Enhancing Large Vision Language Models with Self-Training on Image Comprehension

Yihe Deng[*1], **Pan Lu**[*1,3], **Fan Yin**[1], **Ziniu Hu**[1], **Sheng Shen**[2]
Quanquan Gu[1], **James Zou**[3], **Kai-Wei Chang**[1], **Wei Wang**[1]

[1]University of California, Los Angeles
[2]University of California, Berkeley [3]Stanford University

https://stic-lvlm.github.io/

Abstract

Large vision language models (LVLMs) integrate large language models (LLMs) with pre-trained vision encoders, thereby activating the model's perception capability to understand image inputs and conduct subsequent reasoning for different queries. Improving this capability requires high-quality vision-language data, which is costly and labor-intensive to acquire. Self-training approaches have been effective in single-modal settings to alleviate the need for labeled data by leveraging model's own generation. However, effective self-training remains a challenge regarding the unique visual perception and reasoning capability of LVLMs. To address this, we introduce **S**elf-**T**raining on **I**mage **C**omprehension (**STIC**), which emphasizes a self-training approach specifically for image comprehension. First, the model self-constructs a preference dataset for image descriptions using unlabeled images. Preferred responses are generated through a step-by-step prompt, while dis-preferred responses are generated from either corrupted images or misleading prompts. To further self-improve reasoning on the extracted visual information, we let the model reuse a small portion of existing instruction-tuning data and append its self-generated image descriptions to the prompts. We validate the effectiveness of STIC across seven different benchmarks, demonstrating substantial performance gains of 4.0% on average while using 70% less supervised fine-tuning data than the current method. Further studies investigate various components of STIC and highlight its potential to leverage vast quantities of unlabeled images for self-training. Code and data are made publicly available on GitHub.

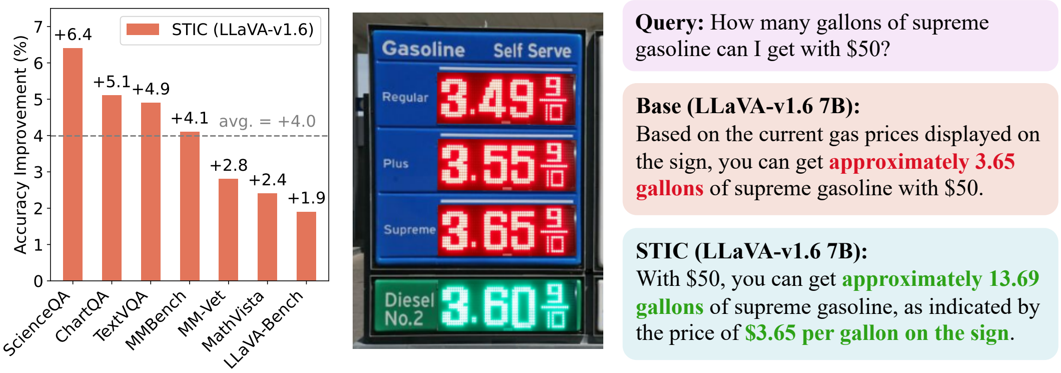
1 Introduction

In recent years, we have witnessed remarkable advancements in large language models (LLMs), such as GPT-4 (OpenAI, 2023a) and the LLaMA family (Touvron et al., 2023a,b). The increasing importance of processing multimodal inputs, including images and text, has significantly driven progress in vision language models (Radford et al., 2021; Jia et al., 2021b; Goel et al., 2022). Leveraging the powerful language understanding and generation capabilities of LLMs, researchers have advanced vision language models into large vision language models (LVLMs). This enhancement is achieved by integrating LLMs with image encoders (Radford et al., 2021; Li et al., 2023a), which were pre-trained on large-scale image-text pairs to ensure alignment between the two domains. For instance, LLaVA (Liu et al., 2023b) integrates a vision encoder from CLIP (Radford et al., 2021) with the LLM Vicuna (Chiang et al., 2023b), which is further fine-tuned on carefully constructed vision-language instructional datasets to activate the model's perception capability of capturing the vision information according to different queries. This recent development has substantially expanded the requirement for large-scale instruction fine-tuning data for LVLMs (Gao et al., 2023b; Bai et al., 2023; Chen et al., 2023b; Gao et al., 2024; Anthropic, 2024; McKinzie et al., 2024).

[*]Equal contribution.

[†]Contribution statement is provided after Section 7.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



Query: How many gallons of supreme gasoline can I get with $50?

Base (LLaVA-v1.6 7B):
Based on the current gas prices displayed on the sign, you can get **approximately 3.65 gallons** of supreme gasoline with $50.

STIC (LLaVA-v1.6 7B):
With $50, you can get **approximately 13.69 gallons** of supreme gasoline, as indicated by the price of **$3.65 per gallon on the sign**.

Figure 1: **Left**: Accuracy improvement of our method, STIC, compared to the original LLaVA-v1.6 (Liu et al., 2024) on seven benchmarks. **Right**: Response examples from the original LLaVA-v1.6 and STIC (LLaVA-v1.6), which enhances image comprehension and subsequent reasoning capabilities.

While LVLMs have shown promising results, a key challenge lies in the acquisition of high-quality fine-tuning data. Obtaining human-curated content at scale is often prohibitively expensive, especially for multi-modal data. Many recent studies resort to GPT-4V (OpenAI, 2023b) for generating or labeling high-quality vision-language fine-tuning data. However, this approach does not significantly reduce the cost (Liu et al., 2023b; Wu et al., 2024). For instance, using GPT-4V to generate $6k$ image descriptions with $1k$ tokens per output would cost approximately $200. There remains a pressing need for cost-effective methods to gather fine-tuning data to further enhance LVLMs.

To tackle the data acquisition bottleneck in multi-modality, we propose **S**elf-**T**raining on **I**mage **C**omprehension (**STIC**). Our method is inspired by the recent success of self-training (Chen et al., 2024; Yuan et al., 2024; Fränken et al., 2024; Rosset et al., 2024) in LLMs, which leverages self-generated data to improve their downstream performance. However, different from the text-only domain, the unique vision modality of LVLMs introduces new challenges, as LVLMs must understand the input image content before reasoning and responding to any related textual queries about the image. Therefore, the proposed STIC approach is a novel two-stage self-training method that targets both *image perception* and *reasoning over images and texts*.

The overall framework is summarized in Figure 2. STIC specifically emphasizes the **image comprehension self-training** of LVLMs where the model generates its own preference dataset focused on image description. The self-generated *dispreferred response* is obtained by gathering model responses from either (1) prompts likely to elicit inaccurate responses or (2) corrupted images. The *preferred responses* are collected via a detailed prompt that guides the model through a step-by-step image description process. Figure 3 shows examples of such generated responses. During fine-tuning, we consider a DPO loss (Rafailov et al., 2023) with an additional regularized term explicitly emphasizing the preferred response. Lastly, we allow the model to self-improve its reasoning ability based on its own extracted image information by reusing a small amount of existing instruction fine-tuning data and appending its self-generated image descriptions to the prompts. We refer to this second stage as **description-infused fine-tuning**. Notably, our STIC approach *does not require pre-labeled information of the images*, which contrasts to the recent works that rely on such information for constructing vision-language preference data (Zhou et al., 2024).

To demonstrate the effectiveness of STIC, we conduct extensive experiments on seven vision-language benchmarks, including ScienceQA (Lu et al., 2022), TextVQA (Singh et al., 2019), ChartQA (Masry et al., 2022), LLaVA-Bench (Liu et al., 2023a), MMBench (Liu et al., 2023c), MM-Vet (Yu et al., 2023), and MathVista (Lu et al., 2024). These benchmarks encompass scientific reasoning, math reasoning, optical character recognition (OCR), and conversation capabilities based on vision inputs, spanning various image sources such as natural, chart, and text-rich images. We employ LLaVA-v1.6 (Liu et al., 2024) as the primary base LVLM for our experiments and unitize $6k$ images from MSCOCO (Lin et al., 2014) to construct the image description preference data. As depicted in Figure 1, STIC achieves consistent and significant performance improvements across these benchmarks, with an average accuracy gain of **4.0%** over the base LVLM and a notable gain of **6.4%** on ScienceQA. We also provide an example of the different responses from the original LVLM and STIC in Figure 1, where STIC successfully identifies the key visual information and accurately reason with it. These results demonstrate the remarkable effectiveness of our image comprehension self-training approach in enhancing the visual perception capabilities of LVLMs.

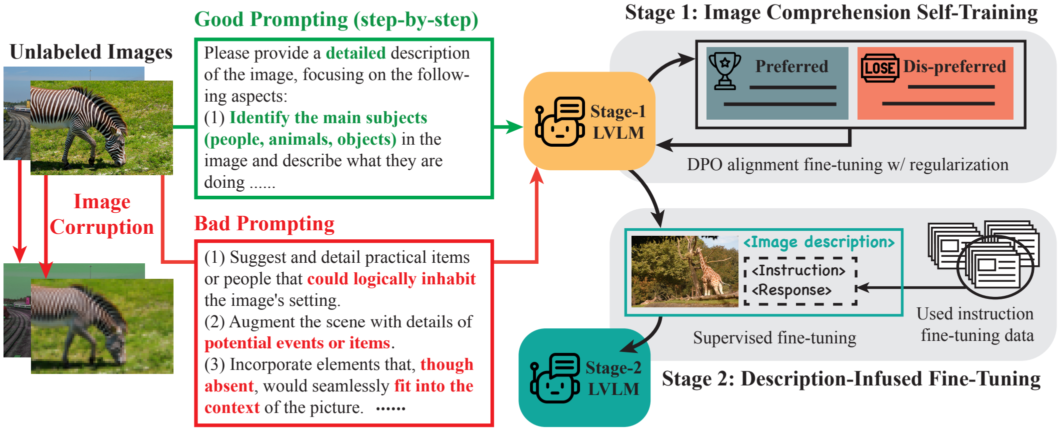

Unlabeled Images
Good Prompting (step-by-step)
Please provide a detailed description of the image, focusing on the following aspects:
(1) Identify the main subjects (people, animals, objects) in the image and describe what they are doing ......
Image Corruption
Bad Prompting
(1) Suggest and detail practical items or people that could logically inhabit the image's setting.
(2) Augment the scene with details of potential events or items.
(3) Incorporate elements that, though absent, would seamlessly fit into the context of the picture. ......
Stage 1: Image Comprehension Self-Training
Stage-1 LVLM
Preferred
LOSE Dis-preferred
DPO alignment fine-tuning w/ regularization
<Image description>
<Instruction>
<Response>
Used instruction fine-tuning data
Supervised fine-tuning
Stage-2 LVLM
Stage 2: Description-Infused Fine-Tuning

Figure 2: Framework overview of STIC, a two-stage self-training algorithm focusing on the image comprehension capability of the LVLMs. In Stage 1, the base LVLM self-constructs its preference dataset for image description using well-designed prompts, poorly-designed prompts, and distorted images. In Stage 2, a small portion of the previously used SFT data is recycled and infused with model-generated image descriptions to further fine-tune the base LVLM.

