RE-ALIGN: Aligning Vision Language Models via Retrieval-Augmented Direct Preference Optimization

Anonymous ACL submission

Abstract

The emergence of large Vision Language Models (VLMs) has broadened the scope and capabilities of single-modal Large Language Models (LLMs) by integrating visual modalities, thereby unlocking transformative cross-modal applications in a variety of real-world scenarios. Despite their impressive performance, VLMs are prone to significant hallucinations, particularly in the form of cross-modal inconsistencies. Building on the success of Reinforcement Learning from Human Feedback (RLHF) in aligning LLMs, recent advancements have focused on applying direct preference optimization (DPO) on carefully curated datasets to mitigate these issues. Yet, such approaches typically introduce preference signals in a brute-force manner, neglecting the crucial role of visual information in the alignment process. In this paper, we introduce RE-ALIGN, a novel alignment framework that leverages image retrieval to construct a dual-preference dataset, effectively incorporating both textual and visual preference signals. We further introduce rDPO, an extension of the standard direct preference optimization that incorporates an additional visual preference objective during fine-tuning. Our experimental results demonstrate that RE-ALIGN not only mitigates hallucinations more effectively than previous methods but also yields significant performance gains in general visual question-answering (VQA) tasks. Moreover, we show that RE-ALIGN maintains robustness and scalability across a wide range of VLM sizes and architectures. This work represents a significant step forward in aligning multimodal LLMs, paving the way for more reliable and effective cross-modal applications.

1 Introduction

The recent emergence of powerful Vision Language Models (VLMs) (Li et al., 2022, 2023a; Liu et al., 2024a; Li et al., 2024b; Meta, 2024; Bai et al., 2023; Wang et al., 2024b; Lu et al., 2024; Wu et al., 2024) has significantly extended the capabilities of Large Language Models (LLMs) (Devlin et al., 2018; Radford et al., 2019; Brown et al., 2020; Team et al., 2023; Roziere et al., 2023; Touvron et al., 2023a,b; Raffel et al., 2020; Yang et al., 2024; Team, 2024) into the visual domain, paving the way for innovative real-world applications that integrate multimodal information (Moor et al., 2023; Li et al., 2024a; Shao et al., 2024; Xing et al., 2024b; Rana et al., 2023; Kim et al., 2024). Despite their promising performance, VLMs remain susceptible to hallucinations—instances where the model produces outputs containing inaccurate or fabricated details about objects, attributes, and the logical relationships inherent in the input image (Rohrbach et al., 2018; Bai et al., 2024). Several factors contribute to this cross-modal inconsistencies, including the separate low-quality or biased training data, imbalanced model architectures, and the disjoint pretraining of the vision encoder and LLM-backbone (Cui et al., 2023; Bai et al., 2024; Zhou et al., 2024a).

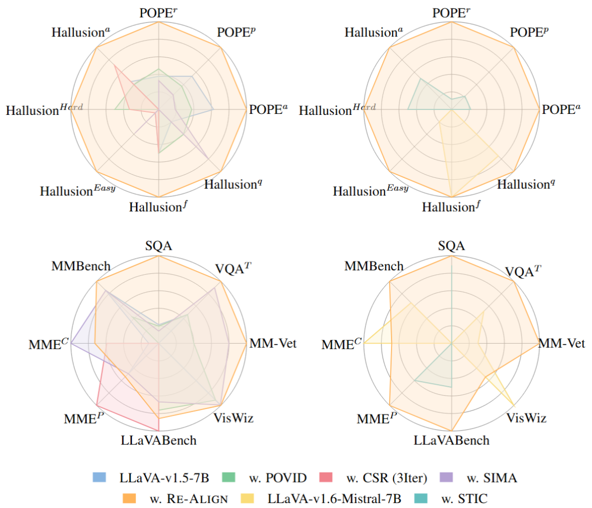

Figure 1: Benchmark performance comparison (min-max normalized).

To mitigate the hallucinations in VLMs, the Directed Preference Optimization (DPO) techniques have been widely adopted (Deng et al., 2024; Zhou et al., 2024a; Fang et al., 2024; Zhou et al., 2024b; Guo et al., 2024; Chen et al., 2024; Wang et al., 2024c; Yu et al., 2024b; Li et al., 2023b; Wang et al., 2024a). This involves constructing datasets enriched with human preference signals specifically targeting hallucinations, and then fine-tuning the models using algorithms like Direct Preference Optimization (DPO) (Rafailov et al., 2024). Existing methods generating the preference data by perturbing the ground truth responses (Zhou et al., 2024a) and corrupting the visual inputs/embeddings (Deng et al., 2024; Amirloo et al., 2024) to generate rejected responses or correcting/refining responses to produce chosen responses (Chen et al., 2024; Yu et al., 2023a). While methods based on response refinement yield the most reliable preference signals, they face scalability challenges due to the significant costs of manual correction processes. Conversely, directly corrupting input visual information or ground truth responses is overly simplistic, as this brute-force approach fails to generate plausible and natural hallucinations in a controlled manner. Moreover, during fine-tuning, directly applying DPO may cause the model to overly prioritize language-specific preferences, which potentially leads to suboptimal performance and an increased propensity for hallucinations (Wang et al., 2024a).

In this paper, we propose **RE-ALIGN**, a novel framework that alleviates VLM hallucinations by integrating image retrieval with direct preference optimization (DPO). Our method deliberately injects controlled hallucinations into chosen responses using image retrieval, generating rejected responses that offer more plausible and natural preference signals regarding hallucinations. Additionally, by incorporating both the retrieved image and the original input image, RE-ALIGN constructs a dual preference dataset. This dataset is then leveraged to finetune VLMs with our proposed **rDPO** objective—an extension of DPO that includes an additional visual preference optimization objective, further enhancing the alignment process with valuable visual preference signals.

2 Preliminaries

To mitigate hallucinations in VLMs, we introduce an alignment framework based on direct preference optimization (DPO) with image retrieval. In this section, we present preliminary definitions and notations for VLMs and preference optimization, which serve as the foundation for our proposed framework.

