EFUF: Efficient Fine-Grained Unlearning Framework for Mitigating Hallucinations in Multimodal Large Language Models

Shangyu Xing   **Fei Zhao**   **Zhen Wu**[*]   **Tuo An**
Weihao Chen   **Chunhui Li**   **Jianbing Zhang**   **Xinyu Dai**
National Key Laboratory for Novel Software Technology, Nanjing University, China
`{xsy, zhaof, ant, chenwh, lich}@smail.nju.edu.cn`
`{wuz, zjb, daixinyu}@nju.edu.cn`

Abstract

Multimodal large language models (MLLMs) have attracted increasing attention in the past few years, but they may still generate descriptions that include objects not present in the corresponding images, a phenomenon known as object hallucination. To eliminate hallucinations, existing methods manually annotate paired responses with and without hallucinations, and then employ various alignment algorithms to improve the alignment capability between images and text. However, they not only demand considerable computation resources during the finetuning stage but also require expensive human annotation to construct paired data needed by the alignment algorithms. To address these issues, we propose an efficient fine-grained unlearning framework (EFUF), which performs gradient ascent utilizing three tailored losses to eliminate hallucinations without paired data. Extensive experiments show that our method consistently reduces hallucinations while preserving the generation quality with modest computational overhead. Our code and datasets are available at https://github.com/starreeze/efuf.

1 Introduction

In the burgeoning field of artificial intelligence, the advent of multimodal large language models (MLLMs) has opened new frontiers in human-computer interaction, data processing, and automated content generation (Zhu et al., 2023; Liu et al., 2023b; Chen et al., 2023; Ye et al., 2023). These sophisticated models, capable of understanding both text and images, have significantly advanced our ability to automate complex tasks.

However, an intriguing and critical phenomenon known as "hallucination" in these models poses unique challenges for current research. Hallucination in MLLMs refers to the generation of inconsistent responses that are not grounded by the

Figure 1: An example of hallucination in MLLM.

multimodal context (Sun et al., 2023). For example, as shown in Figure 1, the caption includes the object "landing gear", but in fact it does not appear in the image. Such hallucinations will lead to misinformation, potentially undermining user trust in numerous downstream applications.

Recent methods for mitigating multimodal hallucination can be divided into two categories: inference-based methods (Lee et al., 2023; Zhou et al., 2023; Yin et al., 2023; Wang et al., 2023; Sicong Leng, 2023; Wang et al., 2024; Chen et al., 2024) and finetuning-based methods (Sun et al., 2023; Yu et al., 2023; Liu et al., 2023a; Zhao et al., 2023; Jiang et al., 2023). Inference-based methods correct or restrict generated content through external expert review, self-reflection or decoding strategies during inference stage. However, they usually require additional inference steps with increased costs and delay (Yu et al., 2023). Furthermore, each task demands specific procedure or prompt (Xu et al., 2024), adding to the complexity

---

[*] Corresponding author.

of implementation. Overcoming these drawbacks, finetuning-based approaches are proposed to adjust the model directly through specialized datasets and preference alignment algorithms. These algorithms, including RLHF (Sun et al., 2023; Liu et al., 2023a), DPO (Yu et al., 2023; Zhao et al., 2023; Zhou et al., 2024) and contrastive learning (Jiang et al., 2023), enhance the congruence between text and images, leading to improved alignment. Although they have achieved good performance, two critical issues emerge:

First, their data demands are substantial, as they require a comprehensive set of paired positive and negative samples for effective finetuning. The alignment algorithms they employed demand paired hallucinated and non-hallucinated responses for each query. Acquiring such specific and varied response sets for each query presents a significant challenge. Recent methodologies in this field predominantly rely on human labor to annotate the output from the MLLM, requiring specialized expertise and incurring considerable expenditure of time and financial resources.

Second, The finetuning of MLLM utilizing these alignment algorithms usually demands considerable computational resources. Most of these techniques are sophisticated and necessitate the simultaneous operation of multiple models to execute preference alignment, thereby escalating the overall cost significantly.

To tackle the above issues, we propose the **E**fficient **F**ine-Grained **U**nlearning **F**ramework (EFUF), which offers the advantage of not necessitating paired data and being more efficient during the finetuning phase. Our method, grounded in the principles of unlearning, mainly relies on performing gradient ascent on negative samples to mitigate hallucinations, eliminating the need for costly manually-annotated paired data. Additionally, it consumes considerably fewer computational resources. Unlike traditional alignment algorithms that require simultaneous operation of multiple models to execute preference alignment, EFUF operates without this requirement.

The key to applying the unlearning algorithm is how to curate positive and negative samples, i.e., distinguish between real and hallucinated objects, in a manner that is both cost-effective and reliable. Intuitively, the similarity between objects and their corresponding images can act as an indicator for hallucinations, since the image contains real objects but not the hallucinated ones. Inspired by

Zhao et al. (2024), we propose to utilize the CLIP model (Radford et al., 2021) to evaluate text-image congruence. Trained on a vast corpus of text-image pairs, CLIP stands as a robust tool to help identify hallucinations.

After ascertaining the capability of CLIP through a preliminary experiment, we curate our dataset manually-free by utilizing CLIP scores, before applying our unlearning-based method to MLLMs. This process enables us to harness the power of unlearning, offering a potent and efficient approach for mitigating hallucinations in MLLMs.

Our contribution can be summarized as follows:
1) To the best of our knowledge, we provide a new perspective to utilize unlearning to mitigate multimodal hallucination in MLLMs.
2) We propose an efficient fine-grained unlearning framework EFUF, which can obtain positive and negative examples separately in a cost-effective and reliable manner.
3) EFUF has good compatibility and can be easily extended to existing MLLMs. Experiments conducted across a range of MLLMs validate the effectiveness of our method.

2 Related Work

In this section, we review the existing studies on Hallucination Mitigation for MLLM and Unlearning algorithm.

