Say It Better: RL-Based Prompt Tuning for Enhancing Open-Vocabulary Recognition

Mikhail Avshalumov[a], Zoya Volovikova[a,b], Dmitry Yudin[a,b], Alexandr Panov[a,b,1]

[a] *Moscow Institute of Physics and Technology, 9 Institutsky per., Dolgoprudny, Russia*
[b] *Artificial Intelligence Research Institute (AIRI), 32 Kutuzovsky Ave., Moscow, Russia*
[c] *Fedral Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia*

Abstract

Referring image recognition using natural language prompts has gained traction with the rise of open-vocabulary vision tasks. We propose an enhancement strategy for open-vocabulary image recognition models through a fine-tuning approach that transforms a user's query into a prompt optimally aligned with the input expectations of a given recognition model. This transformation is performed by an additional Large Language Model, which is further trained to adapt prompts for improved visual grounding. We introduce both online and offline modes for this transformation, leveraging deep learning techniques and model feedback to iteratively refine query formulations. Our approach is evaluated on multiple image recognition models and tested on datasets including RefCOCOg and GOLDg, demonstrating substantial improvements in prediction accuracy. Code and dataset are available at https://github.com/ZoyaV/reframing.

Keywords: Open-vocabulary object detection, Image recognition, Prompt tuning, Large language model

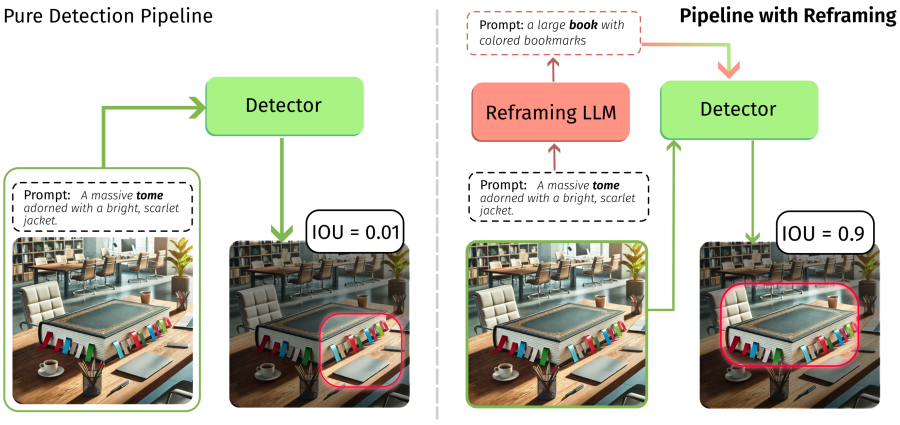


Figure 1: The proposed idea. Based on the characteristics of the LLM detector, it paraphrases the input prompt to improve detection quality.

1. Introduction

Open-vocabulary image recognition has emerged as a pivotal area of research in computer vision, offering the capacity to identify a wide array of visual entities beyond the constraints of a closed, pre-defined label space. In contrast to traditional supervised models, which are inherently limited to a fixed set of training classes, open-vocabulary systems aim to generalize to arbitrary textual queries, thereby enhancing their applicability in open-world scenarios. These capabilities are particularly important for real-world deployments, where it is impractical to anticipate all possible categories of interest in advance. For example, GroundingDINO [1] and YOLO-World [2] models exhibit exceptional proficiency in out-of-domain detection, achieving remarkable results in identifying objects that were previously unobserved.

Despite their strengths, open-vocabulary recognition models remain sensitive to the formulation of the input prompt. Even minor variations in phrasing can result in significant changes in model behavior, which introduces unpredictabil-

*Corresponding author
Email addresses: avshalumov.ms@phystech.edu (Mikhail Avshalumov), yudin@airi.net (Dmitry Yudin)

ity and limits reliability in critical applications. This phenomenon is attributed to the inherent textual biases encoded during training, often influenced by the distribution of captions or labels in large-scale datasets. Consequently, the ability to tailor prompts in a manner that aligns with the model's internal representations becomes a key factor in maximizing performance. This challenge motivates the development of methods that can systematically refine or adapt user queries without requiring manual intervention or predefined class lists.

In this work, we address the above limitations by proposing a targeted prompt adaptation framework for open-vocabulary image recognition models. We introduce a learned mechanism for dynamically reformulating natural language prompts to better match the idiosyncrasies of specific image recognition models.