Vision Language Models VLMs typically consist of three main components: a vision encoder $f_v(\cdot)$, a projector $f_p(\cdot)$, and an LLM backbone $\mathcal{L}(\cdot)$. Given a multimodal input query (x, v), where x is a textual instruction and v is a visual image, VLMs generate a corresponding response $y = [y_1, \cdots, y_m]$ autoregressively. Here, each y_i represents an output token, and m denotes the total number of tokens in the generated response.

Direct Preference Learning Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ziegler et al., 2019) is a key approach for aligning machine learning models with human preferences. Among these techniques, the Direct Preference Optimization (DPO) algorithm (Rafailov et al., 2024) stands out for its popularity and for demonstrating superior alignment performance. We represent a VLM with a policy π, which, given an input query (x, v), generates a response y from the distribution $\pi(\cdot|x, v)$. We denote by π_0 the initial VLM model, fine-tuned on instruction-following VQA data by supervised fine-tuning (SFT). Specifically, we define a preference dataset $\mathcal{D} = \{(x, v, y_w, y_l)\}$, where for each input, the response y_w is preferred to the response y_l. The DPO objective is formulated as follows, leveraging the preference dataset $\mathcal{D}$:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,v,y_w,y_l)\sim\mathcal{D}}$$
$$\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x,v)}{\pi_0(y_w|x,v)} - \beta\log\frac{\pi_\theta(y_l|x,v)}{\pi_0(y_l|x,v)}\right)\right].$$

Compared to deep RL-based methods like Proximal Policy Optimization (PPO) (Schulman et al., 2017; Christiano et al., 2017; Ziegler et al., 2019), DPO is more computationally efficient, easier to tune, and thus more widely adopted (Dong et al., 2024).

Image Retrieval Image retrieval aims to find relevant images from large databases – such as vector databases or indexed corpora – based on semantic similarity criteria. In this paper, we convert all images into vector representations and utilize the cosine similarity metric to evaluate their proximity to a reference image. The similarity between two images, v_1 and v_2, is computed as follows:

$$s = \left\langle \frac{f_p(v_1)}{||f_p(v_1)||}, \frac{f_p(v_2)}{||f_p(v_2)||} \right\rangle,$$

where $< \cdot, \cdot >$ denotes the inner product in l_2 space, $f_p(v_i)$ represents the image embeddings generated by the vision encoder $f_v(\cdot)$ of VLMs. In this paper, we employ the FAISS library (Douze et al., 2024; Johnson et al., 2019) for efficient vector searches, retrieving the top-k most relevant images.

3 Methods

In this paper, we propose RE-ALIGN, a novel framework that integrates preference optimization with image retrieval to improve cross-modal alignment in VLMs.

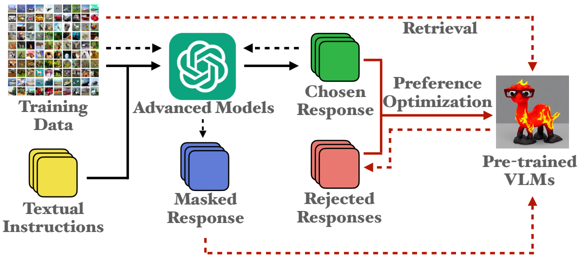

Figure 2: Illustration of RE-ALIGN framework.

As shown in Figure 2, the process begins with an advanced VLM generating chosen responses from input images from the training set. A selective masking process is then applied, strategically omitting segments associated with objects, attributes, or logical relationships identified in the image. Next, leveraging the retrieved image from the same training dataset and the masked responses, the hallucination-prone VLM is prompted to complete the masked elements, obtaining rejected responses. The generated preference pairs (chosen vs. rejected) are then used to fine-tune the VLM with $\mathcal{L}_{\text{rDPO}}$ (eq. (1)), a preference objective that integrates both visual and textual information to penalize hallucinations and reinforce grounded reasoning. Algorithm 1 provides an overview of RE-ALIGN, while the detailed process is explained in the following subsections.

Algorithm 1 Overview of RE-ALIGN

Required:
(1) Unlabeled images $\{v_i\}$ with instructions $\{x_i\}$;
(2) an advanced VLM model $\mathcal{V}$;
(3) caption masking prompt P_m;
(4) masked caption completion prompt P_c;
(5) a text encoder $\mathcal{T}$.
Input: A reference model π_0 with vision encoder $f_v(\cdot)$, VLM π_θ, hyper-parameter k, τ.

$\mathcal{D} \leftarrow \emptyset$ // Init preference dataset
$N \leftarrow |\{v_i\}|$
for $i = 1, \cdots, N$ **do**
  $y_w \leftarrow \mathcal{V}(x_i, v_i)$ // Get preferred response
  $y_m \leftarrow \mathcal{V}(P_m, x_i, v_i)$ // Strategic masking
  $s_i^j = \text{sim}(f_v(v_i), f_v(v_j)), \forall i \neq j$
  // Retrieve top-k similar images
  $s_i^{j1}, \cdots, s_i^{jk} \leftarrow \text{Top}_k(s_i^j)$
  $y_l \leftarrow \text{None}, v_l \leftarrow \text{None}$
  for $t = 1, \cdots, k$ **do**
    // Generate candidate hallucinations
    $y_c \leftarrow \mathcal{V}(P_c, y_m, v_{jt})$
    if $\text{sim}(\mathcal{T}(y_w), \mathcal{T}(y_c)) \geq \tau$ **then**
      // Assign rejected response
      $y_l \leftarrow y_c, v_l \leftarrow v_{j_k}$
  if y_l is None **then**
    continue
  $\mathcal{D} \leftarrow \mathcal{D} \cup \{x_i, v_i, v_l, y_w, y_l\}$
Update π_θ through $\mathcal{L}_{\text{rDPO}}$ (eq. (1))
return π_θ

3.1 Preference Generation

Generating high-quality preference data, which includes both accurate ground-truth responses and controlled hallucinated examples, is crucial for effective preference optimization in pre-trained VLMs. Existing methods construct preference data by perturbing ground-truth responses (Zhou et al., 2024a), corrupting visual inputs/embeddings (Deng et al., 2024; Amirloo et al., 2024) to create rejected responses, or refining responses to obtain chosen responses (Chen et al., 2024; Yu et al., 2023a). Refinement produces high quality preference data but comes at a high cost, whereas direct corruption is more scalable yet tends to generate unrealistic hallucinations and fails to produce plausible, natural ones in a controlled manner. To address these limitations, we introduce a novel image retrieval-based pipeline for preference data construction as shown in Figure 3, which consists of three key stages:

