Beyond Human Data: Aligning Multimodal Large Language Models by Iterative Self-Evolution

Wentao Tan[1,2], **Qiong Cao**[2*†], **Yibing Zhan**[2†], **Chao Xue**[2], **Changxing Ding**[1,3]

[1]South China University of Technology
[2]JD Explore Academy, Beijing
[3]Pazhou Lab, Guangzhou
ftwentaotan@mail.scut.edu.cn, {caoqiong1, xuechao19}@jd.com, zybjy@mail.ustc.edu.cn, chxding@scut.edu.cn

Abstract

Human preference alignment can significantly enhance the capabilities of Multimodal Large Language Models (MLLMs). However, collecting high-quality preference data remains costly. One promising solution is the self-evolution strategy, where models are iteratively trained on data they generate. Current multimodal self-evolution techniques, nevertheless, still need human- or GPT-annotated data. Some methods even require extra models or ground truth answers to construct preference data. To overcome these limitations, we propose a novel multimodal self-evolution framework that empowers the model to autonomously generate high-quality questions and answers using only unannotated images. First, in the question generation phase, we implement an image-driven self-questioning mechanism. This approach allows the model to create questions and evaluate their relevance and answerability based on the image content. If a question is deemed irrelevant or unanswerable, the model regenerates it to ensure alignment with the image. This process establishes a solid foundation for subsequent answer generation and optimization. Second, while generating answers, we design an answer self-enhancement technique to boost the discriminative power of answers. We begin by captioning the images and then use the descriptions to enhance the generated answers. Additionally, we utilize corrupted images to generate rejected answers, thereby forming distinct preference pairs for effective optimization. Finally, in the optimization step, we incorporate an image content alignment loss function alongside the Direct Preference Optimization (DPO) loss to mitigate hallucinations. This function maximizes the likelihood of the above generated descriptions in order to constrain the model's attention to the image content. As a result, model can generate more accurate and reliable outputs. Experiments demonstrate that our framework is competitively compared with previous methods that utilize external information, paving the way for more efficient and scalable MLLMs.

Code — https://github.com/WentaoTan/SENA

Introduction

An effective strategy for enhancing MLLM capabilities is human preference alignment (Amirloo et al. 2024; Li et al. 2023b; Zhou et al. 2024a), which focuses on training models to better align with user preferences using high-quality preference data. This approach can improve the model's resistance to hallucinations and its ability to follow complex instructions. However, collecting high-quality preference data is labor-intensive and costly, often requiring extensive manual annotation (Sun et al. 2023; Yu et al. 2024a) or relying on data generated by advanced models like GPT-4v (Achiam et al. 2023; Li et al. 2023b; Zhou et al. 2024a).

To achieve low-cost human preference alignment, self-evolution methods have emerged, enabling models to construct preference data to train themselves (Deng et al. 2024; Wang et al. 2024b; Ahn et al. 2024; Yu et al. 2024b; Zhu et al. 2024). Fig. 1 (a) shows the exact steps of existing self-evolution methods. First, questions annotated by humans or GPT are collected. Then, the evolved model answers. Some methods generate multiple answers and evaluate their quality through a self-rewarding mechanism (Yuan et al. 2024) or by using extra MLLMs (Yu et al. 2024b), CLIP (Rafailov et al. 2024; Zhou et al. 2024b), or ground-truth answers (Wang et al. 2024b). Other methods initially generate chosen answers, then create rejected responses by modifying image content (Zhu et al. 2024) or using misleading prompts (Deng et al. 2024). Finally, a DPO loss function is applied to align with human preference. Despite their promise, current methods often rely on annotated data and additional models, which increases complexity.

To address the above limitations, we explore the potential of going beyond human data and establish a straightforward and efficient multimodal self-evolution framework: **Only a set of unlabeled images is needed for any model to further improve its performance!**

Our framework, depicted in Fig. 1 (b), operates with a single model and unlabeled images, eliminating the need for annotated data by having the model generate both questions and answers. We tackle three significant challenges in this process. The first challenge is generating reliable questions, as meaningless questions lead to useless training data. We introduce an image-driven self-questioning mechanism where the model verifies the answerability of generated questions based on the image content and regenerates them if necessary. This enhances question quality and facilitates effective learning. Additionally, to fully utilize the diverse visual information in the images, we add descriptive ques-

*Project Lead.
†Corresponding Authors.