2.1 Hallucination Mitigation for MLLM

To mitigate hallucinations for MLLM, various methods have been proposed. According to different phase during which they tackle the hallucinations, their work can be divided into two categories:

(1) Inference-based methods. They employ external experts, self-reflection framework or decoding strategies to constrain or modify generated content during the inference phase, thereby reducing hallucinations. For example, LURE (Zhou et al., 2023) utilizes manually-crafted features to detect hallucinations and therefore revises the generated text. Woodpecker (Yin et al., 2023) proposes to post-edit hallucinations by combining the output of MLLMs and a more accurate expert VQA model using GPT-3.5. VIGC (Wang et al., 2023) iteratively refines the instruction data using generation and correction framework. VOLCANO (Lee et al., 2023) trains the MLLM to give self-feedback, and then performs self-reflection on the original generated text according to the feedback. VCD

(Sicong Leng, 2023) first introduces contrastive decoding in MLLM by disturbing the visual inputs and calculate visual uncertainty to restrict the generation of hallucinated tokens. ICD (Wang et al., 2024) utilizes disturbance on instructions instead of images. HIO (Chen et al., 2024) employs a hallucinated model to further widen the gap between hallucinated and correct tokens, achieving better contrastive outcomes. Although these methods do not need to train the model, they require additional inference steps with increased costs and delay (Yu et al., 2023), and specific procedure and prompt must be designed for each task (Xu et al., 2024).

(2) Finetuning-based methods. Overcoming the potential drawbacks of the first category, these methods involve crafting specific datasets and finetuning the model, aiming for better alignment between images and text. For instance, LLaVA-RLHF (Sun et al., 2023) first adopts RLHF to mitigate hallucinations. Based on this work, RLHF-V (Yu et al., 2023) introduces fine-grained alignment by manually correcting the outputs of MLLMs. Beyond standard RLHF, some works utilize other improved algorithms for better efficiency, e.g., DPO (Zhao et al., 2023; Zhou et al., 2024), instruction tuning (Liu et al., 2023a), and contrastive learning (Jiang et al., 2023). However, these methods require expensive manually annotated paired data, and most of them also demand substantial computational resources during the finetuning stage. Therefore, in this work, we focus on reducing the data and computation requirements.

2.2 Unlearning

Unlearning refers to a technique designed to induce a model to "forget" specific behaviors or data, primarily through the application of gradient ascent methods (Cao and Yang, 2015). Recently, unlearning for LLM is receiving increasing attention. Jang et al. (2023) demonstrate that straightforward gradient ascent can effectively eliminate privacy vulnerabilities in LLMs. Later, Yao et al. (2023) propose the use of random mismatch and restrictions on KL divergence for positive samples, reducing the negative impact of unlearning on the general performance of LLMs.

In our research, we extend the concept of unlearning to the realm of multimodal hallucination mitigation in MLLMs, proposing a novel solution for enhancing model reliability and accuracy in multimodal contexts. In contrast to earlier approaches that apply unlearning across the entirety of a model's responses, our methodology focuses exclusively on the unlearning of hallucinated objects. This precise, fine-grained unlearning strategy allows for a more sophisticated refinement of the model's outputs, ensuring that only inaccuracies are corrected without diminishing the model's capabilities in other areas. To the best of our knowledge, this is the first attempt to adopt unlearning to multimodal large language models.

3 Preliminary Experiment

The initial phase of our research involves confirming the hypothesis that text-image congruence can serve as a reliable indicator of hallucination occurrences. To this end, we designed a preliminary study aimed at validating this premise. Below, we detail the methods and findings of this experiment.

3.1 Hallucinated *v.s.* Non-Hallucinated

Our approach involves employing the CLIP model to assess the similarity between text and corresponding images, with the objective of determining whether there is a discernible difference in the similarity scores of hallucinated versus non-hallucinated content. Following Zhou et al. (2023), we manually annotate 200 image captions generated by MiniGPT (Zhu et al., 2023) and LLaVA (Liu et al., 2023b), labeling objects as either hallucinated or non-hallucinated. Subsequently, we define an object-level image-relevance score by calculating fine-grained CLIP similarities for these objects in relation to their associated image segments, aiming to uncover any significant disparities in score distributions.

Formally, let $V = \{v_1, v_2, ..., v_m\}$ denotes the collection of images, and $T = \{t_1, t_2, ..., t_m\}$ is the corresponding captions generated by the MLLM. For each $t_i \in T$, we manually annotated all the objects in the caption, represented by $O_i = \{o_i^1, o_i^2, ..., o_i^n\}$, and $O = \{O_1, O_2, ..., O_m\}$. After that, we determine whether the object is hallucinated, i.e., whether it appears in the image, assigning each object a binary value $h(o_i^j)$ as follows:

$$h(o) = \begin{cases} 1, & \text{if the object } o \text{ is hallucinated;} \\ 0, & \text{if the object } o \text{ is not hallucinated.} \end{cases}$$

Based on this evaluation, we categorize the objects into two groups: the hallucinated group $H_1 = \{o|o \in O, h(o) = 1\}$ and the non-hallucinated group $H_0 = \{o|o \in O, h(o) = 0\}$. We then calculate the fine-grained CLIP score between each
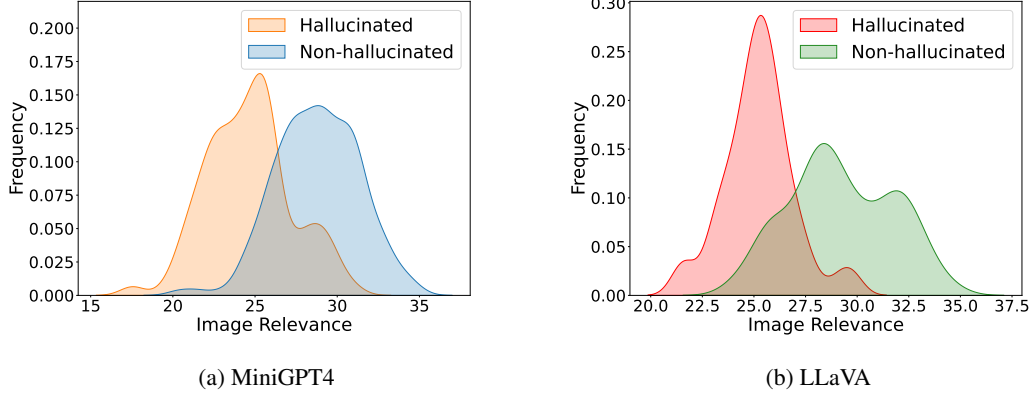
(a) MiniGPT4   (b) LLaVA

Figure 2: Comparison of hallucinated and non-hallucinated objects generated by MiniGPT4 (a) and LLaVA (b) on image-relevance scores.