- **Strategical masking:** Given an input pair (x_i, v_i) and its corresponding chosen response y_w generated by a pretrained VLM, a strategic masking process removes words or segments associated with objects, attributes, or logical relationships inferred from the image, producing the masked response y_m.
- **Image retrieval:** All images $\{v_i\}$ in the training set are embedded using the original vision encoder of the pre-trained VLMs, forming the

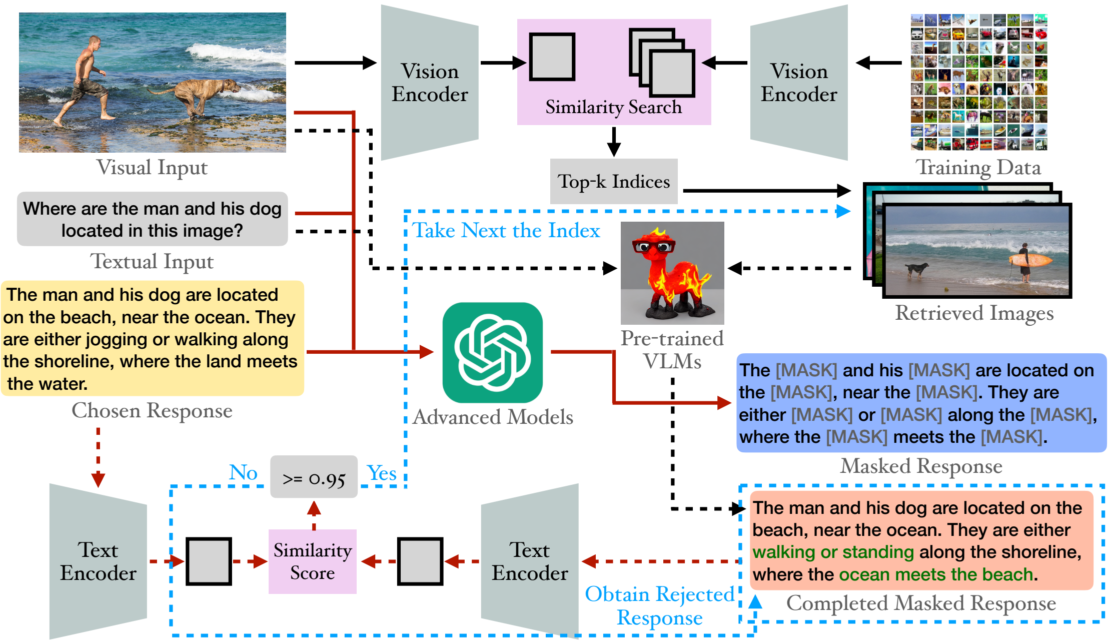

Figure 3: Illustration of the preference generation process, utilizing the original vision encoder from initial VLMs and the SentenceTransformer as the text encoder.

knowledge base $\mathcal{K}$. The top-k most similar images to v_i are then retrieved from $\mathcal{K}$ using a cosine similarity search.

- **Inducing hallucinations:** VLMs are prompted to generate a candidate completion y_m for the masked response conditioned on the instruction x and a retrieved image v_{j_t} where $t \in [1, k]$ denotes the rank of images based on their cosine similarity to the input v_i. Both the chosen response y_w and the reconstructed response y_c are embedded using a SentenceTransformer model. If the cosine similarity between these embeddings falls below 0.95, y_c is designated as the rejected response y_l. Otherwise, the process continues with the next image $v_{j_{t+1}}$ in the similarity-ranked sequence until a suitable candidate is identified or all k retrieved images have been examined.

3.2 Preference Optimization

The curated preference dataset is subsequently used to fine-tune VLMs through direct preference learning. We propose retrieval-augmented direct preference optimization (rDPO), an extension of DPO that integrates an additional visual preference optimization objective. Given a preference dataset $\mathcal{D} = \{x, v, v_l, y_w, y_l\}$, the retrieval-augmented direct preference optimization objective is formulated as follows:

$$\mathcal{L}_{\text{vDPO}} = -\mathbb{E}_{(x,v,v_l,y_w,y_l)\sim\mathcal{D}}$$

$$\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x,v)}{\pi_0(y_w|x,v)} - \beta\log\frac{\pi_\theta(y_w|x,v_l)}{\pi_0(y_w|x,v_l)}\right)\right],$$

where (x, v) denotes the input query of VLMs, (y_w, y_l) represents the preference responses pair, and v_l is the retrieved image for v. The loss function of rDPO is the combination of standard DPO objective and visual preference optimization:

$$\mathcal{L}_{\text{rDPO}} = \mathcal{L}_{\text{DPO}} + \mathcal{L}_{\text{vDPO}}. \tag{1}$$

By incorporating both textual and visual preference signals, our approach allows VLMs to effectively exploit multimodal information during optimization, in contrast to prior alignment methods that depend exclusively on language-based preferences. In contrast to mDPO (Wang et al., 2024a), which introduces image preference by randomly cropping the original input images, rDPO adopts retrieval-augmented generation to integrate visual preference signals in a more coherent and semantically meaningful way.

4 Experiments

We conduct three categories of experiments to empirically validate the effectiveness of our proposed method. First, we evaluate the ability of RE-ALIGN

Methods	POPE^r	POPE^p	POPE^a	Hallusion^q	Hallusion^f	Hallusion^{Easy}	Hallusion^{Hard}	Hallusion^a
LLaVA-v1.5-7B	88.14	87.23	85.10	10.3297	18.2081	41.7582	40.2326	46.3242
w. POVID	88.21	87.16	85.06	10.5495	18.2081	41.5385	40.9302	46.6785
w. CSR (3Iter)	87.83	87.00	85.00	10.1099	18.2081	41.7582	40.6977	46.9442
w. SIMA	88.10	87.10	85.03	10.9890	17.6301	43.0549	40.2326	45.2728
w. RE-ALIGN	**88.65**	**87.43**	**85.16**	**11.2088**	**18.7861**	**45.5165**	**41.6279**	**47.6156**
LLaVA-v1.6-Mistral-7B	88.83	87.93	86.43	13.6264	19.0751	47.4725	33.4884	46.0585
w. STIC	89.03	88.20	86.56	12.9670	17.3410	47.2527	34.1860	46.3242
w. RE-ALIGN	**90.55**	**89.20**	**87.03**	**13.8462**	**19.0751**	**48.3516**	**34.8837**	**46.5899**

Table 1: Impact of RE-ALIGN across hallucination benchmarks for VLMs, and comparisons with baselines.

