FINE-GRAINED VERIFIERS: PREFERENCE MODELING AS NEXT-TOKEN PREDICTION IN VISION-LANGUAGE ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

The recent advancements in large language models (LLMs) and pre-trained vision models have accelerated the development of vision-language large models (VLLMs), enhancing the interaction between visual and linguistic modalities. Despite their notable success across various domains, VLLMs face challenges in modality alignment, which can lead to issues like hallucinations and unsafe content generation. Current alignment techniques often rely on coarse feedback and external datasets, limiting scalability and performance. In this paper, we propose FiSAO (Fine-Grained Self-Alignment Optimization), a novel self-alignment method that utilizes the model's own visual encoder as a fine-grained verifier to improve vision-language alignment without the need for additional data. By leveraging token-level feedback from the vision encoder, FiSAO significantly improves vision-language alignment, even surpassing traditional preference tuning methods that require additional data. Through both theoretical analysis and experimental validation, we demonstrate that FiSAO effectively addresses the misalignment problem in VLLMs, marking the first instance of token-level rewards being applied to such models. Our code is avaliable at `https://anonymous.4open.science/r/FISAO-57F0/`.

1 INTRODUCTION

The advent of large language models (LLMs) (Brown et al., 2020; Touvron et al., 2023; Yang et al., 2024a) and pre-trained vision models (Radford et al., 2021a; Liu et al., 2023c) has propelled vision-language large models (VLLMs) by advancing connections between visual and linguistic modalities through linear projection (Li et al., 2023b) or q-former (Dai et al., 2023b). These VLLMs have demonstrated notable capabilities across diverse domains such as medical applications (Liu et al., 2023b), autonomous driving (Zhou et al., 2023a), and embodied intelligence (Peng et al., 2023). However, challenges remain in precisely aligning vision and language modalities for integrated inference due to their independent pre-training (Jang et al., 2023; Liu et al., 2024a). This pre-training process often results in incompatible modality-specific representations, hindering the formation of a coherent aligned representation space during joint training (Jang et al., 2023). Misalignment between modalities can lead to safety risks such as biased or inappropriate content generation (Gong et al., 2023; Tu et al., 2023) and hallucinations, where outputs are not grounded in visual input (Wang et al., 2023). These risks are particularly concerning in tasks like visual question answering (Cui et al., 2023; Fan et al., 2024), OCR (Shi et al., 2023), and image captioning (Gunjal et al., 2024), where precise alignment is critical.

To address these misalignment issues, recent works have explored strategies such as instruction tuning (Liu et al., 2023a; Chen et al., 2024b), preference tuning (Yu et al., 2023a), and post-processing methods (Zhou et al., 2023b; Yin et al., 2023). However, most prevalent alignment methods rely heavily on external datasets (Zhou et al., 2024a), models (Yin et al., 2023), or costly human annotations (Yu et al., 2023a). Preference tuning, for example, requires extensive manual labeling, either from human experts (Sun et al., 2023; Yu et al., 2023a) or commercial models (Lee et al., 2023; Li et al., 2023b), which imposes significant costs on building reward datasets and limits scalability. Worse still, these alignment methods often rely on coarse feedback, such as sentence-level (Zhou et al., 2024b; Deng et al., 2024) or output-level rewards (Li et al., 2023d), framing the reward modeling task

Table 1: Feature comparison of different preference tuning approaches.

Model Name	Reward Model	Additional Data	GPT-Assisted
Vlfeedback (Li et al., 2023d)	×	✓	✓
Human-Preference (Sun et al., 2023)	✓	✓	×
POVID (Zhou et al., 2024a)	×	✓	✓
FiSAO	×	×	×

as a simple classification problem that scores outputs as desirable or undesirable. Focusing solely on assigning a numerical score for an entire output fails to leverage VLLMs' token-level generation capabilities, limiting their ability to perform detailed reasoning and precise objective identification.

To mitigate the abovementioned limitations, we propose **F**ine-Gra**i**ned **S**elf-**A**lignment **O**ptimization (**FiSAO**), a method for precisely self-aligning modalities in VLLMs using token-level fine-grained feedback from the vision encoder. Our findings indicate that coarse feedback shows a weak correlation with hallucination detection, while fine-grained reward more effectively differentiates between hallucinated and correct outputs (see Section 3.1). In other words, when using hallucination detection as a proxy for alignment measurement, token-level feedback from the vision encoder offers more informative signals for preference tuning compared to coarse scores. Our theoretical analysis further confirms that this fine-grained feedback improves modality alignment (see Section 3.2). Additionally, FiSAO eliminates the need for external annotations or tools by leveraging its vision encoder as a fine-grained verifier, rewarding each generated token based on its alignment with the visual input. As a result, FiSAO effectively harnesses the model's text generation capabilities and demonstrates superior performance compared to preference tuning methods that rely on additional data. We compare FiSAO with other preference tuning approaches in Table 1.

Our primary contributions can be summarized as follows: We first empirically analyze the differences between coarse and fine-grained rewards in addressing misalignment issues, finding that coarse feedback from pre-trained vision encoders, such as sentence-level rewards, shows a weak correlation with hallucination detection, whereas token-level rewards offer more precise signals for modality alignment. Building on these findings, we propose a novel self-training approach, FiSAO, which leverages token-level feedback from the model's own visual encoder, eliminating the need for additional data or external tools. To the best of our knowledge, FiSAO is the first method to introduce token-level rewards for VLLMs. We further demonstrate FiSAO's effectiveness in mitigating misalignment through both empirical results and theoretical analysis.

2 PRELIMINARIES

This section reviews the standard pipeline of preference tuning for VLLMs, as outlined in prior works (Ziegler et al., 2019; Ouyang et al., 2022; Yu et al., 2023a). The process typically consists of three phases: 1) Supervised Fine-Tuning (SFT), 2) Reward Modeling, and 3) Policy Optimization.

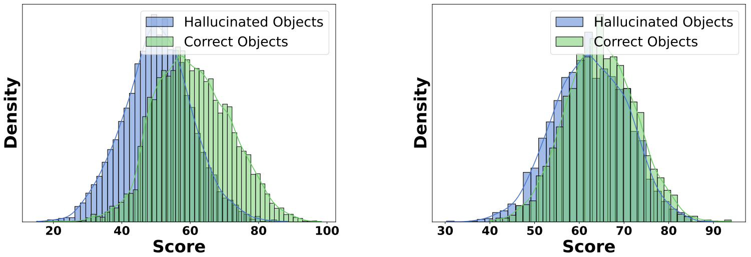
Supervised Fine-Tuning (SFT) Phase. Preference tuning for VLLMs usually begins by jointly training a pre-trained language model and a pre-trained vision encoder on a high-quality instruction dataset (Li et al., 2023b; Dai et al., 2023b), resulting in a SFT model denoted as π_{SFT}.

Reward Modeling Phase. Given text x and visual input v as the prompt, the SFT model π_{SFT} is used to generate a pair of responses $(y_1, y_2) \sim \pi_{\text{SFT}}(y|x, v)$. This pair is then evaluated by humans or AI, with one response marked as preferred y_w and the other as less preferred y_l, denoted as $y_w \succ y_l|x$. This preference is assumed to follow a latent reward model $r^*(y, x, v)$, which is not directly observable. To model this underlying preference, the Bradley-Terry (BT) model is commonly employed to define the preference distribution p^*:

$$p^*(y_w \succ y_l|x) = \frac{\exp(r^*(x, v, y_w))}{\exp(r^*(x, v, y_w)) + \exp(r^*(x, v, y_l))}. \tag{1}$$

Given a static dataset of comparisons $D = \{(x^{(i)}, v^{(i)}, y_w^{(i)}, y_l^{(i)})\}_{i=1}^N$ sampled from p^*, we can parametrize a reward model $r_\phi(x, v, y)$ and estimate its parameters using maximum likelihood estimation. By formulating the estimation of reward model $r_\phi(x, v, y)$ as a binary classification problem, we define the negative log-likelihood loss L_R as follows:

$$L_R(r_\phi, D) = -\mathbb{E}_{(x,v,y_w,y_l)\sim D}[\log \sigma(r_\phi(x, v, y_w) - r_\phi(x, v, y_l))], \tag{2}$$



(a) Distributions of Token-Level Reward (b) Distributions of Sentence-Level Reward

Figure 1: Comparison of token-level (1a) and sentence-level (1b) reward distributions for hallucinated and correct objects in the LLaVA 1.5 model. Further comparisons can be found in Appendix A.2.2.

where σ denotes the logistic function, and reward model $r_\phi(x, v, y)$ is typically initialized from SFT model π_{SFT}, with a linear layer added on top of the final transformer block to produce a scalar output representing the reward prediction (Yu et al., 2023a). Due to the high costs associated with constructing reward model r_ϕ, such as annotation and training, some preference tuning methods employ external models or tools to directly provide rewards (Hessel et al., 2021).

Policy Optimization Phase. The goal of the policy optimization phase is to refine the policy model π_θ using feedback from the reward model r_ϕ, formulated as:

$$\max_{\pi_\theta} \mathbb{E}_{x,v\sim D, y\sim\pi_\theta(y|x,v)}[r_\phi(x, v, y)] - \beta D_{\text{KL}}[\pi_\theta(y|x, v)||\pi_{\text{ref}}(y|x, v)], \tag{3}$$

where β controls the deviation from the reference policy π_{ref} which is initialized as π_{SFT}. This constraint is essential, as it prevents the model from deviating significantly from the original model π_{ref}, maintains generation diversity, and prevents mode collapse to high-reward answers. Due to the discrete nature of language generation, Eqn. 3 is not differentiable. To solve this issue, the standard approach (Ziegler et al., 2019; Ouyang et al., 2022) has been proposed to construct a modified reward function $r(x, v, y) = r_\phi(x, v, y) - \beta(\log \pi_\theta(y|x, v) - \log \pi_{\text{ref}}(y|x, v))$ and then maximize it using Proximal Policy Optimization (PPO) (Schulman et al., 2017).

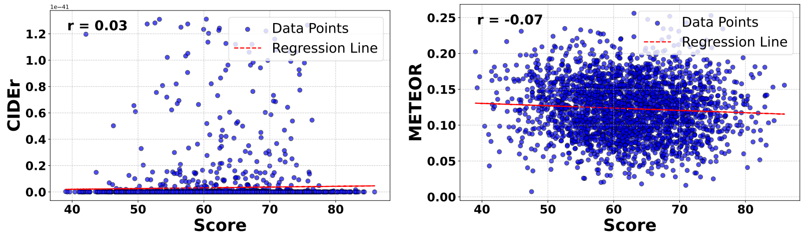
Although the above preference tuning pipeline enhances models with impressive capabilities (Rafailov et al., 2023), it is considerably more complex than supervised learning, incurring significant computational costs. In light of this, recent alignment methods, such as DPO (Rafailov et al., 2023), have been proposed to streamline the process by conducting preference tuning directly on human-preferred responses without the need for a reward model.

3 FiSAO

This section first presents empirical findings (Section 3.1), demonstrating that token-level rewards tend to yield improved alignment in Vision-Language Learning Models (VLLMs) compared to sentence-level rewards. A theoretical justification for the effectiveness of FiSAO is then provided in Section 3.2. Following this, Sections 3.3 and 3.4 detail the two-step preference tuning process employed by FiSAO, consisting of reward modeling and policy optimization. The overall framework of FiSAO is illustrated in Figure 3, while Table 1 compares FiSAO with other preference tuning approaches. Unlike other methods, FiSAO eliminates the need for reward model training, additional data, or high-cost human annotators.

3.1 EMPIRICAL FINDINGS

Hallucinations in VLLMs occur when these models generate content that is not grounded in the input image (Liu et al., 2024a), such as referencing non-existent objects, often indicating weak alignment between the visual and linguistic modalities (Liu et al., 2024a). To investigate vision-language alignment in VLLMs, we examine its relationship to hallucinations. VLLMs commonly extract features using pretrained vision encoders, such as CLIP (Radford et al., 2021a) and Grounding DINO (Liu et al., 2023c). These pretrained vision encoders are trained jointly on vision and language modalities, resulting in a more reliable object recognition (Kuo et al., 2022). Consequently, we



(a) Correlation with CIDEr

(b) Correlation with METEOR

Figure 2: Correlation between the CLIP-based sentence rewards and conventional evaluation metrics: CIDEr (2a) and METEOR (2b). A small Pearson correlation coefficient (r) indicates a weak correlation. More comparison is detailed in Appendix A.2.2.

propose utilizing the vision encoder of the VLLM as a verifier to investigate two distinct types of reward signals: the sentence-level signal, which is commonly employed in prior research (Hessel et al., 2021; Zhou et al., 2024b), and the token-level signal, which has remained largely unexplored.