Model	Hal.	Mean	Std.	p
MiniGPT4	No	28.26	2.74	6.0×10^{-30}
	Yes	25.35	2.70	
LLaVA	No	28.64	2.65	2.5×10^{-12}
	Yes	26.11	2.27	

Table 1: Statistics and significance test on samples generated by MiniGPT4 and LLaVA. Hal. indicates whether the objects are hallucinated, Mean and Std. represent their average and standard deviation of image-relevance scores, and p is the p-value of t-test.

object o_i^j in either group and its corresponding image v_i. Given that most objects cover only a portion of the image, we segment the image into patches and employ a sliding window technique to identify the best match. Thus, the image-relevance score for each object is determined as follows:

$$S(o_i^j) = \max_{w_i \in W_i} \text{CLIP}(o_i^j, w_i), \qquad (1)$$

where W_i represents the set of sliding windows over the patches of the image v_i.

This methodology enables us to obtain two sets of image-relevance scores $S_1 = \{S(o)|o \in H_1\}$ and $S_0 = \{S(o)|o \in H_0\}$. In the next section, we will examine the distributions of these scores and validate our hypothesis that text-image similarity can indicate the likelihood of hallucination.

3.2 Results and Analysis

In our analysis, we applied a two-sample t-test to examine the differences between the score distributions of hallucinated and non-hallucinated objects. The results, as detailed in Table 1, reveal a notable discrepancy between the mean values of these dis-

tributions, as indicated by the p-value. This statistical evidence allows us to confidently reject the null hypothesis that the two distributions have identical means, underscoring the utility of CLIP similarity scores in detecting hallucinations.

To provide a clearer understanding of these differences, we visualized the score distributions through density plots. These plots, illustrated in Figure 2, demonstrate that scores for hallucinated objects typically fall below 32, whereas scores for non-hallucinated objects generally exceed 23 for both the two models. Our quantitative analysis further reveals that among the objects scoring above 32, only 0.6% and 1.6% are hallucinated, and among those below 23, only 2.3% and 1.7% are not hallucinated, for MiniGPT and LLaVA respectively. These findings not only substantiate our hypothesis but also suggest that definitive thresholds can be established to effectively segregate positive and negative samples for the purpose of unlearning.

4 Multimodal Hallucination Mitigation

4.1 Overview

After ascertaining the capability of CLIP through a preliminary experiment, we design EFUF, whose overview is shown in Figure 3. Drawing from established methodologies in prior research (Sun et al., 2023; Yu et al., 2023; Liu et al., 2023a; Zhao et al., 2023; Jiang et al., 2023), our approach is bifurcated into two key stages: dataset construction and the unlearning process itself. Initially, we harness CLIP scores to identify and segregate various samples; after that, unlearning is applied on the model with the curated samples.

Concretely, in constructing the dataset, we first prompt the model to generate captions for given
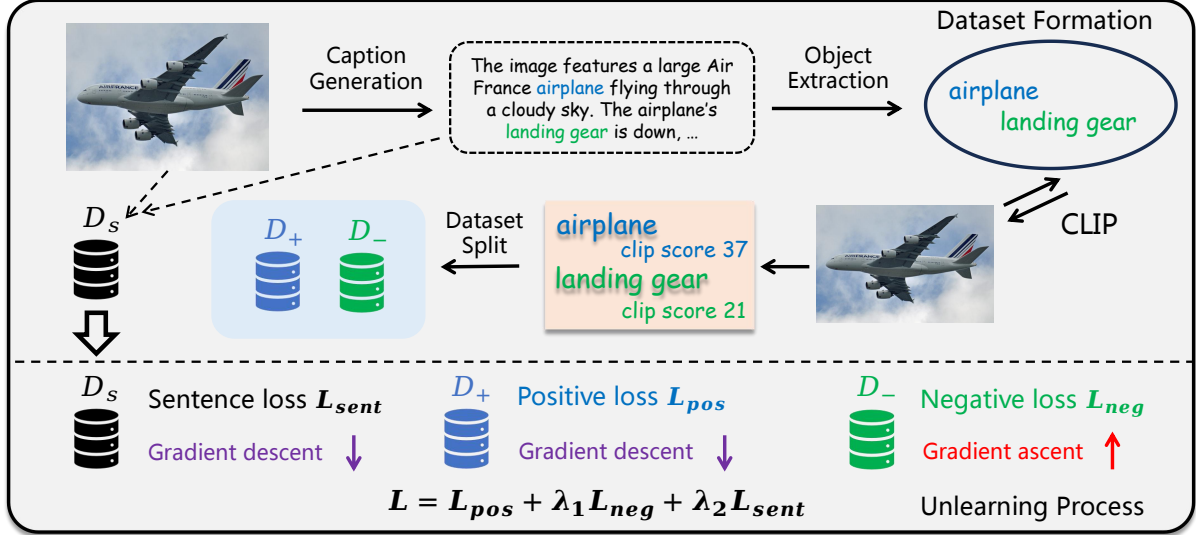
Figure 3: An overview of EFUF. EFUF is divided into two stages: dataset formation and unlearning process. Initially, we extract objects from generated captions and calculate their image relevance utilizing CLIP, followed by the construction of three datasets. Subsequently, three corresponding losses are tailored to finetune the model.

images. After that, we utilize the CLIP model to calculate the fine-grained similarity score of the object phrases in text and the corresponding segments in image. By setting thresholds for the scores, we are able to discern and compile distinct samples from the generated text, forming a dataset for fine-tuning that circumvents the need for labor-intensive manual annotation. During the finetuning phase, we employ an efficient unlearning method, which involves the development of three distinct types of losses. These losses are designed to aid the model in discarding incorrect multimodal alignments that could lead to hallucinations, while preserving the correct alignments essential for tasks. Unlearning generally requires less computation resources compared with conventional alignment algorithms in the finetuning stage, so the computation amount can also be effectively reduced.

4.2 Dataset Formation

Prior to implementing unlearning with MLLMs, it's imperative to define the targets of unlearning and accordingly assemble the requisite positive and negative samples. As evidenced in Section 3.2, specific thresholds can effectively delineate between these samples. Hence, we apply these predetermined image-relevance thresholds to filter the hallucinated and non-hallucinated objects.

Given that a single response may encompass both hallucinated and non-hallucinated objects, a fine-grained approach to unlearning is warranted. Rather than attempting to unlearn an entire re-

sponse wholesale, we opt for a targeted strategy focusing on the subsentences corresponding to the object, delineated by punctuation. Moreover, to preserve the model's overarching sentence comprehension and capabilities, we also compile samples of the complete sentences based on the mean image-relevance scores of all included objects, in addition to the positive and negative subsentences. These three categories of samples collectively form the dataset tailored for the unlearning process, facilitating a more nuanced and effective mitigation of multimodal hallucinations.