Methods	SQA	TextVQA	MM-Vet	VisWiz	LLaVABench	MME^P	MME^C	MMBench	Avg. Rank
LLaVA-v1.5-7B	66.02	58.18	31.6	50.03	64.1	1510.28	357.85	64.60	3.375
w. POVID	65.98	58.18	31.8	49.80	67.3	1495.91	356.07	64.34	3.625
w. CSR (3Iter)	65.46	57.86	31.6	47.02	**68.3**	**1525.44**	365.35	64.08	3.625
w. SIMA	65.83	58.48	32.0	50.04	66.9	1510.33	**371.78**	64.60	2.5
w. RE-ALIGN	**68.10**	**58.55**	**32.1**	**50.06**	67.7	1511.79	367.50	**64.69**	**1.375**
LLaVA-v1.6-Mistral-7B	76.02	63.80	47.6	**59.85**	80.2	1494.22	**323.92**	69.33	2.125
w. STIC	76.42	63.50	47.3	54.21	81.0	1504.91	308.21	69.16	2.625
w. RE-ALIGN	**76.47**	**64.08**	**48.3**	57.27	**81.8**	**1512.09**	318.93	**69.42**	**1.25**

Table 2: Impact of RE-ALIGN across general benchmarks for VLMs, and comparisons with baselines.

to mitigate hallucinations and improve generalizability across diverse VQA tasks, demonstrating its consistent superiority over baseline approaches and achieving state-of-the-art performance. Next, we examine RE-ALIGN's effectiveness in aligning VLMs across various model sizes and architectures, including both text-to-image and unified models, where it delivers substantial performance over vanilla models and existing baselines. Finally, we assess the impact of our proposed rDPO objective in preference optimization, showing that it consistently surpasses standard DPO in aligning VLMs and achieving superior results in both hallucination mitigation and general tasks.

4.1 RE-ALIGN for VLMs Alignment

Datasets We conducted experiments on both hallucination detection and general VQA tasks. Specifically, we assess our method's performance in hallucination detection using the POPE dataset (Li et al., 2023c) and HallusionBench (Guan et al., 2023). For general VQA tasks, we leverage a diverse suite of benchmarks including ScienceQA (Lu et al., 2022), TextVQA (Singh et al., 2019), MM-Vet (Yu et al., 2023b), VisWiz (Gurari et al., 2018), LLaVABench (Liu, 2023), MME (Fu et al., 2023), and MMBench (Liu et al., 2024c).

Beslines We compare our method with several widely adopted alignment frameworks for VLMs, including **POVID** (Zhou et al., 2024a), **CSR** (Zhou et al., 2024b), **SIMA** (Wang et al., 2024c), **STIC** (Deng et al., 2024). For more details on these baselines, please refer to the Appendix.

Methods	POPE^r	POPE^p	POPE^a
Janus-Pro-1B	85.46	85.03	84.13
w. RE-ALIGN	$87.53_{\uparrow 2.07}$	$87.33_{\uparrow 2.30}$	$85.86_{\uparrow 1.73}$
Janus-Pro-7B	88.41	87.30	85.70
w. RE-ALIGN	$89.73_{\uparrow 1.32}$	$88.37_{\uparrow 1.07}$	$86.27_{\uparrow 0.57}$
LLaVA-v1.5-7B	88.14	87.23	85.10
w. POVID	$88.21_{\uparrow 0.07}$	$87.16_{\downarrow 0.07}$	$85.06_{\downarrow 0.04}$
w. CSR (3Iter)	$87.83_{\downarrow 0.31}$	$87.00_{\downarrow 0.23}$	$85.00_{\downarrow 0.10}$
w. SIMA	$88.10_{\downarrow 0.04}$	$87.10_{\downarrow 0.13}$	$85.03_{\downarrow 0.07}$
w. RE-ALIGN	$88.65_{\uparrow 0.51}$	$87.43_{\uparrow 0.20}$	$85.16_{\uparrow 0.06}$
LLaVA-v1.5-13B	88.07	87.53	85.60
w. CSR (3Iter)	$88.38_{\uparrow 0.31}$	$87.90_{\uparrow 0.37}$	$85.46_{\downarrow 0.14}$
w. SIMA	$88.04_{\downarrow 0.03}$	$87.40_{\downarrow 0.13}$	$85.40_{\downarrow 0.20}$
w. RE-ALIGN	$90.03_{\uparrow 1.96}$	$89.20_{\uparrow 1.30}$	$86.20_{\uparrow 0.74}$
LLaVA-v1.6-Vicuna-7B	88.52	87.63	86.36
w. RE-ALIGN	$88.94_{\uparrow 0.42}$	$88.03_{\uparrow 0.40}$	$86.63_{\uparrow 0.27}$
LLaVA-v1.6-Vicuna-13B	88.24	87.70	86.43
w. RE-ALIGN	$88.79_{\uparrow 0.55}$	$88.10_{\uparrow 0.40}$	$86.60_{\uparrow 0.17}$

Table 3: Impact of RE-ALIGN across various model scales on POPE.

