [13, 28, 32]. Notably, the Proximal Policy Optimization algorithm (PPO) [57] enhances model reliability by training an auxiliary reward model to assess response quality and then guide the model in optimizing its outputs based on the reward signals. Moreover, Direct Preference Optimization (DPO) [51] has emerged as a simpler alternative, learning directly from pre-collected feedback without requiring a reward model. The DPO loss is:

$$\mathcal{L}_{\text{DPO}}(\boldsymbol{\theta}) = -\,\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}_w,\boldsymbol{y}_l)\sim D}\left[\log\sigma\left(\beta\log\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}_w|\boldsymbol{x})}{\pi_{\text{ref}}(\boldsymbol{y}_w|\boldsymbol{x})} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}_l|\boldsymbol{x})}{\pi_{\text{ref}}(\boldsymbol{y}_l|\boldsymbol{x})}\right)\right], \tag{1}$$

where $\boldsymbol{x} = [\boldsymbol{v}, \boldsymbol{q}]$.

Here, D represents the preference dataset for learning, σ denotes the sigmoid function, $\pi_{\boldsymbol{\theta}}$ indicates the policy



(a) (b)

Figure 2. **Object position distribution in MLLM hallucination analysis.** (a) illustrates the progressive deterioration of hallucination effects in Multimodal Large Language Models (MLLMs) with increasing description length in the image captioning task, while (b) demonstrates the effectiveness of early-stage intervention in mitigating the propagation of hallucination.

model under training, π_{ref} represents the unchanged reference model, $\boldsymbol{y}_w$ stands for the positive sample, and $\boldsymbol{y}_l$ represents the negative sample, both based on the input $\boldsymbol{x}$, which includes the image $\boldsymbol{v}$ and prompt $\boldsymbol{q}$. The hyperparameter β governs the separation between the policy model and the reference model. Many recent methods [16, 69, 77, 82, 86] leverage DPO to mitigate hallucinations by curating preference data to guide the models. More related works of this study are discussed in Sec. F.

2.2. Motivation

This section outlines the motivations behind this work. The implementation details of related experiments are provided in Sec. A.

Hallucination grows with text length. To better understand the causes of Object Hallucination (OH), we analyze the distributions of hallucinated and factual objects in image captions generated by MLLMs. Specifically, as shown in Fig. 2a, where the horizontal axis represents the normalized position of an object in the caption (as a percentage), while the vertical axis denotes the normalized frequency (probability density), the blue curve represents hallucinated objects, and the orange curve corresponds to objects present in the image. The comparison reveals that as caption length increases, the model becomes more prone to hallucinations, with fewer factual objects described and more hallucinated ones introduced. This trend is further corroborated by sentence-level analysis in Fig. 2b. These findings lead us to hypothesize that *intervening at the initial occurrence of hallucination could be critical in reducing its recurrence in subsequent model outputs*.

Early intervention mitigates hallucinations. To evaluate the effectiveness of early intervention in curbing hallucination propagation, we analyze the impact of addressing hallucinations at the sentence level in image captioning tasks. Specifically, as illustrated in Fig. 2b, eliminating hallucinated objects in the second sentence—compared to vanilla greedy decoding—significantly reduces the likelihood of hallucinated objects in subsequent sentences while increasing the probability of factual objects present in the image. Similar results are observed when addressing hallucinations in the third sentence, as shown in Sec. A.2. These findings underscore the necessity of early intervention to mitigate hallucinations effectively.

To enable early intervention, an open-vocabulary object detector [7, 37] could be employed during inference to verify the presence of the objects generated by the model within the image. While this method effectively reduces hallucinations without sacrificing caption diversity, as demonstrated in Sec. A.2, it is time-consuming; despite the object detector being efficient, the model's sampling process incurs significant computational overhead.

Consequently, we opt for a preference learning strategy during model training, which mitigates hallucinations without compromising the original inference efficiency.

3. Method

3.1. Overview

Existing preference learning methods may use an external model to rewrite sentences or rely on model-generated responses as training data. However, these methods may introduce discrepancies in distribution and expression patterns between the training data and the model's original output. Hence, we propose SENTINEL, which performs sentence-level early intervention to mitigate object hallucinations through preference learning with in-domain data, without manual effort or dependence on extensive LLMs.

As shown in Fig. 3, the proposed SENTINEL method takes six essential steps. Specifically, Sec. 3.2 presents the process of generating the in-domain candidates containing the factual and hallucinated objects. Subsequently, Sec. 3.3 introduces the construction of preference data pairs derived from these in-domain candidates. These two steps can be integrated into the In-domain Preference Data Construction phase (shown in Algorithm 1). Finally, in Sec. 3.4 we elaborate on how SENTINEL leverages the curated preference data to achieve preference learning.

3.2. In-domain Candidate Bootstrapping

To construct positive and negative preference data pairs without relying on external models for rewriting, we perform multiple sampling rounds on the current model and extract objects from the outputs. We then apply a consistency cross-checking method to classify the model's output objects into three categories: *hallucinated*, *uncertain*, and *factual*, which are used to construct preference data in subsequent steps. This process is termed In-domain Candidate