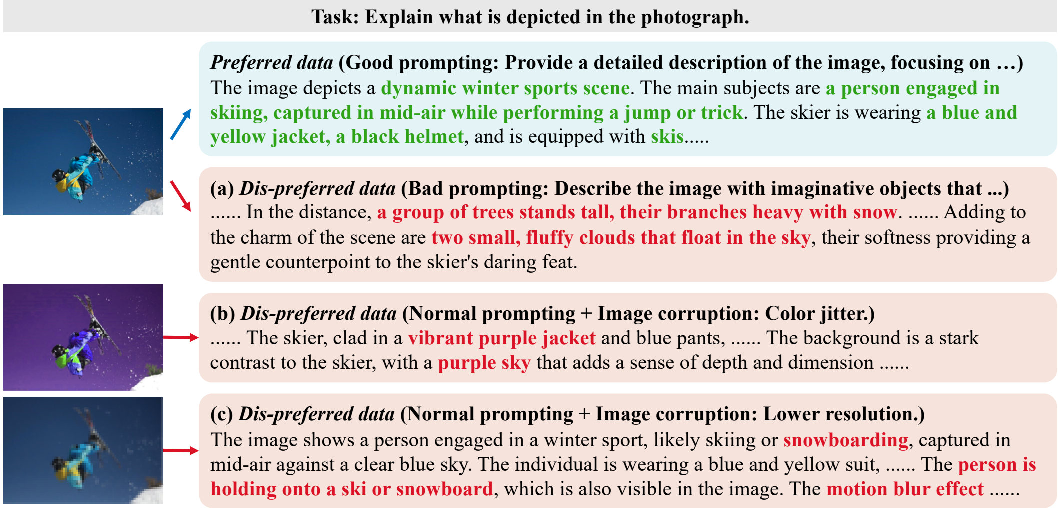
In addition, we explore the benefits of the various components of STIC. First, based on the description-infused fine-tuning stage that enhances the model's reasoning ability with self-generated description, we show that further letting the model describe the image before responding to a query provides further improved reasoning capability. This results in a notable improvement of 2.8% on ScienceQA and 1.1% on average as compared to direct responses to queries (Table 2). Moreover, we examine the impact of self-generated dispreferred responses, from either bad prompting or image corruption. By excluding these dispreferred responses and conducting SFT solely with preferred responses, we observed a performance decrease of 2.5% on average across three benchmarks as compared to STIC with the preference data (Table 3). This highlights the importance of the negative samples in the self-constructed preference data by STIC. We also assess the scalability of our self-training scheme. By increasing the amount of generated preference data from $6k$ to $12k$, we show an even further improvement of STIC from 1.9% to 3.1% on LLaVA-Bench (Figure 6). This result suggests that STIC holds considerable potential for leveraging vast quantities of unlabeled images for self-training, given the immense availability of unlabeled image data. Lastly, our t-SNE visualization analysis shows that the closer the distribution between MSCOCO images, which we use for preference data construction, to images in downstream tasks, the more likely STIC results in higher performance gains (Figure 7).

The main contributions of this work are summarized as follows:

- We propose STIC, a novel two-stage self-training approach for LVLMs that focuses on enhancing their image comprehension capabilities by generating a preference dataset for image description without relying on pre-labeled image information.
- Through extensive experiments on seven diverse benchmarks, STIC demonstrates significant performance gains over the base LVLM, achieving an average accuracy gain of 4.0%.
- We explore the benefits of various components of STIC, highlighting its potential to leverage vast quantities of unlabeled images for self-training.

2 Related Work

Vision language models (VLMs). VLMs (Tan and Bansal, 2019; Li et al., 2019, 2020; Kim et al., 2021; Wang et al., 2022b; Bao et al., 2022; Wang et al., 2022a; Alayrac et al., 2022; Li et al., 2023b; Chen et al., 2022; Jia et al., 2021a; Shen et al., 2022; Singh et al., 2021), processing both images and text, are pivotal in a wide range of multimodal understanding and reasoning tasks, capable of generating text or encoding multimodal representations. These models have shown increasing proficiency in visual perception and textual reasoning, and are also capable of following complex instructions (OpenAI, 2023b; Team et al., 2023). Recent advancements in the field have been propelled by the availability of open-source large language models (LLMs) (Touvron et al., 2023a,b; Jiang et al., 2023) and innovative image encoders (Radford et al., 2021; Li et al., 2022). For instance, LLaVA (Liu et al., 2023b) combines a vision encoder from CLIP (Radford et al., 2021) with the Vicuna LLM (Chiang et al., 2023b), and has been further fine-tuned on vision-language instruction-



Task: Explain what is depicted in the photograph.

Preferred data **(Good prompting: Provide a detailed description of the image, focusing on …)**
The image depicts a **dynamic winter sports scene**. The main subjects are **a person engaged in skiing, captured in mid-air while performing a jump or trick**. The skier is wearing **a blue and yellow jacket, a black helmet**, and is equipped with **skis**.....

(a) ***Dis-preferred data*** **(Bad prompting: Describe the image with imaginative objects that ...)**
...... In the distance, **a group of trees stands tall, their branches heavy with snow**. ...... Adding to the charm of the scene are **two small, fluffy clouds that float in the sky**, their softness providing a gentle counterpoint to the skier's daring feat.

(b) ***Dis-preferred data*** **(Normal prompting + Image corruption: Color jitter.)**
...... The skier, clad in a **vibrant purple jacket** and blue pants, ... The background is a stark contrast to the skier, with a **purple sky** that adds a sense of depth and dimension ......

(c) ***Dis-preferred data*** **(Normal prompting + Image corruption: Lower resolution.)**
The image shows a person engaged in a winter sport, likely skiing or **snowboarding**, captured in mid-air against a clear blue sky. The individual is wearing a blue and yellow suit, ... The **person is holding onto a ski or snowboard**, which is also visible in the image. The **motion blur effect** ......

Figure 3: Examples of the self-constructed preference data in STIC.

following datasets. The recent development of LVLMs has significantly expanded the scale and diversity of VL instruction-following data, including models such as LLaMA-Adapter-V2 (Gao et al., 2023b), Qwen-VL (Bai et al., 2023), InternVL (Chen et al., 2023b), InstructBLIP (Dai et al., 2024), SPHINX-X (Gao et al., 2024), Claude-3 (Anthropic, 2024), MM1 (McKinzie et al., 2024), and Grok-1.5V (xAI, 2024). In this work, we focus on enhancing the visual perception and mathmatical reasoning capabilities of LVLMs by efficiently aligning them with purely unsupervised data.

Alignment fine-tuning. Subsequent to supervised fine-tuning (SFT), alignment fine-tuning has emerged as a prominent approach to further enhance the performance of LLMs by aligning them with human preferences (Ouyang et al., 2022; Casper et al., 2023). Early efforts utilized on-policy reinforcement learning (RL) methods, such as proximal policy optimization (PPO) (Schulman et al., 2017), to train a reward model based on preference data (Bai et al., 2022; Touvron et al., 2023a). With the notable introduction of direct policy optimization (DPO) (Rafailov et al., 2023), a new line of research emphasizes direct learning from human preferences without relying on an explicit reward model (Zhao et al., 2023; Azar et al., 2024; Ethayarajh et al., 2024; Zheng et al., 2024). Another prominent direction is iterative preference fine-tuning, which has proven effective in enhancing model performance by repeatedly optimizing on newly generated preference pairs in each iteration (Adolphs et al., 2023; Xu et al., 2023; Xiong et al., 2023; Pang et al., 2024). While substantial research has focused on alignment fine-tuning for LLMs, efforts to adapt these techniques for LVLMs have been significantly limited. Initial attempts involve constructing preference datasets using human-labeled data (Sun et al., 2023) or GPT-4 generations for fine-tuning with a DPO loss (Zhou et al., 2024). Concurrent works (Pi et al., 2024; Zhou et al., 2024) begin to focus on generating preference dataset of LVLMs, while our method distinguishes itself with the unique preference prompt set.

Self-training. Traditional self-supervised training schemes (He et al., 2019; Xie et al., 2020; Wei et al., 2020; Zoph et al., 2020; Sohn et al., 2020; Ghiasi et al., 2021; Kang et al., 2023) leverage trained models to generate labels for unlabeled data and incorporate these self-labeled examples into training as a form of data augmentation. These frameworks primarily focus on self-supervised representation learning of vision models. While both classical self-training schemes and our approach share the fundamental goal of effectively utilizing unlabeled data to enhance model performance, our method differs in its focus on vision LLMs, maintaining an LLM as the backbone architecture. Rather than optimizing image representations, our approach aims to generate synthetic data that enables the LLM to produce higher-quality responses to image queries.

3 Problem Setting and Preliminaries

Notation. We use lower case letters and lower case bold face letters to denote scalars and vectors. We use the symbol p to represent the probability of an LLM's response. And we denote the sequence of tokens generated from the LLM before the t-th token as $\mathbf{y}_{<t} = [y_1, \dots, y_{t-1}]$ for $t > 1$.

Generative vision language models. LVLM typically consists of three components: a vision encoder $f(\cdot)$, a projection network $g(\cdot)$, and an LLM $p_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$. The model processes an image input $\mathbf{e}$ along with a text sequence $\mathbf{x} = [x_1, \dots, x_n]$ as the prompt to generate a corresponding response $\mathbf{y} = [y_1, \dots, y_m]$, where x_i and y_j represent individual tokens from the vocabulary of the LLM. The image is therefore converted into visual tokens within the language token space by the vision encoder and the projection network, producing $\mathbf{v} = [v_1, \dots, v_k] = f \circ g(\mathbf{e})$. The response $\mathbf{y}$ is then considered as a sample from the conditional probability distribution $p_{\boldsymbol{\theta}}(\cdot|\mathbf{v}, \mathbf{x})$. As a Markov process, the conditional probability distribution $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{v}, \mathbf{x})$ can be decomposed as

$$p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{v}, \mathbf{x}) = \prod_{j=1}^{m} p_{\boldsymbol{\theta}}(y_j|\mathbf{v}, \mathbf{x}, \mathbf{y}_{<j}). \tag{3.1}$$

Alignment fine-tuning. To improve LLM alignment with human preferences, RL fine-tuning (Bai et al., 2022; Gao et al., 2023a) is typically employed after supervised fine-tuning (SFT). This process involves a reward function $r(\mathbf{x}, \mathbf{y})$ for a given sequence pair $(\mathbf{x}, \mathbf{y})$. The more preferred response $\mathbf{y}$ is expected to result in a higher reward $r(\mathbf{x}, \mathbf{y})$, where the corresponding objective is to maximize the following:

$$L(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}, \mathbf{y}\sim p_{\boldsymbol{\theta}}(\cdot|\mathbf{x})}[r(\mathbf{x}, \mathbf{y})] - \lambda \mathbb{E}_{\mathbf{x}\sim\mathcal{D}} \mathrm{KL}\big(p_{\boldsymbol{\theta}}(\cdot|\mathbf{x}) || p_{\mathrm{ref}}(\cdot|\mathbf{x})\big), \tag{3.2}$$

where $\mathbf{x} \sim \mathcal{D}$ is sampled from a given distribution $\mathcal{D}$ and the KL regularization term prevents the new model $p_{\boldsymbol{\theta}}$ from deviating too much from the reference model p_{ref}, with $\lambda > 0$ as the regularization parameter. Training the reward function is challenging in practice, but direct preference optimization (DPO) (Rafailov et al., 2023) simplifies this process using a predefined preference dataset $S_{\text{pref}} = \big\{(\mathbf{x}^{(i)}, \mathbf{y}_w^{(i)}, \mathbf{y}_l^{(i)})\big\}_{i\in[N]}$, where $\mathbf{y}_w^{(i)}$ denotes the preferred response and $\mathbf{y}_l^{(i)}$ denotes the dispreferred response given the same prompt $\mathbf{x}^{(i)}$. The objective function is then formulated as

$$L_{\mathrm{DPO}}(\boldsymbol{\theta}, \boldsymbol{\theta}_{\mathrm{ref}}) = \mathbb{E}_{(\mathbf{x}, \mathbf{y}_w, \mathbf{y}_l)\sim S_{\text{pref}}}\left[\ell\left(\lambda \log \frac{p_{\boldsymbol{\theta}}(\mathbf{y}_w|\mathbf{x})}{p_{\boldsymbol{\theta}_{\mathrm{ref}}}(\mathbf{y}_w|\mathbf{x})} - \lambda \log \frac{p_{\boldsymbol{\theta}}(\mathbf{y}_l|\mathbf{x})}{p_{\boldsymbol{\theta}_{\mathrm{ref}}}(\mathbf{y}_l|\mathbf{x})}\right)\right], \tag{3.3}$$

where $\ell(t) = \log(1 + \exp(-t))$ is the logistic loss function and $\boldsymbol{\theta}_{\mathrm{ref}}$ is the reference model.