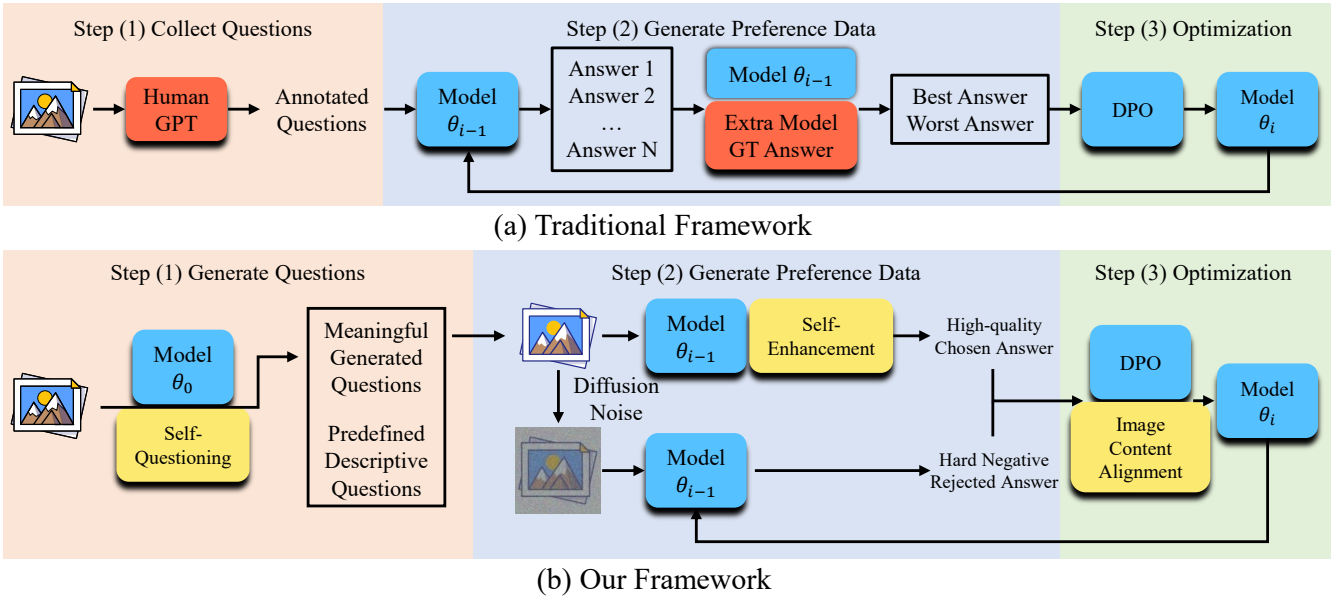


Figure 1: Comparisons between (a) traditional framework and (b) our framework. Our framework combines carefully designed prompt mechanisms and an alignment function, completely eliminating the reliance on annotated data and additional models.

tions (Liu et al. 2024c) that encourage the model to output descriptions of the images. These descriptions will be utilized to address the other two challenges.

The second challenge is generating discriminative answer pairs. Our experiments show that randomly generated answers often exhibit similar quality, rendering them unsuitable for preference alignment. While some methods (Zhu et al. 2024; Deng et al. 2024) use corrupted images to create hard negative answers, simply producing less informative rejected answers is insufficient. We propose enhancing discriminative ability by improving the quality of chosen answers through an answer self-enhancement mechanism. The model generates an initial answer and then refines it using the image description, resulting in a more precise chosen answer. For generating rejected answers, we utilize images augmented with diffusion noise. This approach not only enhances data discriminability and ensures robust preference optimization but also allows the model to verify and correct the chosen answers, ensuring response accuracy.

The third challenge is improving the model's resistance to hallucinations. Since the model uses its generated data for self-training, reducing hallucinations is vital. Research shows that models may generate incorrect answers without referencing the actual image content (Huang et al. 2024; Jiang et al. 2024; Liu et al. 2023a; Zhou et al. 2023; Zhang et al. 2020a,b). Inspired by this, we propose an image content alignment loss function, which maximizes the likelihood of the generated descriptions to shift the model's attention toward actual image content. By combining this with the DPO loss, our approach ensures that as the model learns user preferences, it remains focused on the actual content of the images, resulting in more accurate outputs.

To the best of our knowledge, this is the first multi-modal iterative self-evolution framework that requires no labeled data. We name this framework **SENA** because it effectively integrates image-driven **S**elf-questioning, answer self-**EN**hancement, and image content **A**lignment. In our framework, the model can generate reliable and discriminative preference data, ensuring stable human preference alignment and continuous performance improvement. Experimental results confirm that our approach significantly enhances the model's performance across multiple benchmarks, encompassing both generative and discriminative tasks.

Related Work

Self-Evolution in LLMs. Self-evolution is initially proposed in the realm of LLMs (Calandriello et al. 2024; Rosset et al. 2024; Guo et al. 2024; Swamy et al. 2024; Xiong et al. 2024; Lu et al. 2023), also known as iterative DPO or iterative RLHF (Dong et al. 2024). This approach allows models to align with human preferences using their own-generated data, which significantly reduces annotation costs while improving performance (Yuan et al. 2024). A key challenge in self-evolution is constructing reasonable preference data. One method, self-play (Chen et al. 2024), uses open-source supervised fine-tuning (SFT) data, taking ground truth answers as chosen responses and the model's outputs as rejected ones. This process effectively transforms a weak LLM into a strong one. Another method, self-reward (Yuan et al. 2024), employs a small amount of ranking data to train the model to score its own responses, enabling it to generate reasonable preference data with only 5K labeled examples.

Self-Evolution in MLLMs. This technique is now applied in the MLLM domain (Zhou et al. 2024b; Ahn et al. 2024; Tan et al. 2024, 2023; Wang et al. 2024a, 2022). For

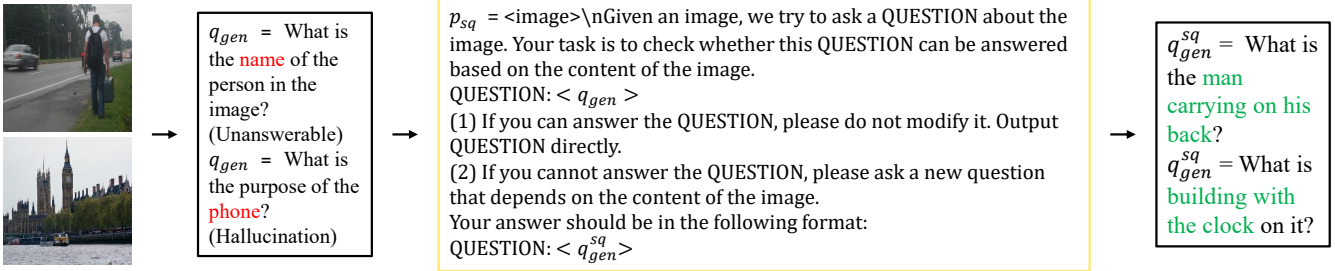


Figure 2: Illustration of the Image-Driven Self-Questioning. SQ checks whether q_{gen} can be answered based on the content of the image. If it cannot, a new question relevant to the image content is generated. The majority of poor-quality questions can be transformed into reliable ones through just one check. Best viewed by zooming in.