Formally, let $D = \{v; x; y\}$ denotes a finetuning dataset for MLLM, where v is the image, x is the text query (prompt), and y is the text answer. The positive subsentence dataset is formulated as

$$D^+ = \left\{ v_i; \mathrm{pre}(o_i^j); \mathrm{cur}(o_i^j) | o_i^j \in O, S(o_i^j) > T_0 \right\},$$

where $\mathrm{cur}(o)$ represents the subsentence where object o situates, $\mathrm{pre}(o)$ represents all the texts before $\mathrm{cur}(o)$, including prompt, and T_0 is the threshold for positive samples. The text that comes after $\mathrm{cur}(o)$ is truncated and unused. Similarly, The negative subsentence dataset is defined as

$$D^- = \left\{ v_i; \mathrm{pre}(o_i^j); \mathrm{cur}(o_i^j) | o_i^j \in O, S(o_i^j) < T_1 \right\},$$

where T_1 is the threshold for negative samples.

To construct a comprehensive dataset featuring complete responses, it is essential to establish a metric for assessing sentence-level hallucinations.

This is achieved by calculating the average image-relevance score across all referenced objects within a response. The formula for this sentence-level image-relevance score is given by

$$S(t_i) = \frac{1}{n} \sum_{j=1}^{n} S(o_i^j). \qquad (2)$$

With this metric, we can curate a dataset of responses by filtering out those responses from the model that meet the specific criterion:

$$D^s = \{v_i; p_i; t_i | t_i \in T, S(t_i) > T_2\},$$

where p_i denotes the prompt for response t_i, and T_2 is the threshold for response samples.

Finally, we take $D_{unlearning} = \{D^+, D^-, D^s\}$ as our unlearning dataset.

4.3 Unlearning for MLLM

After constructing the dataset, the final phase of our approach is the application of unlearning techniques to the model. Prior studies (Eldan and Russinovich, 2023) have shown that employing solely the unlearning loss severely undermines the model's linguistic comprehension, rendering it incapable of producing coherent sentences. Thus, we introduce a dual-faceted fine-grained unlearning approach: applying a negative loss to the subsentences containing hallucinated objects, and a positive loss to those containing non-hallucinated objects. This strategy aims to curtail the production of hallucinated content while encouraging precise object representation, thus diminishing the occurrence of hallucinations. Meanwhile, we also propose a sentence loss, aiming to preserve the model's ability to generate cohesive, long-form text. In the following, we will introduce these losses in detail.

As is indicated by previous works, the core of unlearning is the gradient ascent strategy. Formally, unlearning updates the model parameters by:

$$\Delta \theta = \eta \nabla_\theta L_{ft}(v, x, y; \theta), \qquad (v, x, y) \sim D, \ (3)$$

where θ denotes the model's parameters, η is the (un)learning rate, and L_{ft} signifies the finetuning loss function. In the context of multimodal large language models, the supervised finetuning loss function L is articulated as

$$L_{ft}(v, x, y; \theta) = \frac{1}{|y|} \sum_{i=1}^{|y|} l(f_\theta(v, x, y_{<i}), y_i), \ (4)$$

where f_θ symbolizes the model with parameter θ, and $l(\hat{y}_i, y_i)$ calculates the cross-entropy loss for the predicted and actual values.

To counteract hallucinations while maintaining overall model efficacy, we introduce three distinct losses tailored to the datasets we've constructed. The first, termed *negative loss*, applies gradient ascent to negative subsentences as follows:

$$L_{neg} = -L_{ft}(v, x, y), \qquad (v, x, y) \sim D^-. \ (5)$$

This inversion of the loss function enables gradient ascent. The second, the *positive loss*, aims at encouraging the model to generate correct objects, with its formulation remaining straightforward:

$$L_{pos} = L_{ft}(v, x, y), \qquad (v, x, y) \sim D^+. \ (6)$$

The last, the *sentence loss* is designed to retain model's comprehension and capabilities on full sentences during the unlearning process:

$$L_{sent} = L_{ft}(v, x, y), \qquad (v, x, y) \sim D^s. \ (7)$$

The overall loss equation then becomes a weighted amalgamation of these three components:

$$L = L_{pos} + \lambda_1 L_{neg} + \lambda_2 L_{sent}, \qquad (8)$$

where λ_1 and λ_2 represent the unlearning weight and the sentence weight respectively.

During training, we perform concurrent sampling from the three datasets, individual loss computation, and aggregation to derive the final loss metric. By doing so, we effectively mitigate hallucinations and preserve the model's proficiency in processing extensive sentences.

5 Experiments

5.1 Experimental Settings

Dataset. We adopt MSCOCO (Lin et al., 2014) as our dataset. Since our approach necessitates only the images themselves, their annotations are used exclusively for evaluation. Details of our dataset can be found in Appendix A.2.

Evaluation Metrics. Following Yu et al. (2023), our assessment encompasses two dimensions: trustworthiness measured by the degree of hallucination, and helpfulness determined by the quality of the generated text. To quantify hallucinations, we utilize CHAIR (Rohrbach et al., 2018), MHumanEval (Yu et al., 2023) and POPE (Fu et al., 2023). For

Model	Hallucination Rate					Generation Quality				
	$Chair_S\downarrow$	$Chair_I\downarrow$	$Human_S\downarrow$	$Human_I\downarrow$	POPE↑	Bleu1↑	Bleu2↑	Bleu4↑	Info.↑	ppl.↓
MiniGPT4	45.9	23.2	69.0	27.3	81.0	43.8	29.5	15.5	86.7	0.134
+ *EFUF*	38.9	21.1	45.0	12.7	82.3	45.6	31.1	16.7	87.5	0.121
LLaVA	52.8	22.8	42.0	14.7	85.3	43.2	29.0	15.2	93.7	0.139
+ *EFUF*	41.9	18.7	24.0	7.7	85.9	45.3	31.0	16.8	93.5	0.129
mPLUG-owl	71.1	33.5	60.0	24.1	88.5	43.3	29.1	15.1	91.1	0.129
+ *EFUF*	40.5	23.2	46.0	17.7	90.7	52.3	35.3	19.9	90.0	0.139
ShareGPT4V	46.8	22.3	31.0	9.9	87.8	43.3	29.2	15.4	89.6	0.157
+ *EFUF*	36.9	18.4	14.0	5.4	88.1	46.9	32.5	18.1	91.1	0.159

Table 2: Performance comparison of various MLLMs with and without EFUF. Hallucination is assessed using CHAIR ($Chair_S$, $Chair_I$), MHumanEval ($Human_S$, $Human_I$), and POPE metrics. Quality is evaluated based on consistency with ground truth (Bleu1, Bleu2), informativeness (Info.), and fluency (ppl.). A downward arrow (↓) indicates that lower values are better, whereas an upward arrow (↑) signifies that higher values are preferable.

generation quality, we leverage the BLEU (Papineni et al., 2002) score for assessing the consistency with ground truth, evaluate informativeness through GPT-4's judgment (OpenAI, 2023), and use GPT-2's perplexity score (Radford et al., 2019) to determine text fluency. Details on the evaluation metrics are provided in Appendix A.3.