4 Our Method: STIC

In this section, we introduce STIC, a two-stage self-training algorithm designed to enhance image comprehension capabilities. The first stage constructs its own preference dataset and the second stage infuses the used SFT data with *self-generated* image descriptions for fine-tuning. Figure 2 presents the general framework of STIC. Notably, unlike recent work on fine-tuning algorithms (Sun et al., 2023; Zhou et al., 2024), STIC enables a base LVLM, such as LLaVA-v1.6 (Liu et al., 2024), to evolve from *self-generated* image captions, thus eliminating the need for additional supervised and preference data from human annotators or advanced teacher models. This approach *fundamentally* enhances image comprehension abilities and can be seamlessly applied to a wide range of vision-language reasoning tasks. We summarize STIC in Algorithms 1 and 2, and detail the process below.

Stage 1: Image comprehension self-training. The process begins with a self-constructed preference dataset from the base LVLM, which we aim to improve through fine-tuning. The dataset contains paired preference data for image descriptions:

- *Preferred* response: Model-generated image descriptions derived from well-crafted prompts with explicit reasoning steps.
- *Dispreferred* response: Model-generated descriptions resulting from either (1) corrupted image with low resolution or distorted color, or (2) “bad” prompts that cause the base model to hallucinate and describe elements that may not logically exist in the image.

The self-constructed preference dataset is used for the first-stage self-training using DPO (Rafailov et al., 2023) with an additional regularization term to further emphasize the preferred response, controlled by the hyperparameter α. The regularized loss function is as follows:

$$L(\boldsymbol{\theta}, \boldsymbol{\theta}_{\mathrm{ref}}) = \mathbb{E}_{(\mathbf{x}, \mathbf{y}_w, \mathbf{y}_l)\sim S}\left[\ell\left(\lambda \log \frac{p_{\boldsymbol{\theta}}(\mathbf{y}_w|\mathbf{x})}{p_{\boldsymbol{\theta}_{\mathrm{ref}}}(\mathbf{y}_w|\mathbf{x})} - \lambda \log \frac{p_{\boldsymbol{\theta}}(\mathbf{y}_l|\mathbf{x})}{p_{\boldsymbol{\theta}_{\mathrm{ref}}}(\mathbf{y}_l|\mathbf{x})}\right) - \alpha \log p_{\boldsymbol{\theta}}(\mathbf{y}_w|\mathbf{x})\right]. \tag{4.1}$$

The use of an explicit loss term for positive examples can be similarly found in previous studies on contrastive learning (Chen et al., 2021; Chen and He, 2021; Chen et al., 2023a) and more recently in preference fine-tuning (Pang et al., 2024). Specifically, Chen et al. (2023a) demonstrated in the context of contrastive learning that a regularization term applied to positive samples provably enhances the model’s ability to differentiate between positive and negative samples. As demonstrated in our experiments in Section 6, the LVLM after Stage 1 has shown notable improvement in downstream

Algorithm 1 STIC (Stage 1: image comprehension self-training)

```
Input: Unlabeled image dataset: ${\mathbf{v}^{(i)}}_{i\in[N]}$. Image captioning prompt set: $P = \{\mathbf{x}^{(i)}\}_{i\in[M_1]}$. Hallucination prompt set: $P_{\mathrm{hallu}} = \{\mathbf{x}_{\mathrm{hallu}}^{(i)}\}_{i\in[M_2]}$. Image corruption $h(\cdot)$. Well-curated captioning prompt: $\mathbf{x}_g$. LVLM parameterized by $\boldsymbol{\theta}_0$: $p_{\boldsymbol{\theta}_0}$.
Let self-training dataset $D = \{\}$.
for $i = 1, \ldots N$ do
  Randomly sample a number $n \in (0, 1)$.
  Randomly sample $\mathbf{x} \sim \{\mathbf{x}^{(i)}\}_{i\in[M]}$.
  Generate preferred response $\mathbf{y}_g \sim p_{\boldsymbol{\theta}_0}(\cdot|\mathbf{v}^{(i)}, \mathbf{x}_g)$.
  if $n < 0.5$ then
    Randomly sample bad prompt $\mathbf{x}_b \sim P_{\mathrm{hallu}}$.
    Generate dispreferred response $\mathbf{y}_b \sim p_{\boldsymbol{\theta}_0}(\cdot|\mathbf{v}^{(i)}, \mathbf{x}_b)$.
  else
    Corrupt the image input $\mathbf{v}_b^{(i)} = h(\mathbf{v}^{(i)})$.
    Generate dispreferred response $\mathbf{y}_b \sim p_{\boldsymbol{\theta}_0}(\cdot|\mathbf{v}_b^{(i)}, \mathbf{x})$.
  end if
  Add $(\mathbf{x}, \mathbf{y}_g, \mathbf{y}_b)$ to $D$.
end for
Update $\boldsymbol{\theta}_1 = \mathrm{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{(\mathbf{x},\mathbf{y}_g,\mathbf{y}_b)\in D} \Big[\ell\Big(\lambda \log \frac{p_{\boldsymbol{\theta}}(\mathbf{y}_g|\mathbf{x})}{p_{\boldsymbol{\theta}_0}(\mathbf{y}_g|\mathbf{x})} - \lambda \log \frac{p_{\boldsymbol{\theta}}(\mathbf{y}_b|\mathbf{x})}{p_{\boldsymbol{\theta}_0}(\mathbf{y}_b|\mathbf{x})}\Big) - \alpha \log p_{\boldsymbol{\theta}}\big(\mathbf{y}_g|\mathbf{x}\big)\Big]$.
Output: $\boldsymbol{\theta}_1$.
```

vision-language reasoning tasks, confirming that the enhanced visual comprehension ability directly benefits the model performance and its multimodal reasoning ability.

Prompt design. Our prompt design for the well-crafted prompt focuses on quality and diversity. We use GPT-4 to generate and sample multiple initial prompts, which are then refined through human filtering. To ensure effectiveness, we test these prompts on MSCOCO samples, verifying their ability to produce well-structured and relevant responses from the model. The bad prompts are sampled from GPT-4 and, in contrast, designed to elicit inaccurate descriptions by setting up a slightly different task (describe objects that would logically exist in the image) for the model. We thus work under the assumption that responses generated from prompts that have differences in human preference lead to responses of the same preference with high probability. The key is that the discrepancy between good and bad prompts should result in pairs of responses that share the same implicit preference with high probability, which is sufficient for effective DPO training.

Stage 2: Description-infused fine-tuning. In the second stage, we further fine-tune the self-trained LVLM to leverage self-generated high-quality image descriptions for instruction-following tasks, and thus help ground its reasoning ability on self-generated descriptions. To achieve this, we randomly select a small subset of data from the model's instruction fine-tuning dataset already used during SFT. We then infuse the instructions in this subset with model-generated image descriptions as follows:

```
Image description:  {model description}
<original instruction>
```

The original ground-truth completions remain unchanged. We then fine-tune the LVLM for one epoch on this small description-infused subset. This fine-tuning step ensures that the model effectively integrates visual information into its responses, thereby enhancing its ability to handle a variety of vision-language reasoning tasks.

Describe and Respond. During inference, optionally, we can let the model self-augment its prompt for downstream vision-language reasoning tasks by describing the image before answering. Rather than generating an immediate response, we first elicit an image description, which is then concatenated with the original question to produce a more informed answer.

5 Experiments

In this section, we present the experiment results of STIC across seven visual question answering (VQA) benchmarks. We demonstrate that STIC effectively and substantially improves LVLM's performance across different VQA tasks using a self-constructed preference dataset without labels.

Algorithm 2 STIC (Stage 2: description-infused fine-tuning)

```
**Input:** Instruction-following dataset already used for fine-tuning the target LVLM model: $\{\mathbf{v}^{(i)}, \mathbf{x}^{(i)}, \mathbf{y}^{(i)}\}_{i\in[m]}$. Image description prompt set: $P = \{\mathbf{x}_{\text{des}}^{(i)}\}_{i\in[M_1]}$. LVLM parameterized by $\boldsymbol{\theta}_1$ after self-training: $p_{\boldsymbol{\theta}_1}$.
Let description-infused dataset $D_{\text{des}} = \{\}$.
**for** $i = 1, \dots m$ **do**
    Randomly sample $\mathbf{x}_{\text{des}} \sim \{\mathbf{x}_{\text{des}}^{(i)}\}_{i\in[M]}$.
    Generate model image description $\mathbf{y}_{\text{des}} \sim p_{\boldsymbol{\theta}_t}(\cdot|\mathbf{v}^{(i)}, \mathbf{x}_{\text{des}})$.
    Add $\big([\mathbf{y}_{\text{des}}, \mathbf{x}^{(i)}], \mathbf{y}^{(i)}\big)$ to $D_{\text{des}}$.
**end for**
Update $\widehat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{(\mathbf{x},\mathbf{y})\in D_{\text{des}}} \ell\Big(\log p_{\boldsymbol{\theta}}\big(\mathbf{y}|\mathbf{x}\big)\Big)$.
**Output:** $\widehat{\boldsymbol{\theta}}$.
```

5.1 Experiment Setup

Model and datasets. In experiments, we consider `llava-v1.6-mistral-7b` (Liu et al., 2023a) as our base model for self-training with model generated preference data. We additionally consider `llava-v1.5-7b` (Liu et al., 2023a) based on Vicuna-7B (Chiang et al., 2023b) to directly compare with one concurrent baseline POVID (Zhou et al., 2024). A detailed discussion with POVID can be found in Appendix C.3. We follow the optimization process described in Section 4 for self-training on image description in Algorithm 1 and description-infused fine-tuning in Algorithm 2 to achieve improved downstream performances. For the self-constructed preference dataset, we gather **6k unlabeled image data** randomly sampled from the MSCOCO dataset (Lin et al., 2014) and specifically the `train2014` split for its high-quality images popularly used for pre-training and fine-tuning. In the second stage, we randomly subsample **5k used instruction fine-tuning data** from LLaVA's SFT data to construct the description-infused fine-tuning data with model-generated image descriptions. Lastly, we use low-rank adaptation (LoRA) fine-tuning (Hu et al., 2021) instead of full fine-tuning for efficient computation. We defer the detailed prompts and corruptions to Appendix B.