example, RLAIF-V (Yu et al. 2024b) assesses responses generated by the 7B LLaVA-1.5 model (Liu et al. 2024a) using the 34B LLaVA-NEXT model (Liu et al. 2024b) to obtain preference data pairs. SIMA (Wang et al. 2024b) classifies responses based on ground truth answers. However, these methods often rely on external models or ground-truth answers, complicating the framework. Other approaches generate chosen answers conventionally and create rejected answers using corrupted images or misleading prompts. For instance, SeVa (Zhu et al. 2024) employs images contaminated with diffusion noise to prompt the model to generate responses that differ from the actual content. Moreover, STIC (Deng et al. 2024) uses misleading prompts to induce hallucinated responses as rejected answers. Current multimodal methods depend heavily on open-source data, which limits their flexibility and scalability. In contrast, our proposed SENA allows the model to autonomously generate open-ended questions, effectively addressing these limitations and filling a significant gap in the MLLMs.

Method

Our framework encourages the model to autonomously generate reliable questions and discriminative answers for unlabeled images, iteratively enhancing its capabilities through human preferences alignment. The overall process is illustrated in Fig. 1 (b). Given an initial model θ_0 and a database of images D, we plan to evolve N times with M images in each iteration. Thus, we randomly select $N \times M$ images from D.

Generate Questions

Traditional methods rely on human- or GPT-annotated questions, allowing the model only to generate answers. Ideally, the model should be capable of generating both questions and answers simultaneously. Therefore, we use the model θ_0 to generate questions for each sampled image x with the prompt p_{base}: *"Please look at the image and generate a question related to the content of the image."*. The generated questions are denoted as $q_{gen} \sim \theta_0(x, p_{base})$.

Image-Driven Self-Questioning. While p_{base} prompts the model to generate questions based on the image content, it sometimes produces nonsensical or irrelevant q_{gen}, as shown in Fig. 2. These flawed questions negatively impact subsequent answers and hinder model optimization. To address this, we introduce an Image-Driven Self-Questioning (SQ) mechanism. The model evaluates whether q_{gen} can be answered based on the image content using the prompt p_{sq}. If it cannot, a new question, $q_{gen}^{sq} \sim \theta_0(x, q_{gen}, p_{sq})$, is generated. This process ensures reliable questions and lays a solid foundation for self-evolution.

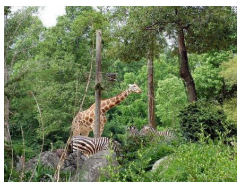
Moreover, we find that some q_{gen}^{sq} tend to focus on the prominent objects in the image. To help the model learn about other visual details, we introduce a descriptive question q_{des}, which is randomly sampled from a set of generic prompts P_{des} (Liu et al. 2024c). The set P_{des} can be found in the Supplementary Materials, with one example being *"Describe the image concisely."*. This addition results in $N \times M$ image-question triplets, represented as $(x, q_{des}, q_{gen}^{sq})$. Once the images and questions are prepared, the model iteratively evolves as described below.

Generate Preference Data

At the start of the i-th iteration ($1 \leq i \leq N$), the model θ_{i-1} generates chosen and rejected answers for the image-question dataset. Some methods utilize θ_{i-1} to randomly generate multiple answers, labeling them as chosen or rejected based on evaluations from self-rewarding mechanisms (Yuan et al. 2024), additional MLLMs (Yu et al. 2024b), or ground-truth answers (Wang et al. 2024b). However, as shown in the Supplementary Materials, these randomly generated answers often have similar quality, making it challenging to create discriminative preference pairs.

A more effective approach involves using original images for generating chosen answers and content-distorted images for rejected answers (Zhu et al. 2024; Leng et al. 2024). For a data point $(x, q_{des}, q_{gen}^{sq})$, the original image x and the question are input into the model to produce the chosen answer y_w. The rejected answer y_l is generated using the noisy image x', obtained by adding T times of diffusion noise to x. Formally, $y_w \sim \theta_{i-1}(x, q)$ and $y_l \sim \theta_{i-1}(x', q)$, where $q \in \{q_{des}, q_{gen}^{sq}\}$. While this method constructs preference data effectively, it may not yield sufficiently discriminative pairs, hindering the learning efficiency. Although more aggressive image corruption techniques could be employed, they may lead to unstable training (Zhu et al. 2024).