5.2 Baselines

To affirm the robustness of EFUF across a spectrum of MLLMs, we conducted evaluations against a suite of state-of-the-art base models. These include MiniGPT4 (Zhu et al., 2023), mPLUG-owl (Ye et al., 2023), LLaVA (Liu et al., 2023b), and ShareGPT4V (Chen et al., 2023), which are pretrained on extensive multimodal datasets and subsequently finetuned on high-quality instructions. In our experiments, we integrate EFUF into them to obtain the enhanced model.

6 Results and Analysis

6.1 Main Results

As is shown in Table 2, we evaluate EFUF across a variety of MLLMs, assessing both the hallucination rate and generation quality.

Hallucination Rate. Based on the results, our approach demonstrates a consistent reduction in hallucination rates across all four MLLMs, with an average improvement of approximately 15% and 5% on the $Chair_S$ and $Chair_I$ metric, 18% and 8% on the $Human_S$ and $Human_I$ metric, and 1% on the POPE metric. These findings validate the effectiveness and adaptability of our method, emphasizing its capacity to notably lower hallucination rates across cutting-edge models.

Generation Quality. Table 2 also highlights the improvements of EFUF in generation quality. Results show that our method not only reduces the hallucination rate but also enhances overall generation quality. Specifically, it improves BLEU-1 by 4%, BLEU-2 by 3%, BLEU-4 by 2%, informativeness by 1%, and fluency by 1%, across the four models. These enhancements stem from two main factors: the unlearning strategy which promotes accurate object generation, and the sentence loss design which enhances fluency.

6.2 Ablation Study

Without loss of generality, we select the MiniGPT4 model for the ablation study to investigate the effects of different modules of our proposed method. As outlined in Section 4.3, our approach is fundamentally comprised of two key elements: the sentence loss and the unlearning mechanism, which itself includes the negative loss and the positive loss. In order to quantify the contribution of each component, we contrast EFUF against the following configurations: (1) vanilla unlearning: a strategy employing the coarse-grained unlearning, leveraging both positive and negative entire sentences identified based on their sentence-level image relevance scores; (2) fine-grained unlearning: the unlearning strategy applied in EFUF, but without the sentence loss; (3) sentence-loss-only method: a method that solely applies the sentence loss of EFUF, omitting the unlearning aspects. The subsequent content details the outcomes and insights derived from these experimental comparisons.

Effects of Unlearning. As shown in Table 3, we observe marginal improvements in hallucination

Method	Hallucination Rate					Generation Quality				
	Chair$_S$↓	Chair$_I$↓	Human$_S$↓	Human$_I$↓	POPE↑	Bleu1↑	Bleu2↑	Bleu4↑	Info.↑	ppl.↓
MiniGPT4	45.9	23.2	69.0	27.3	81.0	43.8	29.5	15.5	86.7	0.134
+ *unlearn.*	42.4	22.7	56.0	17.3	82.0	44.2	29.8	15.6	87.6	0.120
+ *f.g. unlearn.*	36.1	17.9	39.0	9.7	82.7	47.3	32.8	17.1	87.2	0.170
+ *sentence loss*	44.1	29.8	58.0	17.0	81.7	43.6	29.1	16.0	86.8	0.120
+ *EFUF*	38.9	21.1	45.0	12.7	82.3	45.6	31.1	16.7	87.5	0.121

Table 3: Performance comparison of EFUF with vanilla unlearning strategy (*unlearn.*), fine-grained unlearning strategy (*f.g. unlearn.*), and sentence-loss-only method (%). Although fine-grained unlearning achieves the lowest hallucination rate, it drastically sacrifices fluency, making the generated content difficult for humans to read.

Method	Hallucination Rate					Generation Quality				
	Chair$_S$↓	Chair$_I$↓	Human$_S$↓	Human$_I$↓	POPE↑	Bleu1↑	Bleu2↑	Bleu4↑	Info.↑	ppl.↓
LLaVA	52.8	22.8	42.0	14.7	85.3	43.2	29.0	15.2	**93.7**	0.139
+ *RLHF*	60.2	24.8	40.0	12.7	**87.0**	39.8	25.8	12.6	<u>93.5</u>	**0.126**
+ *HADPO*	52.3	21.6	<u>28.0</u>	10.8	84.2	43.8	29.6	<u>15.7</u>	91.4	0.148
+ *POVID*	**41.3**	<u>19.2</u>	29.0	<u>8.3</u>	<u>86.3</u>	<u>44.5</u>	<u>30.0</u>	15.1	86.8	0.233
+ *EFUF*	<u>41.9</u>	**18.7**	**24.0**	**7.7**	85.9	**45.3**	**31.0**	**16.8**	<u>93.5</u>	<u>0.129</u>

Table 4: Performance comparison of different hallucination mitigation methods for LLaVA on metrics measuring hallucination rate and generation quality. Best scores are in bold and second bests are underlined.

Method	MME	GQA	SQA	QBench
LLaVA	**1491**	<u>63.0</u>	66.9	<u>59.2</u>
+ *RLHF*	1212	48.4	65.4	53.0
+ *HADPO*	1441	61.2	<u>67.2</u>	58.6
+ *POVID*	1438	61.9	**68.4**	<u>59.2</u>
+ *EFUF*	<u>1468</u>	**63.2**	66.4	**59.3**

Table 5: Performance comparison of different hallucination mitigation methods for LLaVA on metrics measuring VQA and reasoning capability.

rate reduction and BLEU score enhancement, when the method of vanilla unlearning and sentence loss are applied. However, these gains are trivial compared to those achieved by fine-grained unlearning and the complete EFUF, highlighting the essential role fine-grained unlearning plays in mitigating hallucinations and generating correct objects.