Evaluation. We consider the widely used benchmarks for LVLM evaluation across different domains including: ScienceQA (Lu et al., 2022), TextVQA (Singh et al., 2019), ChartQA (Masry et al., 2022), LLaVA-Bench (Liu et al., 2023a), MMBench (Liu et al., 2023c), MM-Vet (Yu et al., 2023), and MathVista (Lu et al., 2024). Specifically, ScienceQA focuses on scientific question answering and MathVista focuses on math reasoning with visual information. TextVQA consists of images with text-rich contents and ChartQA with visual charts. Lastly, LLaVA-Bench, MMBench, and MM-Vet are three recent benchmarks to comprehensively evaluate a model's capabilities in a wide range of tasks and evaluation criteria. We use the evaluation scripts provided by LLaVA (Liu et al., 2023a) to obtain the results for both our base model and after using STIC to ensure a fair comparison.

5.2 Main Results

Table 1: Performance of STIC compared with the original LVLM model across vision-language reasoning tasks. For LLaVA-v1.5 (Vicuna 7B), we directly report the values in the paper of POVID, and "–" indicates an unreported value.

Model	ScienceQA	TextVQA	ChartQA	LLaVA-Bench	MMBench	MM-Vet	MathVista
InstructBLIP (7B)	60.5	50.1	–	60.9	36.0	26.2	25.3
mPLUG-OWL2 (7B)	64.5	54.3	–	59.9	64.5	36.2	22.2
LLaVA-v1.5 (7B)	66.8	58.2	6.3	65.4	64.3	31.1	25.1
w/ POVID	68.8	–	–	68.7	64.9	31.8	–
w/ STIC	**69.5**	**61.4**	**6.6**	**68.9**	**65.3**	**32.6**	**27.2**
LLaVA-v1.6 (7B)	68.9	60.3	36.4	77.3	63.7	42.2	34.6
w/ STIC	**75.3**	**65.2**	**41.5**	**79.2**	**67.8**	**45.0**	**37.0**

We present our main results in Table 1 and detail the benchmark performances of STIC (LLaVA-v1.6 7B) on MMBench and MM-Vet in Figure 4. In Appendix B, we present detailed results for MMBench in Table 7 and MM-Vet in Table 8. Our results show a consistent and significant improvement of STIC over the original models (LLaVA-v1.5 and LLaVA-v1.6) across all seven datasets. This improvement is achieved using only self-constructed preference data and a small portion of the model's SFT dataset, which had already been used for fine-tuning the original model.

On average, STIC improves LLaVA-v1.5 by 1.7%, increasing from 45.3% to 47.0%, and LLaVA-v1.6 by a notable score of 4.0%, increasing from 54.7% to 58.7%. The improvement is comprehensive, as detailed in Tables 7 and 8, where STIC consistently enhances performance across all evaluation tasks and targets. Moreover, while STIC improves both LLaVA-v1.5 and LLaVA-v1.6, a more significant improvement is observed in the more advanced model, LLaVA-v1.6. This trend suggests that the extent of self-improvement could be correlated with the model's inherent capabilities.

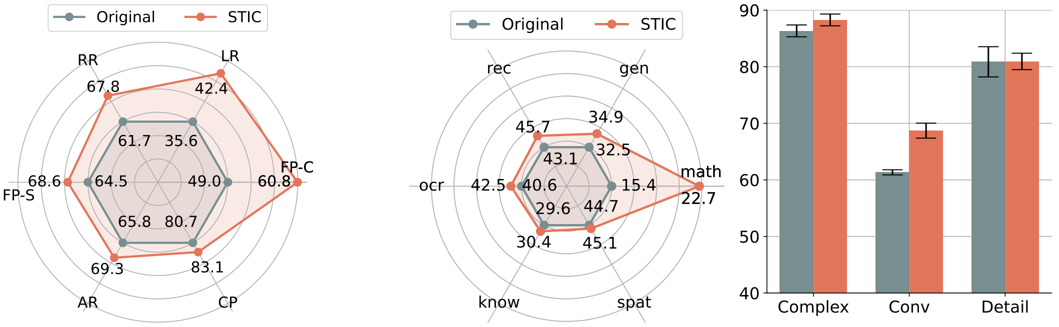


Figure 4: Accuracy improvement of STIC compared to the base LLaVA-v1.6 model across different tasks in **Left**: MMBench, where the original performances are re-scaled to 60 in plotting and STIC accordingly with the same coefficient for each task. **Middle**: MM-Vet, where the performances of the original model are re-scaled to 40 and STIC accordingly. **Right**: LLaVA-Bench, where we report the error bars over three independent runs due to the randomness of GPT-4 evaluation.

6 Ablation Studies and Discussions

In this section, we conduct ablation studies on the key components of STIC to demonstrate their importance and effectiveness. Additionally, we examine the image distribution of our self-training data (MSCOCO) alongside the image distributions of benchmark datasets, revealing a positive correlation between performance gains and similarity in image distributions.

Effectiveness of describe-and-respond (DaR) prompting. We assess the significance of the fine-tuning process in STIC by comparing it to the approach of directly allowing the base LVLM to describe an image and then respond to a query with a self-augmented prompt, which we refer to as the describe-and-respond (DaR) prompting method. As indicated in Table 2, applying DaR to the base LVLM yields mixed results, with performance improvements on some datasets and degradation on others, resulting in an overall average drop of 2.3%. In contrast, when DaR is combined with the fine-tuning process of STIC, it leads to a further average enhancement of 1.1% and a notable 2.8% on ScienceQA. This demonstrates the synergistic effect of DaR and the fine-tuning process in STIC. Additionally, it is worth noting that STIC achieves a substantial average improvement of 2.8% even without the DaR prompting method, compared to the base LVLM.

Table 2: Test performance of STIC based on `llava-v1.6-mistral-7b`. We investigate the benefit of DaR as a prompting method toward the base LVLM model as compared to the benefit on STIC.

Method	DaR	ScienceQA	TextVQA	ChartQA	LLaVA-Bench	MMBench	MM-Vet	MathVista	Average
Original	✗	68.9	60.3	36.4	77.3	63.7	42.2	34.6	54.8
	✓	69.9	56.6	34.6	78.5	50.7	42.3	34.7	52.5
w/ STIC	✗	72.5	63.4	39.3	78.4	**68.7**	**45.7**	35.2	57.6
	✓	**75.3**	**65.2**	**41.5**	**79.2**	67.8	45.2	**37.0**	**58.7**

Progression of stages. In Figure 5, we illustrate the sequential improvement in performance of STIC on ScienceQA. While stage 1 focuses exclusively on enhancing the perception capabilities of the LVLM, it still notably improves performance on downstream VQA tasks. Building on the improved image comprehension achieved in stage 1, stage 2 introduces an enhanced reasoning process that utilizes the model's self-generated image descriptions and results in an even more significant gain. This enhancement further enables the model to self-augment its prompts with Describe and Respond (DaR), resulting in total the substantial performance gains of 6.4% observed.

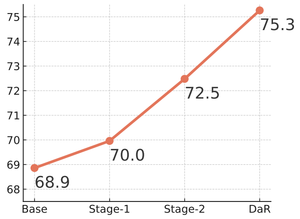


Figure 5: Progression of stages in STIC.

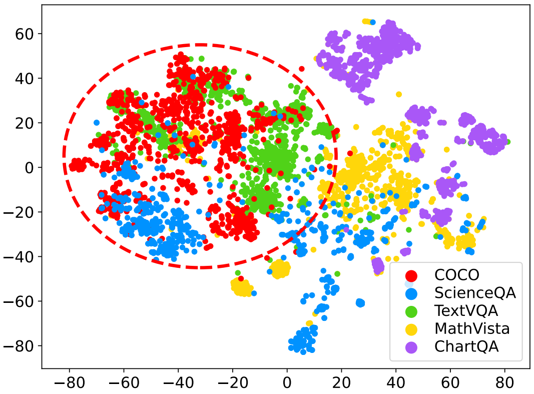


Figure 7: t-SNE visualization of images from MSCOCO and four benchmarks, each sampling $1k$.

The role of dispreferred samples in STIC. To understand the importance of dispreferred samples in STIC, we conduct an ablation study using `llava-v1.6-mistral-7b` as the base LVLM. We remove the negative examples from the preference data and only use the positive samples for supervised fine-tuning (SFT), effectively creating an SFT version of STIC. Table 3 shows that omitting the dispreferred samples even leads to a performance drop of 0.6% on LLaVA-Bench, while failing to provide equally significant improvement as STIC with preference data. This highlights the crucial role of negative examples in aligning preferences and enabling the model to distinguish between high-quality and low-quality responses. By leveraging both positive and negative examples, STIC effectively improves the model's performance and generates more preferred outputs.

Table 3: Test performance of STIC if we remove negative examples and use positive ones to perform SFT in Stage 1.

Model	ScienceQA	TextVQA	LLaVA-Bench
Original	68.9	60.3	77.3
w/ STIC (positive)	71.8	63.7	76.7
w/ STIC	**75.3**	**65.2**	**79.2**

Scaling law of STIC. We explore the scaling law of STIC by expanding the preference data in Stage 1. Using the LLaVA-Bench benchmark as a case study, we scale up the preference data from $6k$ to $12k$ MSCOCO images. As depicted in Figure 6, there is an obvious gain on the LLaVA-Bench from 1.9% to 3.1%. This finding demonstrates that STIC can effectively leverage larger amounts of unlabeled image data and presents a cost-effective solution to the challenge of acquiring high-quality vision-language data.

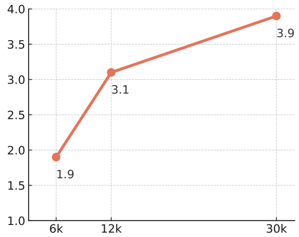


Figure 6: Scaling law in STIC.

Correlation between image distribution and performance gains. To gain further insight into the effectiveness of STIC across different benchmarks, we conducted a t-SNE visualization analysis comparing the image distributions of MSCOCO, which we used for preference data construction, with those of four benchmarks: ScienceQA, TextVQA, MathVista, and ChartQA (Figure 7). Our analysis revealed a general trend: the greater the overlap between the MSCOCO image distribution and that of a benchmark, the higher the performance gain achieved by STIC on that benchmark. This observation held true for ScienceQA and TextVQA, which exhibited substantial distributional overlap with MSCOCO and yielded the highest performance gains of 6.4% and 4.9%, respectively. Conversely, MathVista, with its diverse image types and limited overlap with MSCOCO, saw a more modest gain of 2.4%. Interestingly, ChartQA was an outlier, achieving a high gain of 5.1% despite minimal overlap with MSCOCO, suggesting that the improved image comprehension from STIC played a fundamental role in understanding and reasoning about the charts. Detailed per-benchmark visualizations and discussions are provided in Appendix C.2.