Token-level rewards are obtained by calculating the similarity between the token-level text embeddings and the image embeddings from the VLLM's vision encoder. Sentence-level rewards, on the other hand, are computed based on the similarity between the embedding of the entire sentence and the image embedding. We conduct two experiments: (1) we plot the distribution of scores across the sentence-level and token-level signals for both hallucinated and correctly identified objects, and (2) we examine the relationship between sentence-level rewards and conventional evaluation metrics for VLLMs, such as CIDEr and METEOR. The scores are obtained by calculating the dot product of the text and image embeddings derived from the pretrained vision encoder within the VLLM. We generate captions for 5,000 images randomly sampled from the COCO training dataset and utilize the widely recognized CHAIR hallucination benchmark (Rohrbach et al., 2018) to identify correctly identified and hallucinated objects. We present our observations as follows:

Token-level rewards differentiate objects better than sentence-level rewards. Figure 1 presents a comparison of score distributions for hallucinated and correct objects generated by LLaVA-1.5 using two types of rewards: token-level and sentence-level. In the token-level reward distribution (Figure 1a), we observe that hallucinated objects are generally associated with lower scores compared to correct objects. In contrast, in the sentence-level reward distribution (Figure 1b), the two distributions largely overlap, with both hallucinated and correct objects peaking around the same score range (60-70). This indicates that, at the sentence level, the reward signal struggles to distinguish between hallucinated and correct objects.

Sentence-level rewards show a weak correlation with conventional metrics. Figure 2 illustrates the relationship between CLIP scores and conventional evaluation metrics BLEU and ROUGE for the generated captions. The scatter plots for BLEU (left) and ROUGE (right) depict the distribution of data points and their corresponding regression lines. From these figures, it is evident that there is a very weak correlation between the scores and both BLEU and ROUGE, with correlation coefficients of $r = -0.01$ for each. Specifically, a high sentence-level score does not necessarily indicate a high-quality sentence. This observation suggests that sentences-level rewards may not be reliable indicators of model performance.

3.2 THEORETICAL FRAMEWORK FOR INCORPORATING FEEDBACK FROM PRE-TRAINED VISION MODELS INTO VLLM TRAINING

In this section, we present a theoretical framework to demonstrate how integrating feedback from pre-trained vision models can enhance the performance of Vision-Language Large Models (VLLMs). We show that under certain assumptions, utilizing vision feedback leads to improved quality of model outputs compared to relying solely on supervised fine-tuning (SFT).

Problem Setup and Notation. We consider a VLLM that processes inputs consisting of an image $v \in \mathbb{R}^{d_v}$ and text $t \in \mathbb{R}^{d_t}$, forming the combined input $x = (v, t)$. Although text data is typically composed of discrete tokens, following prior work (Nakada et al., 2023; Chen et al., 2023), we model these tokens as continuous random vectors for analytical tractability. The data generative model for the image v and text t is defined as:

$$v = U_v z_v + \xi_v, \quad \text{and} \quad t = U_t z_t + \xi_t, \tag{4}$$

where $U_v \in \mathbb{O}^{d_v \times r}$ and $U_t \in \mathbb{O}^{d_t \times r}$ are orthonormal matrices ($U_v^\top U_v = I_r$, $U_t^\top U_t = I_r$) representing decoders that map latent variables to high-dimensional representations. $z_v, z_t \in \mathbb{R}^r$ are latent (low-dimensional) signals capturing the underlying content. $\xi_v \in \mathbb{R}^{d_v}$ and $\xi_t \in \mathbb{R}^{d_t}$ are sub-Gaussian noise vectors with bounded covariance norms.

The ground truth response $y_{\text{truth}} \in \mathbb{R}^{d_t}$ (e.g., the desired textual output) is generated as:

$$y_{\text{truth}} = V_1^* v + V_2^* t + \epsilon_y, \tag{5}$$

where $V_1^* \in \mathbb{R}^{d_t \times d_v}$ and $V_2^* \in \mathbb{R}^{d_t \times d_t}$ are weight matrices, and $\epsilon_y \in \mathbb{R}^{d_t}$ is a noise term. Our goal is to model the conditional distribution $\pi_{\theta_t}(y \mid x)$ of the output y given the input $x = (v, t)$. We assume $\pi_{\theta_t}(y \mid x)$ follows a Gaussian distribution:

$$\pi_{\theta_t}(y \mid x) \propto \exp\left(-\frac{1}{2\sigma^2}\|y - (V_1 v + V_2 t)\|^2\right), \tag{6}$$

where $V_1 \in \mathbb{R}^{d_t \times d_v}$, $V_2 \in \mathbb{R}^{d_t \times d_t}$ are model parameters, and $\sigma > 0$ is the standard deviation.

Assumptions. To facilitate our theoretical analysis, we make several simplifying assumptions:

Linear Transformations: We assume linear relationships between latent variables and observed data as specified in the generative model.

Orthogonal Projection Matrices: The matrices U_v and U_t are orthonormal, projecting latent variables z_v and z_t into high-dimensional spaces.

Sub-Gaussian Noise and Gaussian Likelihood: The noise terms ξ_v and ξ_t are sub-Gaussian. The likelihood for the output y is modeled as Gaussian.

Feedback from Pre-trained Vision Encoders. Building on prior work (Nakada et al., 2023), with abundant image-text pairs, the learned CLIP embeddings converge to: $\mathcal{F}_I(v) \approx U_v^\top v$, and $\mathcal{F}_T(t) \approx U_t^\top t$. These embeddings represent the projections of v and t onto the latent space. We define the feedback from pre-trained vision encoders as: $R_I(y) = \langle U_v^\top v, U_t^\top y \rangle$, which measures the alignment between the image and the generated text in the latent space.

Merged Scores Definition. The supervised fine-tuning (SFT) score is defined as:

$$R_{\text{sft}}(y) = -\|y - y_{\text{truth}}\|^2 = -\|y - (V_1^* v + V_2^* t)\|^2, \tag{7}$$

quantifying the negative squared error between the model's output and the ground truth. We introduce the merged score combining SFT and vision feedback:

$$R(y) = (1 - \lambda) R_{\text{sft}}(y) + \lambda R_I(y), \tag{8}$$

where $\lambda \in [0, 1]$ balances the contributions of SFT and vision feedback. When $\lambda = 0$, only SFT is used; a positive λ incorporates vision feedback.

Theoretical Result. To evaluate the quality of the generated text y, we consider a target variable $z \in \mathbb{R}$ defined as: $z = {\beta^*}^\top y_{\text{truth}}$, with $\beta^* \in \mathbb{R}^{d_t}$ being a fixed vector. The loss function assessing y is: $L(y) = \min_{\beta \in \mathbb{R}^{d_t}} \mathbb{E}\left[(z - \beta^\top y)^2\right]$, which measures the expected squared error in predicting z from y. Our main result is encapsulated in the following theorem:

Theorem 3.1. *Suppose the optimal distribution $\pi^*_{\theta_t}(y \mid x)$ lies within the model class $\{\pi_\theta(y \mid x) : \theta \in \Theta\}$. Then, there exists some $\lambda > 0$ such that:*

$$\mathbb{E}_{\pi_{\theta(\lambda)}(y|x)}[L(y)] < \mathbb{E}_{\pi_{\theta(0)}(y|x)}[L(y)], \tag{9}$$

where $\pi_{\theta(\lambda)}(y \mid x)$ is the model trained with the merged score $R(y)$ using parameter λ, and $\pi_{\theta(0)}(y \mid x)$ is the model trained using only supervised fine-tuning ($\lambda = 0$).

The proof is provided in Appendix A.3.1. This theorem indicates that incorporating feedback from pre-trained vision encoders ($\lambda > 0$) can improve the performance of VLLMs in terms of the loss $L(y)$, compared to relying solely on supervised fine-tuning.

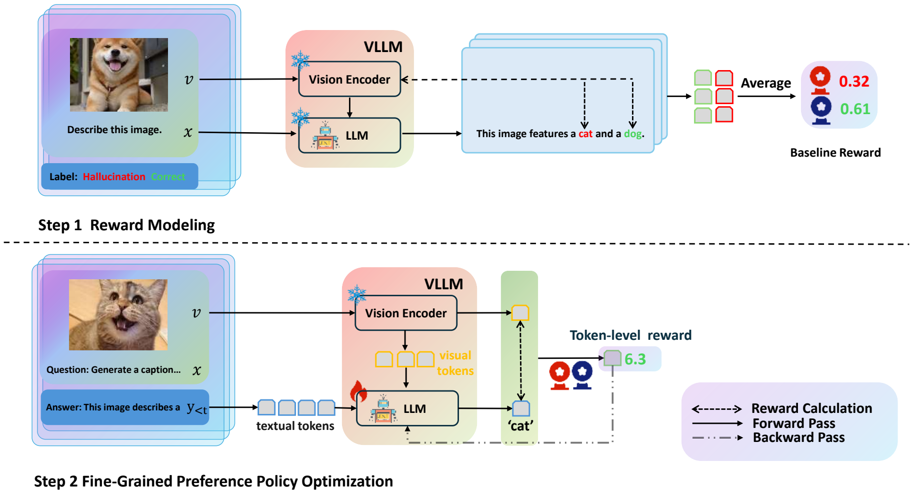


Figure 3: The overall framework of FiSAO. We employ two steps to achieve self-alignment from fine-grained feedback: (1) calculate the fine-grained reward based on the baseline score obtained from correct and hallucinated tokens. (2) optimize the preference policy using this reward to align the model's responses during training.

3.3 REWARD MODELING FOR FISAO

3.3.1 GENERATION FROM THE PERSPECTIVE OF SEQUENTIAL DECISION-MAKING

In this section, we introduce a novel perspective on preference tuning for VLLMs, conceptualizing it as a decision-making process that takes next-token prediction. As discussed in Section 3.1, it is more appropriate to utilize token-level feedback from the fine-grained verifier. Therefore, we consider preference tuning as a decision-making process undertaken by an agent. In this context, after observing the input text and image, a VLLM policy π_θ takes actions by predicting the next token. Here, we consider a standard finite state Markov decision process (MDP) for VLLMs (Puterman, 2014), represented as a tuple $M = (S, A, P, \gamma, R)$. In this context, S is the set of states s, representing the current context or history of generated tokens in the VLLM. The set A denotes the actions a, which correspond to the possible next tokens that the VLLM can generate. The transition probabilities $P \in \Delta(S)_{S\times A}$ indicate the probability of transitioning from one state to another given an action. The discount factor $\gamma \in (0, 1]$ is typically set to 1 in our case, focusing on the undiscounted scenario. Lastly, R is a bounded reward function $R : S \times A \times S \to \mathbb{R}$, providing feedback or reward for the VLLM π_θ taking action a in state s and transitioning to a new state.

Given an appropriate reward function in M, the optimal policy $\pi^*_M \in \Pi$ is the solution to the optimization problem of maximizing the expected discounted total future reward:

$$\max_{\pi\in\Pi} \mathbb{E}_{a_t\sim\pi}\left[\sum_{t=0}^{T} \gamma^t R(s_t, a_t, s_{t+1})\right]. \tag{10}$$

This perspective highlights how fine-grained rewards can be applied to enhance and guide VLLMs, enhancing the vision-langauge alignment in VLLMs.