Answer Self-Enhancement. To address this issue, we fo-



q_{des} = Describe the image concisely.
y_{des} = A zebra lies on the ground beside a giraffe, close to a tree. (0.3059)
q^{sc}_{gen}= What are the animals in the forest?
y_{gen} = Giraffe. (0.2761)

p_{se} = <image>\nGiven the image, the description of the image, a question about the image and the answer, your task is to improve the quality of the answer.
Image Description: < y_{des} >
Question: < q_{des} > or < q^{sq}_{gen} >
Current Answer: < y_{des} > or < y_{gen} >
Based on the Image Description, please analyze the question further and generate a more precise and concise response.
Your answer should be formatted as:
Improved Answer: < y^{se}_{des} > or < y^{se}_{gen} >

y^{si}_{des} = A giraffe and three zebras are in a lush green forest, the giraffe is eating leaves from a tree. (0.3567)
y^{si}_{gen} = Giraffe and zebras in a forest. (0.3254)

Figure 3: Illustration of the Answer Self-Enhancement techniques. SE analyzes the previous question-and-answer pairs with the help of the image description and enhances the responses. The values in parentheses represent the CLIP scores of the answer-image pairs, which we use to indicate the quality of the answers. Best viewed by zooming in.

cus on enhancing the quality of y_w. Fortunately, the answer to question q_{des} serves as a valuable image description, providing prior knowledge that helps the model better understand questions, analyze existing answers, and ultimately improve answer quality. Therefore, we propose an Answer Self-Enhancement (SE) technique, as detailed in Fig. 3. For clarity, we denote the answers y_w to q_{des} and q^{sq}_{gen} as y_{des} and y_{gen}, respectively. The SE technique uses y_{des} to enhance both answers: $y^{se}_{des} \sim \theta_{i-1}(x, y_{des}, q_{des}, y_{des}, p_{se})$ and $y^{se}_{gen} \sim \theta_{i-1}(x, y_{gen}, q^{sq}_{gen}, y_{des}, p_{se})$. The enhanced answers y^{se}_{des} and y^{se}_{gen} become the new chosen answers y^{se}_{w}, along with y_l to create more discriminative preference data. Although it may seem unusual for SE to enhance y_{des} using itself, this strategy allows the model to reassess the question and generate better answers. Since each image has two questions, we ultimately generate a total of $2M$ samples for subsequent human preference alignment.

Optimization

We utilize the widely adopted Direct Preference Optimization (DPO) loss to align the model θ_{i-1} with human preferences. DPO will first construct a reference model θ_{ref}, which is initialized with the parameters from θ_{i-1} and kept frozen. The goal of DPO is to ensure that as evolution progresses, the current model θ_{i-1} is more likely to generate high-quality answers y_w than θ_{ref}, and less likely to generate rejected answers y_l compared to θ_{ref}. Specifically, given the input image x, question q, and output sequence y, the likelihood $\pi_{\theta_{ref}}(y|x,q)$ is computed as:

$$\pi_{\theta_{ref}}(y|x,q) = \prod_{s=1}^{|y|} P_{\theta_{ref}}(y|x,q,y_{<s}), \tag{1}$$

where $|y|$ represents the token length of y. The DPO loss function is then defined as:

$$L_{DPO} = -\log\sigma\Bigg(\beta\log\frac{\pi_{\theta_{i-1}}(y^{se}_w|x,q)}{\pi_{\theta_{ref}}(y^{se}_w|x,q)} - \beta\log\frac{\pi_{\theta_{i-1}}(y_l|x,q)}{\pi_{\theta_{ref}}(y_l|x,q)}\Bigg), \tag{2}$$

where σ is the sigmoid function, β is a hyperparameter that adjusts the loss sensitivity to preference differences, and $q \in \{q_{des}, q^{sq}_{gen}\}$. Notably, although we use the noisy image x' for generating rejected responses, the likelihood is still calculated based on the original image x.

Image Content Alignment. As the model relies on generated data for training, improving its resistance to hallucinations is essential for ongoing evolution. Research indicates that models may produce incorrect answers without referencing actual image content (Huang et al. 2024). To address this, we design an Image Content Alignment (CA) function to steer the model's focus towards the image content:

$$L_{align} = -\frac{1}{|y^{se}_{des}|}\log\pi_{\theta_{i-1}}(y^{se}_{des}|x,q_{des}). \tag{3}$$

By maximizing the generation probability of the image content y^{se}_{des}, we guide the model's attention to the images, facilitating better understanding and interpretation. As the model's descriptions become more precise, it can produce higher-quality answers through SE, creating a positive feedback loop that continuously improves performance. Now the final optimization function is given by:

$$L_{total} = L_{DPO} + L_{align}. \tag{4}$$

Ultimately, we have completed one iteration and advanced the model from θ_{i-1} to θ_i. SENA will repeat Steps (2) and (3) until $i = N$. The three key designs are complementary; SQ and SE generate high-quality training data that aids CA's learning, while CA enhances the model's generative abilities, enabling SE to function more effectively.

Experiments

Implementation Details

The dataset D is sourced from the LLaVA665k SFT dataset (Liu et al. 2024a), which includes COCO (Lin et al. 2014), GQA (Hudson and Manning 2019), TextVQA (Singh et al. 2019), OCRVQA (Mishra et al. 2019), and Visual Genome (Krishna et al. 2017), totaling 665K images. We conduct three iterations of evolution, setting N = 3. And we plan to use approximately 1% of D per iteration, amounting to M = 6K images. Only images are used without annotation, resulting in a final random sample of 18K images.