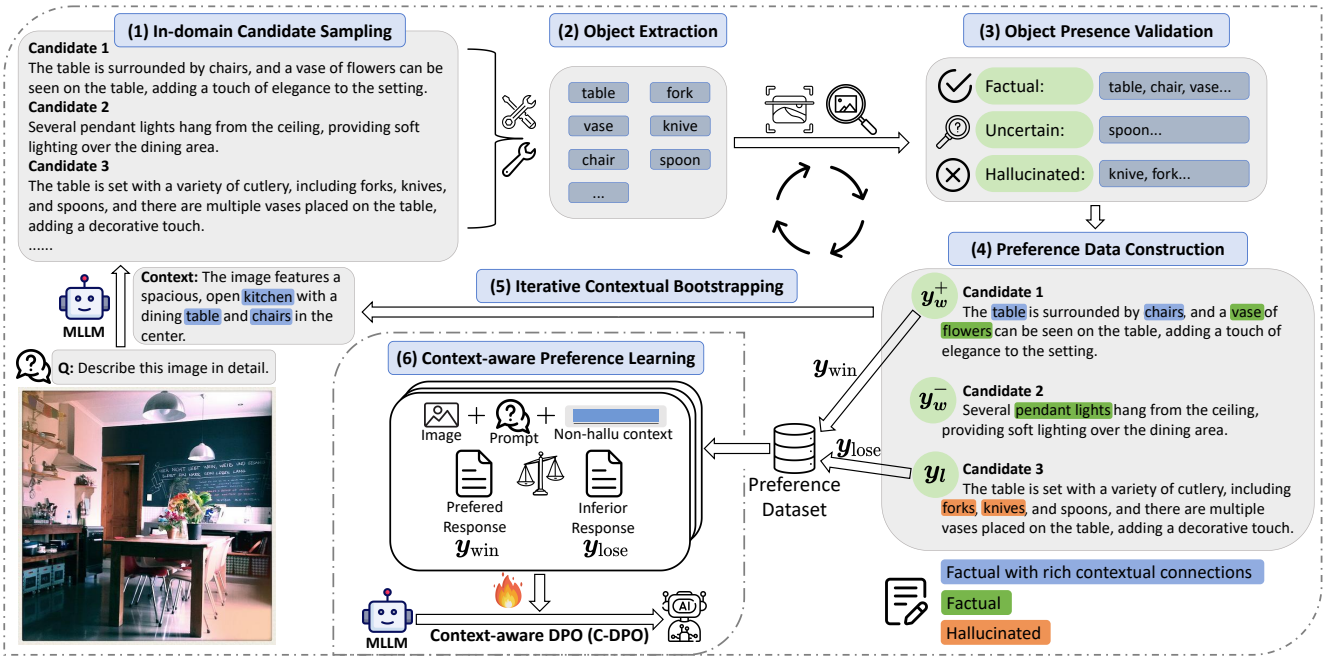


Figure 3. **The overview of SENTINEL.** The proposed SENTINEL takes six essential steps: **(1)** Generate multiple in-domain responses conditioned on the input image, prompt, and context $\boldsymbol{c}$. **(2)** Identify and extract all mentioned objects from each generated sentence. **(3)** Utilizing two object detectors to validate the existence of extracted objects through cross-referencing. **(4)** Categorize generated sentences into hallucinated and non-hallucinated groups based on detection results. **(5)** Extend the generation context with verified non-hallucinated sentences to guide subsequent outputs. **(6)** Fine-tune the model using the context-aware DPO (C-DPO) loss with the in-domain, style-consistent, and context-varying preference data.

Booststrapping, as illustrated in Fig. 3 (1)-(3).

In-domain candidate sampling. In our approach, we use sampling-based decoding to obtain n candidate samples. This ensures that the positive ($\boldsymbol{y}_w$) and negative ($\boldsymbol{y}_l$) samples are drawn from the same distribution as the current model, preserving consistency in textual styles and linguistic structures. The generation halts upon sentence completion (e.g., detection of a period), at which point sentences are automatically segmented for subsequent discrimination.

Object extraction. After generating candidate sentences, we extract the mentioned objects from the text for hallucination detection. To achieve this, we utilize the Scene-GraphParser [30] model to transform the textual descriptions into a series of triplet-based scene graphs. By parsing these scene graphs, we identify specific noun entities from the subjects and objects, which are subsequently used as candidate objects for existence verification.

Object presence validation. Following object extraction, we apply cross-checking to validate the presence of candidate objects in the image. Specifically, we utilize two open-vocabulary object detectors, GroundingDINO [37] and Yolo World [7], for cross-validation. This approach demonstrates superior performance compared to using a single detector, as shown in Fig. 8 of the ablation study.

The cross-checking results are categorized into three types: (1) *hallucinated* (both models confirm absence), (2) *factual* (both models confirm presence), and (3) *uncertain* (conflicting results). Sentences containing hallucinated objects are tagged as "*hallucinated*", whereas those only containing factual objects are tagged as "*non-hallucinated*", forming positive-negative sample pairs for preference learning. To ensure data quality and minimize detector bias, we ignore uncertain objects.

Algorithm 1 In-domain Preference Data Construction

Input: Image $\boldsymbol{v}$, prompt $\boldsymbol{q}$, context $\boldsymbol{c}$ (initially empty)
Output: Training samples $(\boldsymbol{v}, \boldsymbol{q}, \boldsymbol{c}, \boldsymbol{y}_w^+, \boldsymbol{y}_l)$

while Model M does not generate </s> **do**
  Sample n in-domain candidates s_i using $\boldsymbol{v}$, $\boldsymbol{q}$, and $\boldsymbol{c}$
  for each sample s_i **do**
    Extract entities from the sample
    Validate the presence of entities using object detectors
  Select $\boldsymbol{y}_w^+$ as context-coherent non-hallucinated sample
  Select $\boldsymbol{y}_l$ as hallucinated sample
  Construct preference samples $(\boldsymbol{v}, \boldsymbol{q}, \boldsymbol{c}, \boldsymbol{y}_w^+, \boldsymbol{y}_l)$
  Append a non-hallucinated sample $\boldsymbol{y}_w^+$ to the context $\boldsymbol{c}$