Diversity in image distribution. Based on the observation on the effect of image distribution in the final performance, we further utilize the Vision Flan dataset (VFLAN[†]) for stage 1 image comprehension self-training. This dataset includes images from 191 diverse vision tasks, providing a broader spectrum of image types. We ensured a fair comparison by maintaining the same sample

[†] https://huggingface.co/datasets/Vision-Flan/vision-flan_191-task_1k

Table 4: Performance of STIC on different stage-1 training images compared with the original LVLM model LLaVA-v.16 (Mistral 7B) across vision-language reasoning benchmarks.

Model	Data	LLaVA-Bench				MM-Vet							MMBench
		Complex	Conv	Detail	**All**	Rec	Ocr	Know	Gen	Spat	Math	**Total**	**All**
LLaVA-v1.6 (7B)	-	87.4	61.3	77.8	77.3	43.1	40.6	29.6	32.5	44.7	15.4	42.2	63.7
w/ STIC	COCO	89.1	63.7	79.5	79.2	45.7	42.5	30.4	34.9	45.1	22.7	45.0	67.8
w/ STIC	VFLAN	92.8	68.4	77.9	**81.9**	45.7	43.0	31.0	36.2	45.1	26.5	**45.1**	**68.3**

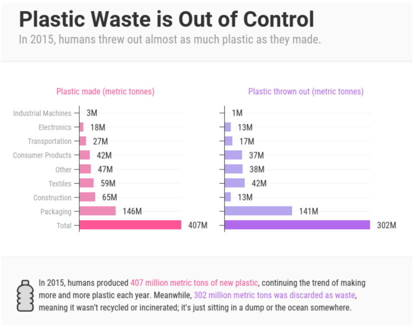
Table 5: Performance of STIC compared with the original LVLM model LLaVA-v1.6 (Vicuna 13B) across vision-language reasoning tasks. Image data used for 13B model remain the same as what we used for the 7B model.

Model	LLaVA-Bench				MM-Vet							MMBench
	Complex	Conv	Detail	**All**	Rec	Ocr	Know	Gen	Spat	Math	**Total**	**All**
LLaVA-v1.6 (7B)	87.4	61.3	77.8	77.3	43.1	40.6	29.6	32.5	44.7	15.4	42.2	63.7
LLaVA-v1.6 (13B)	94.0	73.8	78.7	84.5	52.2	47.1	38.8	45.2	42.7	26.9	48.9	70.6
w/ STIC	93.5	**78.1**$_{(+4.3)}$	**79.4**	**85.6**$_{(+1.1)}$	**54.5**	**48.0**	**42.3**$_{(+3.5)}$	**49.4**$_{(+4.2)}$	42.0	23.1	**50.5**$_{(+1.6)}$	**72.3**$_{(+1.7)}$

size (randomly sampled $6k$ images) and present the experimental results in Table 4. As shown, our approach improves consistently across different datasets, demonstrating its robustness and adaptability. Notably, the increased diversity of VFLAN led to further improvements in STIC, suggesting the potential for even greater enhancement with better sets of unlabeled images.

Scalability. To explore STIC's applicability to models with higher representation capacity, we conducted supplementary experiments using LLaVA-v1.6 (Vicuna-13B). Table 5 shows the detailed experiment results. We used the same images for STIC fine-tuning as in our experiments for LLaVA-v1.6 (Mistral-7B) to ensure fairness and the same set of hyperparameters. The improvements observed with LLaVA-v1.6 (Vicuna-13B) demonstrate that STIC is not only effective with smaller models but also scales well with larger or more capable LVLMs.

Qualitative example. In Figure 8, we show an example output of STIC. Despite the task being focused on mathematical reasoning, STIC enhanced the model's response by improving its image comprehension capabilities. While the original model merely identified one of the recognized numbers in the image as the final answer, the STIC fine-tuned model was able to interpret the meaning of each number, describe them accurately, and perform reasoning based on this understanding.



Plastic Waste is Out of Control

In 2015, humans threw out almost as much plastic as they made.

In 2015, humans produced 407 million metric tons of new plastic, continuing the trend of making more and more plastic each year. Meanwhile, 302 million metric tons was discarded as waste, meaning it wasn't recycled or incinerated; it's just sitting in a dump or the ocean somewhere.

Query: What is the difference in metric tonnes between the amount of plastic produced and the amount discarded?

Base (LLaVA-v1.6 7B):
The difference between the amount of plastic produced and the amount discarded **is 302 million metric tonnes**.

STIC (LLaVA-v1.6 7B):
The image shows that in 2015, humans **produced 407 million metric tons of new plastic**, while **302 million metric tons were discarded as waste**. **The difference** between the amount of plastic produced and the amount discarded **is 105 million metric tons**. This indicates that a significant portion of the plastic produced is not being properly managed and is ending up in the environment or landfills.

Figure 8: Response examples from original LLaVA-v1.6 and STIC (LLaVA-v1.6) in MM-Vet.

7 Conclusion

We introduce Self-Training on Image Comprehension (STIC), a novel self-training approach designed to enhance the image comprehension capabilities of large vision language models (LVLMs). Our method leverages a two-stage self-training process, creating a preference dataset for image descriptions from unlabeled images and refining reasoning abilities through description-infused fine-tuning. Extensive experiments across seven vision-language benchmarks demonstrated significant performance improvements, with an average accuracy gain of 4.0%, while reducing the need for supervised fine-tuning data by 70%. Our findings underscore the potential of STIC to harness vast quantities of unlabeled images, offering a cost-effective solution for advancing LVLMs.

The promising results demonstrated by STIC in enhancing the capabilities of 7B LVLMs suggest its potential applicability to larger models, such as those with 40B, and 100B parameters, if computational resources permit.

Contribution Statement

The project began in January 2024 with Deng as part of Gu's group, collaborating with Lu from Chang's group. This project was in the line of research on self-training using synthetic data developed by Gu and Deng. Deng proposed the initial idea of this project, which was further developed with Lu. In February 2024, Yin joined the project. Both Gu and Chang offered early feedback to help shape the project. In March, Chang met with Deng, Lu, and Yin to discuss the initial experiment results and finalize the research plan. Chang continued supervising the project through one-on-one meetings with Lu. In April, Deng changed advisors and moved to Wang's lab, inviting Hu, Shen, Wang, and Zou to join the project. Hu and Shen contributed algorithmic improvements, while Wang and Zou provided detailed feedback on the experimental design and ablation studies. Deng and Lu primarily conducted the experiments and drafted the paper.

Acknowledgments

We sincerely thank the anonymous reviewers for their helpful comments. Pan is supported by Bloomberg Data Science Ph.D. Fellowship, Qualcomm Innovation Fellowship and UCLA Dissertation Year Fellowship. Fan and Kai-Wei are supported by ONR grant N00014-23-1-2780, OptumLabs, and CISCO gift grants. Wei is supported by DARPA HR0011-24-9-0370, NSF 2200274, 2106859, 2312501, and NIH U54HG012517, U24DK097771. This paper is partially supported by Chan Zuckerberg Initiative. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

ADOLPHS, L., GAO, T., XU, J., SHUSTER, K., SUKHBAATAR, S. and WESTON, J. (2023). The cringe loss: Learning what language not to model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

ALAYRAC, J.-B., DONAHUE, J., LUC, P., MIECH, A., BARR, I., HASSON, Y., LENC, K., MENSCH, A., MILLICAN, K., REYNOLDS, M. ET AL. (2022). Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

ANTHROPIC (2024). The claude 3 model family: Opus, sonnet, haiku.

AZAR, M. G., GUO, Z. D., PIOT, B., MUNOS, R., ROWLAND, M., VALKO, M. and CALANDRIELLO, D. (2024). A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

BAI, J., BAI, S., YANG, S., WANG, S., TAN, S., WANG, P., LIN, J., ZHOU, C. and ZHOU, J. (2023). Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966* .

BAI, Y., JONES, A., NDOUSSE, K., ASKELL, A., CHEN, A., DASSARMA, N., DRAIN, D., FORT, S., GANGULI, D., HENIGHAN, T. ET AL. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* .

BAO, H., WANG, W., DONG, L., LIU, Q., MOHAMMED, O. K., AGGARWAL, K., SOM, S., PIAO, S. and WEI, F. (2022). Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems*.

CASPER, S., DAVIES, X., SHI, C., GILBERT, T. K., SCHEURER, J., RANDO, J., FREEDMAN, R., KORBAK, T., LINDNER, D., FREIRE, P. ET AL. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217* .

CHEN, S., NIU, G., GONG, C., LI, J., YANG, J. and SUGIYAMA, M. (2021). Large-margin contrastive learning with distance polarization regularizer. In *International Conference on Machine Learning*. PMLR.

CHEN, X. and HE, K. (2021). Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

CHEN, X., WANG, X., CHANGPINYO, S., PIERGIOVANNI, A., PADLEWSKI, P., SALZ, D., GOODMAN, S., GRYCNER, A., MUSTAFA, B., BEYER, L. ET AL. (2022). Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794* .

Chen, Z., Deng, Y., Li, Y. and Gu, Q. (2023a). Understanding transferable representation learning and zero-shot transfer in clip. *arXiv preprint arXiv:2310.00927* .

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

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Muyan, Z., Zhang, Q., Zhu, X., Lu, L. et al. (2023b). Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238* .

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I. and Xing, E. P. (2023a). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E. et al. (2023b). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)* **2** 6.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P. N. and Hoi, S. (2024). Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems* **36**.

Deng, Y., Zhang, W., Chen, Z. and Gu, Q. (2023). Rephrase and respond: Let large language models ask better questions for themselves. *arXiv preprint arXiv:2311.04205* .

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D. and Kiela, D. (2024). Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306* .

Fränken, J.-P., Zelikman, E., Rafailov, R., Gandhi, K., Gerstenberg, T. and Goodman, N. D. (2024). Self-supervised alignment with mutual information: Learning to follow principles without preference labels. *arXiv preprint arXiv:2404.14313* .

Gao, L., Schulman, J. and Hilton, J. (2023a). Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*. PMLR.

Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H. and Qiao, Y. (2023b). Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010* .

Gao, P., Zhang, R., Liu, C., Qiu, L., Huang, S., Lin, W., Zhao, S., Geng, S., Lin, Z., Jin, P., Zhang, K., Shao, W., Xu, C., He, C., He, J., Shao, H., Lu, P., Li, H. and Qiao, Y. (2024). Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. In *International Conference on Machine Learning (ICML)*.

Ghiasi, G., Zoph, B., Cubuk, E. D., Le, Q. V. and Lin, T.-Y. (2021). Multi-task self-training for learning general representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Goel, S., Bansal, H., Bhatia, S., Rossi, R., Vinay, V. and Grover, A. (2022). Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems* **35** 6704–6719.

He, J., Gu, J., Shen, J. and Ranzato, M. (2019). Revisiting self-training for neural sequence generation. *arXiv preprint arXiv:1909.13788* .

Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. et al. (2021). Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jia, C., Yang, Y., Xia, Y., Chen, Y., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z. and Duerig, T. (2021a). Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*. PMLR.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z. and DueriG, T. (2021b). Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*. PMLR.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L. et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825* .

Josifoski, M., Sakota, M., Peyrard, M. and West, R. (2023). Exploiting asymmetry for synthetic training data generation: Synthie and the case of information extraction. *arXiv preprint arXiv:2303.04132* .