We employ LLaVA-1.5-vicuna-7B (Liu et al. 2024a) as the initial model θ_0. All outputs are generated using greedy

Method	Component			Iteration	Generative Task								Discriminative Task		
	SQ	SE	CA		LLaVA$^{\text{W}}$	MM-VET	MMHal		AMBER-Gen.				AMBER-Dis.		MMBench
							Score	Rate↓	CHAIR↓	Cover	Hal↓	Cog↓	Accuracy	F1	
θ_0					59.6	31.7	1.90	0.61	7.6	**51.8**	35.1	4.3	71.7	74.3	64.6
θ_1^{Base}				1	62.8	33.7	1.88	0.62	7.4	51.0	34.4	3.4	70.4	72.4	65.1
θ_1^{SQ}	✓			1	64.9	33.9	2.01	0.59	6.3	50.8	30.4	3.0	71.6	73.9	**65.3**
θ_1^{SE}		✓		1	65.0	34.4	2.08	0.56	6.3	50.7	31.4	2.9	72.2	75.1	65.2
θ_1^{CA}			✓	1	66.7	33.2	2.17	0.56	6.5	51.5	30.8	3.5	74.8	77.9	**65.3**
θ_1	✓	✓	✓	1	66.9	34.2	2.28	0.54	5.6	51.6	25.2	1.9	75.1	79.8	65.0
θ_2	✓	✓	✓	2	**67.5**	35.1	**2.40**	**0.51**	5.3	50.6	23.3	**1.6**	75.7	80.2	65.2
θ_3	✓	✓	✓	3	67.4	**35.8**	2.33	0.52	**4.9**	49.4	**20.5**	1.7	**79.4**	**83.6**	64.8

Table 1: Ablation study on each key component SQ, SE and CA. θ_0 is the LLaVA-1.5-7B model.

Apply SE on		LLaVA$^{\text{W}}$	AMBER		
y_{des}	y_{gen}		CHAIR↓	Accuracy	F1
		62.8	7.4	70.4	72.4
✓		63.9	6.6	71.3	74.1
	✓	63.7	6.8	72.0	74.4
✓	✓	**65.0**	**6.3**	**72.2**	**75.1**

Table 2: Impact of applying Self-Enhancement on different answers.

Apply AL on		LLaVA$^{\text{W}}$	AMBER		
y_{des}	y_{gen}		CHAIR↓	Accuracy	F1
		62.8	7.4	70.4	72.4
	✓	63.3	6.5	73.5	77.1
✓	✓	64.7	**6.4**	**75.0**	**78.2**
✓		**66.7**	6.5	74.8	77.9

Table 3: Impact of applying alignment loss (AL) on different answers.

decoding. Each iteration consists of 1 epoch with a batch size of 128 and a learning rate of 2e-6. The number of diffusion noise additions, T, is set to 600, and the scaling parameter β in DPO is fixed at 0.1.

Evaluation Benchmarks

We evaluate our model's generative and discriminative capabilities. Generative capability assesses the model's ability to produce detailed and accurate content, using benchmarks such as LLaVA$^{\text{W}}$ (Liu et al. 2024c), MM-Vet (Yu et al. 2023), MMHal-Bench (Sun et al. 2023), and AMBER (Wang et al. 2023). Discriminative capability measures the model's ability to distinguish between relevant and irrelevant information, leveraging benchmarks like AMBER (Wang et al. 2023) and MMBench (Liu et al. 2023b). Detailed descriptions of these benchmarks are available in the Supplementary Materials. Some benchmarks require scoring using the GPT-4 API. To ensure fair comparisons, we use the GPT-4-1106-preview version for all tests, as there can be significant performance variability across different API versions. Each test will be conducted three times, with the average score taken as the final result to minimize randomness and ensure reliable evaluation.

Ablation Study

We conduct a comprehensive ablation study using the LLaVA-1.5-7B (θ_0) model. Initially, θ_0 undergoes one evolution round under the baseline framework, resulting in θ_1^{Base}. We then add methods SQ, SE, and CA individually to the baseline, each undergoing one evolution round, creating θ_1^{SQ}, θ_1^{SE}, and θ_1^{CA}, respectively. This allows us to assess each component separately. Lastly, we combine all methods into a complete framework and conduct three evolution rounds starting from θ_0, resulting in models θ_1, θ_2, and θ_3. All results are listed in Table 1.

(a) Baseline Framework: The baseline framework allows the model to self-generate questions and answers without labeled data or extra models. It improves some benchmarks, but θ_1^{Base} struggles with hallucination-related benchmarks like MMHal-bench due to noise in the generated data, indicating a need for content refinement. **(b) Self-Questioning:** The SQ mechanism refines the quality of generated questions by identifying and regenerating meaningless ones. The performance of θ_1^{SQ} shows noticeable improvement over θ_1^{Base}. **(c) Self-Enhancement:** SE enhances the quality of the chosen answers utilizing image description. The model θ_1^{SE} achieves significant progress compared to θ_1^{Base}, while also broadly enhancing the generation and recognition abilities of θ_0. This indicates that SE effectively helps the model learn valuable knowledge from the generated data. **(d) Content Alignment:** CA is designed to enhance the model's focus on image content, greatly improving the model's generative ability and discriminative power. For example, θ_1^{CA} achieves a high score of 66.7 in LLaVA$^{\text{W}}$ and an excellent F1 score of 77.9 in the AMBER-Discriminative task. **(e) Overall Framework Effectiveness:** First, the performance of θ_1 clearly exceeded that of the previous models, indicating that the three key designs are complementary. Furthermore, the performance of the model evolved in each iteration generally improves compared to the previous iteration, and all three models significantly outperform the θ_0 model. This strongly validates the effectiveness of our framework. Notably, all these enhancements are achieved using unlabeled images, highlighting the scalability and great practical value of our approaches.