3.3. Context-aware Preference Data Generation

With sentences labeled as "*hallucinated*" or "*non-hallucinated*" from Sec. 3.2, this section introduces context-aware preference data generation. As illustrated in Fig. 3

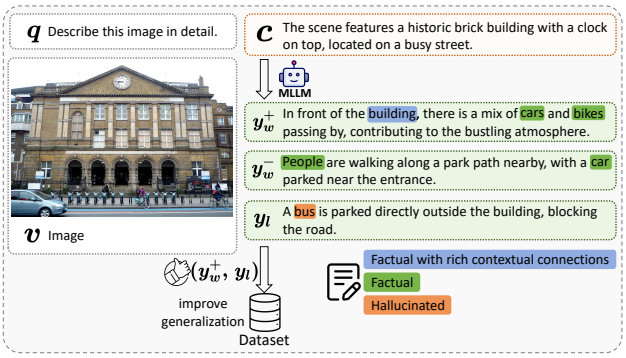


Figure 4. **Categories of in-domain candidates.** The in-domain candidates fall into three types. Employing non-hallucinated, context-coherent descriptions (y_w^+) as positive samples, paired with hallucinated descriptions (y_l), enhances the model's generalization performance and robustness.

(4)-(5), this process extracts contextually relevant data, ensuring the training data better represents the model's output distribution. The specifics are elaborated below.

Preference data construction. The preference data is typically composed of the image, the corresponding prompt, the positive sample, the negative sample, and the context (i.e., all generated sentences excluding the current one). In the construction of sample pairs, positive samples y_w are selected from the *non-hallucinated* sentences, while negative samples y_l are derived from the *hallucinated* sentences.

Subsequently, we partition the positive samples y_w into two categories: (1) the context-coherent positive sample y_w^+, wherein some of the described objects are explicitly referenced in the context, and (2) the context-agnostic positive sample y_w^-, where none of the objects are mentioned in the context. In essence, the objects described in y_w^+ exhibit a strong correlation with the context, while those in y_w^- display a weaker or negligible correlation. Illustrative examples are provided in Figs. 3 and 4.

We observe that the context-coherent sample y_w^+ can effectively mitigate hallucinations without compromising the model's generalization capabilities, and incorporating y_w^- as the positive samples results in performance reduction, as shown in Tab. 3. This observation underscores the importance of contextual signals in guiding the model's generation process. Specifically, the richer contextual information in y_w^+ samples appears to enhance the model's ability to preserve contextual coherence and prioritize salient content, resulting in performance improvements [15].

Iterative Contextual Bootstrapping (ICB). The proposed SENTINEL framework is designed to enable early intervention for mitigating hallucinations in generative models. Given the context c, which represents the hallucination-free content preceding the current output, the model is trained to distinguish between a non-hallucinated positive sample y_w^+ and a hallucinated negative sample y_l. To enhance robustness across diverse contexts, we introduce the Iterative Contextual Bootstrapping (ICB) strategy, as depicted in Fig. 5.

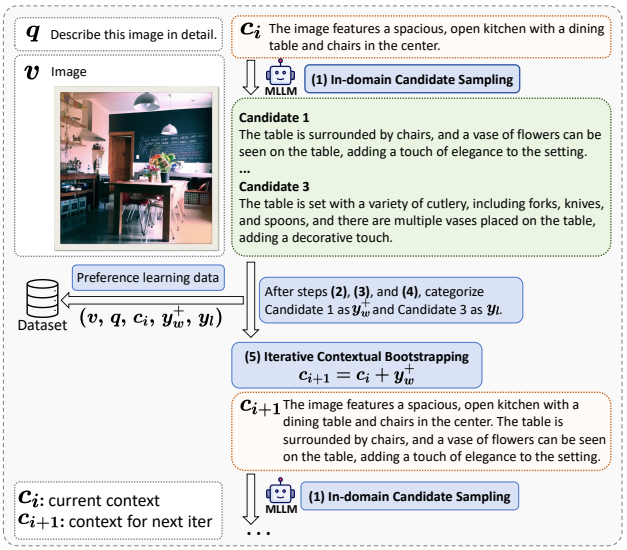


Figure 5. **Visualization of the Iterative Contextual Bootstrapping (ICB) framework.** Given an input image and corresponding question, this pipeline iteratively generates diverse contextual samples, enabling robust hallucination mitigation across varying contexts and significantly improving model generalization.

Specifically, given the query q, the input image v, and the current context c_i, we generate multiple candidate outputs by repeatedly sampling from the MLLM. These candidates are then processed through a structured pipeline consisting of (2) object extraction, (3) object presence validation, and (4) preference data construction, as illustrated in Fig. 3. This pipeline is designed to identify a non-hallucinated positive sample y_w^+ and a hallucinated negative sample y_l. By aggregating v, q, c_i, y_w^+ and y_l, we construct a preference data pair (v, q, c_i, y_w^+, y_l), which is subsequently appended to the dataset for preference learning.

Furthermore, to bootstrap the preference data with different hallucination-free contexts, we construct $c_{i+1} = c_i + y_w^+$ for the next iteration by appending the positive sample y_w^+ to the current context c_i. The updated context c_{i+1} is then processed through the same procedure as described above to generate a new preference data pair. This iterative approach ensures that the preference data is enriched with progressively more complex and varied contexts, enabling the model to generalize its hallucination mitigation capabilities across different scenarios. The effectiveness of this pipeline is validated and discussed in Sec. B.2.