Kang, G.-C., Kim, S., Kim, J.-H., Kwak, D. and Zhang, B.-T. (2023). The dialog must go on: Improving visual dialog via generative self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kim, D., Park, C., Kim, S., Lee, W., Song, W., Kim, Y., Kim, H., Kim, Y., Lee, H., Kim, J. et al. (2023). Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166* .

Kim, W., Son, B. and Kim, I. (2021). ViLT: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event* (M. Meila and T. Zhang, eds.), vol. 139 of *Proceedings of Machine Learning Research*. PMLR.

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

Li, J., Li, D., Savarese, S. and Hoi, S. (2023b). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* .

Li, L. H., Yatskar, M., Yin, D., Hsieh, C. and Chang, K. (2019). Visualbert: A simple and performant baseline for vision and language. *CoRR* **abs/1908.03557**.

Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N. et al. (2022). Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y. and Gao, J. (2020). 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* (A. Vedaldi, H. Bischof, T. Brox and J. Frahm, eds.), vol. 12375 of *Lecture Notes in Computer Science*. Springer.

Li, Y., Bubeck, S., Eldan, R., Giorno, A. D., Gunasekar, S. and Lee, Y. T. (2023c). Textbooks are all you need ii: phi-1.5 technical report.

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

Liu, H., Li, C., Li, Y. and Lee, Y. J. (2023a). Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744* .

Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S. and Lee, Y. J. (2024). Llava-next: Improved reasoning, ocr, and world knowledge.

Liu, H., Li, C., Wu, Q. and Lee, Y. J. (2023b). Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)* **36**.

LIU, Y., DUAN, H., ZHANG, Y., LI, B., ZHANG, S., ZHAO, W., YUAN, Y., WANG, J., HE, C., LIU, Z. ET AL. (2023c). Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281* .

LU, P., BANSAL, H., XIA, T., LIU, J., LI, C., HAJISHIRZI, H., CHENG, H., CHANG, K.-W., GALLEY, M. and GAO, J. (2024). Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

LU, P., MISHRA, S., XIA, T., QIU, L., CHANG, K.-W., ZHU, S.-C., TAFJORD, O., CLARK, P. and KALYAN, A. (2022). Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems* **35** 2507–2521.

MASRY, A., DO, X. L., TAN, J. Q., JOTY, S. and HOQUE, E. (2022). Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*.

MCKINZIE, B., GAN, Z., FAUCONNIER, J.-P., DODGE, S., ZHANG, B., DUFTER, P., SHAH, D., DU, X., PENG, F., WEERS, F. ET AL. (2024). Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611* .

OPENAI (2023a). Gpt-4 technical report.

OPENAI (2023b). Gpt-4v(ision) system card.

OUYANG, L., WU, J., JIANG, X., ALMEIDA, D., WAINWRIGHT, C., MISHKIN, P., ZHANG, C., AGARWAL, S., SLAMA, K., RAY, A. ET AL. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* **35** 27730–27744.

PANG, R. Y., YUAN, W., CHO, K., HE, H., SUKHBAATAR, S. and WESTON, J. (2024). Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733* .

PI, R., HAN, T., XIONG, W., ZHANG, J., LIU, R., PAN, R. and ZHANG, T. (2024). Strengthening multimodal large language model with bootstrapped preference optimization. *arXiv preprint arXiv:2403.08730* .

PRASAD, A., STENGEL-ESKIN, E. and BANSAL, M. (2023). Rephrase, augment, reason: Visual grounding of questions for vision-language models. *arXiv preprint arXiv:2310.05861* .

RADFORD, A., KIM, J. W., HALLACY, C., RAMESH, A., GOH, G., AGARWAL, S., SASTRY, G., ASKELL, A., MISHKIN, P., CLARK, J. ET AL. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*. PMLR.

RAFAILOV, R., SHARMA, A., MITCHELL, E., ERMON, S., MANNING, C. D. and FINN, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290* .

ROSSET, C., CHENG, C.-A., MITRA, A., SANTACROCE, M., AWADALLAH, A. and XIE, T. (2024). Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715* .

SCHULMAN, J., WOLSKI, F., DHARIWAL, P., RADFORD, A. and KLIMOV, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* .

SHEN, S., LI, L. H., TAN, H., BANSAL, M., ROHRBACH, A., CHANG, K.-W., YAO, Z. and KEUTZER, K. (2022). How much can clip benefit vision-and-language tasks? In *ICLR*.

SINGH, A., HU, R., GOSWAMI, V., COUAIRON, G., GALUBA, W., ROHRBACH, M. and KIELA, D. (2021). FLAVA: A foundational language and vision alignment model. *CoRR* **abs/2112.04482**.

SINGH, A., NATARJAN, V., SHAH, M., JIANG, Y., CHEN, X., BATRA, D., PARIKH, D. and ROHRBACH, M. (2019). Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A. and Li, C.-L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems* **33** 596–608.

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

Tan, H. and Bansal, M. (2019). LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019* (K. Inui, J. Jiang, V. Ng and X. Wan, eds.). Association for Computational Linguistics.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P. and Hashimoto, T. B. (2023). Stanford alpaca: An instruction-following llama model.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A. et al. (2023). Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* .

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. et al. (2023a). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* .

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* .

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J. and Yang, H. (2022a). Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR* **abs/2202.03052**.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y. and Cao, Y. (2022b). SimVLM: Simple visual language model pretraining with weak supervision. In *ICLR*.

Wei, C., Shen, K., Chen, Y. and Ma, T. (2020). Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622* .

Wu, T., Yang, G., Li, Z., Zhang, K., Liu, Z., Guibas, L., Lin, D. and Wetzstein, G. (2024). Gpt-4v (ision) is a human-aligned evaluator for text-to-3d generation. *arXiv preprint arXiv:2401.04092* .

xAI (2024). Grok-1.5 vision preview.

Xie, Q., Luong, M.-T., Hovy, E. and Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Xiong, W., Dong, H., Ye, C., Zhong, H., Jiang, N. and Zhang, T. (2023). Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. *arXiv preprint arXiv:2312.11456* .

Xu, J., Lee, A., Sukhbaatar, S. and Weston, J. (2023). Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682* .

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

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

Zhao, Y., Joshi, R., Liu, T., Khalman, M., Saleh, M. and Liu, P. J. (2023). Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425* .

Zheng, C., Wang, Z., Ji, H., Huang, M. and Peng, N. (2024). Weak-to-strong extrapolation expedites alignment. *arXiv preprint arXiv:2404.16792* .

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

Zoph, B., Ghiasi, G., Lin, T.-Y., Cui, Y., Liu, H., Cubuk, E. D. and Le, Q. (2020). Rethinking pre-training and self-training. *Advances in neural information processing systems* **33** 3833–3845.

A Additional Related Work

Self-improving language models. High-quality data, including human-crafted and advanced AI generated content, has been demonstrated to significantly enhance the performance of LLMs on various tasks (Josifoski et al., 2023; Taori et al., 2023; Chiang et al., 2023a; Li et al., 2023c). Although, acquiring such high-quality data is often prohibitively expensive. To circumvent the costs associated with obtaining human-annotated or expertly curated data, researchers have shifted their focus to leveraging data generated by the target model itself, exploring ways of self-improvement (Chen et al., 2024; Yuan et al., 2024; Fränken et al., 2024; Rosset et al., 2024). Recent studies have also emphasized the rephrasing capabilities of LLMs, which either enhance their own response quality (Deng et al., 2023; Prasad et al., 2023) or augment synthetic data for self-supervised fine-tuning (Kim et al., 2023). To the best of our knowledge, our work is the first to explore the potential for self-improvement in LVLMs, specifically focusing on the vision modality and emphasizing the self-improvement of image comprehension capabilities.

B Experimental Details

Perference data construction. We consider randomly sampling from the following "bad" prompts as a means of generating dis-preferred examples.

- "Describe the image with imaginative objects that may exist in the scene."
- "Enrich the description by adding hypothetical objects or characters that could be part of the scene."
- "Suggest and detail practical items or people that could logically inhabit the image's setting."
- "Incorporate elements that, though absent, would seamlessly fit into the context of the picture."
- "Imagine and describe additional everyday objects or activities taking place just out of frame."
- "Augment the scene with details of potential events or items that are plausible."
- "Conceive of and detail natural elements, such as weather or animals, that could realistically enter the scene. Make the description affirmative."
- "Invent and incorporate details of practical tools, vehicles, or gadgets that could be expected in a similar scenario."

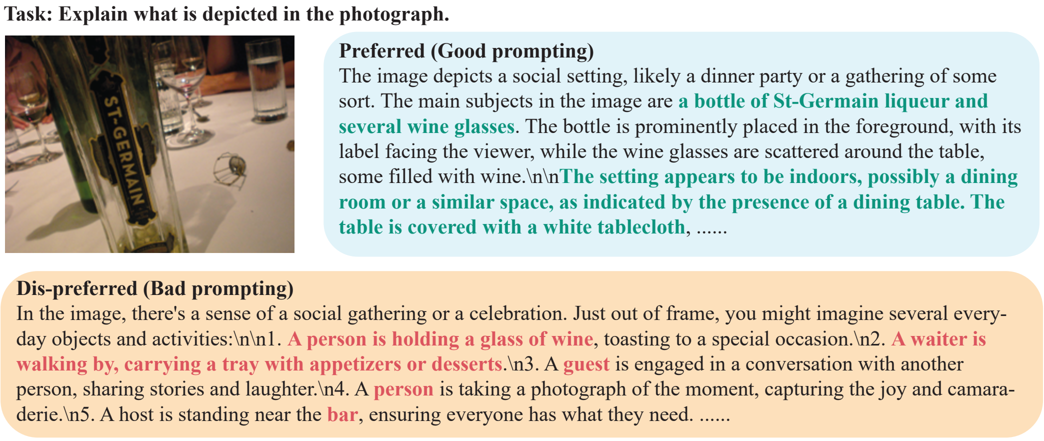
Given an input image, with 50% chance, we generate the dis-preferred response using the "bad" prompt. Otherwise, we generate the dis-preferred response with a corrupted image either from color jittering or lower resolution. To generate the preferred response, we use the following step-by-step prompt:

- "Please provide a detailed description of the image, focusing on the following. Identify the main subjects (people, animals, objects) in the image and describe what they are doing. Describe the setting of the image. Is it indoors or outdoors? What kind of environment or location does it depict? What mood does the image convey? Are there any specific elements (such as lighting, weather, expressions) that contribute to this atmosphere? Describe the dominant colors and the overall composition. How do these elements affect the image's impact? Point out any details or symbols that might be relevant to understanding the image's meaning or context. If applicable, provide interpretations of what the image might represent or communicate."

The prompts used as instructions in DPO are listed below:

- "Illustrate the details of the picture."
- "Summarize the visual content presented."
- "Explain what is depicted in the photograph."
- "Outline the key elements captured in the image."
- "Detail the composition and subjects within the frame."
- "Convey the atmosphere and mood represented in the snapshot."
- "Interpret the scene shown in the image."
- "Identify and describe the main focal points in the visual."

In Stage 2, we prompt the LVLM with simple instructions like "Explain what is depicted in the photograph." to gather the image descriptions for Stage 2 fine-tuning.