Impact of SE on Answers. SE uses image descriptions to

Method	Generative Task								Discriminative Task		
	LLaVA^{W}	MM-VET	MMHal		AMBER-Gen.				AMBER-Dis.		MMBench
			Scores	Rate↓	CHAIR↓	Cover	Hal↓	Cog↓	Accuracy	F1	
BLIP-2 (Li et al. 2023a)	38.1	22.4	-	-	-	-	-	-	-	-	-
MiniGPT-4 (Zhu et al. 2023)	-	22.1	-	-	13.6	63.0	65.3	11.3	63.6	64.7	30.9
InstructBLIP-7B (Dai et al. 2023)	60.9	26.2	2.10	0.58	8.8	52.2	38.2	4.4	76.5	81.7	38.4
Shikra-13B (Chen et al. 2023)	-	-	-	-	-	-	-	-	-	-	58.8
Qwen-VL-7B (Bai et al. 2023)	60.9	26.2	-	-	8.8	52.2	38.2	4.4	76.5	81.7	38.4
mPLUG-Owl2 (Ye et al. 2024)	59.9	36.2	-	-	10.6	52.0	39.9	4.5	75.6	78.5	63.5
LLaVA-1.5-7B† (Liu et al. 2024a)	59.6	31.7	1.90	0.61	7.6	51.8	35.1	4.3	71.7	74.3	64.6
with annotated data or extra models:											
+ SeVa† (Zhu et al. 2024)	63.3	**37.0**	2.12	0.57	7.3	**54.0**	37.3	2.9	79.3	**83.6**	**65.6**
+ STIC†† (Deng et al. 2024)	63.0	31.8	2.07	0.56	7.6	52.1	35.8	4.4	71.6	74.2	64.3
+ SIMA† (Wang et al. 2024b)	60.1	32.4	2.11	0.55	6.4	47.4	26.1	3.2	73.4	76.4	65.0
+ RLAIF-V† (Yu et al. 2024b)	62.8	29.2	**2.95**	**0.34**	**2.9**	50.2	**16.0**	**1.0**	54.2	73.7	63.5
+ CSR† (Zhou et al. 2024b)	65.7	32.2	2.07	0.60	3.8	45.0	16.9	1.4	73.1	76.0	64.1
without annotated data or extra models:											
+ SENA (Ours)	**67.4**	35.8	2.33	0.52	4.9	49.4	20.5	1.7	**79.4**	**83.6**	64.8

Table 4: Comparisons with multiple MLLMs and various self-evolution frameworks. † indicates evaluation results based on the models released by the authors, while †† indicates evaluation results based on the code released by the authors.

enhance the quality of all chosen answers. As shown in Table 2, even improving just one type of answer, such as y_{gen} or y_{des}, can boost model performance. This clearly demonstrates the effectiveness of SE. **Impact of CA on Answers.** CA enhances the model's ability to focus on images by maximizing the log-likelihood of descriptive answers, thereby reducing hallucinations. This loss can also be applied to generated answers. As shown in Table 3, applying L_{Align} to generated answers yields positive results as well. However, its performance is optimal when applied to descriptive answers. This is because descriptive answers encompass most of the image content, while generated answers may only focus on specific objects within the image.

Comparison with SOTA

We name the model θ_3 as SENA and compare it with other models. The comparison results are summarized in Table 4.

Existing methods heavily rely on annotations. For instance, SeVa (Zhu et al. 2024) uses instruction data from TextVQA and OCRVQA to improve the model's OCR capabilities. However, SeVa focuses on constructing hard negative rejected answers without addressing potential hallucination issues, resulting in a notable hallucination problem. Additionally, some methods require the ground truth answers to the questions. STIC (Deng et al. 2024) mixes these ground truth answers with generated data for supervised fine-tuning, while SIMA (Wang et al. 2024b) uses them for answer selection, leading to significant annotation costs.

Some methods also necessitate extra models. RLAIF-V (Yu et al. 2024b) shows strong anti-hallucination performance by using the 34B LLaVA-NEXT model to filter responses from the 7B LLaVA-1.5 model. However, this can lead to a preference for shorter, less detailed responses, affecting performance on generative benchmarks like MM-VET. CSR (Zhou et al. 2024b) uses CLIP scores to select answers, with higher-scoring responses being chosen and lower-scoring ones being rejected. This approach improves the model's performance on image captioning tasks. Take the examples shown in Fig. 3, when a query involves multiple objects in an image, a response like "giraffe and zebras" is more precise than just "giraffe," resulting in a higher CLIP score. However, there are instances where CLIP scores can be misleading. For questions targeting a single object, CLIP tends to favor responses that mention multiple objects, which may not align with the actual query. Consequently, CSR might sometimes generate answers that do not match the instructions, negatively affecting its ability to accurately follow instructions on tasks like MM-VET.

In contrast, SENA uses only unlabeled images to generate unique questions, combined with general descriptive questions to provide diverse instructions. It also introduces three technical solutions to improve question generation, answer generation, and preference optimization, thereby enhancing the model's generative and discriminative abilities while maintaining strong resistance to hallucinations.

Qualitative analysis

In this section, we conduct a qualitative analysis of our model's evaluation results on MMHal-Bench (Sun et al. 2023), exploring which specific capabilities of the model have been enhanced during the self-evolution process. The results are presented in Fig. 4.

(a) Generate Accurate and Comprehensive Descriptions. The LLaVA-1.5-7B model often provides detailed descriptions but may lead to hallucinations, such as incorrectly mentioning other people in the image. In contrast, our models deliver accurate descriptions. Notably, as self-evolution progresses, our model increasingly focuses on finer details

Question: Describe the image in detail.