3.4. Context-aware Preference Learning

The preference data generated through the processes outlined in Sec. 3.2 and Sec. 3.3 can be formally represented as (x, c, y_w^+, y_l), where x is the input, including the image v and the prompt q, c denotes the context, y_w^+ is the context-

Model	Method	Hallucination benchmarks						General benchmarks			
		Object HalBench [55]		AMBER [63]			HallusionBench [12]	VQAv2 [10]	TextVQA [59]	ScienceQA [41]	MM-Vet [78]
		Resp.↓	Ment.↓	CHAIR↓	Hal.↓	Cog.↓	Question Acc.↑	Acc.↑	Acc.↑	Image Acc.↑	Overall↑
LLaVA-v1.5-7B	baseline	52.7	28.0	8.4	35.5	4.0	46.86	**78.5**	**58.2**	66.8	31.0
	VCD [27]	51.3	25.9	9.1	39.8	4.2	-	77.0	56.1	68.7	29.8
	OPERA [19]	45.3	22.9	6.5	28.5	3.1	-	78.2	**58.2**	68.2	30.3
	DoLa [8]	44.0	25.1	6.2	27.7	2.9	-	76.3	56.6	67.5	30.8
	EFUF [70]	39.3	22.6	5.8	28.2	3.1	47.03	78.1	57.2	66.4	31.2
	HA-DPO [82]	37.0	20.9	6.7	30.9	3.3	**47.74**	77.6	56.7	**69.7**	30.6
	POVID [86]	33.4	16.6	5.3	28.7	3.0	46.59	77.2	56.6	68.8	31.8
	CLIP-DPO [45]	-	-	3.7	16.6	1.3	-	-	56.4	67.6	-
	RLAIF-V [77]	7.8	4.2	**2.8**	15.7	**0.9**	35.43	75.2	55.1	68.2	29.9
	TPO [16]	5.6	3.2	3.6	20.5	1.6	40.12	75.9	55.3	67.1	25.7
	Ours	**4.3**	**2.6**	2.9	**14.6**	1.2	47.56	78.4	**58.2**	69.2	**32.6**
LLaVA-v1.5-13B	baseline	46.0	23.0	6.9	31.9	3.3	46.43	**80.0**	**61.2**	71.6	36.0
	VCD [27]	43.7	21.6	7.8	36.2	3.7	-	78.5	59.5	72.0	33.7
	vanilla-DPO [69]	6.7	3.6	2.8	15.5	1.6	46.41	79.2	60.4	71.8	35.0
	HSA-DPO [69]	5.3	3.2	**2.1**	13.4	1.2	46.14	78.3	60.0	71.3	33.7
	Ours	**3.3**	**1.9**	2.7	**11.7**	**0.9**	**46.77**	79.9	61.0	**72.8**	**36.2**

Table 1. **Comparison of hallucination mitigation methods in MLLMs: effectiveness and general capabilities.** This evaluation highlights the best and second-best results in **bold** and underlined, respectively. All comparisons are performed under identical model size constraints. "Resp." and "Ment." denote response-level and mention-level hallucination rates, while "Hal." and "Cog." represent the Hallucination Score and Cognitive Score, respectively. More evaluation details are provided in Sec. D.

coherent positive sample, and $\boldsymbol{y}_l$ is the negative sample.

The learning objective is to guide the model, conditioned on the input $\boldsymbol{x}$ and the context $\boldsymbol{c}$, to maximize the likelihood of generating the contextually coherent positive sample $\boldsymbol{y}_w^+$ while minimizing the likelihood of producing the negative sample $\boldsymbol{y}_l$. To achieve this, we adapt the Direct Preference Optimization (DPO) loss by incorporating the context $\boldsymbol{c}$ as part of the input. We term this modified loss as context-aware DPO (C-DPO), which is formulated as follows:

$$\mathcal{L}_{\text{C-DPO}}(\boldsymbol{\theta}) = -\mathbb{E}_{(\boldsymbol{x}',\boldsymbol{y}_w^+,\boldsymbol{y}_l)\sim D}\left[\log\sigma\Big(\beta\log\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}_w^+|\boldsymbol{x}')}{\pi_{\text{ref}}(\boldsymbol{y}_w^+|\boldsymbol{x}')} - \beta\log\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}_l|\boldsymbol{x}')}{\pi_{\text{ref}}(\boldsymbol{y}_l|\boldsymbol{x}')}\Big)\right], \quad \text{where } \boldsymbol{x}' = [\boldsymbol{x},\boldsymbol{c}] = [\boldsymbol{v},\boldsymbol{q},\boldsymbol{c}]. \tag{2}$$

In C-DPO, the context $\boldsymbol{c}$ is excluded from the loss computation, and gradients are only derived from the discrimination between $\boldsymbol{y}_w^+$ and $\boldsymbol{y}_l$. This design ensures that the model focuses on learning the contextual coherence of the positive sample without being directly influenced by the context during gradient updates. Further discussions and comparisons between the proposed C-DPO and the standard DPO are provided in Sec. C.3.

4. Experiments

In this section, we conduct comprehensive experiments to evaluate the effectiveness of our SENTINEL in reducing hallucinations while improving the general abilities of the model. We first introduce the experimental setup in Sec. 4.1, then present the main results in Sec. 4.2, and finally conduct ablation studies in Sec. 4.3 to analyze our method's effectiveness. More results are in Secs. C and D.

4.1. Experimental Setup

Training. To ensure a fair comparison, we follow the settings of prior works [16, 19, 26, 27, 34, 45, 56, 66, 69, 69, 70, 77, 79, 82, 86], using LLaVA-v1.5 as the reference model across all experiments. For data collection, we prompt the model with detailed image descriptions [76] to generate training data, with images sourced from the Visual Genome dataset [23]. Model training is conducted using C-DPO (Eq. (2)) in combination with LoRA [17], and optimized with AdamW [40]. The 7B and 13B models are trained for one epoch on 8.6K and 7.0K samples, respectively, with learning rates of 2×10^{-7} and 3×10^{-7}. Additional training details are provided in Sec. C.