Experimental Setup We sample 11k images from the LLaVA-Instruct-150K dataset (Liu et al., 2024a) to construct preference data, as illustrated in Figure 3. These images are initially used to generate QA piars based on image captions and simple VQA tasks using GPT-4o mini (OpenAI, 2024). Furthermore, the images are encoded using `clip-vit-large-patch14` (Radford et al., 2021a) to construct the knowledge base for image retrieval. For rejected responses, we use GPT-4o mini to mask the chosen response and `all-mpnet-base-v2` (Reimers and Gurevych, 2019) to compute the similarity between the completed masked response and the original chosen response. We use LLaVA-v1.5-7B (Liu et al., 2024a) and LLaVA-v1.6-Mistral-7B (Li et al., 2024b) as our backbone models and perform RE-ALIGN finetuning for 1 epoch. All evaluations are conducted with a temperature setting of 0, and baseline results are reproduced using the publicly available model weights.

Results Table 1 shows the performance of RE-ALIGN compared to baseline methods on hallucination benchmarks. Notably, RE-ALIGN achieves the best among the evaluated methods on both POPE and HallusionBench for LLaVA-v1.5-7B (Liu et al., 2024a) and LLaVA-v1.6-Mistral-7B (Li et al., 2024b), highlighting the effectiveness of our approach in mitigating hallucinations of VLMs. As demonstrated in Table 2, RE-ALIGN can provide generally on-par or better performance than the vanilla models and baseline alignment methods on each evaluated general VQA task, ultimately achieving the best overall results. This finding indicates that RE-ALIGN can enhance hallucination mitigation without compromising general performance.

4.2 Scalability and Generalizability

Experimental Setup The experimental setup follows the same setting as VLMs alignment experiments, except for the backbone models, where we employ a diverse array of VLMs varying in size and architecture:

- **Image-to-Text models**: the typical architecture of VLMs, where a vision encoder is integrated with an LLM to enable cross-modal understanding. In this section, we evaluate RE-ALIGN on LLaVA-v1.5-7B (Liu et al., 2024a), LLaVA-v1.5-13B (Liu et al., 2024a), LLaVA-v1.6-Vicuna-7B (Li et al., 2024b), and LLaVA-v1.6-Vicuna-13B (Li et al., 2024b).
- **Unified Models**: encoder-decoder architecture that decouples visual encoding for multimodal understanding and generation. In this section, we evaluate RE-ALIGN on Janus-Pro-1B (Chen et al., 2025) and Janus-Pro-7B (Chen et al., 2025).

Results Table 3 presents the performance of RE-ALIGN using both standard image-to-text and unified VLM backbones across model sizes from 1B to 13B on the POPE benchmark (Li et al., 2023c). In experiments with the LLaVA-v1.5 series (Liu et al., 2024a), none of the baseline approaches consistently improve performance for either the 7B or the 13B models, highlighting the limited scalability of these methods. In contrast, RE-ALIGN achieved substantial performance gains, outperforming both the baseline models and the vanilla version—most notably on the LLaVA-v1.5-13B variant. Similarly, experiments with the LLaVA-v1.6-Vicuna series (Li et al., 2024b) revealed the same trend, further underscoring RE-ALIGN's superior scalability. For unified vision-language models, especially Janus-Pro, integrating RE-ALIGN yields a significant performance boost. Notably, Janus-Pro-1B experiences the greatest improvement, underscoring RE-ALIGN's robustness across different model architectures. However, Janus-Pro-1B, being the smallest among the evaluated VLMs, also exhibits the poorest overall performance on POPE, suggesting a correlation between model size and the propensity for hallucinations.

4.3 Effects of rDPO

Dataset Due to budget constraints and the need for reproducibility, we have excluded benchmarks that require evaluation by GPT-4 (Achiam et al., 2023). Instead, we focus on the following tasks: ScienceQA (Lu et al., 2022), TextVQA (Singh et al., 2019), MM-Vet (Yu et al., 2023b), VisWiz (Gurari et al., 2018), LLaVABench (Liu, 2023), MME (Fu et al., 2023), MMBench (Liu et al., 2024c), and POPE (Li et al., 2023c).

Experimental Setup The experimental setup follows the same setting as VLMs alignment experiments, with the exception of the direct optimization objectives. To further explore the impact of our proposed rDPO, we conduct experiments on the same

Methods	SQA	TextVQA	MM-Vet	MME^P	MME^C	MMBench	POPE^r	POPE^p	POPE^a
LLaVA-v1.5-7B	66.02	58.18	31.6	1510.28	357.85	64.60	88.14	87.23	85.10
w. RE-ALIGN (DPO)	66.26	58.24	30.9	1506.49	357.85	64.52	88.18	87.30	85.23
w. RE-ALIGN (rDPO)	68.10	58.55	32.1	1511.79	367.50	64.69	88.65	87.43	85.16
LLaVA-v1.6-Mistral-7B	76.02	63.80	47.6	1494.22	323.92	69.33	88.83	87.93	86.43
w. RE-ALIGN (DPO)	76.07	63.88	46.8	1505.85	316.79	69.24	88.93	88.03	86.47
w. RE-ALIGN (rDPO)	76.47	64.08	48.3	1512.09	318.93	69.42	90.55	89.20	87.03

Table 4: Impact of rDPO across general and hallucination benchmarks for VLMs, and comparisons with baselines.

constructed preference dataset using the standard DPO (Rafailov et al., 2024) during the one-epoch finetuning process.

Results Table 4 summarizes the performance of RE-ALIGN when using both standard DPO and rDPO as the direct optimization objectives, evaluated on general VQA and hallucination tasks with LLaVA-v1.5-7B (Liu et al., 2024a) and LLaVA-v1.6-Mistral-7B (Li et al., 2024b) as backbones. The results indicate that employing rDPO as the finetuning objective consistently yields superior performance over standard DPO across both task categories, highlighting the benefits of incorporating visual preference signals during the alignment process for VLMs. Notably, even when solely employing DPO, RE-ALIGN not only achieves performance gains over the vanilla models but also outperforms the baselines evaluated in the VLM alignment experiments on several tasks. This underscores the effectiveness of our image retrieval-based preference data construction.