At the core of our method is an auxiliary large language model (LLM), which is fine-tuned to generate alternative phrasings of user-specified prompts. These reformulations are conditioned not only on the input query but also on performance feedback obtained from the image recognition model—such as class confidence scores, similarity ranks, or metric-based evaluations like Intersection over Union (IoU). Through this feedback loop, the system incrementally identifies and prioritizes linguistic formulations that are more likely to activate the model's latent visual concepts, effectively aligning user intent with model response. This optimization can be conducted in two distinct modes: (i) an offline setting, where training is guided by aggregated statistics from previously collected prompts; and (ii) an online setting, where the adaptation is performed in real-time during inference.

To rigorously evaluate our approach, we conduct experiments using multiple state-of-the-art image recognition models. We include benchmarks on two datasets that are complementary in their purpose and design. First, we employ RefCOCOg, a widely adopted benchmark in the field of referring expression comprehension. This dataset is especially valuable due to its natural language complexity and its prominence in the evaluation of open-vocabulary visual grounding tasks, making it a strong proxy for real-world language vari-

ability. Second, we test on GOLD-A, a newly curated dataset tailored to better reflect the distributional characteristics of the models under study. Since the recognition models we experiment with have been trained on data drawn from GOLD-A or similar sources, it provides a more controlled environment for understanding prompt sensitivity and model-specific biases.

The main contributions of our study are as follows:

1. We propose a novel prompt adaptation method for open-vocabulary image recognition, which leverages an auxiliary LLM to reformulate natural language queries in alignment with model-specific textual biases.
2. We introduce and formalize two modes of adaptation—online and offline—based on whether model feedback is integrated dynamically or statically during prompt optimization.
3. We validate our method across various recognition architectures, using RefCOCOg and the GOLD-A dataset.

2. Related Work

2.1. Open-vocabulary detection algorithms

Open-vocabulary object detection is frequently trained using existing bounding box annotations to identify arbitrary classes through language generalization. OV-DETR[3] leverages image and text embeddings, encoded with the CLIP[4] model, as queries to decode specific categories within the DETR[5] framework. ViLD[6] utilizes the teacher CLIP model to enhance an R-CNN-like detector with knowledge, ensuring that the learned region embeddings incorporate language semantics. GLIP[7] approaches object detection as a grounding task, utilizing additional grounding data to achieve aligned semantics at phrase and region levels, illustrating its ability to exceed the rigorous performance standards in fully supervised detection benchmarks. DetCLIP[8] combines large-scale image datasets with captions and employs generated pseudo-labels to improve the detectors' knowledge base, significantly boosting their

generalization capabilities. GroundingDINO[9] advances beyond the closed-set detector DINO[1] by integrating vision-language modality fusion at various stages, including feature enhancement, language-guided query selection, and cross-modality decoding, underscoring the importance of correct prompt selection for each detector's effectiveness.

Furthermore, some methods are trained to merge multimodal information from the ground up, not dependent on a closed-set detector as a base. For example, the ONE-PEACE[10] method processes text and images through separate adapters, transforming the input into uniform features, which are then fed into a transformer-based modality fusion encoder. This model is trained using Intra-Modal Denoising Contrastive Learning.

The You Only Look Once(YOLO) [11] series, known for its efficiency, is constrained to predefined categories, limiting its applicability in open-world scenarios. YOLO-World [2] addresses this limitation by employing an open vocabulary in conjunction with vision-language modeling and extensive pre-training. It introduces a Reparameterizable Vision-Language Pathway Aggregation Network (RepVL-PAN) and a region-text contrastive loss to enhance the integration of visual and linguistic data.

Despite the diverse approaches and methods employed by these detectors, selecting the appropriate prompt is crucial for each one's effectiveness.

2.2. Prompt Tuning Algorithms

Various strategies have been developed for prompt tuning within detection models. CoOP[12], for instance, introduces context optimization to streamline the engineering of prompts for few-shot classification, treating the context of prompts as continuous representations that are learned in an end-to-end manner from a limited dataset. ViLD, on the other hand, employs a static prompt template, such as 'a photo of [CATEGORY] in the scene'. Built upon ViLD, DetPro[13] expands the capabilities of CoOP to facilitate Open-Vocabulary Object Detection. Unlike ViLD, DetPro innovates with fine-grained automatic prompt learning and a unique approach to background interpretation, aiming

to identify the most effective prompts. However, a common limitation among these methods is their reliance on a predefined set of classes to create prompts, with the aspiration that these methods for generating prompts will successfully generalize to novel classes.

2.3. Language models fine-tuning algorithms

Various methodologies exist for fine-tuning language models, with reinforcement learning-based approaches gaining prominence. One notable method is RLHF[14][15], which leverages PPO and human evaluations to improve model performance, standing out as a prevalent technique in today's landscape. Another approach is ReST[16], which employs offline RL algorithms to enhance language models further. Despite their advancements, these approaches often grapple with computational complexity and instability. Additionally, there are methods that conceptualize language model fine-tuning as fulfilling constraints within a target probability distribution, symbolizing the ideal language model[17][18]. Direct Preference Optimization(DPO) addresses the complexities associated with RLHF, obviating the need for explicit constraints. One can also find an example of tuning a language model for integration with an RL agent [19].