Evaluation benchmarks. We evaluate the hallucination extent and general capabilities of our SENTINEL method across multiple benchmarks. For hallucination evaluation, we use widely adopted benchmarks, including Object HalBench [55], AMBER [63], and HallusionBench [12]. To assess general capabilities, we employ VQAv2 [10], TextVQA [59], ScienceQA [41], and MM-Vet [78]. Further details of these benchmarks are provided in Sec. D.1.

Baselines. To show the effectiveness of our method, we compare SENTINEL with several state-of-the-art (SOTA) methods. Specifically, VCD [27], OPERA [19], and DoLa [8] focus on enhanced decoding strategies, while HA-DPO [82], POVID [86], CLIP-DPO [45], RLAIF-V [77], and TPO [16] leverage preference training. Additionally, Vanilla DPO applies the original DPO objective Eq. (1) using training data from HSA-DPO, while EFUF [70] is an unlearning-based approach. Details are in Sec. D.2.

4.2. Main Results

Comparison with recent SOTAs. As shown in Tab. 1, we compare our method with baseline methods across several

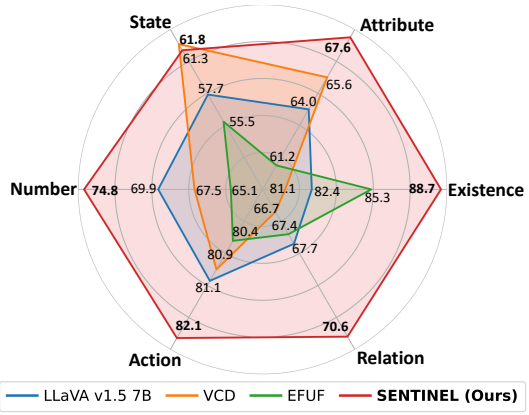


Figure 6. **Impact on different hallucination types.** Comparison between multiple methods shows that our method reduces hallucination in all six hallucination types.

benchmarks. The results demonstrate that SENTINEL significantly reduces the models' hallucination rate. Specifically, for the 7B model, our method achieves a **4.3** response-level (Resp.) and a **2.6** mention-level (Ment.) hallucination rate. Compared to the previous SOTA method, TPO [16], which achieves a 5.6 response-level and 3.2 mention-level hallucination rate, our proposed SENTINEL surpasses it by further reducing hallucinations by a total of 24% on Object Halbench. Furthermore, even on the 13B model, compared to the baseline, which achieves 6.9 CHAIR, 31.9 Hallucination score (Hal), and 4.0 Cognitive score (Cog), our proposed SENTINEL significantly improves performance, achieving 2.7 CHAIR, 11.7 Hal score, and 0.9 Cog score, respectively. These results demonstrate that our method is also effective on larger models.

Comprehensive hallucination mitigation. To further evaluate the effect of our method on various hallucination types, we conducted experiments on the discriminative part of the AMBER [63] benchmark and report the F1 scores for each hallucination type. As shown in Fig. 6, LLaVA-v1.5 with SENTINEL outperforms the baseline across all six hallucination types, demonstrating the effectiveness of our method in mitigating various hallucination issues. Notably, for the *Existence* hallucination type, our method improves the 7B model by **6.3** and the 13B model by **7.6** compared to the baseline. Detailed results are provided in Sec. D.4.

Improved general capabilities. As shown in Tab. 1, SENTINEL enhances the general capabilities of the model on multiple benchmarks. Specifically, SENTINEL demonstrates stable performance on VQAv2 and TextVQA, whereas previous methods designed for hallucination mitigation suffer from significant performance degradation. Moreover, on ScienceQA and MM-Vet, our method not only surpasses previous hallucination mitigation methods but also outperforms the vanilla LLaVA-v1.5. These results show that SENTINEL not only significantly mitigates hallucinations but also improves general capability.

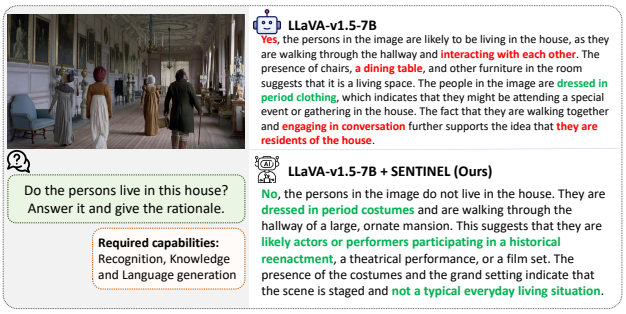


Figure 7. **Qualitative results of SENTINEL.** Our method can effectively eliminate hallucinations in MLLMs while enhancing the model's general capabilities.

Method	Object HalBench		AMBER		MM-Vet
	Resp. ↓	Ment. ↓	Acc ↑	F1 ↑	Overall ↑
LLaVA-v1.5-7B	52.7	27.9	71.5	74.1	31.1
Ours (8.6K $(\boldsymbol{y}_w^+, \boldsymbol{y}_l)$)	**4.3**	**2.6**	**76.1**	**79.3**	**32.6**
Ours (8.6K Rewrited $(\boldsymbol{y}_w^+, \boldsymbol{y}_l)$)	4.8$_{\uparrow 0.5}$	2.9$_{\uparrow 0.3}$	75.0$_{\downarrow 1.1}$	78.0$_{\downarrow 1.3}$	31.3$_{\downarrow 1.3}$

Table 2. **Effects of rewritten samples.** Rewriting the preference training samples $(\boldsymbol{y}_w^+, \boldsymbol{y}_l)$ results in performance reduction.