Reference Answer: In this image, I see a book, a pen, and a television on a table. The TV screen shows a man wearing a tie and a suit speaking in a park.
LLaVA-1.5-7B: The image features a man wearing glasses and a suit, standing in front of a television. He appears to be the main focus of the scene, as he is positioned in the center of the image. The television is located on the left side of the frame, with the man standing in front of it. There are a few other people in the background, but they are not the main subject of the image. The man in the suit is the central figure, and the television serves as the main point of interest in the scene.
θ_1 : A man is standing in front of a television, which is displaying a picture of him. He is wearing a suit and tie, and appears to be the main subject of the image.
θ_2 : A man wearing glasses is displayed on a television screen, looking straight ahead. He is wearing a suit and tie, and is standing in front of a tree. The television screen takes up a significant portion of the image, with the man being the main focus.
θ_3 : A man wearing glasses and a tie is displayed on a television screen. He is standing outside, in front of a bridge and a tree. The television screen takes up a significant portion of the image, with the man being the main focus. There is also a pen and a book in front of the television.

(a) Generate accurate and comprehensive descriptions

Question: Describe the person playing tennis in this photo.

Reference Answer: No, there is only a tennis racket and a ball on the ground. No one is playing tennis in this photo.
LLaVA-1.5-7B: The person playing tennis in this photo is a woman who is holding a tennis racket and a tennis ball. She is standing on a tennis court, ready to play.
CSR: The person playing tennis in this photo is a woman who is holding a tennis racket and a tennis ball, preparing to hit the ball during a game of tennis.
SIMA: The person playing tennis in this photo is a woman.
STIC: The person playing tennis in this photo is a woman who is holding a tennis racket and a tennis ball. She is standing on a tennis court, preparing to play or practice her game.
SENA (Ours): There is no person playing tennis in this photo. The image features a tennis racket with a ball on the court.

Question: How much is it per hour to park at the parking meter?
Reference Answer: It is $4 per hour to park at the parking meter.
LLaVA-1.5-7B: It costs $4.25 per hour to park at the parking meter.
CSR: The parking meter charges $4.72 per hour to park.
SIMA: It costs $4.72 per hour to park at the parking meter.
STIC: It costs $7.20 per hour to park at the parking meter.
SENA (Ours): $4.00 per hour.

(c) Accurately recognize text in images

Question: How many mobile phones are there in the image?
Reference Answer: There is one mobile phone in the image.
LLaVA-1.5-7B: There are two mobile phones in the image.
CSR: There are two mobile phones in the image.
SIMA: There are two mobile phones in the image.
STIC: There are two mobile phones in the image.
SENA (Ours): 1

(d) Accurately counting objects

Figure 4: Comparison of outputs from various models on different visual tasks in MMHal-Bench.

in the images. For instance, θ_2 shows greater attention to the trees in the image compared to θ_1. Furthermore, the θ_3 model not only accurately describes the man but also notes a pen and a book in front of the television, showcasing its improved focus on key elements. This enhancement is attributed to our CA loss, which helps the model better attend to the content of the images. **(b) Accurately Answer Misleading Questions.** Some models struggle with misleading questions as they inaccurately describe a woman playing tennis despite the reference indicating no one is playing. Conversely, the SENA model correctly notes the absence of players and mentions a racket and ball, demonstrating better interpretation of misleading questions. **(c) Accurately Recognize Text in Images.** The SENA model excels in text recognition. For example, when asked about the hourly parking fee, LLaVA-1.5-7B states $4.25, while SENA accurately identifies it as $4.00, aligning with the reference answer. **(d) Accurately Counting Objects.** Accurate object counting is crucial for visual understanding. The LLaVA-1.5-7B miscounts two phones in the image, while the SENA correctly identifies a single phone, highlighting its enhanced ability to locate and count objects in images.

Conclusion and Limitations

This paper introduces SENA, a multi-model self-evolution framework that differs significantly from traditional methods as it does not require arbitrary annotations. This framework is supported by three mechanisms: image-driven self-questioning, answer self-enhancement, and an image content alignment function. These mechanisms address key challenges in generating reliable questions, constructing discriminative preferences data, and optimizing the model to reduce hallucinations. Experimental results and qualitative analysis indicate that the SENA model significantly outperforms the baseline model across various tasks, excelling in generating accurate descriptions, answering misleading questions, recognizing text in images, and counting objects.

However, our framework still has limitations. For instance, its performance on certain benchmarks lags behind self-evolution methods that utilize annotated data, and there is a performance plateau after three rounds of evolution. These challenges motivate our ongoing efforts to refine and improve the framework in future work.

Acknowledgements

This work was partially supported by the Major Science and Technology Innovation 2030 “New Generation Artificial Intelligence” key project (No. 2021ZD0111700), the National Natural Science Foundation of China under Grants 62476099 and 62076101, the Guangdong Basic and Applied Basic Research Foundation under Grants 2024B1515020082 and 2023A1515010007, the Guangdong Provincial Key Laboratory of Human Digital Twin under Grant 2022B1212010004, the TCL Young Scholars Program, and the 2024 Tencent AI Lab Rhino-Bird Focused Research Program.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ahn, D.; Choi, Y.; Kim, S.; Yu, Y.; Kang, D.; and Choi, J. 2024. i-SRT: Aligning Large Multimodal Models for Videos by Iterative Self-Retrospective Judgment. *arXiv preprint arXiv:2406.11280*.