A recent advancement in reinforcement learning-based fine-tuning is Group Relative Policy Optimization (GRPO) [20], which is built upon PPO by eliminating the need for a separate value function estimator. Instead, it evaluates multiple responses generated for the same prompt, computing standardized scores to assess their relative quality within a group. This approach simplifies the training pipeline and enhances stability by focusing on relative performance rather than absolute reward values.

3. Methodology

Our approach is based on tuning an additional large language model which will rephrase the user's input query to make it more suitable for a specific

detection model. Figure 1 shows the pipeline of the detection system with the language model. The system receives an image and a text query for object search in the image. The query is then passed to the paraphrasing network, which rewrites it in a format more suitable for the detection model. The updated query is then sent to the detection model. In this pipeline, model must be specially trained to work with the specific detector network.

Our framework offers three training modes: **offline**, **online-PPO** and **online-GRPO**.

In offline mode, training is conducted on pre-labeled data containing information on the quality of object recognition for various queries for a specific detection model.

Online modes include pre-training, where Supervised Fine-Tuning (SFT) is used in the first stage, and then the language model is fine-tuned using RL algorithms, utilizing feedback from the detector.

SFT is employed to shift the language model's probability distribution toward generating outputs that are better suited for the target task. By training on a curated dataset of original queries, their rephrasings, and corresponding detector evaluations, SFT encourages the model to produce rephrasings that are likely to lead to improved object recognition results.

However, SFT alone is limited by the distribution of textual patterns present in the training data. To explore and generate novel textual formulations that may not be represented in the SFT dataset, a second stage of training with Reinforcement Learning (RL) is introduced. In this stage, algorithms such as PPO or GRPO are used to fine-tune the model based on direct feedback from the detector, allowing it to discover and optimize new query variations that further enhance detection performance.

All three of our approaches require working with a dataset of a specific format, including the original user queries, the rephrased queries, the target objects, and the evaluation of the detector's performance on the rephrased queries (e.g., IOU, Score).

3.1. Dataset generation

Dataset format We propose a new dataset based on the GOLD [21] dataset, where images and descriptions are augmented to enhance diversity. We chose this dataset as the basis because it was used to train both of the detection models we are using. We assume that since the model was trained on these data, the model's response to them should be maximized, and therefore, they constitute the data with the targeted bias of the detector language model. Our dataset comprises a collection of images, along with bounding boxes and descriptions for the objects depicted in these images. The dataset generation process involves several steps to ensure that data maintains high quality and relevance to the original dataset. The structure of our dataset is detailed in Table 1. We refer this new dataset as GOLD-A and use it for both: offline and online modes of our approach.

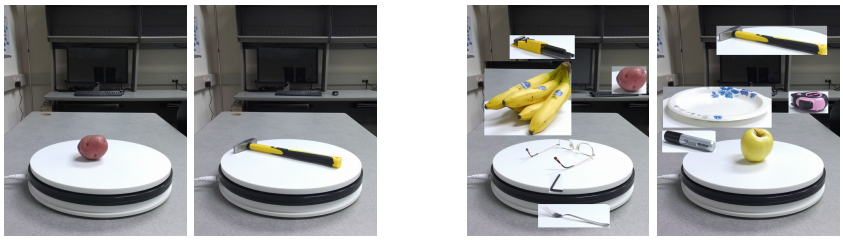
Descriptions generation. We utilize a pretrained model for sentence paraphrasing using a zero-shot approach, where the model is not specifically fine-tuned for our dataset but is prompted to generate paraphrased sentences. The prompt format employed is "Paraphrase sentence: sentence [/INST]", which instructs the model to rephrase the input while preserving its original meaning. The outputs are then extracted for the paraphrased content. Additionally, a quality control mechanism ensures that each sentence ends with a single period, lacks any question marks, exclamation marks, colons, or semicolons, and is longer than five characters to maintain the coherence and relevance of the descriptions.

The paraphrasing process is applied to a series of base descriptions linked with specific object IDs and bounding box information. For each entry, the system produces multiple paraphrased versions using the defined prompt, and the results include paraphrased sentences along with a correctness score. The augmented dataset is periodically saved to secure progress.

Images augmentation. In the original GOLD dataset, all images are uniform, with the target object centered and sharply contrasting with the surrounding objects. Consequently, SOTA open-world detection models often perform

Table 1: Dataset structure.