3.3.2 ESTIMATION OF BASELINE SCORES FOR GROUND TRUTH AND HALLUCINATED DISTRIBUTIONS

To fairly evaluate the model's performance using feedback from the fine-grained verifier, it is crucial to establish a baseline score. In this section, we estimate the baseline reward for the reward calculation process. Assume that the model generates a set of responses $Y = \{y^1, y^2, \ldots, y^s\}$ in response to visual inputs and queries $(x^1, v^1), \ldots, (x^s, v^s)$ from the training dataset. Object tokens of these responses can be divided into two subsets: Y_{gt} and Y_{hal}. Here, Y_{gt} represents the object tokens that are correctly aligned with the corresponding visual input , determined by the ground truth labels, while Y_{hal} consists of the tokens that are identified as hallucinated or misaligned with the corresponding visual input. For each correct object set O^i and hallucinated object set $\tilde{O}^i$ in i-th response, we calculate a score $S(\cdot,\cdot)$ using the dot product between the features of object token and the visual input v^j, derived from the fine-grained verifier. Finally, the average scores for correct objects μ_{gt} and hallucinated objects μ_{hal} are calculated as follows:

$$\mu_{\text{gt}} = \frac{1}{\sum_{i=1}^{s} ||O^i||} \sum_{i=1}^{s} \sum_{o_j \in O^j} S(o_i^j, v^i), \quad \mu_{\text{hal}} = \frac{1}{\sum_{i=1}^{s} ||\tilde{O}^i||} \sum_{i=1}^{s} \sum_{o_j \in \tilde{O}^i} S(o_j^i, v^i), \tag{11}$$

where $||\cdot||$ denotes cardinality of a set. Eqn. 11 can help define the boundary used to calculate the final reward for fine-grained preference policy optimization.

3.3.3 FINE-GRAINED REWARD CALCULATION

In this section, we present the calculation of fine-grained rewards used for preference tuning, aiming to enhance the alignment between the generated tokens and the visual input v. Formally, let the model's response to a query x with visual input v be denoted as $\{y_1, y_2, \ldots, y_T\}$. To effectively select tokens suitable for feedback, we focus on common objects present in the dataset. We construct an entity set C by collecting labels from Detic (Zhou et al., 2022) and COCO (Lin et al., 2015), then expanding it to include similar words and plural forms. Detailed information is provided in Appendix A.1.1. To incorporate feedback from the fine-grained verifier, we calculate negative and positive reward thresholds based on baseline scores of correct and hallucinated responses, as described in Section 3.3.2. Specifically, we use the average scores μ_{gt} and μ_{hal} for correct and hallucinated tokens, respectively, and introduce a margin λ to define thresholds for reward assignment. The fine-grained reward $R = R(s_t, a_t, s_{t+1})_{t=1}^{T}$ for each token is calculated using the following formula:

$$R(s_t, a_t, s_{t+1}) = \begin{cases} \mathcal{N}(S(y_t, v), (\mu_{\text{hal}} - \lambda)) - \xi D_{\text{KL}}[\pi_{\text{ref}}(x, y_{<t}, v) || \pi_\theta(x, y_{<t}, v)], \\ \text{if } y_t \in C \,\&\, S(y_t, v) < \mu_{\text{hal}} - \lambda \\ \mathcal{N}(S(y_t, v), (\mu_{\text{gt}} + \lambda)) - \xi D_{\text{KL}}[\pi_{\text{ref}}(x, y_{<t}, v) || \pi_\theta(x, y_{<t}, v)], \\ \text{if } y_t \in C \,\&\, S(y_t, v) > \mu_{\text{gt}} + \lambda \\ 0, \qquad \text{otherwise} \end{cases} \tag{12}$$

where $S(y_t, v)$ is the dot product score between the embedding of token y_t and the visual input v computed by the pre-trained vision encoder. The function $\mathcal{N}(\cdot,\cdot)$ denotes a normalization function applied to the score based on the threshold. The parameters μ_{gt} and μ_{hal} represent the average scores of correct and hallucinated tokens, respectively, and λ is the margin used to create thresholds for reward assignment. The term ξ is a scaling factor for the Kullback-Leibler (KL) divergence penalty, and $D_{\text{KL}}[\pi_{\text{ref}} \cdot || \cdot \pi_\theta]$ is the KL divergence between the reference policy π_{ref} and the current policy π_θ, promoting consistency with the reference. More details on the calculation and implementation can be found in Appendix A.1.5.

3.4 FINE-GRAINED PREFERENCE POLICY OPTIMIZATION FOR FISAO

Following (Ouyang et al., 2022; Yu et al., 2023a), our approach employs a clipped-PPO method to train the model. This method involves cutting the probability ratios to mitigate large updates, ensuring stable and reliable training. Unlike standard PPO, our approach learns from fine-grained feedback at the token level for each state. By incorporating fine-grained preference signals, FiSAO ensures better vision-language alignment in VLLMs. The objective function is defined as:

$$L(\theta) = \mathbb{E}_{a_t \sim \pi}\left[\sum_{t=1}^{T} \min\left\{r_t(\theta), \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\right\} R(s_t, a_t, s_{t+1})\right], \tag{13}$$

where $r_t(\theta)$ is the probability ratio, R_t is the advantage estimate and ϵ is a hyperparameter that determines the clipping range, and clip$(\cdot)$ is a clipping function that constrains the value of $r_t(\theta)$. The probability ratio $r_t(\theta)$ is calculated as:

$$r_t(\theta) = \frac{\pi_\theta(y_t | x, y_{<t}, v)}{\pi_{\text{ref}}(y_t | x, y_{<t}, v)}, \tag{14}$$

where π_{ref} and π_θ are the policies before and after the update, respectively. We show the detailed process of FiSAO in Algorithm 1.

Algorithm 1 FiSAO

```
Require: Dataset: $\mathcal{D} = \{(x^i, v^i)\}_{i=1}^{N}$; Reference model: $\pi_{\text{ref}}$; Policy model: $\pi_\theta$; PPO training epochs $e$
Ensure: Updated policy model $\pi_\theta$
for each $(x, v) \in \mathcal{D}$ do
    Generate the response from query and image $\{y_1, y_2, \ldots, y_T\} = \pi_\theta(x, v)$
    for each state $y_t$ in $\{y_0, y_1, \ldots, y_T\}$ do
        Compute the score $R(s_t, a_t, s_{t+1})$ from Eqn. 12
    for each epoch in $e$ do
        Calculate probability ratio $r_t(\theta)$ from Eqn. 14
        Update $\pi_\theta$ using Eqn. 13
return $\pi_\theta$
```

Table 2: The performance of FiSAO across all benchmarks is presented. The best result is bolded.

	Comprehensive Benchmark					VQA			Hallucination Benchmark		
Method	MME^{P} ↑	MME^{C} ↑	SEED ↑	MMB↑	MM-Vet ↑	SQA^{I} ↑	POPE↑	GQA↑	Cap_val↑	CHAIR_{S} ↓	CHAIR_{I} ↓
LLaVA-1.5	1510.7	348.2	58.6	64.3	30.5	66.8	85.9	62.0	56.6	54.3	11.3
+ Vlfeedback	1432.7	321.8	59.3	64.0	**31.2**	66.2	83.7	**63.2**	54.8	40.3	13.2
+ Human-Prefer	1490.6	335.0	58.1	63.4	31.1	65.8	81.5	61.3	50.4	38.7	11.3
+ POVID	1452.8	325.3	60.2	**64.9**	31.8	68.8	**86.9**	61.7	57.3	**35.2**	**8.3**
+ FiSAO	**1522.6**	**349.0**	**60.6**	64.8	30.7	**69.3**	85.7	62.0	**61.2**	39.9	9.9
InstructBlip	1237.5	292.1	38.5	36.0	26.0	43.5	**84.8**	48.0	65.5	60.3	11.9
+ Vlfeedback	1241.3	298.9	40.4	**37.7**	26.6	44.6	78.5	47.7	64.0	56.5	9.7
+ Human-Prefer	1250.9	304.2	39.3	37.2	26.6	44.1	79.0	47.5	64.8	51.2	10.8
+ POVID	1255.1	301.8	38.3	37.2	26.3	43.4	84.6	**48.3**	66.5	51.5	10.5
+ FiSAO	**1398.0**	**318.9**	**38.9**	37.4	**26.9**	**46.3**	84.7	48.2	**66.7**	**42.2**	**8.8**

4 EXPERIMENT

In this section, we evaluate FiSAO on the modality alignment of Vision-Language Large Models (VLLMs), showcasing its effectiveness in enhancing models' performance. Our investigation aims to answer the following questions: (1) Does FiSAO enhance the visual understanding capabilities of VLLMs compared to previous approaches? (2) How does the primary component of FiSAO contribute to performance across different benchmarks? (3) Does our method modify the reward distribution of objects in the model's output before and after training?

4.1 EXPERIMENTAL SETUP

Implementation Details. We employ LLaVA-1.5 7B (Liu et al., 2024b) and InstructBLIP (Dai et al., 2023b) as the backbone models. During the preference tuning process, we adapt Low-Rank Adaptation (LoRA) (Hu et al., 2021) fine-tuning. We select the first 8k data from the LLaVA-Instruct 150k dataset (Li et al., 2023b). As both InstructBLIP and LLaVA are trained using the LLaVA-Instruct 150k dataset, no additional data is introduced into our model training. Training is conducted over one epoch, with Proximal Policy Optimization (PPO) being applied for four epochs per sample, utilizing four A100 80GB GPUs. Fine-tuning LLaVA-1.5 7B takes approximately six hours, while fine-tuning InstructBLIP 13B requires around ten hours. For more detailed information on training hyperparameters and training data, please refer to Appendix A.1.5.

Evaluation Benchmarks. We conduct evaluations on three types of benchmarks: comprehensive benchmarks, general VQA benchmarks and COCO benchmarks. Specifically, these include: (1) Comprehensive benchmarks (MME (Fu et al., 2024a), SEEDbench (Li et al., 2023a), MMbench (Liu et al., 2024c), MM-Vet (Yu et al., 2023b)); (2)VQA (ScienceQA (SQA) (Lu et al., 2022), POPE (Li et al., 2023e), GQA (Hudson & Manning, 2019)); (3) Caption benchmark (Bo Li* & Liu, 2024) (Average score of BLEU, ROUGE-L and CIDER), CHAIR (Rohrbach et al., 2019)). The detailed information is in Appendix A.1.3.

Baselines. We compare FiSAO with previous preference tuning approaches, including Silkie (VLFeedback)(Li et al., 2023d), which uses preference distillation via DPO on a multi-modal dataset from 12 LVLMs; LLaVA-RLHF (Human-preference)(Sun et al., 2023), which integrates RLHF to

Table 3: Comparison of FiSAO and other open-sourced state-of-the-art VLLMs.

Method	MME^{P} ↑	MME^{C}↑	SEED ↑	MMB↑	MM-Vet↑	SQA^{I} ↑	GQA ↑
BLIP-2	1293.8	290.0	46.4	38.1	22.4	61.0	41.0
Qwen-VL-Chat	1487.6	360.7	58.2	60.6	47.3	68.2	57.5
mPLUG-Owl2	1450.2	313.2	57.8	64.5	36.2	68.7	56.1
LLaVA-1.5	1510.7	348.2	58.6	64.3	30.5	66.8	62.0
FiSAO (LLaVA-1.5)	1522.6	349.0	60.6	64.8	30.7	69.3	62.0
InstructBlip	1237.5	292.1	38.5	36.0	26.0	43.5	48.0
FiSAO (InstructBlip)	1398.0	318.9	38.9	37.4	26.9	46.3	48.2

align visual faithfulness with human expectations; and POVID (Zhou et al., 2024a), which generates non-preferred responses to optimize VLLMs through DPO (Rafailov et al., 2023). Furthermore, we compare FiSAO with other state-of-the-art open-source VLLMs, including BLIP-2 (Li et al., 2023c), InstructBLIP (Dai et al., 2023a), Qwen-VL-Chat (Bai et al., 2023), mPLUG-Owl2 (Ye et al., 2023b). More details can be seen in Appendix A.1.4.

4.2 EXPERIMENTAL RESULTS ON BENCKMARKS (RQ1)

Comparison with Other Preference Tuning Approaches. As shown in Table 2, our method demonstrates clear advantages over other preference tuning approaches, which often require training reward models or incur high data costs. The superiority of FiSAO lies in its use of fine-grained verifier, which more effectively captures the intrinsic preferences of VLLMs and achieves stronger modality alignment between the pre-trained vison models and LLMs. Additionally, on the LLaVA backbone, FiSAO surpasses existing approaches, delivering an average performance improvement of 8.7%. This underscores FiSAO's effectiveness in leveraging fine-grained token-level rewards to align visual and textual modalities seamlessly.