Qualitative results. To further demonstrate the effectiveness of our method, we conduct case studies. As shown in Fig. 7, the baseline model misinterprets the image content, leading to an incorrect conclusion. In contrast, our model effectively understands image content and provides a more detailed and precise description. This example highlights how our approach effectively reduces hallucinations while simultaneously enhancing the model's overall capability. We conduct more case studies in Sec. G.

4.3. Ablation Studies

In this section, we conduct a series of ablation experiments to further analyze the effectiveness of SENTINEL. More discussions can be found in Sec. D.5.

Effectiveness of data style consistency. To analyze the effect of preference data style, we train the model using rewritten data for comparison. Specifically, we instructed GPT-4 [1] to rewrite $(\boldsymbol{y}_w^+, \boldsymbol{y}_l)$ while ensuring coherence with the context $\boldsymbol{c}$. As shown in Tab. 2, the rewriting results show performance degradation in reducing hallucinations and general ability. This highlights the advantage of our approach in preserving data style consistency. Furthermore, we conduct a detailed analysis in Sec. D.5, which shows that models trained on in-domain data converge to a lower preference optimization loss and achieve better differentiation between positive and negative samples, whereas training with rewritten data provides less improvements.

Effectiveness of cross-checking. To validate the effectiveness of cross-checking for object presence, we conduct experiments using only the Grounding DINO or YOLO World for detection. In this setting, if the model determines that an

Method	Data Scale	Object HalBench		TextVQA	ScienceQA	MM-Vet
		Resp. ↓	Ment. ↓	Acc	I-Acc↑	Overall ↑
LLaVA-v1.5-7B	-	52.7	27.9	**58.2**	66.8	31.1
$\boldsymbol{y}_w^+$ 100%	**8.6K**	**4.3**	**2.6**	**58.2**	**69.2**	**32.6**
$\boldsymbol{y}_w^+$ 50% + $\boldsymbol{y}_w^-$ 50%	10.0K$_{\uparrow 1.4K}$	4.8$_{\uparrow 0.5}$	2.9$_{\uparrow 0.3}$	58.1$_{\downarrow 0.1}$	69.0$_{\downarrow 0.2}$	32.0$_{\downarrow 0.6}$
$\boldsymbol{y}_w^-$ 100%	14.0K$_{\uparrow 5.4K}$	4.6$_{\uparrow 0.3}$	3.0$_{\uparrow 0.4}$	58.1$_{\downarrow 0.1}$	68.7$_{\downarrow 0.5}$	31.6$_{\downarrow 1.0}$

Table 3. **Comparison between context-coherent samples $\boldsymbol{y}_w^+$ and context-agnostic samples $\boldsymbol{y}_w^-$.** This table reveals that incorporating context-coherent samples $\boldsymbol{y}_w^+$ yields better performance.

Method	Object HalBench[55]		AMBER[63]		
	Resp. ↓	Ment. ↓	CHAIR ↓	Hal ↓	Cog ↓
LLaVA-v1.5-7B	52.7	27.9	8.4	35.5	4.0
Non-hallucinated context	**4.3**	**2.6**	**2.9**	**14.6**	**1.2**
Natural context	8.6	4.7	3.3	15.6	1.5
Hallucinated context	14.3	7.1	3.9	18.6	1.8

Table 4. **Comparison between different new context formation strategies during the iterative contextual bootstrapping pipeline.** Appending non-hallucinated sample $\boldsymbol{y}_w^+$ to the existing context $\boldsymbol{c}_i$ yields superior performance compared to incorporating hallucinated samples $\boldsymbol{y}_l$ or greedy decoding contexts, highlighting the effectiveness of our proposed approach.

object is absent, it is directly classified as hallucinated. As shown in Fig. 8, leveraging two object detectors for cross-validation significantly outperforms using a single model, effectively reducing the hallucination rate.

Effect of different $\boldsymbol{y}_w$ types on model performance. As shown in Tab. 3, we conduct a detailed study on the impact of different types and proportions of the positive data $\boldsymbol{y}_w$ on model performance. The results show that $\boldsymbol{y}_w^+$ samples, which contain richer contextual information, enhance the model's generalization ability while achieving similar hallucination reduction with less data.

Effect of non-hallucinated sentences as context $\boldsymbol{c}$. To analyze the impact of using non-hallucinated sentences as context $\boldsymbol{c}$, we evaluate three different settings for generating new context: selecting a hallucinated sentence, selecting a non-hallucinated sentence, or directly using a model-generated sentence from greedy decoding. As shown in Tab. 4, using a non-hallucinated sentence as context improves the model's ability to distinguish hallucinations and significantly reduces their occurrence in the output. This further demonstrates that intervening at the first instance of hallucination is critical for minimizing its recurrence.

Effect of data scale. To analyze the impact of the training data scale on our method, we train the model using different dataset sizes (1k/2k/4k/6k/8k) and evaluate its performance on Object Halbench. As shown in Fig. 8, our method further mitigates model hallucinations as data scale up. This demonstrates the potential and scalability of SENTINEL. Furthermore, since our method does not rely on ultra-large proprietary models or human annotators for dataset construction, it can efficiently collect more training data.

Integrating with existing preference learning methods.

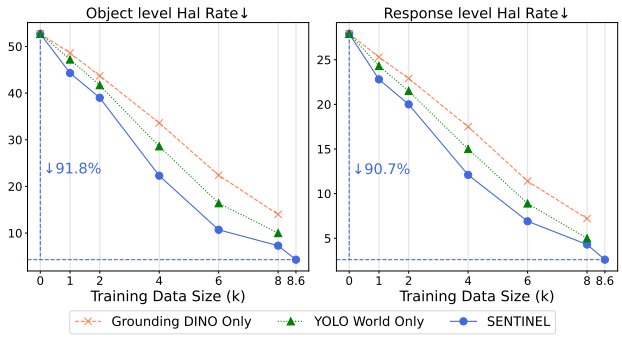


Figure 8. **Impact of training data quantity on hallucination rate in Object Halbench [55].** The results show that SENTINEL demonstrates better efficiency, effectiveness, and scalability, while effectively reducing hallucination rates across varying data scales.