5 Discussions

Discussion with mDPO In this section, we detail the differences between our proposed rDPO and mDPO (Wang et al., 2024a). In mDPO, a conditional preference optimization objective is introduced to force the model to determine the preference label based on visual information:

$$\mathcal{L}_{\text{CoDPO}} = -\mathbb{E}_{(x,v,y_w,y_l)\sim\mathcal{D}}$$
$$\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x,v)}{\pi_0(y_w|x,v)} - \beta\log\frac{\pi_\theta(y_w|x,v_c)}{\pi_0(y_w|x,v_c)}\right)\right],$$

where v_c denotes a randomly cropped image of the original input image v. Specifically, visual preference signals are generated by randomly masking 20% of the input visual tokens to encourage the model to capture preferences based on visual cues.

In contrast, RE-ALIGN extends and enhances this approach by incorporating a more semantically meaningful visual preference pair. Instead of relying solely on random crops, RE-ALIGN retrieves a relevant image from the same dataset that corresponds to the original input. This retrieval-based augmentation provides a stronger contrastive signal, improving the model's ability to discern fine-grained visual details and reducing spurious correlations. Moreover, beyond mitigating hallucinations in VLMs, RE-ALIGN has been demonstrated that it also significantly enhances performance on general VQA tasks.

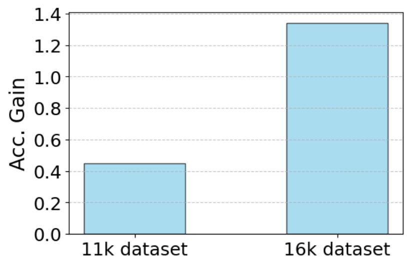

Figure 4: Performance gains of RE-ALIGN with LLaVA-v1.6-Mistral-7B as the backbone on ScienceQA with respect to the size of preference data.

Segment-level Preference Building on the findings of (Yu et al., 2024b), we generate preference data by inducing hallucinations at the segment level than at the sentence level (as seen in approaches such as POVID (Zhou et al., 2024a), STIC (Deng et al., 2024), and CSR (Zhou et al., 2024b)), to provide robust supervision signals during the alignment process. This finer-grained preference modeling yields clearer and more precise learning signals, enabling the model to better distinguish between subtle hallucinations and ground truth responses. To further investigate these segment-level preference signals, we expanded the finetuning dataset from $11k$ to $16k$ image samples. As illustrated in Figure 4, when using LLaVA-v1.6-Mistral-7B as the backbone with ScienceQA as the case study, RE-ALIGN achieved a significant performance improvement—from 0.45 to 1.34—demonstrating the effectiveness of our approach.

6 Related Work

Reinforcement Learning from Human Feedback Reinforcement Learning from Human Feedback (RLHF) has emerged as a crucial technique for incorporating human preference signals into machine learning methods and models (Dong et al., 2024). RLHF frameworks can be broadly categorized into deep RL-based approaches and direct preference learning approaches. In deep RL-based methods, a reward model is first constructed, after which Proximal Policy Optimization (PPO) (Schulman et al., 2017; Christiano et al., 2017; Ziegler et al., 2019) is employed to optimize the reward signals with KL regularization (Ouyang et al., 2022; Touvron et al., 2023b). While the direct preference learning approaches optimize a designed loss target on the offline preference dataset directly, eliminating the need for a separate reward model(Rafailov et al., 2024; Azar et al., 2024; Tang et al., 2024; Ethayarajh et al., 2024).

Vision Language Models Large Vision Language Models (VLMs) (Li et al., 2022, 2023a; Liu et al., 2024a; Li et al., 2024b; Meta, 2024; Bai et al., 2023; Wang et al., 2024b; Lu et al., 2024; Wu et al., 2024) extended the understanding and reasoning capabilities of Large Language Models (LLMs) (Devlin et al., 2018; Radford et al., 2019; Brown et al., 2020; Team et al., 2023; Roziere et al., 2023; Touvron et al., 2023a,b; Raffel et al., 2020; Yang et al., 2024; Team, 2024) into the visual domain. By integrating vision encoders, such as CLIP (Radford et al., 2021b), image patches are first converted into embeddings and then projected to align with text embedding space, unlocking unprecedented cross-modal applications in the real world, such as biomedical imaging (Moor et al., 2023; Li et al., 2024a), autonomous systems (Shao et al., 2024; Tian et al., 2024; Sima et al., 2023; Xing et al., 2024b), and robotics (Rana et al., 2023; Kim et al., 2024).

Alignment of Vision Language Models Current VLMs often suffer from hallucinations, producing inaccurate or misleading information that fails to accurately represent the content of the provided image (Zhu et al., 2024; Bai et al., 2024). Such misalignments can have catastrophic consequences when these models are deployed in real-world scenarios (Xing et al., 2024a). To address cross-modality hallucinations, recent research has primarily focused on applying direct preference optimization (Deng et al., 2024; Zhou et al., 2024a; Fang et al., 2024; Zhou et al., 2024b; Guo et al., 2024; Chen et al., 2024; Wang et al., 2024c; Yu et al., 2024b; Li et al., 2023b; Wang et al., 2024a) or contrastive learning (Sarkar et al., 2024) on the curated datasets with preference signals, and utilizing model editing techniques (Liu et al., 2024b; Yu et al., 2024a).

7 Conclusion

In this paper, a novel framework, RE-ALIGN, for aligning VLMs to mitigate hallucinations is proposed. Our approach leverages image retrieval to deliberately induce segment-level hallucinations, thereby generating plausible and natural preference signals in a controlled manner. By integrating the retrieved images, a dual-preference dataset that encompasses both textual and visual cues is curated. Furthermore, we propose the rDPO objective, an extension of DPO that includes an additional visual preference optimization objective, to enhance the alignment process with valuable visual preference signals. Comprehensive empirical results from a range of general VQA and hallucination benchmarks demonstrate that RE-ALIGN effectively reduces hallucinations in VLMs while enhancing their overall performance. Moreover, it demonstrates superior scalability across various model architectures and sizes.

Limitations

Although RE-ALIGN has demonstrated superior performance on both hallucination and general VQA benchmarks, it does not always achieve state-of-the-art results on general tasks; in some cases, its performance is even worse than that of vanilla VLMs. Future research could explore strategies to eliminate this alignment tax or or identify an optimal balance for this trade-off.

The potential risks of this work align with the general challenges of RLHF alignment. As more powerful alignment techniques are developed, they may inadvertently empower adversarial approaches that exploit these models, potentially leading to unfair or discriminatory outputs. Meanwhile, these adversarial strategies can be used to generate negative samples, which can ultimately contribute to the development of more robust and reliable VLMs over time.