Comparison with Other Open-Sourced VLLMs. Table 3 compares FiSAO with other state-of-the-art VLLMs. Our method, implemented on the LLaVA-1.5 architecture, achieves competitive results across multiple benchmarks, demonstrating its effectiveness in various tasks such as vision question answering and image captioning. This highlights FiSAO's capability in integrating fine-grained token-level rewards to enhance modality alignment in VLLMs.

4.3 ANALYSIS (RQ2&RQ3)

Ablation Study. Table 4 summarizes the results of the ablation study conducted on FiSAO. Each row represents a different configuration: the presence (✓) or absence (×) of fine-grained rewards and PPO training. When fine-grained rewards are not used regardless of PPO training, performance metrics are notably lower across all benchmarks compared to configurations where fine-grained rewards are employed. Introducing PPO training alone shows an improvement, but the most significant gains are observed when both fine-grained rewards and PPO training are utilized. This combination achieves the highest scores, demonstrating the effectiveness of integrating both strategies in enhancing model performance and alignment across various evaluation tasks. These findings underscore the importance of fine-grained token-level rewards in optimizing VLLMs such as FiSAO for multimodal tasks.

How does Reward Margin Effect Model's Performance? We present how different reward margins impact the model's performance across various benchmarks in Table 5. The table highlights how varying the reward margin λ affects the performance of LLaVA-1.5 + FiSAO across multiple benchmarks. The results indicate notable variations in performance metrics based on the choice of reward margin. Specifically, when the margin is either too small or too large, a decline is observed in metrics such as CHAIR_I, suggesting diminishing returns with extreme reward margins. Although overall performance remains relatively stable, these findings underscore the importance of optimizing the reward margin to balance precision and generalization in FiSAO for enhancing the performance of VLLMs.

How does FiSAO Alter the Reward Distribution of Objects in the Model's Output before and after Training? To better demonstrate how our method enhances vision-language alignment and ensures the generation of high-scoring objects, we visualize the reward distribution of generated objects on the CHAIR benchmark, as depicted in Figure 4.

Table 4: Ablation study results. Each row illustrates a different configuration, indicating the presence (✓) or absence (×) of fine-grained rewards and PPO training. The best results in each column are highlighted in bold.

		Comprehensive Benchmark					VQA			Hallucination Benchmark		
Fine-grained	PPO	$\text{MME}^{\text{P}}\uparrow$	$\text{MME}^{\text{C}}\uparrow$	SEED ↑	MMB ↑	MM-Vet ↑	$\text{SQA}^{\text{I}}\uparrow$	POPE ↑	GQA ↑	Cap_val ↑	$\text{CHAIR}_{\text{S}}\downarrow$	$\text{CHAIR}_{\text{I}}\downarrow$
×	×	1431.9	340.0	59.6	64.0	30.6	67.7	85.7	61.4	54.5	54.0	11.0
×	✓	1509.3	**350.4**	59.5	64.1	30.5	67.5	**85.9**	60.9	56.6	55.3	11.4
✓	✓	**1522.6**	349.0	**60.6**	**64.8**	**30.7**	**69.3**	85.7	**62.0**	**61.2**	**39.9**	**9.9**

Table 5: Performance of FiSAO with varying margins. The best results in each column are highlighted in bold.

	Comprehensive Benchmark					VQA			Hallucination Benchmark		
λ	$\text{MME}^{\text{P}}\uparrow$	$\text{MME}^{\text{C}}\uparrow$	SEED ↑	MMB ↑	MM-Vet ↑	$\text{SQA}^{\text{I}}\uparrow$	POPE ↑	GQA ↑	Cap_val ↑	$\text{CHAIR}_{\text{S}}\downarrow$	$\text{CHAIR}_{\text{I}}\downarrow$
5	1509.3	**350.4**	60.4	64.1	30.6	67.5	84.4	61.7	57.1	53.3	10.8
10	**1522.6**	349.0	**60.6**	**64.8**	**30.7**	**69.3**	**85.7**	**62.0**	**61.2**	**39.9**	**9.9**
20	1501.4	348.6	59.2	64.5	31.0	67.9	85.1	61.6	59.7	56.5	13.6

Before training, VLLMs tend to generate objects with lower scores, indicating that the outputs are not well-aligned with the preferences encoded by the visual encoder. After applying our method, the reward distribution shifts to the right, signifying that the model generates objects with consistently higher rewards. This shift demonstrates improved alignment between the visual and language components of the model, as well as an enhanced ability to produce outputs that better match the encoded preferences. By presenting this change, we aim to highlight the effectiveness of fine-grained feedback in improving both vision-language alignment and the overall performance of VLLMs.

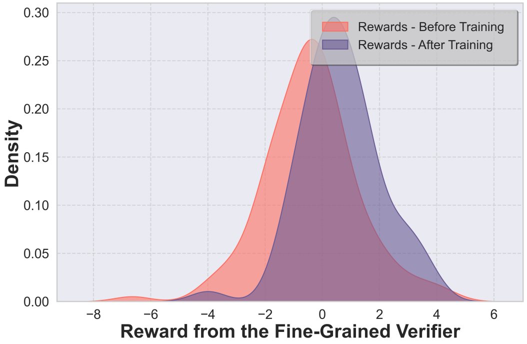
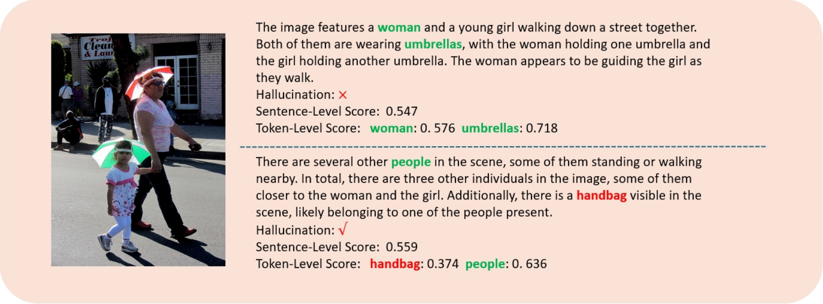


Figure 4: Comparison of reward distributions for generated objects on LLaVA-1.5 before and after Training.

Case Study on Sentence-Level Reward and Token-Level Reward. In this section, we conduct a case study where two sentences from an image are selected for evaluation using both token-level and sentence-level scoring. From Figure 5, we can observe that the sentence-level score is not sensitive to hallucinatory sentences, as it assigns similar scores to both sentences. In contrast, token-level scoring more effectively identifies hallucinatory objects.



The image features a **woman** and a young girl walking down a street together. Both of them are wearing **umbrellas**, with the woman holding one umbrella and the girl holding another umbrella. The woman appears to be guiding the girl as they walk.
Hallucination: ×
Sentence-Level Score: 0.547
Token-Level Score: **woman**: 0. 576 **umbrellas**: 0.718

There are several other **people** in the scene, some of them standing or walking nearby. In total, there are three other individuals in the image, some of them closer to the woman and the girl. Additionally, there is a **handbag** visible in the scene, likely belonging to one of the people present.
Hallucination: √
Sentence-Level Score: 0.559
Token-Level Score: **handbag**: 0.374 **people**: 0. 636

Figure 5: Case study on sentence-level reward and token-level reward.

5 CONCLUSION

In this study, we addressed the alignment issues prevalent in Vision-Language Large Models (VLLMs) by investigating the integration of pre-trained vision encoders with large language models. Through comprehensive analysis, we introduced a novel self-training method using fine-grained Proximal Policy Optimization (PPO) that does not rely on additional data. This method leverages the model's visual encoder as a reward model to enhance alignment at the token level, demonstrating superior performance compared to existing preference tuning approaches.

ETHICS STATEMENT

This paper aims to enhance vision-language alignment for Vision-Language Large Models (VLLMs) and obey the ICLR code of ethics.

REPRODUCIBILITY STATEMENT

All the results in this work are reproducible. We provide detailed settings for our experiments in Table 7. Our code is available at `https://anonymous.4open.science/r/FISAO-57F0/`, and it is shared anonymously.

REFERENCES

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

Kaichen Zhang* Fanyi Pu* Xinrun Du Yuhao Dong Haotian Liu Yuanhan Zhang Ge Zhang Chunyuan Li Bo Li*, Peiyuan Zhang* and Ziwei Liu. Lmms-eval: Accelerating the development of large multimoal models, March 2024. URL `https://github.com/EvolvingLMMs-Lab/lmms-eval`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Beitao Chen, Xinyu Lyu, Lianli Gao, Jingkuan Song, and Heng Tao Shen. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *arXiv preprint arXiv:2405.15356*, 2024a.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 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*, pp. 14239–14250, 2024b.

Zixiang Chen, Yihe Deng, Yuanzhi Li, and Quanquan Gu. Understanding transferable representation learning and zero-shot transfer in clip. *arXiv preprint arXiv:2310.00927*, 2023.

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

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

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

Ailin Deng, Zhirui Chen, and Bryan Hooi. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *arXiv preprint arXiv:2402.15300*, 2024.

Yue Fan, Jing Gu, Kaiwen Zhou, Qianqi Yan, Shan Jiang, Ching-Chen Kuo, Xinze Guan, and Xin Eric Wang. Muffin or chihuahua? challenging large vision-language models with multipanel vqa. *arXiv preprint arXiv:2401.15847*, 2024.

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. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024a.

Deqing Fu, Tong Xiao, Rui Wang, Wang Zhu, Pengchuan Zhang, Guan Pang, Robin Jia, and Lawrence Chen. Tldr: Token-level detective reward model for large vision language models. *arXiv preprint arXiv:2410.04734*, 2024b.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. Vdgd: Mitigating lvlm hallucinations in cognitive prompts by bridging the visual perception gap. *arXiv preprint arXiv:2405.15683*, 2024.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. *arXiv preprint arXiv:2311.05608*, 2023.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *CoRR*, abs/2104.08718, 2021. URL `https://arxiv.org/abs/2104.08718`.

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

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Jiho Jang, Chaerin Kong, Donghyeon Jeon, Seonhoon Kim, and Nojun Kwak. Unifying vision-language representation space with single-tower transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 980–988, 2023.

Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023a.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023c.

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

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pp. 121–137. Springer, 2020.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023a.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024b.

Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *arXiv preprint arXiv:2310.17956*, 2023b.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023c.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024c.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

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

Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. In *International Conference on Artificial Intelligence and Statistics*, pp. 4348–4380. PMLR, 2023.

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

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021a.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021b.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://arxiv.org/abs/2305.18290`.

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

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Yongxin Shi, Dezhi Peng, Wenhui Liao, Zening Lin, Xinhong Chen, Chongyu Liu, Yuyi Zhang, and Lianwen Jin. Exploring ocr capabilities of gpt-4v (ision): A quantitative and in-depth evaluation. *arXiv preprint arXiv:2310.16809*, 2023.

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

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

Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. How many unicorns are in this image? a safety evaluation benchmark for vision llms. *arXiv preprint arXiv:2311.16101*, 2023.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023.

Yuancheng Xu, Udari Madhushani Sehwag, Alec Koppel, Sicheng Zhu, Bang An, Furong Huang, and Sumitra Ganesh. Genarm: Reward guided generation with autoregressive reward model for test-time alignment. *arXiv preprint arXiv:2410.08193*, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.

Kailai Yang, Zhiwei Liu, Qianqian Xie, Jimin Huang, Erxue Min, and Sophia Ananiadou. Selective preference optimization via token-level reward function estimation. *arXiv preprint arXiv:2408.13518*, 2024b.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023a.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, 2023b.

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

Eunseop Yoon, Hee Suk Yoon, SooHwan Eom, Gunsoo Han, Daniel Wontae Nam, Daejin Jo, Kyoung-Woon On, Mark A Hasegawa-Johnson, Sungwoong Kim, and Chang D Yoo. Tlcr: Token-level continuous reward for fine-grained reinforcement learning from human feedback. *arXiv preprint arXiv:2407.16574*, 2024.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*, 2023a.

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

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

Xingcheng Zhou, Mingyu Liu, Bare Luka Zagar, Ekim Yurtsever, and Alois C Knoll. Vision language models in autonomous driving and intelligent transportation systems. *arXiv preprint arXiv:2310.14414*, 2023a.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022.