Effects of the Sentence Loss. Compared to EFUF, the fine-grained unlearning approach results in a slightly lower hallucination rate but at the cost of informativeness and fluency. In this scenario, BLEU scores fall short of capturing this issue, as they only measure n-gram matches. The decline in fluency is highlighted by a significant increase in perplexity, rendering the responses largely unreadable by humans. Manual examination further reveals that the generated content often consists fragmented and incoherent sentences. Conversely, method employing only the sentence loss

and EFUF do not exhibit these flaws, emphasizing the vital function of sentence loss in maintaining high-quality text generation.

In summary, our analysis confirms the necessity of integrating both fine-grained unlearning and sentence loss to effectively reduce hallucinations without compromising the model's proficiency in generating comprehensive, fluent sentences. This combined approach ensures model performance while notably reduces hallucinations.

6.3 Comparison with Other Methods

To further evaluate the performance of EFUF, we compare it with other methods tailored to hallucination mitigation. These include LLaVA-RLHF (Sun et al., 2023), HA-DPO (Zhao et al., 2023), and POVID (Zhou et al., 2024), which are all evaluated using their officially released checkpoints. We benchmark EFUF against these methods on the LLaVA model, since their checkpoints are all based on LLaVA.

Hallucination Rate & Generation Quality. We measure EFUF's generation quality along with hallucination rate in Table 4. Compared to other hallucination mitigation methods, EFUF demonstrates comparable or superior performance, while requiring minimal data construction cost and training resources among all. Additionally, our improvements in generation quality are on par with RLHF-based methods, which typically demand expensive human
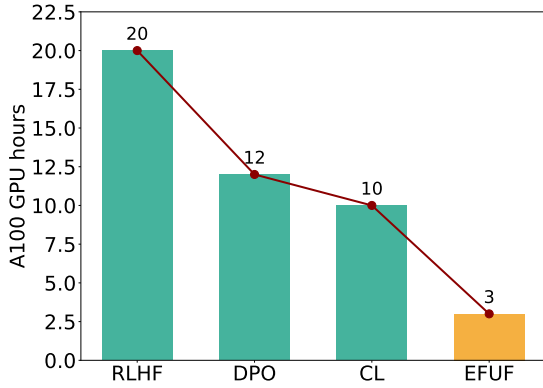
Figure 4: Training time comparison of EFUF with other finetuning-based methods (A100 GPU hours).

annotations and significant computations. These outcomes highlight our method's effectiveness and efficiency.

VQA & Reasoning Capability. To provide a more holistic evaluation of EFUF, we also assessed its performance on VQA and reasoning tasks. We employed benchmarks such as MME (Fu et al., 2024), GQA (Hudson and Manning, 2019), ScienceQA (Lu et al., 2022), and QBench (Wu et al., 2024). Table 5 reports the results for the baseline model, EFUF, and competing methods. EFUF demonstrates modest performance fluctuation across these benchmarks compared to other hallucination mitigation strategies, indicating that our method does not negatively affect VQA and reasoning capabilities.

6.4 Training Cost

EFUF distinguishes itself from conventional finetuning approaches to hallucination mitigation through its markedly lower end-to-end training costs. A key advantage of EFUF lies in its dataset construction process, which obviates the need for costly human annotations. Traditional methods typically rely on extensive human-labeled datasets, often comprising around 10,000 samples at expenses surpassing $3,000 (Sun et al., 2023; Yu et al., 2023). Otherwise, they create the dataset with the assistance of GPT-4, involving up to 500,000 samples pre-screened before manual review, incurring costs for around 200 million tokens equivalent to $2,000 (Liu et al., 2023a; Jiang et al., 2023).

In stark contrast, EFUF's resource efficiency extends to its training demands. As depicted in Figure 4, EFUF's training on an A100 GPU for a

MiniGPT4 model requires merely 3 GPU hours, a fraction of the resources needed by other methods. For comparison, RLHF-based finetuning typically consumes 20 GPU hours (Sun et al., 2023), DPO ranges from 8 (Yu et al., 2023) to 16 (Zhao et al., 2023) GPU hours, and contrastive learning method requires around 10 GPU hours (Jiang et al., 2023).

This substantial reduction on resource requirements in both dataset construction and training stage not only makes EFUF a cost-effective approach but also enhances its scalability and accessibility for broader applications in hallucination mitigation within the realm of multimodal large language models.

6.5 Additional Analyses

To further substantiate the effectiveness of EFUF, we provide extensive supplementary analyses in the appendices. As presented in Appendix B, EFUF complements and enhances the performance of existing hallucination mitigation strategies. We also explore the impact of varying weights as hyperparameters in Appendix C. Finally, a case study detailed in Appendix D quantitatively evaluates the generated text under different methods, showcasing the distinct advantages of our proposed solution.

7 Conclusion

In this paper, we find that text-image similarity is helpful for identifying multimodal hallucinations, and propose a novel unlearning framework to mitigate hallucinations in MLLM. Specifically, we first curate different samples utilizing the image-relevance score derived from CLIP similarity, and then design three distinct losses to perform unlearning on the curated samples. Extensive experiments on different baselines show that our method effectively reduces multimodal hallucinations while retaining the general performance of the model.

Limitations

The limitations of our work mainly contain two aspects. Firstly, the exploration of alternative methods for assessing text-image similarity presents an avenue for further research. Our findings affirm the utility of text-image relevance in constructing datasets for the unlearning process, with the relevance scores derived using the CLIP model. Additional methodologies for determining text-image relevance warrant exploration, which may further optimize the construction of unlearning datasets.

Secondly, in line with most preceding research, our investigation primarily addresses object hallucinations, gauged by the presence or absence of the depicted object in the corresponding image. The exploration of other varieties of hallucinations, including but not limited to the attributes or positioning of objects within the image, represents a significant area for future work.

Acknowledgements

References

Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480.

Beitao Chen, Xinyu Lyu, Lianli Gao, Jingkuan Song, and Heng Tao Shen. 2024. Alleviating hallucinations in large vision-language models through hallucination-induced optimization.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *CoRR*, abs/2311.12793.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *CoRR*, abs/2310.02238.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 14389–14408, Toronto, Canada. Association for Computational Linguistics.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2023. Hallucination augmented contrastive learning for multimodal large language model. *CoRR*, abs/2312.06968.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2023. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. *CoRR*, abs/2311.07362.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 292–305. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. 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*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *CoRR*, abs/2304.08485.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

NVIDIA, Péter Vingelmann, and Frank H.P. Fitzek. 2020. Cuda, release: 10.2.89.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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, Brussels, Belgium, October 31 - November 4, 2018*, pages 4035–4045. Association for Computational Linguistics.