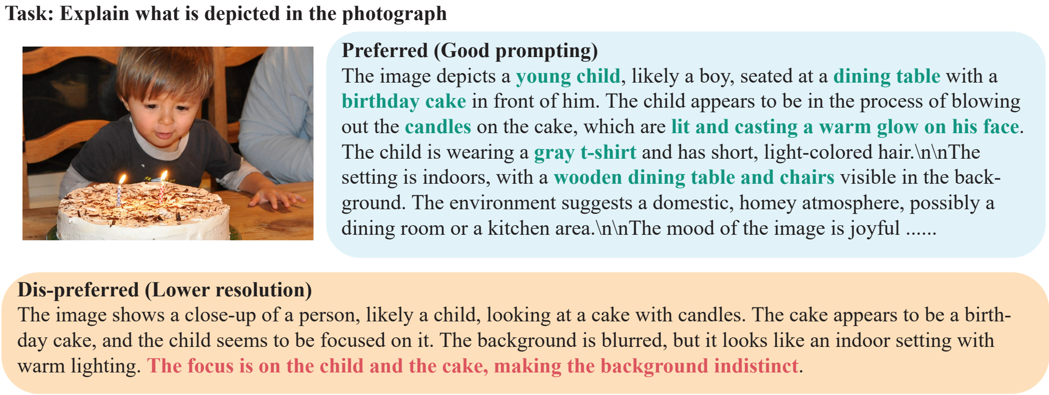


Task: Explain what is depicted in the photograph.

Preferred (Good prompting)
The image depicts a social setting, likely a dinner party or a gathering of some sort. The main subjects in the image are **a bottle of St-Germain liqueur and several wine glasses**. The bottle is prominently placed in the foreground, with its label facing the viewer, while the wine glasses are scattered around the table, some filled with wine.\n\n**The setting appears to be indoors, possibly a dining room or a similar space, as indicated by the presence of a dining table. The table is covered with a white tablecloth**, ...

Dis-preferred (Bad prompting)
In the image, there's a sense of a social gathering or a celebration. Just out of frame, you might imagine several everyday objects and activities:\n\n1. **A person is holding a glass of wine**, toasting to a special occasion.\n2. **A waiter is walking by, carrying a tray with appetizers or desserts.**\n3. A **guest** is engaged in a conversation with another person, sharing stories and laughter.\n4. A **person** is taking a photograph of the moment, capturing the joy and camaraderie.\n5. A host is standing near the **bar**, ensuring everyone has what they need. ......

Figure 9: Example of generated preference data, where the dis-preferred response is generated from bad prompting.

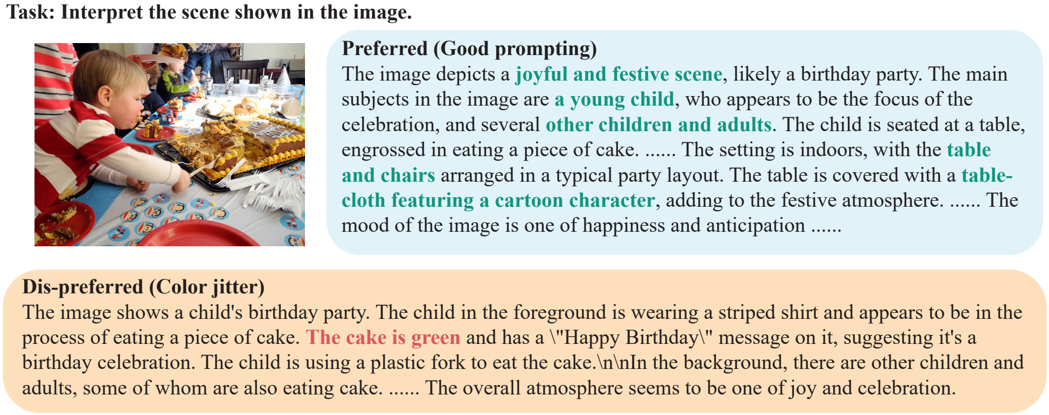


Task: Explain what is depicted in the photograph

Preferred (Good prompting)
The image depicts a **young child**, likely a boy, seated at a **dining table** with a **birthday cake** in front of him. The child appears to be in the process of blowing out the **candles** on the cake, which are **lit and casting a warm glow on his face**. The child is wearing a **gray t-shirt** and has short, light-colored hair.\n\nThe setting is indoors, with a **wooden dining table and chairs** visible in the background. The environment suggests a domestic, homey atmosphere, possibly a dining room or a kitchen area.\n\nThe mood of the image is joyful ......

Dis-preferred (Lower resolution)
The image shows a close-up of a person, likely a child, looking at a cake with candles. The cake appears to be a birthday cake, and the child seems to be focused on it. The background is blurred, but it looks like an indoor setting with warm lighting. **The focus is on the child and the cake, making the background indistinct**.

Figure 10: Example of generated preference data, where the dis-preferred response is generated from images with lower resolution.



Task: Interpret the scene shown in the image.

Preferred (Good prompting)
The image depicts a **joyful and festive scene**, likely a birthday party. The main subjects in the image are **a young child**, who appears to be the focus of the celebration, and several **other children and adults**. The child is seated at a table, engrossed in eating a piece of cake. ...... The setting is indoors, with the **table and chairs** arranged in a typical party layout. The table is covered with a **tablecloth featuring a cartoon character**, adding to the festive atmosphere. ...... The mood of the image is one of happiness and anticipation ......

Dis-preferred (Color jitter)
The image shows a child's birthday party. The child in the foreground is wearing a striped shirt and appears to be in the process of eating a piece of cake. **The cake is green** and has a \"Happy Birthday\" message on it, suggesting it's a birthday celebration. The child is using a plastic fork to eat the cake.\n\nIn the background, there are other children and adults, some of whom are also eating cake. ...... The overall atmosphere seems to be one of joy and celebration.

Figure 11: Example of generated preference data, where the dis-preferred response is generated from images with color jittering.

Fine-tuning details. We train for 1 epoch in each stage, including the image comprehension self-training stage and the description-infused fine-tuning stage. We use the same hyperparameters for LoRA fine-tuning in both stages, with `lora_r` $= 128$, `lora_alpha` $= 256$, and `lora_target` $=$ `all`. The fine-tuning hyperparameters for Stage 1 are presented in Table 6. The parameters remain the same for Stage 2 fine-tuning, with the only differences being a learning rate of $2e-5$ and a global batch size of 64.

Table 6: Fine-tuning hyperparameters.

Learning rate	$1e-7$
Optimizer	`AdamW`
Global batch size	4
Regularization coefficient α	$1/1024$
`mm_projector_lr`	$2e-5$
`mm_projector_type`	`mlp2x_gelu`
`gradient_accumulation_steps`	1
`image_aspect_ratio`	`pad`
`group_by_modality_length`	`True`
`weight_decay`	0
`warmup_ratio`	0.03
`lr_scheduler_type`	`cosine`
`model_max_length`	1024

Describe and Respond.

```
User:  <image>\n Detail the composition and subjects within the frame.
Model:  <image description>
User:  <image>\nImage description:\n<image description>\n<question>
Model:  <response>
```

Evaluation details. We use the evaluation scripts provided by LLaVA (Liu et al., 2023a) for all evaluations. It is important to note that the potential new evaluation scripts and prompts used to report the results for LLaVA-v1.6 have not been released at the time of writing this manuscript. This may cause discrepancies in the evaluation results of the original model.

Compute resources. Experiments were conducted on NVIDIA RTX A6000 GPU clusters. The entire self-training process for LLaVA v1.5 (7B) and LLaVA v1.6 (7B), using $6k$ image data and $5k$ reused instruction fine-tuning data, takes approximately 6 hours on 4 GPUs. The evaluation process of STIC for the benchmarks typically varies from 2 to 8 hours, mainly depending on the test set size.

C Experimental Results

Table 7: Detailed performance of STIC compared with the original VLM model on the MM-Bench dev set.

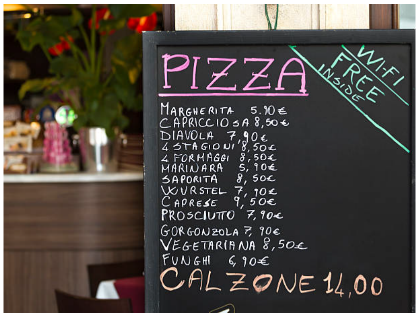
Model	MMBench					
	LR	AR	RR	FP-S	FP-C	CP
Original	35.6	65.8	61.7	64.5	49.0	80.7
w/ STIC	**42.4**	**69.3**	**67.8**	**68.6**	**60.8**	**83.1**

Table 8: Detailed performance of STIC compared with the original VLM model on the MM-Vet benchmark.

Model	MM-Vet					
	rec	ocr	know	gen	spat	math
Original	43.1	40.6	29.6	32.5	44.7	15.4
w/ STIC	**45.7**	**42.5**	**30.4**	**34.9**	**45.1**	**22.7**

C.1 Example Outputs

In Figure 12 and 13, we demonstrate more output examples of the original LLaVA-v1.6 and STIC.

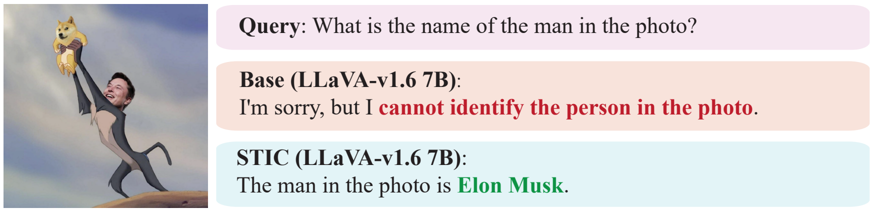


Query: I am getting one funghi pizza and one prosciutto pizza. How much should I pay in total?

Base (LLaVA-v1.6 7B):
Based on the menu prices, the total cost for one funghi pizza and one prosciutto pizza **would be 14,00€**.

STIC (LLaVA-v1.6 7B):
Based on the menu displayed in the image, **the price for a funghi pizza is 6,90€, and the price for a prosciutto pizza is 7,90€.** Therefore, the total cost for one funghi pizza and one prosciutto pizza **would be 14,80€**.

Figure 12: Response examples from original LLaVA-v1.6 and STIC (LLaVA-v1.6) in MM-Vet.



Query: What is the name of the man in the photo?

Base (LLaVA-v1.6 7B):
I'm sorry, but I **cannot identify the person in the photo**.

STIC (LLaVA-v1.6 7B):
The man in the photo is **Elon Musk**.

Figure 13: Response examples from original LLaVA-v1.6 and STIC (LLaVA-v1.6) in LLaVA-Bench.

C.2 t-SNE Visualization Analysis

ScienceQA, TextVQA, MathVista, and ChartQA were chosen because there are at least $1,000$ images in the test set, providing enough data points for analysis.

MSCOCO vs ScienceQA. The gain achieved by STIC on ScienceQA was 6.4%, the highest across all seven benchmarks. As evident from Figure 14 (a), the images in ScienceQA have substantial overlap with those in MSCOCO. This suggests that the image comprehension capabilities developed by STIC through self-training on MSCOCO translated effectively to the scientific reasoning tasks in ScienceQA.

MSCOCO vs TextVQA. STIC yielded a gain of 4.9% on TextVQA, one of the higher gains observed across the benchmarks. Figure 14 (b) shows a significant overlap between the image distributions of TextVQA and MSCOCO. This indicates that the enhanced image understanding from self-training on MSCOCO proved beneficial for the text-based visual question answering tasks in TextVQA.