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

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

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

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

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

A APPENDIX

A.1 EXPERIMENTAL SETTINGS

A.1.1 DETAILS OF ENTITY SET

First, we construct an entity set using the labels from Detic (Zhou et al., 2022) and COCO (Lin et al., 2015). We present the case of these datasets' labels in Table 6 Then, we expand the original set to C by including similar words and plural forms using the `inflect` library and the `wordnet` module from the `nltk` library. The expanded set C contains 5678 words compared to the original set, which contains 1204 words. The `inflect` library is used to generate plural and singular forms of the original labels, while the `wordnet` module from `nltk` is employed to find synonyms. This method allows us to create a comprehensive entity set by considering various linguistic forms, thus enhancing the robustness of our dataset.

A.1.2 OVERVIEW OF THE BACKBONE MODELS

LLaVA-1.5 is a multimodal model designed for general-purpose visual and language understanding. It integrates a vision encoder with the Vicuna language model, making it capable of processing images and generating text-based responses. The model is an open-source chatbot that has been fine-tuned on multimodal instruction-following data generated by GPT. It is built upon the transformer architecture, specifically leveraging the LLaMA/Vicuna foundation models.

InstructBLIP is a sophisticated vision-language model designed to follow detailed instructions. It is built upon the BLIP-2 architecture, incorporating a vision encoder, a language model, and a Query

Original Word	Expanded Words
apple	apples
handbag	bag, handbags, pocketbook, purse
suitcase	grip, suitcases
bagel	bagels
boat	boats, sauceboat, boat
bob	dock, cork, bobs
bread	breads, lucre, lolly, staff of life
cat	purge, chuck, cats
chair	moderate, chairs, chairperson, lead, chairman
duck	ducks, duck, dip, douse
jar	jars, clash, shock
person	someone, person, individual, somebody, people, soul
shirt	shirts
taco	tacos, greaser, wetback, taco
wheel	cycle, wheels, roll

Table 6: Cases of original Words and their expanded forms.

Transformer (Q-Former) that bridges the two components. The Q-Former module is specifically enhanced to handle instruction text tokens, allowing it to extract task-relevant features from images effectively.

A.1.3 DETAILS OF EVALUATION BENCHMARK

- MME (Fu et al., 2024a) is a comprehensive benchmark for evaluating the performance of LVLMs in multimodal tasks. It measures models' capabilities across two key areas: perception and cognition, using 14 specially designed subtasks that test interpretative and analytical skills.
- SEED-Bench (Li et al., 2023a) focuses on evaluating the generative comprehension abilities of LVLMs. It includes a dataset of 19K multiple-choice questions with detailed human annotations, spanning 12 evaluation dimensions that cover both spatial and temporal understanding in image and video modalities.
- MMBench (Liu et al., 2024c) employs a dual approach: it provides an extensive dataset that broadens the range and variety of evaluation questions, and introduces the innovative CircularEval strategy, which uses ChatGPT to convert free-form predictions into structured choices.
- MM-Vet (Yu et al., 2023b) is a benchmark created to evaluate the diverse competencies of LVLMs. It organizes complex multimodal tasks into 16 unique integrations based on six core vision-language capabilities, offering a detailed analysis of model performance across various question types and answer styles.
- ScienceQA (Lu et al., 2022) is a multimodal benchmark aimed at assessing and diagnosing AI systems' multi-hop reasoning and interpretability in the science domain. It includes a dataset of around 21K multiple-choice questions across various scientific topics, complete with detailed answer annotations, related lectures, and explanations.
- GQA (Hudson & Manning, 2019) is a dataset designed for advanced visual reasoning in real-world scenarios, using scene graph-based structures to generate 22 million diverse, semantically-programmed questions. It features a novel set of evaluation metrics focused on consistency, grounding, and plausibility, setting a high standard for vision-language task assessment.
- POPE (Li et al., 2023e) is an evaluation method for examining object hallucination in LVLMs. It transforms the evaluation into a binary classification task, asking LVLMs simple Yes-or-No questions to identify hallucinated objects. POPE employs various object sampling strategies to reveal model tendencies towards hallucination.
- The COCO-caption benchmark assesses image captioning models using BLEU, ROUGE, and CIDEr scores, providing a comprehensive measure of caption quality. We calculate the average

of these scores and multiply by 100 to obtain the final score. This benchmark utilizes the COCO dataset, emphasizing the accuracy and relevance of generated captions. Detailed evaluation methodology and task specifics can be found in the `lmms_eval` repository, specifically under the `tasks/coco2017_cap_val` directory.[1]

- CHAIR (Rohrbach et al., 2019) is a well-known tool for evaluating object hallucination in image captioning tasks. It includes two variants: CHAIRI and CHAIRS, which assess object hallucination at the instance and sentence levels, respectively. Specifically, we randomly sampled 500 images from the COCO (Lin et al., 2015) validation set and evaluated object hallucination using the CHAIR metric.

A.1.4 DETAILS OF BASELINES

- Silkie (Vlfeed- back) (Li et al., 2023d) focuses on improving large vision language models (LVLMs) by using preference distillation. The authors created a vision-language feedback (VLFeedback) dataset, consisting of multi-modal instructions and responses generated by 12 different LVLMs. The model pool includes prominent models like GPT-4V and LLaVA-series. By applying direct preference optimization (DPO) on this dataset, they developed the Silkie model, which shows significant improvements in perception and cognition capabilities.
- LLaVA-RLHF (Human-preference) (Sun et al., 2023) explores the integration of reinforcement learning with human feedback (RLHF) to enhance vision-language models. The LLaVA series, built on Vicuna models and fine-tuned with GPT-4 generated multi-modal data, is further improved by aligning visual faithfulness and human preferences. This approach aims to ensure that the generated responses are more aligned with human expectations and the visual content they describe, providing a more reliable and contextually accurate output
- POVID (Zhou et al., 2024a) is a framework for generating non-preferred responses in Vision-Language Large Models (VLLMs) aimed at preference optimization. The framework employs two strategies: hallucination text responses and noisy image responses at token and instance levels. This approach helps in understanding and optimizing VLLMs by intentionally producing outputs that are less preferred, thus identifying areas for improvement in model performance and user interaction.

A.1.5 HYPERPARAMETER DETAILS

In this section, we show the detailed information on training hyperparameters and training data in Table 7. Specifically, for the normalized function $\mathcal{N}(\cdot,\cdot)$, we calculate the score for correct objects as $\frac{S(y_t,v)-(\mu_{gt}+\lambda)}{S_{\max}-(\mu_{gt}+\lambda)}$, and for hallucinated objects as $\frac{S(y_t,v)-(\mu_{hal}-\lambda)}{(\mu_{gt}-\lambda)-S_{\min}}$. $S_{\min}$ and $S_{\max}$ represent the minimum and maximum possible scores, respectively. In this way, we constrain the reward within the range of -1 to 1.

A.2 ADDITIONAL ANALYSIS

A.2.1 DETAILED ANALYSIS ON COCO-CAPTION BENCHMARK

Table 8 provides a comprehensive comparison of various methods evaluated on COCO-caption benchmark. Our method, denoted as FiSAO, demonstrates significant improvements across multiple metrics, highlighting its efficacy in enhancing caption generation quality. On the LLaVA backbone, FiSAOconsistently outperforms the baseline and other preference-tuning methods across all BLEU metrics, as well as METEOR, ROUGE L, and CIDEr scores. These results underscore the robustness of FiSAOin capturing nuanced textual and visual features, achieving superior alignment and coherence in the generated captions. Similarly, for the InstructBLIP backbone, FiSAOmaintains a competitive

---

[1] `https://github.com/EvolvingLMMs-Lab/lmms-eval/tree/main/lmms_eval/tasks`

Backbone Model	LLaVA-1.5	InstructBLIP
Parameter	7B	13B
Reward Model	CLIP-ViT-L-334	CLIP-ViT-L
Dataset	LLaVA-Instruct	LLaVA-Instruct
Fine-Tuning Method	LoRA	LoRA
Number of Epochs	1	1
PPO Training Epochs	4	4
GPUs Used	4 A100 80GB GPUs	4 A100 80GB GPUs
Training Time	~6 hours	~10 hours
LoRA r	128	128
LoRA Alpha	256	256
Learning Rate	5e-7	4e-6
LoRA Parameter	all linear	all linear
ξ	0.2	0.2
λ	10	10

Table 7: Training parameters for LLaVA-1.5 7B and InstructBLIP 13B models.

edge, achieving high scores across the evaluation metrics and outperforming other preference-tuning approaches. The improvements observed with FiSAOhighlight its effectiveness in leveraging fine-grained token-level rewards to enhance the alignment between visual and textual modalities.

Table 8: Evaluation results on COCO-caption benchmark.

Method	**Bleu 1**	**Bleu 2**	**Bleu 3**	**Bleu 4**	**METEOR**	**ROUGE L**	**CIDEr**
LLaVA	0.7312	0.5641	0.4150	0.2976	0.2929	0.5559	1.1038
+ Vlfeedback	0.7149	0.5487	0.3734	0.2788	0.2835	0.5398	1.0969
+ Human-Prefer	0.6741	0.5047	0.3613	0.2519	0.2864	0.5329	0.9142
+ POVID	0.7360	0.5680	0.4197	0.3030	0.2954	0.5601	1.1305
+ FiSAO	0.7925	0.6259	0.4681	0.3407	0.2811	0.5774	1.1970
InstructBLIP	0.8220	0.6682	0.5199	0.3973	0.2982	0.5984	1.3498
+ Vlfeedback	0.7919	0.6346	0.4886	0.3689	0.3000	0.5874	1.3055
+ Human-Prefer	0.8034	0.6431	0.5068	0.3759	0.3104	0.6012	1.2902
+ POVID	0.8204	0.6671	0.5198	0.3977	0.3009	0.6002	1.3619
+ FiSAO	0.8239	0.6707	0.5231	0.4008	0.2985	0.5994	1.3526

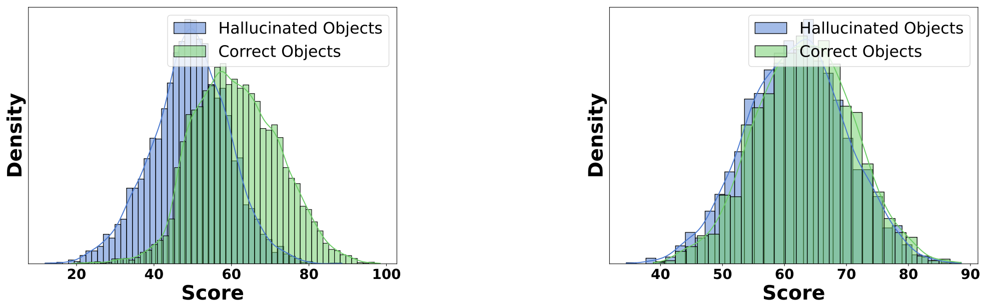
A.2.2 ADDITIONAL ANALYSIS ON SENTENCE-LEVEL REWARD

We present the sentence-level rewards of the generated captions on InstructBLIP in Figure 6. We can observe the low distinction between correct and hallucinated captions. We also show comparison of Fine-Grained and sentence-level reward distribution in Figure 7 and Figure 8, where the sentence-level reward shows no explicit correlation with traditional evaluation scores. This comparison highlights that the Fine-Grained reward distribution tends to be more useful, offering a detailed view of the model's performance. These analyses further demonstrate that using Fine-Grained rewards is more effective than sentence-level rewards.

We also calculate the average sum of token-level rewards in a sentence and explore its relationship with conventional evaluation metrics in Figure 12. We observe that, compared to sentence-level rewards, token-level rewards exhibit strong correlation with conventional evaluation metrics.