Amirloo, E.; Fauconnier, J.-P.; Roesmann, C.; Kerl, C.; Boney, R.; Qian, Y.; Wang, Z.; Dehghan, A.; Yang, Y.; Gan, Z.; et al. 2024. Understanding Alignment in Multimodal LLMs: A Comprehensive Study. *arXiv preprint arXiv:2407.02477*.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Calandriello, D.; Guo, D.; Munos, R.; Rowland, M.; Tang, Y.; Pires, B. A.; Richemond, P. H.; Lan, C. L.; Valko, M.; Liu, T.; et al. 2024. Human alignment of large language models through online preference optimisation. *arXiv preprint arXiv:2403.08635*.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Chen, Z.; Deng, Y.; Yuan, H.; Ji, K.; and Gu, Q. 2024. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Deng, Y.; Lu, P.; Yin, F.; Hu, Z.; Shen, S.; Zou, J.; Chang, K.-W.; and Wang, W. 2024. Enhancing Large Vision Language Models with Self-Training on Image Comprehension. *arXiv preprint arXiv:2405.19716*.

Dong, H.; Xiong, W.; Pang, B.; Wang, H.; Zhao, H.; Zhou, Y.; Jiang, N.; Sahoo, D.; Xiong, C.; and Zhang, T. 2024. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*.

Guo, S.; Zhang, B.; Liu, T.; Liu, T.; Khalman, M.; Llinares, F.; Rame, A.; Mesnard, T.; Zhao, Y.; Piot, B.; et al. 2024. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*.

Jiang, C.; Xu, H.; Dong, M.; Chen, J.; Ye, W.; Yan, M.; Ye, Q.; Zhang, J.; Huang, F.; and Zhang, S. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27036–27046.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13872–13882.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, L.; Xie, Z.; Li, M.; Chen, S.; Wang, P.; Chen, L.; Yang, Y.; Wang, B.; and Kong, L. 2023b. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; and Wang, L. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024c. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2023b. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Lu, J.; Zhong, W.; Huang, W.; Wang, Y.; Mi, F.; Wang, B.; Wang, W.; Shang, L.; and Liu, Q. 2023. Self: Language-driven self-evolution for large language model. *arXiv preprint arXiv:2310.00533*.

Mishra, A.; Shekhar, S.; Singh, A. K.; and Chakraborty, A. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, 947–952. IEEE.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Rosset, C.; Cheng, C.-A.; Mitra, A.; Santacroce, M.; Awadallah, A.; and Xie, T. 2024. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Swamy, G.; Dann, C.; Kidambi, R.; Wu, Z. S.; and Agarwal, A. 2024. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*.

Tan, W.; Ding, C.; Jiang, J.; Wang, F.; Zhan, Y.; and Tao, D. 2024. Harnessing the Power of MLLMs for Transferable Text-to-Image Person ReID. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17127–17137.

Tan, W.; Ding, C.; Wang, P.; Gong, M.; and Jia, K. 2023. Style interleaved learning for generalizable person re-identification. *IEEE Transactions on Multimedia*.

Wang, G.; Ding, C.; Tan, W.; and Tan, M. 2024a. Decoupled Prototype Learning for Reliable Test-Time Adaptation. *arXiv preprint arXiv:2401.08703*.

Wang, J.; Wang, Y.; Xu, G.; Zhang, J.; Gu, Y.; Jia, H.; Yan, M.; Zhang, J.; and Sang, J. 2023. An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation. *arXiv preprint arXiv:2311.07397*.

Wang, P.; Ding, C.; Tan, W.; Gong, M.; Jia, K.; and Tao, D. 2022. Uncertainty-aware clustering for unsupervised domain adaptive object re-identification. *IEEE Transactions on Multimedia*, 25: 2624–2635.

Wang, X.; Chen, J.; Wang, Z.; Zhou, Y.; Zhou, Y.; Yao, H.; Zhou, T.; Goldstein, T.; Bhatia, P.; Huang, F.; et al. 2024b. Enhancing Visual-Language Modality Alignment in Large Vision Language Models via Self-Improvement. *arXiv preprint arXiv:2405.15973*.

Xiong, W.; Dong, H.; Ye, C.; Wang, Z.; Zhong, H.; Ji, H.; Jiang, N.; and Zhang, T. 2024. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*.

Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; and Huang, F. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13040–13051.

Yu, T.; Yao, Y.; Zhang, H.; He, T.; Han, Y.; Cui, G.; Hu, J.; Liu, Z.; Zheng, H.-T.; Sun, M.; et al. 2024a. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13807–13816.

Yu, T.; Zhang, H.; Yao, Y.; Dang, Y.; Chen, D.; Lu, X.; Cui, G.; He, T.; Liu, Z.; Chua, T.-S.; et al. 2024b. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Yuan, W.; Pang, R. Y.; Cho, K.; Sukhbaatar, S.; Xu, J.; and Weston, J. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.-S.; and Sun, Q. 2020a. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33: 655–666.

Zhang, D.; Zhang, H.; Tang, J.; Wang, M.; Hua, X.; and Sun, Q. 2020b. Feature pyramid transformer. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, 323–339. Springer.

Zhou, Y.; Cui, C.; Rafailov, R.; Finn, C.; and Yao, H. 2024a. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*.

Zhou, Y.; Cui, C.; Yoon, J.; Zhang, L.; Deng, Z.; Finn, C.; Bansal, M.; and Yao, H. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Zhou, Y.; Fan, Z.; Cheng, D.; Yang, S.; Chen, Z.; Cui, C.; Wang, X.; Li, Y.; Zhang, L.; and Yao, H. 2024b. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zhu, K.; Zhao, L.; Ge, Z.; and Zhang, X. 2024. Self-Supervised Visual Preference Alignment. *arXiv preprint arXiv:2404.10501*.