References

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

Elmira Amirloo, Jean-Philippe Fauconnier, Christoph Roesmann, Christian Kerl, Rinu Boney, Yusu Qian, Zirui Wang, Afshin Dehghan, Yinfei Yang, Zhe Gan, et al. 2024. Understanding alignment in multimodal llms: A comprehensive study. *arXiv preprint arXiv:2407.02477*.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2024. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14239–14250.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024. Rlhf workflow: From reward modeling to online rlhf, 2024. *URL https://arxiv. org/abs/2405.07863*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Yunhao Fang, Ligeng Zhu, Yao Lu, Yan Wang, Pavlo Molchanov, Jan Kautz, Jang Hyun Cho, Marco Pavone, Song Han, and Hongxu Yin. 2024. Vila ^{2}: Vila augmented vila. *arXiv preprint arXiv:2407.17453*.

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. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. 2024. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

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

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

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

Haotian Liu. 2023. Llava-bench.

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

Shi Liu, Kecheng Zheng, and Wei Chen. 2024b. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

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

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)*.

Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

OpenAI. 2023. Gpt-4v(ision) system card.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. 2023. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. In *7th Annual Conference on Robot Learning*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. 2024. Mitigating object hallucination via data augmented contrastive tuning. *arXiv preprint arXiv:2405.18654*.

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

Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. 2024. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15120–15130.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. 2023. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Rémi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Ávila Pires, and Bilal Piot. 2024. Generalized preference optimization: A unified approach to offline alignment. *arXiv preprint arXiv:2402.05749*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. 2024. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

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

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Shuo Xing, Hongyuan Hua, Xiangbo Gao, Shenzhe Zhu, Renjie Li, Kexin Tian, Xiaopeng Li, Heng Huang, Tianbao Yang, Zhangyang Wang, Yang Zhou, Huaxiu Yao, and Zhengzhong Tu. 2024a. AutoTrust: Benchmarking Trustworthiness in Large Vision Language Models for Autonomous Driving. *arXiv*.

Shuo Xing, Chengyuan Qian, Yuping Wang, Hongyuan Hua, Kexin Tian, Yang Zhou, and Zhengzhong Tu. 2024b. Openemma: Open-source multimodal model for end-to-end autonomous driving. *arXiv*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin

Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Runpeng Yu, Weihao Yu, and Xinchao Wang. 2024a. Attention prompting on image for large vision-language models. In *European Conference on Computer Vision*, pages 251–268. Springer.

Tianyu Yu, Jinyi Hu, Yuan Yao, Haoye Zhang, Yue Zhao, Chongyi Wang, Shan Wang, Yinxv Pan, Jiao Xue, Dahai Li, et al. 2023a. Reformulating vision-language foundation models and datasets towards universal multimodal assistants. *arXiv preprint arXiv:2310.00653*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024b. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

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

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

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

Tinghui Zhu, Qin Liu, Fei Wang, Zhengzhong Tu, and Muhao Chen. 2024. Unraveling cross-modality knowledge conflicts in large vision-language models. *arXiv preprint arXiv:2410.03659*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Details of the Evaluated Baselines

We compare our proposed method with following alignment frameworks for VLMs:

- **POVID** (Zhou et al., 2024a): constructing preference data by prompting GPT-4V (OpenAI, 2023) to generate hallucinations while intentionally injecting noise into image inputs, followed by fine-tuning VLMs using DPO.
- **CSR** (Zhou et al., 2024b): iteratively generates candidate responses and curates preference data using a self-rewarding mechanism, followed by fine-tuning VLMs via DPO.
- **SIMA** (Wang et al., 2024c): self-generates responses and employs an in-context self-critic mechanism to select response pairs for preference data construction, followed by fine-tuning with DPO.
- **STIC** (Deng et al., 2024): self-generates chosen responses and constructs preference data by introducing corrupted images or misleading prompts, followed by fine-tuning with regularized DPO.

B Prompts used for Preference Data Construction

During the construction of the preference dataset for RE-ALIGN, we employed GPT-4o mini (OpenAI, 2024) to mask the chosen response using the following prompt.

Strategic Masking

Please mask any words of the segments related to the objects, attributes, and logical relationships of the input image in the following description by replacing them with [MASK].

Then, we instruct the VLMs to produce a candidate completion for the masked response to generate the final rejected response using the following prompt.

Masking Completion

Please complete the following sentence based on the input image by filling in the masked segments.

C Examples of Preference Pair

Table 5 and 6 provide examples of the constructed preference data for the VQA and image captioning, and each data sample contains textual instruction, input image, retrieved image, chosen response, and rejected response.

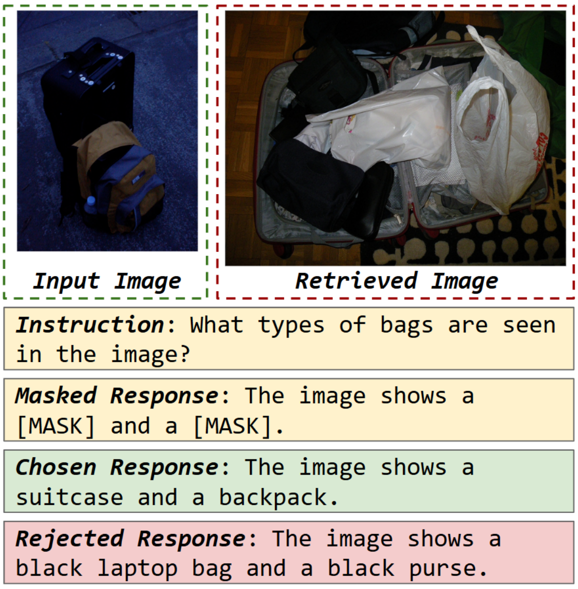

Figure 5: Example preference pair for VQA generated using RE-ALIGN.

D Response Examples

Figure 7 presents example responses from both the original LLaVA-v1.5-7B model and RE-ALIGN as evaluated on LLaVABench. Notably, the original model's response exhibits server object hallucinations, while RE-ALIGN delivers a clearer and more accurate description of the image.