MSCOCO vs MathVista. On MathVista, STIC achieved a gain of 2.4%, which, while still notable, was lower compared to other benchmarks. As Figure Figure 14 (c) illustrates, the images in MathVista have limited overlap with those in MSCOCO. Moreover, MathVista features diverse image types and mathematical reasoning tasks that pose additional challenges beyond image comprehension. These factors likely contributed to the more modest performance gain.

MSCOCO vs ChartQA. STIC attained a high gain of 5.1% on ChartQA, despite the images in ChartQA having minimal overlap with those in MSCOCO, as shown in Figure 14 (d). This seemingly contradictory result can be explained by considering that the improved image comprehension capabilities developed by STIC play a fundamental role in understanding and reasoning about the charts in ChartQA. Thus, even with limited distributional similarity, the enhanced perception skills proved valuable for this benchmark.

C.3 Discussion with POVID.

We detail the differences between STIC and POVID. In POVID, the dispreferred response is generated either by adding Gaussian noise to the original image or by manually injecting hallucinations into the ground truth completion, using the labeled object information of the images. In contrast, STIC (1) specifically targets the image description task, (2) constructs preference datasets exclusively from *unlabeled* images using self-generated content for both preferred and dispreferred responses, (3) employs an automatic model generation process without manual injections or modifications, and (4) utilizes only a small portion of SFT data for *instruction-following fine-tuning* with uniquely infused model descriptions. Lastly, we compare the data types and scales used in POVID and STIC in Figure 15.

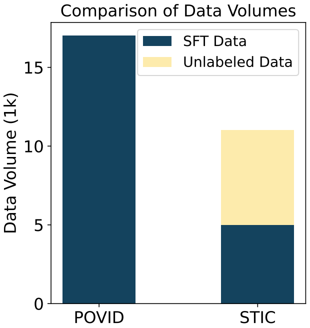
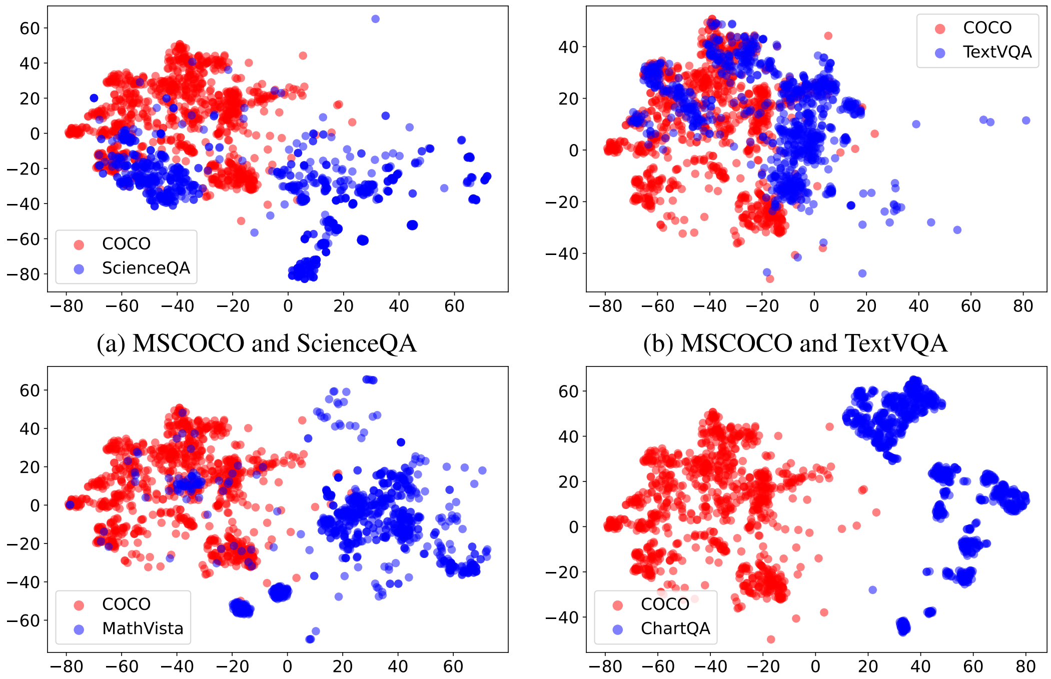


Figure 15: Data comparison.

C.4 Investigation of Prompt Quality

Table 9 presents an additional experiments using DaR to demonstrate prompt quality. We compared normal prompts from our main paper (e.g., "Illustrate the details of the picture.") with the hallucination prompts and well-curated prompts used for DPO pair generation. The results show an expected discrepancy in QA performance: hallucination prompts significantly decreased performance, while well-curated prompts maintained a decent performance. We also included results based on a prompt proposed by Llama-3 8B and filtered using the same restrictions. The performance difference between GPT-4 and Llama-3 8B prompts underscores the quality of GPT-4's proposals.



(a) MSCOCO and ScienceQA

(b) MSCOCO and TextVQA

(c) MSCOCO and MathVista

(d) MSCOCO and ChartQA

Figure 14: t-SNE visualization of images from MSCOCO and benchmarks.

Table 9: Test performance of `llava-v1.6-mistral-7b` using various prompts with DaR. We evaluate prompt quality using DaR as a prompting method. DaR=`None` represents the original LVLM model's performance. Normal prompt refers to the simple prompt we used for DaR in our paper. GPT-4's well-curated prompt refers to the prompt we used for preferred response generation, and we include Mistral 7B's curated prompt for additional comparison.

Model	DaR	LLaVA-Bench	MM-Vet	MMBench
LLaVA-v1.6 (7B)	None	77.3	42.2	63.7
	Normal Prompt	78.5$_{(+1.2)}$	42.3$_{(+0.1)}$	50.7$_{(-13.0)}$
	Hallu Prompt	73.7$_{(-3.6)}$	40.5$_{(-1.7)}$	40.7$_{(-23.0)}$
	Well-curated (Llama-3 8B)	77.2$_{(+0.1)}$	40.0$_{(-2.2)}$	60.1$_{(-3.6)}$
	Well-curated (GPT-4)	79.1$_{(+2.1)}$	42.9$_{(+0.7)}$	60.9$_{(-2.8)}$

D Limitations and Future Work

While STIC demonstrates significant performance gains across a diverse set of benchmarks, there are still some limitations to be addressed in future work. First, STIC achieves a relatively small accuracy gain on MathVista compared to other benchmarks. This is likely because MathVista features various mathematical reasoning abilities across a wide range of image types, from elementary school to college level problems, which go beyond the scope of image comprehension. In contrast, STIC uses MSCOCO, containing only natural images, for constructing its preference data. To further improve performance on tasks like MathVista, future work could expand the source image types and generate task-specific image description data that better aligns with the target benchmark.

Second, while our approach of using good prompts for positive examples and bad prompts or corrupted images for negative examples is straightforward and effective, this preference data may be insufficient for scenarios requiring nuanced understanding of image details. More sophisticated strategies for preference data construction could potentially further boost fine-tuning performance with the DPO loss.

Additionally, our current method relies on a two-stage process of first training on the preference data and then performing description-infused fine-tuning. An interesting direction for future work would be to explore integrating these stages into a single end-to-end training process, which could potentially lead to even greater synergies and performance gains.

Lastly, while we have demonstrated the scalability of STIC by doubling the amount of preference data, leading to further improvements, we have not yet explored the upper limits of this scaling. It is possible that even larger amounts of self-training data could lead to diminishing returns at some point. Characterizing the scaling behavior of STIC more fully is an important direction for future research.

Despite these limitations, we believe STIC represents an important step towards leveraging the vast amounts of unlabeled image data available to enhance the image comprehension capabilities of large vision-language models in a cost-effective manner.

E Broader Impacts

The development of STIC, our self-training approach for enhancing the image comprehension capabilities of large vision language models (LVLMs), presents several potential societal impacts. Positively, our method can democratize access to advanced vision-language models by significantly reducing the cost and effort required for fine-tuning, making state-of-the-art LVLMs more accessible to researchers and organizations with limited resources. This can accelerate advancements in healthcare, education, and environmental monitoring, where improved image comprehension can lead to better diagnostic tools, personalized learning experiences, and more effective environmental protection measures. Additionally, by encouraging the reuse and recycling of existing data, STIC aligns with sustainable AI practices, promoting efficient use of computational and data resources.

However, there are potential negative societal impacts that must be considered. Enhanced LVLM capabilities could be misused for generating disinformation, creating fake profiles, or conducting unauthorized surveillance, contributing to the spread of misinformation and erosion of public trust. Fairness considerations are crucial, as biased training data may lead to outputs that disproportionately impact specific groups, resulting in unfair treatment or discrimination. Privacy concerns also arise from using self-generated data, particularly if models are trained on sensitive or personal visual content without proper consent. To mitigate these risks, strategies such as gated release of models, robust fairness audits, diverse data inclusion, and enhanced transparency about the technology's limitations and risks are essential to ensure responsible use.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction focus on the novel self-training method for image comprehension, description-infused fine-tuning, and the significant performance gains across several benchmarks. These contributions are accurately reflected and supported by the theoretical and experimental results presented in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the proposed method are thoroughly discussed in Appendix D, where the authors address potential challenges and constraints, including domain alignment issues and possible failure cases, providing a clear reflection on the robustness and scope of their approach.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Section 3 provides the problem setting and preliminaries, outlining the assumptions and theoretical foundations supporting our proposed approach. Theoretical assumptions are clearly stated and supported by relevant literature.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5 and the Appendix B provide detailed information on the experimental setup, datasets, and evaluation metrics, ensuring reproducibility. Additionally, we commit to releasing the entire codebase, including data generation and fine-tuning procedures, to facilitate replication of our results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper commits to providing open access to the code, including data generation and fine-tuning procedures. We discuss the open benchmarks used for evaluation in Section 5.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 provides details on the experimental setup, Section 6 covers the ablation studies, and Appendix B includes additional experimental details such as data splits, hyperparameters, and the type of optimizer used.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: On LLaVA-bench, where the evaluations are done with GPT-4, we reported the error bars in Figure 4. However, most evaluations in this work are done using greedy decoding, resulting in a deterministic process that does not incur error bars. Therefore, the reported values of the baseline method that we compare to similarly did not report error bars. Lastly, due to the computational expense of fine-tuning large language models multiple times, we did not include error bars. However, we acknowledge this limitation and ensure that the reported results are consistent and reliable based on our experimental setup.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B provides detailed information on the compute resources used for experiments, including the type of compute workers (GPU) and execution time.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper adheres to the NeurIPS Code of Ethics, ensuring ethical standards are maintained in data usage, experimental procedures, and reporting of results. The paper preserves anonymity and considers the implications of the research on society.

   Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix E discusses both the potential positive societal impacts, such as advancements in vision-language understanding and applications in education, and the potential negative societal impacts, including concerns about misuse and ethical considerations. The paper also addresses possible mitigation strategies for negative impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our approach does not involve releasing models or data, so there are no high-risk elements requiring safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use open-sourced models and datasets for our experiments, all of which are properly cited and used under their original licenses.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not introduce any new assets, so this question is not applicable.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects, so this question is not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects, so this question is not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.