Method	Object HalBench [55]		AMBER [63]		HallusionBench [12]	TextVQA [59]	MM-Vet [78]
	Resp. ↓	Ment. ↓	Acc↑	F1↑	Question Acc↑	Acc↑	Overall ↑
LLaVA-v1.5-7B	52.7	28.0	71.5	74.1	46.86	58.2	31.0
HA-DPO [82]	37.0$_{\downarrow 29.8\%}$	20.9$_{\downarrow 25.4\%}$	74.2$_{\uparrow 2.7}$	78.0$_{\uparrow 3.9}$	47.74$_{\uparrow 0.88}$	56.7$_{\downarrow 1.5}$	30.6$_{\downarrow 0.4}$
HA-DPO + Ours (6K)	**8.0** $_{\downarrow 78.4\%}$	**4.6**$_{\downarrow 78.0\%}$	**76.6**$_{\uparrow 2.4}$	**84.2**$_{\uparrow 6.2}$	**48.72**$_{\uparrow 0.98}$	57.1$_{\downarrow 0.4}$	**33.5**$_{\uparrow 2.9}$

Table 5. **Effectiveness of combining the proposed SENTINEL with HA-DPO.** Only a subset of our training data is needed to reduce hallucinations while enhancing generalization effectively.

To further demonstrate SENTINEL's generalization, we explore integrating with previous hallucination mitigation approaches. As shown in Tab. 5, incorporating a subset of our data into the GPT-generated dataset collected by HA-DPO [82] effectively mitigates hallucinations while significantly enhancing the model's generalization. This highlights SENTINEL's complementarity with other preference learning methods and its potential for broader applicability.

5. Concluding Remarks

Summary. In this work, we address the critical challenge of hallucinations in multimodal large language models (MLLMs). While prior methods have shown promise, they often introduce significant computational overhead, rely on costly resources, or create distributional discrepancies. To tackle these issues, we propose SENTINEL, a framework that intervenes early at the onset of hallucinations by leveraging in-domain preference learning. SENTINEL employs an in-domain candidate bootstrapping strategy, context-aware preference data generation, and a context-aware DPO (C-DPO) loss to effectively curb the propagation of hallucinations while preserving the model's intrinsic distribution. Experimental results across multiple benchmarks demonstrate the superiority of SENTINEL, establishing it as a scalable, efficient, and model-agnostic solution for enhancing the reliability of MLLMs.

Limitation. Currently, as SENTINEL lacks the capability to incorporate spatiotemporal information, it might not be able to effectively address the hallucination issues that require long-term reasoning in video MLLMs. This limitation highlights the need for further research in this area.

Acknowledgments. This work is supported by the Shenzhen Science and Technology Innovation Program (JCYJ20240813105901003, KJZD20240903102901003), Guangdong Basic and Applied Basic Research Foundation (2025A1515011546), and National Key R&D Program of China (2024YFE0215300).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 7, 6, 8, 9

[2] 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. 1, 9

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 8, 9

[4] 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, 2, 9

[5] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024. 1

[6] 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, 9

[7] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3, 4

[8] 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. 1, 2, 6, 7, 9

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1, 9

[10] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 2, 6, 5, 9

[11] Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. Mask-dpo: Generalizable fine-grained factuality alignment of llms. *arXiv preprint arXiv:2503.02846*, 2025. 4

[12] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 6, 8, 5, 9

[13] 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*, 2024. 2, 7

[14] Yudong Han, Liqiang Nie, Jianhua Yin, Jianlong Wu, and Yan Yan. Visual perturbation-aware collaborative learning for overcoming the language prior problem. *arXiv preprint arXiv:2207.11850*, 2022. 9

[15] Zongbo Han, Zechen Bai, Haiyang Mei, Qianli Xu, Changqing Zhang, and Mike Zheng Shou. Skip\n: A simple method to reduce hallucination in large vision-language models. *arXiv preprint arXiv:2402.01345*, 2024. 5

[16] Lehan He, Zeren Chen, Zhelun Shi, Tianyu Yu, Jing Shao, and Lu Sheng. A topic-level self-correctional approach to mitigate hallucinations in mllms. *arXiv preprint arXiv:2411.17265*, 2024. 3, 6, 7

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022. 6

[18] Mingzhe Hu, Shaoyan Pan, Yuheng Li, and Xiaofeng Yang. Advancing medical imaging with language models: A journey from n-grams to chatgpt. *arXiv preprint arXiv:2304.04920*, 2023. 1

[19] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multimodal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2, 6, 7, 9

[20] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 7, 8

[21] Liqiang Jing and Xinya Du. Fgaif: Aligning large vision-language models with fine-grained ai feedback. *arXiv preprint arXiv:2404.05046*, 2024. 2, 7

[22] Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*, 2023. 9

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 2017. 6, 3

[24] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 9

[25] Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*, 2024. 2

[26] Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*, 2023. 6

[27] 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*, 2024. 1, 2, 6, 7, 9

[28] 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. 2

[29] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv:2403.18814*, 2023. 9

[30] Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. Factual: A benchmark for faithful and consistent textual scene graph parsing. *arXiv preprint arXiv:2305.17497*, 2023. 4, 1, 2

[31] Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Scott Yih, and Xilun Chen. Flame: Factuality-aware alignment for large language models. *Advances in Neural Information Processing Systems*, 2024. 2