A.2.3 ADDITIONAL ANALYSIS ON REWARD DISTRIBUTION OF OBJECTS

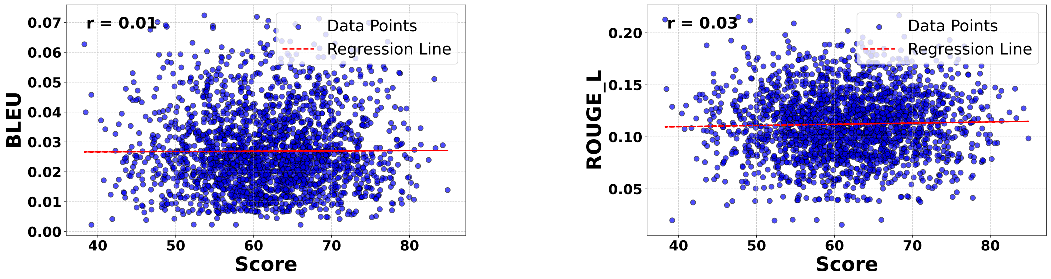
To further illustrate how our method enhances the alignment between visual encoders and VLLMs, we present the reward distribution of hallucinated objects in Figure Figure 10. The figure shows that, before training, the reward distribution for hallucinated objects in both LLaVA and InstructBLIP is



(a) Fine-Grained Reward

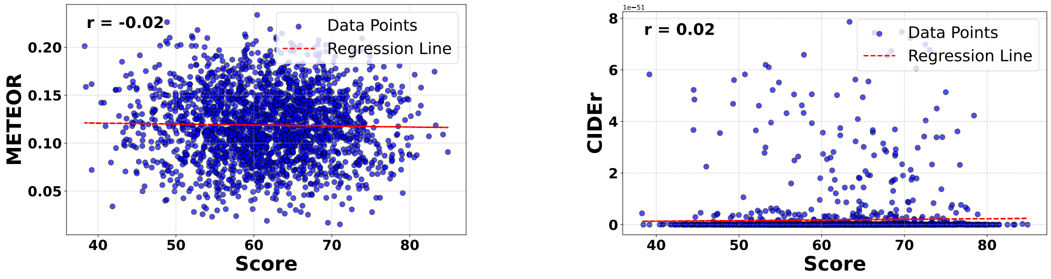
(b) Sentence-Level Reward

Figure 6: Comparison of fine-grained and sentence-level reward distributions in InstructBLIP.



(a) BLEU

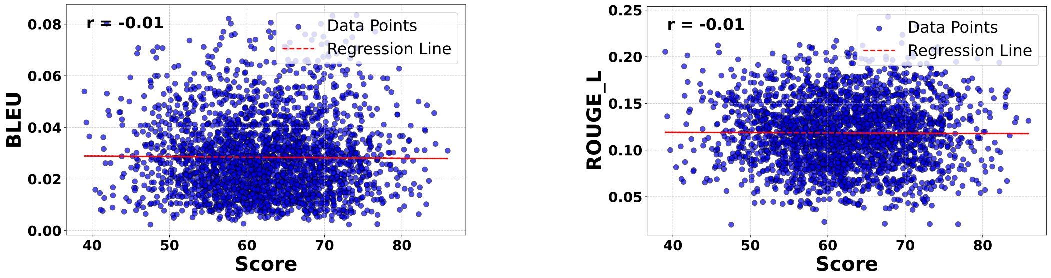
(b) ROUGE



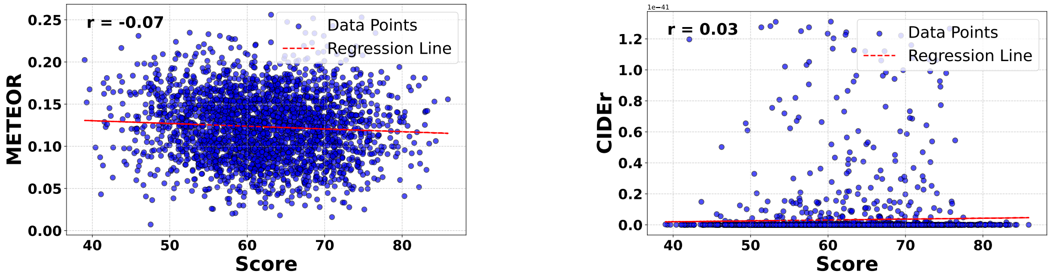
(c) METEOR

(d) CIDEr

Figure 7: Correlation between sentence reward and conventional evaluation metrics on InstructBLIP.

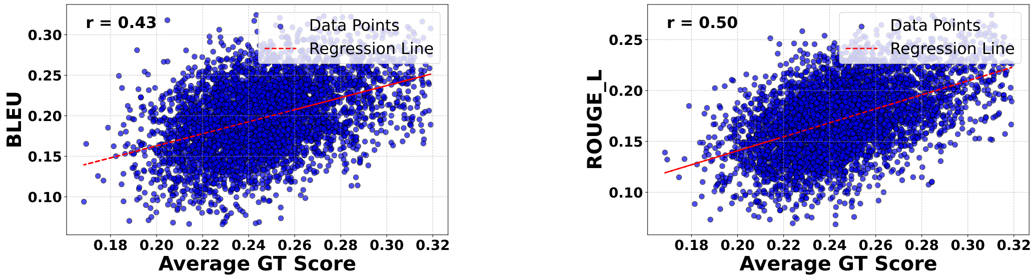


(a) BLEU (b) ROUGE

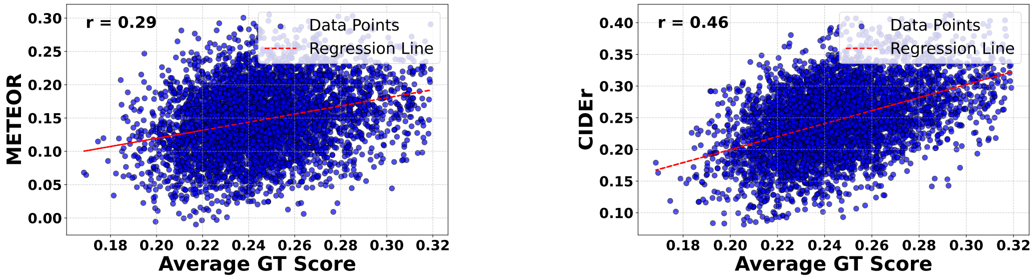


(c) METEOR (d) CIDEr

Figure 8: Correlation between sentence reward and conventional evaluation metrics on LLaVA



(a) BLEU (b) ROUGE



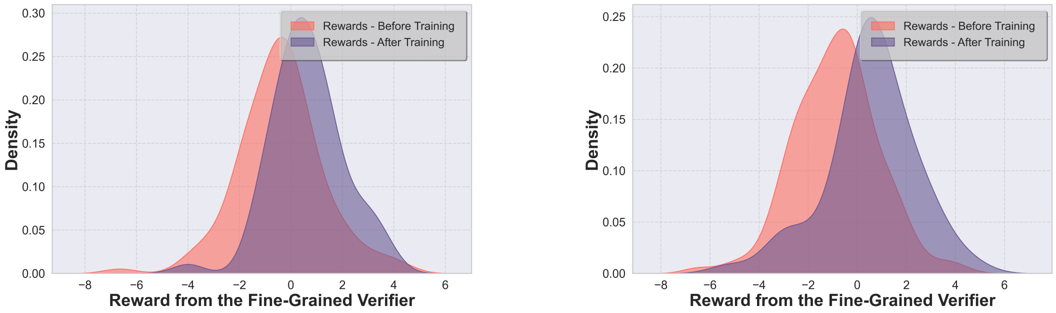
(c) METEOR (d) CIDEr

Figure 9: Correlation between average sum of token-level rewards and conventional evaluation metrics on LLaVA.

Method	MME^{P} ↑	MME^{C} ↑	SEED ↑	Cap_val ↑	CHAIR_{S} ↓	CHAIR_{I} ↓	SciQA ↑
LLaVA-1.5	1510.7	348.2	58.6	56.6	54.3	11.3	66.8
+ Vlfeedback	1432.7	321.8	59.3	54.8	40.3	13.2	66.2
+ Human-Prefer	1490.6	335.0	58.1	50.4	38.7	11.3	65.8
+ POVID	1452.8	325.3	60.2	57.3	35.2	**8.3**	68.8
+ FiSAO (Dino)	**1542.6**	**351.1**	**60.3**	**61.5**	**37.4**	9.3	**68.7**

Table 9: The performance of FiSAO across benchmarks. The best result is bolded.

more scattered and less aligned with the visual encoder's preferences. After applying our method, the reward distribution shifts to the right, indicating improved alignment and consistency with the visual encoder. This shift demonstrates that the model's rewards now more accurately reflect the visual encoder's evaluations, thereby enhancing the overall performance of vision-language alignment.



(a) LLaVA

(b) InstructBLIP

Figure 10: Reward distribution comparison before and after training.

A.2.4 CASE STUDIES

In this section, we present detailed case studies comparing the outputs of our model with LLaVA 1.5. The case studies highlight the strengths of FiSAOin generating detailed image descriptions. As shown in Figure 11, FiSAOfocuses on providing a comprehensive overview, including contextual details such as the environment and the placement of objects (e.g., handbag, table settings). This approach ensures that the description covers all relevant aspects of the scene. LLaVA 1.5 includes specific interactions and objects that enhance the vividness of the scene. However, it sometimes generates objects that are not actually present in the images.

A.2.5 ADDITIONAL EXPERIMENTS USING DIFFERENT VISION ENCODERS AS FINE-GRAINED VERIFIERS.

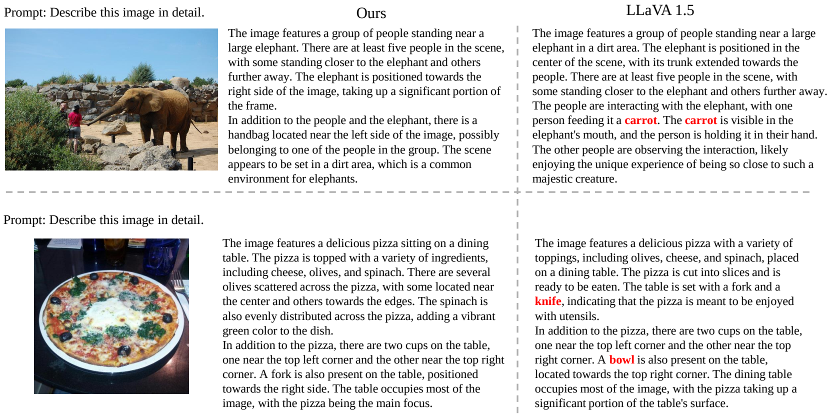
To explore the potential for adapting our approach to MLLMs with multiple visual encoders, such as SigLip and DINO-v2. we conduct two experiments:

- **a. Using DINO-v2 as the reward model.** We adopt DINO-v2 as the reward model to provide fine-grained feedback for training LLaVA. The results are presented in Table 9.

  The results demonstrate that using DINO-v2 as the reward model achieves consistent improvements across benchmarks, validating its effectiveness in providing fine-grained feedback for training LLaVA.

- **b. Visualizing SigLip's differentiation capabilities.** We visualize the distributions of hallucinated and correct tokens using SigLip scores, as shown in Figure 12.

  Our observations show that SigLip demonstrates a significant differentiation between hallucinated and correct tokens compared to CLIP. This finding highlights that our approach is potentially highly adaptable when applied to MLLMs that incorporate multiple visual encoders, opening avenues for future exploration.



Prompt	Ours	LLaVA 1.5
Prompt: Describe this image in detail.	The image features a group of people standing near a large elephant. There are at least five people in the scene, with some standing closer to the elephant and others further away. The elephant is positioned towards the right side of the image, taking up a significant portion of the frame. In addition to the people and the elephant, there is a handbag located near the left side of the image, possibly belonging to one of the people in the group. The scene appears to be set in a dirt area, which is a common environment for elephants.	The image features a group of people standing near a large elephant in a dirt area. The elephant is positioned in the center of the scene, with its trunk extended towards the people. There are at least five people in the scene, with some standing closer to the elephant and others further away. The people are interacting with the elephant, with one person feeding it a **carrot**. The **carrot** is visible in the elephant's mouth, and the person is holding it in their hand. The other people are observing the interaction, likely enjoying the unique experience of being so close to such a majestic creature.
Prompt: Describe this image in detail.	The image features a delicious pizza sitting on a dining table. The pizza is topped with a variety of ingredients, including cheese, olives, and spinach. There are several olives scattered across the pizza, with some located near the center and others towards the edges. The spinach is also evenly distributed across the pizza, adding a vibrant green color to the dish. In addition to the pizza, there are two cups on the table, one near the top left corner and the other near the top right corner. A fork is also present on the table, positioned towards the right side. The table occupies most of the image, with the pizza being the main focus.	The image features a delicious pizza with a variety of toppings, including olives, cheese, and spinach, placed on a dining table. The pizza is cut into slices and is ready to be eaten. The table is set with a fork and a **knife**, indicating that the pizza is meant to be enjoyed with utensils. In addition to the pizza, there are two cups on the table, one near the top left corner and the other near the top right corner. A **bowl** is also present on the table, located towards the top right corner. The dining table occupies most of the image, with the pizza taking up a significant portion of the table's surface.