E Licenses

The LLaVA-Instruct-150K dataset (Liu et al., 2024a) which is used to construct preference data is released under CC BY 4.0 license and it should abide by the policy of OpenAI[1].

For the hallucination benchmarks, POPE (Li et al., 2023c) and HallusionBench (Guan et al., 2023) are released under MIT and BSD-3-Clause licenses.

For the general VQA benchmarks, ScienceQA (Lu et al., 2022), TextVQA (Singh et al., 2019), MM-Vet (Yu et al., 2023b), VisWiz (Gurari et al., 2018), LLaVABench (Liu, 2023), and MMBench (Liu et al., 2024c) are released under MIT, CC BY 4.0, Apache-2.0, CC BY 4.0, Apache-2.0, and Apache-2.0 licenses respectively. While MME (Fu et al., 2023) was released without an accompanying license.

[1] https://openai.com/policies/terms-of-use

F Experimental Cost

The cost for curation the preference dataset by using GPT-4o mini (OpenAI, 2024) cost approximately $90 in total.The evaluation of HallusionBench and LLaVABench using GPT-4 (Achiam et al., 2023) incurred an approximate total cost of $30.

G Computational Cost

All finetuning and evaluation experiments were executed on four NVIDIA A6000ada GPUs. Table 5 details the time required for RE-ALIGN to finetune each model.

Models	Required Time
Janus-Pro-1B	50 min
Janus-Pro-7B	93 min
LLaVA-v1.5-7B	35 min
LLaVA-v1.5-13B	45 min
LLaVA-v1.6-Mistral-7B	30 min
LLaVA-v1.6-Vicuna-7B	46 min
LLaVA-v1.6- Vicuna-13B	72 min

Table 5: Time required for finetuning VLMs with RE-ALIGN.

H Hyperparameter Setting

For all the experiments, we finetuning VLMs with RE-ALIGN for 1 epoch. We deploy LoRA finetuning with `lora_r=128`, `lora_alpha=256`, `target_module=all`, and hyperparameters as presented in Table 6.

Hyperparameter	Setting
β	0.1
Learning rate	1e-5
`weight_decay`	0.0
`warmup_ratio`	0.03
`lr_scheduler_type`	`cosine`
`mm_projector_lr`	2e-5
`mm_projector_type`	mlp2x_gelu
`gradient_accumulation_steps`	8
`per_device_train_batch_size`	1
`bf16`	True
Optimizer	AdamW

Table 6: Hypeterparameter setting for finetuning.

I Social Impacts

Our proposed novel alignment framework for VLMs, RE-ALIGN, not only significantly mitigates the hallucinations of VLMs but also ele-

Instruction: Describe this image in detail.

Masked Response: The scene captures a group of [MASK] playing [MASK] on a [MASK]. They are spread across the [MASK] in various [MASK], either preparing for a [MASK] or actively participating in the [MASK]. In particular, a young [MASK] is getting ready to kick the [MASK], which is positioned almost at the center of the image. Additionally, there are a few [MASK] parked along the edge of the [MASK] that can be seen in the background. Furthermore, two other [MASK] are located within the scene, one of them being closer to the right side, and the other appearing further to the top middle part of the image.

Chosen Response: The scene captures a group of young people playing soccer on a field. They are spread across the field in various positions, either preparing for a play or actively participating in the game. In particular, a young girl is getting ready to kick the soccer ball, which is positioned almost at the center of the image. Additionally, there are a few cars parked along the edge of the field that can be seen in the background. Furthermore, two other sports balls are located within the scene, one of them being closer to the right side, and the other appearing further to the top middle part of the image.

Rejected Response: The scene captures a group of young boys playing soccer on a grassy field. They are spread across the field in various positions, either preparing for a kick or actively participating in the game. In particular, a young boy is getting ready to kick the soccer ball, which is positioned almost at the center of the image. Additionally, there are a few chairs parked along the edge of the field that can be seen in the background. Furthermore, two other boys are located within the scene, one of them being closer to the right side, and the other appearing further to the top middle part of the image.

Figure 6: Example preference pair for image captioning generated using RE-ALIGN.

Question: Describe this photo in detail.

LLaVA-v1.5-7B:
The image features a wooden table with a basket of fruit, including a variety of fruits such as apples, oranges, and bananas. There are three apples on the table, with one placed near the left side, another in the middle, and the third one on the right side. In addition to the apples, there are two oranges on the table, one located in the middle and the other towards the right side. A banana can be seen on the left side of the table, and a cluster of grapes is also present, with some grapes placed near the center and others on the right side of the table. The assortment of fruits creates a colorful and appetizing display.

LLaVA-v1.5-7B with Re-Align:
The image features a wooden table with a basket filled with four ripe, dark purple fruits. These fruits are cut open, revealing their juicy interiors. The fruits are placed in various positions on the table, with one on the left side, another in the center, and two more on the right side. The close-up view of the fruits highlights their vibrant colors and delicious appearance.

Figure 7: Example responses generated by LLaVA-v1.5-7B and RE-ALIGN.

vates their generalization capabilities across diverse multimodal tasks. These advancements hold far-reaching societal implications, particularly in advancing the development of trustworthy, ethically aligned AI systems capable of reliable real-world deployment. To elucidate these implications, we provide a comprehensive overview of potential transformative outcomes:

- **Enhancing trustworthiness:** RE-ALIGN significantly enhances the reliability of AI-generated content by reducing hallucinated outputs and improving factual grounding. This ensures that users and regulatory bodies can place increased confidence in AI-driven decisions and recommendations.
- **Safety-critical applications:** By reducing erratic outputs and improving contextual awareness, RE-ALIGN enables safer deployment of VLMs in high-stakes domains such as healthcare diagnostics, autonomous vehicles, and disaster response systems, where error margins are near-zero and algorithmic trust is paramount.
- **Democratizing access to robust AI:** Our method can democratize access to advanced mutimodal AI models under low-resource or data-scarce settings, which empowers researchers and practitioners with limited computational resources to participate in cutting-edge AI development, ultimately contributing to a more equitable and diverse AI ecosystem.