[32] 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. 2

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 2023. 1, 9

[34] 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*, 2024. 6, 1, 7

[35] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024. 1, 8, 9

[36] 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, 9

[37] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024. 3, 4, 1

[38] Yijun Liu, Jiequan Cui, Zhuotao Tian, Senqiao Yang, Qingdong He, Xiaoling Wang, and Jingyong Su. Typicalness-aware learning for failure detection. *arXiv preprint arXiv:2411.01981*, 2024. 9

[39] Edward Loper and Steven Bird. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*, 2002. 2

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6, 3

[41] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 2022. 2, 6, 5, 9

[42] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 2024. 9

[43] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 9

[44] OpenAI. GPT-4V(ision) system card, 2023. 1, 6, 9

[45] Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In *European Conference on Computer Vision*, 2024. 6

[46] Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*, 2024. 9

[47] Shangpin Peng, Weinong Wang, Zhuotao Tian, Senqiao Yang, Xing Wu, Haotian Xu, Chengquan Zhang, Takashi Isobe, Baotian Hu, and Min Zhang. Omni-dpo: A dual-perspective paradigm for dynamic preference learning of llms. *arXiv preprint arXiv:2506.10054*, 2025. 9

[48] Tianyuan Qu, Longxiang Tang, Bohao Peng, Senqiao Yang, Bei Yu, and Jiaya Jia. Does your vision-language model get lost in the long video sampling dilemma? *arXiv preprint arXiv:2503.12496*, 2025. 9

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021. 3

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 9

[51] 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*, 2023. 2, 4, 9

[52] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2020. 3

[53] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023. 1, 2, 9

[54] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 1

[55] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 2, 6, 8, 4, 5, 7, 9

[56] Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. Data-augmented phrase-level alignment for mitigating object hallucination. *arXiv preprint arXiv:2405.18654*, 2024. 6, 7

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

[58] Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *European Conference on Computer Vision*, pages 139–156. Springer, 2024. 9

[59] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019. 2, 6, 8, 4, 5, 9

[60] Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[61] Zhuotao Tian, Michelle Shu, Pengyuan Lyu, Ruiyu Li, Chao Zhou, Xiaoyong Shen, and Jiaya Jia. Learning shape-aware embedding for scene text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4234–4243, 2019. 9

[62] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[63] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. 2, 6, 7, 8, 5

[64] Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. Declip: Decoupled learning for open-vocabulary dense perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14824–14834, 2025. 9

[65] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 9

[66] Kai Wu, Boyuan Jiang, Zhengkai Jiang, Qingdong He, Donghao Luo, Shengzhi Wang, Qingwen Liu, and Chengjie Wang. Noiseboost: Alleviating hallucination with noise perturbation for multimodal large language models. *arXiv preprint arXiv:2405.20081*, 2024. 6

[67] Yike Wu, Yu Zhao, Shiwan Zhao, Ying Zhang, Xiaojie Yuan, Guoqing Zhao, and Ning Jiang. Overcoming language priors in visual question answering via distinguishing superficially similar instances. In *Proceedings of the 29th International Conference on Computational Linguistics*, 2022. 9

[68] Zhenyu Wu, Ziwei Wang, Xiuwei Xu, Jiwen Lu, and Haibin Yan. Embodied task planning with large language models. *arXiv preprint arXiv:2307.01848*, 2023. 1

[69] 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. *arXiv preprint arXiv:2404.14233*, 2024. 2, 3, 6, 7, 9

[70] Shangyu Xing, Fei Zhao, Zhen Wu, Tuo An, Weihao Chen, Chunhui Li, Jianbing Zhang, and Xinyu Dai. Efuf: Efficient fine-grained unlearning framework for mitigating hallucinations in multimodal large language models. *arXiv preprint arXiv:2402.09801*, 2024. 6, 7, 9

[71] Senqiao Yang, Jiaming Liu, Ray Zhang, Mingjie Pan, Zoey Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Yandong Guo, and Shanghang Zhang. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. *arXiv preprint arXiv:2312.14074*, 2023. 9

[72] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023.

[73] Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. *arXiv preprint arXiv:2412.04467*, 2024. 9

[74] Senqiao Yang, Zhuotao Tian, Li Jiang, and Jiaya Jia. Unified language-driven zero-shot domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23407–23415, 2024. 9

[75] 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*, 2024. 2, 1, 7

[76] 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*, 2024. 2, 6, 7

[77] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source

ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024. 3, 6, 5, 7, 9

[78] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 2, 6, 8, 4, 9

[79] Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *arXiv preprint arXiv:2402.14545*, 2024. 6, 1

[80] Mengxi Zhang, Wenhao Wu, Yu Lu, Yuxin Song, Kang Rong, Huanjin Yao, Jianbo Zhao, Fanglong Liu, Haocheng Feng, Jingdong Wang, et al. Automated multi-level preference for mllms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2, 7

[81] Tiancheng Zhao, Peng Liu, and Kyusong Lee. Omdet: Large-scale vision-language multi-dataset pre-training with multimodal detection network. *IET Computer Vision*, 2024. 3

[82] 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. 2, 3, 6, 8, 1, 4, 7, 9

[83] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Association for Computational Linguistics, 2024. 8

[84] Zhisheng Zhong, Chengyao Wang, Yuqi Liu, Senqiao Yang, Longxiang Tang, Yuechen Zhang, Jingyao Li, Tianyuan Qu, Yanwei Li, Yukang Chen, et al. Lyra: An efficient and speech-centric framework for omni-cognition. *arXiv preprint arXiv:2412.09501*, 2024. 9

[85] 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. 9

[86] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024. 2, 3, 6, 7

[87] 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. 1, 9