Figure 11: Case studies on LLaVA 1.5.

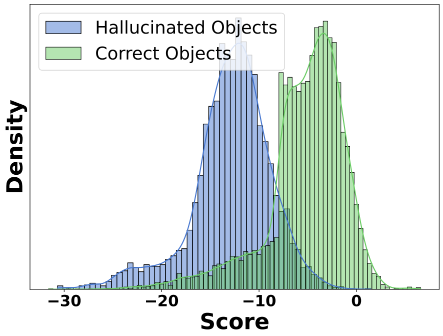


Figure 12: Correlation between the average sum of token-level rewards and conventional evaluation metrics.

A.2.6 ANALYSIS OF THE EXPANDED SET C AND ITS IMPACT ON PERFORMANCE

In this section, we analyze the impact of expanding the object set C on the performance of our approach. The results, presented in Table 12, demonstrate that expanding C leads to noticeable improvements in the performance of LVLMs across multiple benchmarks.

The expanded set C provides a more comprehensive set of object labels from Open Image[2] and Objects365[3], allowing the model to robustly identify and align object tokens. This enhancement improves the contextual understanding of the generated descriptions, contributing to better overall performance.

However, as shown in Table 12, we further expand C by adding additional object labels from, which results in only marginal performance gains. This observation suggests that the original set C already covers a sufficiently comprehensive range of common objects.

A.2.7 COMPUTATIONAL OVERHEAD: TOKEN-LEVEL VS. SENTENCE-LEVEL REWARDS

To analyze the computational trade-offs between token-level and sentence-level rewards, we conducted experiments comparing their efficiency. The results are presented in Table 10.

Metric	**Token-Level**	**Sentence-Level**
Total Samples	5000	5000
Current Speed (s/sample)	0.1883	0.1631

Table 10: Comparison of computational overhead for token-level and sentence-level rewards.

The results indicate that token-level rewards introduce only a minor increase in computational overhead compared to sentence-level rewards. To further explore the overall efficiency, we measured the total runtime across different methods, as shown in Table 11.

Method	**Time per Sample (seconds)**	**Throughput (samples/second)**
Inference Only	2.63	0.38
Ours (w/o Inference)	1.17	0.85
Ours (w Inference)	3.80	0.26
POVID	2.15	0.47
VLFeedback	2.34	0.43

Table 11: Total runtime comparison across different methods.

Excluding inference time, our approach demonstrates higher efficiency with faster reward computation compared to other methods. When considering overall runtime (including inference), the efficiency of our method remains comparable to existing approaches such as POVID and VLFeedback. Importantly, the use of token-level rewards enables more fine-grained performance improvements, highlighting the practical benefits of our approach while maintaining computational feasibility.

Method	**MME^P**	**MME^C**	**SEED**	**MMB**	**MM-Vet**	**SQA^I**	**POPE**	**GQA**	**Cap_val**	**CHAIR$_S$**	**CHAIR$_I$**
Before expansion of C	1515.3	348.7	59.1	64.2	30.5	68.2	85.9	61.9	58.2	47.7	11.8
Original C	**1522.6**	**349.0**	**60.6**	64.8	30.7	**69.3**	85.7	62.0	**61.2**	39.9	9.9
After expansion of C	**1536.1**	**351.1**	**60.9**	64.7	30.6	**69.8**	85.8	62.1	**61.0**	41.1	10.2

Table 12: Performance comparison before and after the expansion of C.

---

[2] `https://storage.googleapis.com/openimages/web/index.html`

[3] `https://www.objects365.org/overview.html`

A.3 WHY DOES FEEDBACK FROM PRETRAINED VISION ENCODERS CONTRIBUTE TO THE MODEL'S PERFORMANCE - THEORETICAL ANALYSIS

A.3.1 PROOF OF THEOREM 3.1

We begin by considering the distribution of the generated response y_p given by $\pi^*_{\theta_t}(y \mid x)$. Since $y_p = \arg\max_y R(y)$, this distribution is a point mass. The global minimizer will converge to $\pi^*_{\theta_t}(y \mid x)$.

By our method, we have

$$y_p = \arg\max_y \left[(1-\lambda)\left(-\|y - (V_1^* v + V_2^* t)\|^2\right) + \lambda \langle U_v^\top v, U_t^\top y \rangle\right].$$

Simplifying, we rewrite the optimization problem as

$$y_p = \arg\min_y \left[\|y - (V_1^* v + V_2^* t)\|^2 - \gamma \langle U_v^\top v, U_t^\top y \rangle\right],$$

where $\gamma = \frac{\lambda}{1-\lambda}$. Taking the derivative with respect to y and setting it to zero yields

$$2\left(y - (V_1^* v + V_2^* t)\right) - \gamma U_t U_v^\top v = 0.$$

Solving for y_p, we obtain

$$y_p = (V_1^* v + V_2^* t) + \frac{\gamma}{2} U_t U_v^\top v.$$

This shows that integrating vision feedback effectively increases the weight on the visual input.

Next, we consider the loss function

$$L(y) = \min_{\beta \in \mathbb{R}^{d_t}} \mathbb{E}\left[\left(z - \beta^\top y\right)^2\right], \tag{15}$$

where $z = \beta^{*\top} y_{\text{truth}}$ and $y_{\text{truth}} = V_1^* v + V_2^* t + \epsilon_y$.

Substituting the expressions for y_p and y_{truth}, we have

$$L(y_p) = \min_\beta \mathbb{E}\left[\left(\beta^{*\top} y_{\text{truth}} - \beta^\top y_p\right)^2\right]. \tag{16}$$

Expanding, we get

$$L(y_p) = \min_\beta \mathbb{E}\left[\left((\beta^{*\top} - \beta^\top)(V_1^* v + V_2^* t) - \beta^\top\left(\frac{\gamma}{2} U_t U_v^\top v\right) + \beta^{*\top}\epsilon_y\right)^2\right]. \tag{17}$$

We introduce an assumption that ϵ_y contains a component that can be estimated via vision feedback. Suppose

$$\epsilon_y = \kappa U_t U_v^\top v + \tilde{\epsilon}, \tag{18}$$

where $\tilde{\epsilon}$ is noise independent of v, and κ is a scalar.

Therefore,

$$y_{\text{truth}} = V_1^* v + V_2^* t + \kappa U_t U_v^\top v + \tilde{\epsilon}. \tag{19}$$

Now, since

$$y_p^{(\lambda)} = V_1^* v + V_2^* t + \frac{\gamma}{2} U_t U_v^\top v, \tag{20}$$

the vision feedback term helps to estimate part of ϵ_y.

We define the mean squared error:

$$\mathrm{MSE}_\lambda = \mathbb{E}\left[\left\|y_p^{(\lambda)} - y_{\text{truth}}\right\|^2\right]. \tag{21}$$

Substituting,

$$\mathrm{MSE}_\lambda = \mathbb{E}\left[\left\|\left(\frac{\gamma}{2} - \kappa\right) U_t U_v^\top v - \tilde{\epsilon}\right\|^2\right]. \tag{22}$$

For $\lambda = 0$,

$$\mathrm{MSE}_0 = \mathbb{E}\left[\left\|-\kappa U_t U_v^\top v - \tilde{\epsilon}\right\|^2\right]. \tag{23}$$

The difference is

$$\Delta\mathrm{MSE} = \mathrm{MSE}_\lambda - \mathrm{MSE}_0 = \left[\left(\frac{\gamma}{2} - \kappa\right)^2 - \kappa^2\right] \mathbb{E}\left[\left\|U_t U_v^\top v\right\|^2\right]. \tag{24}$$

Setting $\gamma = 2\kappa$ (which implies $\lambda = \frac{2\kappa}{2\kappa+1} > 0$), we have

$$\Delta\mathrm{MSE} = -\kappa^2 \mathbb{E}\left[\left\|U_t U_v^\top v\right\|^2\right] < 0. \tag{25}$$

Thus, there exists $\lambda > 0$ such that

$$\mathbb{E}_{\pi_{\theta(\lambda)}(y|x)}[L(y)] < \mathbb{E}_{\pi_{\theta(0)}(y|x)}[L(y)]. \tag{26}$$

This proves the theorem.

By selecting a suitable $\lambda > 0$, we have demonstrated that integrating vision feedback can reduce the expected loss. Therefore, incorporating vision feedback helps the model to predict the output more accurately, which proves Theorem 3.1.

A.3.2 THEORETICAL FRAMEWORK FOR INCORPORATING PRE-TRAINED VISION MODELS' FEEDBACK INTO MODEL TRAINING

In this section, we present a theoretical framework demonstrating how integrating fine-grained feedback from pre-trained vision models can enhance the performance of VLLMs.

To assess the quality of the text output y, following the seeting in Section 3.2, we approach it as a regression problem where there is an associated outcome z linked to the ground-truth text output y_{truth}: $z = \beta^{*\top} y_{\text{truth}}$, with $\beta^* \in \mathbb{R}^{d_t}$. The quality of y is evaluated using the loss function

$$L(y) = \min_{\beta \in \mathbb{R}^{d_t}} \mathbb{E}[(z - \beta^\top y)^2].$$

Note that in this context, a lower value of $L(y)$ indicates better quality of the text output y. We define two specific types of expected loss:

- The *coarse-grained expected loss*:

  $$L_{\text{coarse}} = L(y_p^{\text{coarse}}) = \mathbb{E}\left[\left(z - \beta^{*\top} y_p^{\text{coarse}}\right)^2\right],$$

  where y_p^{coarse} is the coarse-grained prediction.

- The *fine-grained loss*:

  $$L_{\text{fine}} = L(y_p^{\text{fine}}) = \mathbb{E}\left[\left(z - \beta^{*\top} y_p^{\text{fine}}\right)^2\right],$$

  where y_p^{fine} is the fine-grained prediction.

Based on the definitions and loss formulations above, we establish the following theorem.

Theorem A.1. *Fine-grained feedback enables more precise and accurate adjustments to y than coarse-grained feedback. Consequently, the expected loss associated with fine-grained feedback is strictly lower than that of coarse-grained feedback, leading to a positive difference in expected losses:*

$$\Delta L = L_{coarse} - L_{fine} > 0.$$

Proof

We aim to prove that $\Delta L = L_{\text{coarse}} - L_{\text{fine}} > 0$.

1. Setup and Notation

- **True Output**:

$$y_{\text{truth}} = V_1^* v + V_2^* t + \epsilon_y,$$

where ϵ_y is a zero-mean noise term with covariance Σ_ϵ.

- **Model Predicted Outputs**: - **Coarse-Grained Prediction**:

$$y_p^{\text{coarse}} = V_1^* v + V_2^* t + \frac{\gamma_{\text{coarse}}}{2} U_t U_v^\top v.$$

- **Fine-Grained Prediction**:

$$y_p^{\text{fine}} = V_1^* v + V_2^* t + \frac{\gamma_{\text{fine}}}{2} \sum_{i=1}^{n} u_{t,i} \left(u_{v,i}^\top v \right),$$

where n is the number of components for which fine-grained feedback is provided,

- **Loss Function**: For a prediction y_p, the loss function is defined as:

$$L(y_p) = \mathbb{E}\left[\left(z - \beta^{*\top} y_p \right)^2 \right],$$

where $z = \beta^{*\top} y_{\text{truth}}$ and β^* is the optimal coefficient vector.

Our goal is to compute $\Delta L = L_{\text{coarse}} - L_{\text{fine}}$ and prove that $\Delta L > 0$.