Guanzheng Chen Xin Li Shijian Lu Chunyan Miao Lidong Bing Sicong Leng, Hang Zhang. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning large multimodal models with factually augmented RLHF. *CoRR*, abs/2309.14525.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu,

Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, and Conghui He. 2023. VIGC: visual instruction generation and correction. *CoRR*, abs/2308.12714.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding.

Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. 2024. Q-bench: A benchmark for general-purpose foundation models on low-level vision.

Ziwei Xu, Sanjay Jain, and Mohan S. Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *CoRR*, abs/2401.11817.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023. Large language model unlearning.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *CoRR*, abs/2310.16045.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2023. RLHF-V: towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *CoRR*, abs/2312.00849.

Fei Zhao, Taotian Pang, Chunhui Li, Zhen Wu, Junjie Guo, Shangyu Xing, and Xinyu Dai. 2024. Aligngpt: Multi-modal large language models with adaptive alignment capability.

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

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. Aligning modalities in vision large language models via preference fine-tuning.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *CoRR*, abs/2310.00754.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

A Details on Experiment Settings

A.1 Implementation Details

For dataset construction, in order to efficiently obtain the object set O, we prompt the LLaMA-2-70b (Touvron et al., 2023) model to extract all the objects from the response text. During training, we only tune each model's multimodal mapping layers, i.e., ones that map image feature to text token embedding. We train each model for a fixed 1 epoch with AdamW (Loshchilov and Hutter, 2019) as the optimizer, and report their performance on test set. We implement all the models with the PyTorch framework (Paszke et al., 2019), and run experiments on an NVIDIA A100 GPU (NVIDIA et al., 2020). For hyperparameters, we set the weight of unlearning loss λ_1 to 0.3, the weight of sentence loss λ_2 to 0.2, the learning rate η to 1e-5, weight decay to 0.05. Based on the analysis in Section 3, the threshold for normal object T_0 and hallucinated object T_1 is set to 32 and 23, respectively. Besides, to ensure that the number of the entire sentence samples is similar to that of the positive and negative subsentences, we set the threshold for entire sentence T_2 to 27.5.

A.2 Dataset

MSCOCO (Lin et al., 2014) is a comprehensive dataset, encompassing over 300,000 images across more than 80 categories, each meticulously annotated. Our approach, which leverages text image congruence for alignment, necessitates only the images themselves and their associated prompts, omitting any need for annotations. Following Zhou et al. (2023); Liu et al. (2023a), we randomly select 3,200 images with annotation for validation and testing, ensuring no overlap with the training images to maintain the integrity of our experimental conditions.

A.3 Evaluation Metrics

A.3.1 Metrics on Hallucination Rate

To quantify the rate of hallucinations, we utilize CHAIR (Rohrbach et al., 2018) and MHumanEval (Yu et al., 2023), which allow us to measure hallucinations at both the sentence and instance levels for model-generated content. Additionally, POPE (Fu et al., 2023) is incorporated into our evaluation to directly assess the models via VQA. Details of these metrics are given below.

(1) **CHAIR.** Caption Hallucination Assessment with Image Relevance (CHAIR, Rohrbach et al., 2018) is a widely-used metric for evaluating hallucination. It quantifies hallucination by calculating the ratio of non-existent objects referenced in the model's response to the total number of objects mentioned. It features two variations: CHAIR_S for sentence-level and CHAIR_I for instance-level. Both aim to measure object hallucination, albeit from different perspectives:

$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects}\}|}, \qquad (9)$$

$$\text{CHAIR}_S = \frac{|\{\text{hallucinated responses}\}|}{|\{\text{all responses}\}|}, \quad (10)$$

where hallucinated responses refer to the responses containing at least one hallucinated objects.

(2) **MHumanEval.** Recognizing the limitations of CHAIR in covering only a set of pre-defined object categories, we also incorporate human judgment into our evaluation. Following (Yu et al., 2023), we select a random subset of 100 responses for expert review to identify hallucinated and non-hallucinated objects. Similar to CHAIR, we report hallucination rates at both the object level and the response level, offering a holistic view of the model's accuracy in depicting real-world objects.

(3) **POPE.** Consistent with prior studies (Zhao et al., 2023; Jiang et al., 2023), our evaluation incorporates the Polling-based Object Probing Evaluation (POPE) methodology (Li et al., 2023). POPE leverages an automated segmentation tool to delineate objects within images, subsequently querying the model regarding their presence, as well as introducing random non-existent objects. We present the F1 scores, offering insights into the model's image perception capabilities.

A.3.2 Metrics on Generation Quality

Our evaluation of the generated content's quality by MLLM hinges on three key metrics: informativeness, consistency with human responses, and

fluency. These metrics collectively assess the output's relevance, alignment, and readability.

(1) **Informativeness.** Inspired by (Yu et al., 2023), this metric assesses the extent to which the generated captions encapsulate the primary elements depicted in the image. Utilizing the rich annotations provided by the COCO dataset, we engage GPT-4 (OpenAI, 2023) to compare the annotated objects, the ground-truth caption, and the model-generated caption, subsequently assigning a coverage score. This process ensures that the evaluation focuses on the caption's ability to highlight significant image details.

(2) **Consistency to human response.** The fidelity of model-generated content to human-crafted responses is gauged using the BLEU (Papineni et al., 2002) score, which measures the linguistic similarity between the machine's output and expert-written ground truth captions. This metric serves as an indicator of how well the model's responses align with human expectations and standards.

(3) **Fluency.** The smoothness and natural flow of the text produced by the model are evaluated through its perplexity when processed by a pre-trained GPT-2 (Radford et al., 2019) model. A lower perplexity score signifies higher text fluency, indicating that the generated narrative is coherent and easily comprehensible, mirroring the linguistic quality of the text.

B EFUF is beneficial to other hallucination mitigation methods

EFUF stands out not only for its effectiveness and efficiency in dataset construction and training but also for its compatibility with existing hallucination mitigation strategies, such as RLHF and instruction tuning. This compatibility suggests that MLLMs already enhanced with such techniques can further benefit from the integration of EFUF, potentially leading to additional performance improvements.