2. Expressing the Losses

(a) Coarse-Grained Expected Loss

Compute the error:

$$z - \beta^{*\top} y_p^{\text{coarse}} = \beta^{*\top} y_{\text{truth}} - \beta^{*\top} y_p^{\text{coarse}} = \beta^{*\top} \left(\epsilon_y - \delta_{\text{coarse}} \right),$$

where:

$$\delta_{\text{coarse}} = \frac{\gamma_{\text{coarse}}}{2} U_t U_v^\top v.$$

Therefore, the coarse-grained loss is:

$$L_{\text{coarse}} = \mathbb{E}\left[\left(\beta^{*\top} \left(\epsilon_y - \delta_{\text{coarse}} \right) \right)^2 \right].$$

(b) Fine-Grained Expected Loss

Similarly, compute the error:

$$z - \beta^{*\top} y_p^{\text{fine}} = \beta^{*\top} y_{\text{truth}} - \beta^{*\top} y_p^{\text{fine}} = \beta^{*\top} \left(\epsilon_y - \delta_{\text{fine}} \right),$$

where:

$$\delta_{\text{fine}} = \frac{\gamma_{\text{fine}}}{2} \sum_{i=1}^{n} u_{t,i} \left(u_{v,i}^\top v \right).$$

Therefore, the fine-grained loss is:

$$L_{\text{fine}} = \mathbb{E}\left[\left(\beta^{*\top}\left(\epsilon_y - \delta_{\text{fine}}\right)\right)^2\right].$$

3. Computing the Loss Difference ΔL

$$\Delta L = L_{\text{coarse}} - L_{\text{fine}} = \mathbb{E}\left[\left(\beta^{*\top}\left(\epsilon_y - \delta_{\text{coarse}}\right)\right)^2\right] - \mathbb{E}\left[\left(\beta^{*\top}\left(\epsilon_y - \delta_{\text{fine}}\right)\right)^2\right].$$

Expanding the squared terms:

$$\begin{aligned}\Delta L = &\left(\mathbb{E}\left[\left(\beta^{*\top}\epsilon_y\right)^2\right] - 2\mathbb{E}\left[\beta^{*\top}\epsilon_y \cdot \beta^{*\top}\delta_{\text{coarse}}\right] + \mathbb{E}\left[\left(\beta^{*\top}\delta_{\text{coarse}}\right)^2\right]\right) \\ &- \left(\mathbb{E}\left[\left(\beta^{*\top}\epsilon_y\right)^2\right] - 2\mathbb{E}\left[\beta^{*\top}\epsilon_y \cdot \beta^{*\top}\delta_{\text{fine}}\right] + \mathbb{E}\left[\left(\beta^{*\top}\delta_{\text{fine}}\right)^2\right]\right).\end{aligned}$$

Simplifying:

$$\Delta L = \left(\mathbb{E}\left[\left(\beta^{*\top}\delta_{\text{coarse}}\right)^2\right] - \mathbb{E}\left[\left(\beta^{*\top}\delta_{\text{fine}}\right)^2\right]\right) - 2\mathbb{E}\left[\beta^{*\top}\epsilon_y \cdot \beta^{*\top}\left(\delta_{\text{coarse}} - \delta_{\text{fine}}\right)\right].$$

Since ϵ_y has zero mean and is independent of v and t, the cross term vanishes:

$$\mathbb{E}\left[\beta^{*\top}\epsilon_y \cdot \beta^{*\top}\left(\delta_{\text{coarse}} - \delta_{\text{fine}}\right)\right] = 0.$$

Therefore,

$$\Delta L = \mathbb{E}\left[\left(\beta^{*\top}\delta_{\text{coarse}}\right)^2\right] - \mathbb{E}\left[\left(\beta^{*\top}\delta_{\text{fine}}\right)^2\right].$$

4. Computing $\mathbb{E}\left[\left(\beta^{*\top}\delta\right)^2\right]$

Compute $\beta^{*\top}\delta_{\textit{coarse}}$:

$$\beta^{*\top}\delta_{\text{coarse}} = \frac{\gamma_{\text{coarse}}}{2}\beta^{*\top}U_t U_v^\top v = \frac{\gamma_{\text{coarse}}}{2}\sum_{i=1}^{r}\left(\beta^{*\top}u_{t,i}\right)\left(u_{v,i}^\top v\right).$$

Compute $\beta^{*\top}\delta_{\textit{fine}}$:

$$\beta^{*\top}\delta_{\text{fine}} = \frac{\gamma_{\text{fine}}}{2}\sum_{i=1}^{n}\left(\beta^{*\top}u_{t,i}\right)\left(u_{v,i}^\top v\right).$$

Let

$$a_i = \left(\beta^{*\top}u_{t,i}\right)\left(u_{v,i}^\top v\right).$$

Assuming $\gamma_{\text{coarse}} = \gamma_{\text{fine}} = \gamma$, we have:

$$\beta^{*\top}\delta_{\text{coarse}} = \frac{\gamma}{2}\sum_{i=1}^{r} a_i, \quad \beta^{*\top}\delta_{\text{fine}} = \frac{\gamma}{2}\sum_{i=1}^{n} a_i.$$

Compute the Variances:

Assuming v is a zero-mean random vector with covariance Σ_v, and $u_{v,i}$ are orthonormal vectors, we have:

$$\mathbb{E}\left[a_i\right] = 0, \quad \text{Var}\left[a_i\right] = \left(\beta^{*\top}u_{t,i}\right)^2 \cdot \text{Var}\left(u_{v,i}^\top v\right).$$

Assuming $\text{Var}\left(u_{v,i}^\top v\right) = \sigma_v^2$ (a constant), we get:

$$\text{Var}\left[a_i\right] = \sigma_v^2\left(\beta^{*\top}u_{t,i}\right)^2.$$

Therefore,

$$\mathbb{E}\left[\left(\beta^{*\top}\delta_{\text{coarse}}\right)^2\right]=\left(\frac{\gamma}{2}\right)^2\sum_{i=1}^{r}\text{Var}\left[a_i\right]=\left(\frac{\gamma}{2}\right)^2\sigma_v^2\sum_{i=1}^{r}\left(\beta^{*\top}u_{t,i}\right)^2.$$

Similarly,

$$\mathbb{E}\left[\left(\beta^{*\top}\delta_{\text{fine}}\right)^2\right]=\left(\frac{\gamma}{2}\right)^2\sigma_v^2\sum_{i=1}^{n}\left(\beta^{*\top}u_{t,i}\right)^2.$$

Thus, the loss difference is:

$$\Delta L=\left(\frac{\gamma}{2}\right)^2\sigma_v^2\left(\sum_{i=1}^{r}\left(\beta^{*\top}u_{t,i}\right)^2-\sum_{i=1}^{n}\left(\beta^{*\top}u_{t,i}\right)^2\right)=\left(\frac{\gamma}{2}\right)^2\sigma_v^2\sum_{i=n+1}^{r}\left(\beta^{*\top}u_{t,i}\right)^2.$$

5. Theoretical Implications and Conclusion

Since each term $\left(\beta^{*\top}u_{t,i}\right)^2\geq 0$, and unless $\beta^{*\top}u_{t,i}=0$ for all $i=n+1,\ldots,r$, the sum is positive. Therefore,

$$\Delta L>0.$$

This shows that the expected loss using coarse-grained feedback is greater than that using fine-grained feedback:

$$L_{\text{coarse}}>L_{\text{fine}}.$$

Our theoretical analysis extends Theorem 3.1 by demonstrating that fine-grained feedback, which provides more precise adjustments to y, leads to a lower expected loss compared to coarse-grained feedback. This supports the effectiveness of integrating detailed feedback from pre-trained vision models into the training of VLLMs, enhancing their performance over models that rely solely on less precise feedback mechanisms.

A.4 RELATED WORK

A.4.1 VISION-LARGE LANGUAGE MODEL

Recently, the development of large language models (Brown et al., 2020; Touvron et al., 2023)and pre-trained vision models (Radford et al., 2021a), has paved the way for Vision-Large Language Model(VLLMs). These advanced models, which can comprehend both text and images, have greatly enhanced our capacity to automate complex tasks accross various areas such as medical application (Liu et al., 2023b), autonomous driving (Zhou et al., 2023a) and embodied agent (Peng et al., 2023). The fundamental architecture of VLLMs typically integrates both language and vision models. This integration involves aligning the embedding spaces of both modalities using Qformer or a simple fully connected layer (Zhu et al., 2023; Ye et al., 2023a; Li et al., 2023b). However, Vision-Language Large Models (VLLMs) still face the problem of misalignment, as both models are typically pre-trained independently before being aligned through vision-language joint training. This misalignment can lead to several issues, such as safety concerns, where the model may produce inappropriate or biased content (Gong et al., 2023; Tu et al., 2023), hallucinations in VLLMs, where the model generates information not grounded in the images, thus deviating from observable reality (Wang et al., 2023), and deficiencies in logical reasoning (Ghosh et al., 2024), where the model fails to coherently integrate visual and textual information, resulting in inaccurate outputs.

A.4.2 VISION-LANGUAGE ALIGNMENT

Traditional vision-language models (VLMs) have primarily aimed to enhance image-text alignment using methods such as the co-attention framework (Lu et al., 2019), anchor points (Li et al., 2020), and contrastive learning (Radford et al., 2021b). With the significant advancements in large language models (LLMs), recent approaches have explored novel directions to integrate visual encoders with

LLMs, enabling better comprehension of vision-language multi-modal tasks. Aligning visual and linguistic modalities can primarily be categorized into two approaches: alignment from training data and alignment from feedback. Alignment from training data involves using high-quality datasets for SFT (Supervised Fine-Tuning) training, including diverse instructions and dataset compression. This method relies on the diversity and quality of the training data to improve the model's performance. Alignment from feedback focuses on fine-tuning the model using feedback of human (Sun et al., 2023; Yu et al., 2023a) or other models like CLIP (Zhou et al., 2024a) and large models (Li et al., 2023d; Zhao et al., 2023). Two primary methods for learning from feedback in VLLMs are Proximal Policy Optimization (PPO) (Sun et al., 2023) and Direct Preference Optimization (DPO) (Zhao et al., 2023; Li et al., 2023d; Chen et al., 2024a). However, these methods encounter challenges. They may generate out-of-distribution data that fails to significantly enhance the model's performance and entail significant expenses in dataset construction.

A.4.3 TOKEN-LEVEL REWARDS IN LLMS AND VLMS

Recent studies have explored token-level rewards as an alternative to sentence-level scoring, aiming to achieve finer-grained control and improved alignment. In Large Language Models(LLMs), token-level reward models such as TLCR (Yoon et al., 2024) and selective preference optimization methods (Yang et al., 2024b) have demonstrated the benefits of rewarding individual tokens for fine-grained reinforcement learning, enabling better performance on tasks requiring nuanced token-wise feedback. In the context of Vision-Language Models(VLMs), concurrent works such as TLDR (Fu et al., 2024b) and GenARM (Xu et al., 2024) have proposed token-level reward strategies for improving alignment and reducing hallucinations in large vision-language models. Despite their potential, existing feedback-based methods face challenges such as high costs in dataset construction and the need for external tools.

A.5 LIMITATIONS

One limitation of FiSAO is its dependency on the quality and robustness of the pre-trained vision models. While pre-trained vision encoders are highly effective at object recognition, they are generally less adept at capturing more complex aspects such as actions or spatial relationships [1,2]. As a result, FiSAO may not fully address hallucinations related to these areas, as they require more advanced contextual reasoning capabilities that current pre-trained encoders are not specifically optimized for.

Furthermore, if the visual encoder contains inherent biases or inaccuracies, these issues can be propagated through the reward model, potentially affecting the overall alignment process. Addressing these limitations requires either enhancing the visual encoder's capabilities to handle complex relational reasoning or exploring additional strategies to incorporate contextual reasoning into the alignment framework.

Additionally, Our method faces challenges in tasks like MM-Vet, which primarily involve simple QA scenarios with single-word ground truths, limiting the benefits of token-level rewards.

A.6 BROADER IMPACTS

The proposed enhancement in Vision-Language Large Models (VLLMs) through fine-grained policy optimization presents several significant broader impacts across various fields and societal dimensions. FiSAOcontributes to the field of AI by providing a novel approach to self-training without the need for additional data. This can inspire further research into data-efficient training methods, fostering innovation and reducing the environmental impact associated with large-scale data collection and processing. Besides, enhanced vision-language alignment can significantly improve the performance of assistive technologies, such as screen readers and automated transcription services, making digital content more accessible to people with disabilities. This aligns with global efforts to promote inclusivity and equal access to information and technology.