To validate this proposition, we conduct incremental experiments, selecting models enhanced with RLHF (LLaVA-RLHF, Sun et al., 2023) and instruction tuning (LRV, Liu et al., 2023a) as our new baseline for comparison. These models are then incrementally trained with EFUF. Results, detailed in Table 6, indicate a notable reduction in hallucination rates post-EFUF application, without compromising the quality of the generated text. This outcome underscores EFUF's value as an additive method, capable of augmenting the performance of MLLMs already subjected to advanced hallucination mitigating techniques.

C Effects of different weight

In this segment, we delve into the effects of varying the weight assigned to the negative loss λ_1 and sentence loss λ_2 on the performance outcomes of ShareGPT4V model when trained using our EFUF strategy. The investigation is aimed at understanding how adjustments in these parameters influence both the reduction in hallucination rates and the overall quality of generated content, with results reported on validation set.

(1) Effects of negative loss weight λ_1 As summarized in Table 7, as λ_1 is incremented from 0.1 to 0.4, we initially note enhancements in both hallucination reduction and generation quality metrics, up until a value of 0.2. Beyond this threshold and past the value of 0.3, a new trend emerges: while the rate of hallucinations continues to decline, a noticeable degradation in generation quality become apparent. This is particularly evident in the metrics assessing informativeness and fluency, with the most pronounced effects observed once λ_1 exceeds 0.4. Our case study further reveals the model's diminishing capacity to construct lengthy, informative sentences at the value of 0.4, suggesting an overly aggressive unlearning weight might inadvertently impair the model's foundational knowledge and capabilities.

Given these findings, a value of 0.3 for λ_1 is identified as the optimal balance point, effectively minimizing hallucinations without compromising the integrity of generation quality.

(2) Effects of sentence loss weight λ_2 Contrastingly, the impact of λ_2 generally mirrors the inverse of λ_1's effects. A value of 0.1 yields reduced fluency, suggesting that such a low sentence loss weight fails to exert sufficient influence. Conversely, elevating λ_2 to 0.3 incites an increase in the hallucination rate. This phenomenon can be attributed to an overly dominant sentence loss weight, which biases the model towards learning entire sentence patterns at the expense of neglecting to unlearn hallucinated content. Consequently, a value of 0.2 for λ_2 is identified as the optimal setting, striking a balance between minimizing hallucinations and maintaining high-quality sentence generation.

Models	Hallucination Rate					Generation Quality				
	Chair$_S$↓	Chair$_I$↓	Human$_S$↓	Human$_I$↓	POPE↑	Bleu1↑	Bleu2↑	Bleu4↑	Info.↑	ppl.↓
LLaVA-RLHF	60.2	24.8	40.0	12.7	87.0	39.8	25.8	12.6	93.5	0.126
+ *EFUF*	59.7	24.7	38.0	12.4	88.8	40.1	26.1	12.9	93.4	0.126
LRV	39.4	19.9	46.0	16.0	85.1	51.8	36.6	20.5	88.4	0.129
+ *EFUF*	37.3	19.5	45.0	15.1	85.1	51.2	36.3	20.7	87.7	0.118

Table 6: Performance comparison of EFUF added on other hallucination mitigating approaches (%).

Parameter		Hallucination Rate					Generation Quality				
		Chair$_S$↓	Chair$_I$↓	Human$_S$↓	Human$_I$↓	POPE↑	Bleu1↑	Bleu2↑	Bleu4↑	Info.↑	ppl.↓
λ_1	0.1	46.3	22.1	30.0	10.2	87.7	43.2	29.2	15.4	89.5	0.155
	0.2	38.5	19.2	20.0	7.3	88.1	44.5	30.2	16.1	91.2	0.129
	0.3	36.9	18.6	18.0	5.2	88.2	47.5	33.1	18.4	90.9	0.154
	0.4	21.0	12.5	13.0	5.9	88.0	63.5	47.0	18.1	88.5	0.243
λ_2	0.1	35.7	17.7	16.0	4.3	88.4	48.6	34.1	17.9	90.6	0.187
	0.2	36.9	18.6	18.0	5.2	88.2	47.5	33.1	18.4	90.9	0.154
	0.3	39.4	19.6	30.0	7.8	87.9	45.9	31.7	16.8	91.0	0.152

Table 7: Performance of EFUF on the ShareGPT4V model with different negative loss weight λ_1 and sentence loss weight λ_2 (validation set).

D Case Study

In this part, we present a comparative analysis through a case study, aiming to elucidate the distinct advantages of our method EFUF. This comparison involves the baseline MiniGPT4 model, a version subjected solely to sentence loss, and the model enhanced with our EFUF strategy.

The case study, as depicted in Figure 5, highlights a scenario where the base MiniGPT4 model erroneously predicts non-existent elements, such as "large windows" and "bookshelves". This error is a clear instance of multimodal hallucination, where the generated content includes objects not present in the input image. The sentence-loss-only approach, while attempting to better align the model with multimodal contexts, falls short of completely correcting these hallucinations. This shortfall is attributed to finetuning's inherent limitation: it lacks a mechanism to explicitly signal to the model which objects are inaccurately generated and thus should be excluded from the output.

In contrast, our EFUF approach successfully addresses this challenge. By integrating a fine-grained unlearning strategy, EFUF effectively discourages the generation of objects with low relevance to the given image. This direct intervention ensures that the model refrains from including hallucinated objects in its outputs, showcasing a significant improvement over the baseline and sentence-loss-only method.

Please describe the image in detail.

MiniGPT4

The image shows a large room with wooden floors and white walls. There are several tables and chairs set up in the room, with people sitting at them working on their laptops. The room has large windows on two sides, allowing natural light to pour in. There are also several bookshelves along the walls, filled with books and other materials. The overall atmosphere of the room is one of productivity and focus, with people working diligently on their tasks.

MiniGPT4 + *sentence loss*

The image shows a large room with wooden tables and chairs. There are several people sitting at the tables, working on their laptops. The walls are painted white and there are windows on one side of the room that let in natural light. The floor is made of hardwood and there are rugs scattered throughout the room. The room is well lit and there are several lamps on the tables. The atmosphere is quiet and focused, with people concentrating on their work.

MiniGPT4 + *EFUF*

The image shows a large room with several tables and chairs. There are several people sitting at the tables, working on their laptops. The walls are made of wood and there are several paintings on them, adding to the ambiance of the room. The floor is made of hardwood and there are several rugs on it, providing a comfortable surface for people to work on. The ceiling is high and the room is well lit. The overall atmosphere is productivity and focus, with people concentrating on their tasks.

Figure 5: Responses of MiniGPT4 with different methods.