Column name	Description	Example
item id	The class name for the object depicted in the image. Unambiguously determines which image the description pertains to	*fork_1_2*
true bbox	ground truth bbox for object on image	[236, 286, 231, 328]
description	Description of the object: can either be sourced from a base dataset or generated by a language model.	*A fork designed for picking up and eating food from a container is used to insert food into the mouth.*
score	Score from detection model, used for filtration and ranking	0.49



(a) GOLD images

(b) GOLD-A images

Figure 2: Example of images from GOLD and GOLD-A datasets

well regardless of the input question, which is unrealistic in real-world scenarios. To address this, we augmented the original images by cutting out the target object and placing it in a random location within the image. We also added up to five other objects from the same dataset, ensuring their classes do not match the target and that the objects do not overlap. This approach increases image diversity and ensures text queries remain significant. Examples of images from GOLD and GOLD-A are shown in the figure 4.

Dataset evaluation. We expand the augmented dataset through feedback from the detection model, adding additional fields: a bounding box found by the target detector, IOU, which shows how well the detector's predictions align with the current prompt, and a prediction score indicating its confidence in the accuracy of the bounding box.

3.2. Offline Traning Mode

In an offline setup, we utilize a pre-prepared dataset and train using Direct Policy Optimization (DPO) [22] loss. This method does not allow for the direct incorporation of reward information, so the dataset is reformatted into a preference dataset. This dataset consists of prompts and pairs of randomly sampled descriptions. During this preference dataset generation process, three descriptions are selected for each item: one as the prompt and the other two for detailed evaluation to assess their suitability for the detector. These descriptions are categorized as either 'chosen' or 'rejected'. The ranking occurs within each set of three descriptions, where one is the prompt, and the other two are evaluated against it to determine their relative effectiveness. In the first column, the prompts always have a higher reward compared to those in the second column.

The core idea behind DPO is encapsulated in the following gradient formula for the loss function L_{DPO}:

$$\nabla_\theta \mathcal{L}_{DPO}(\pi_\theta; \pi_{ref}) =$$

$$- \mathrm{E}_{(x,y_w,y_l)\sim D} \left[\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w)) \left(\nabla_\theta \log \pi_\theta(y_w|x) - \nabla_\theta \log \pi_\theta(y_l|x)\right)\right], \quad (1)$$

where $\hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}$ represents the reward implicitly defined by the language model π_θ and a reference model π_{ref}. The function σ is a logistic sigmoid that assigns higher weights when the reward estimate is incorrect. The gradient of the loss function L_{DPO} aims to increase the likelihood of preferred completions y_w and decrease the likelihood of dispreferred completions y_l. This weighting is crucial for adjusting the model in a direction that favors more desirable outcomes, while taking into account the strength of the Kullback-Leibler (KL) constraint. The pseudocode of offline approach is shown in Algorithm 1.

3.3. Online Mode

This approach consists of two stages. First, we carry out supervised fine-tuning to adjust the language model's token probability distribution for the

Algorithm 1: Propoosed offline mode

Input: Dataset $\mathcal{D}$, Number of iterations N for DPO, pretrained language model π_θ;

Generate dataset in DPO format;

DPO:

$\pi_{ref} \leftarrow \pi_\theta$;

for $i = 1$ **to** N **do**

  foreach *tuple* (p, y_w, y_l) *in* $\mathcal{D}_{DPO}$ **do**

    $\nabla_\theta \mathcal{L}_{DPO}(\pi_\theta; \pi_{ref}) \leftarrow$

    $\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w)) \left(\nabla_\theta \log \pi_\theta(y_w|x) - \nabla_\theta \log \pi_\theta(y_l|x)\right)$,

  end

end

specific task.For this step we use a subset D_t from the dataset D. To select samples, we calculate the 80th percentile of the rewards in D_t and then choose only those rows where the reward is above this threshold. Next, we use reinforcement learning with Proximal Policy Optimization(PPO) [23], enabling the model to discover new, previously unseen responses that enhance the detection model's performance. Then the model engages in a cyclical process. It sequentially generates responses, evaluates them with the detection model, and uses the feedback from these evaluations to update its weights via PPO. The PPO loss function for our approach is:

$$\nabla_\theta L^{CLIP}(\theta) = E_t \left[\nabla_\theta \left(\min \left(r_t(\theta)\hat{A}_t, \mathit{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \right) \right) \right], \qquad (2)$$

where: $r_t(\theta)$ is the probability ratio between the new policy and the old policy. $\hat{A}_t$ is the estimated advantage function. ϵ is a hyperparameter that controls the clipping range. This formula ensures that the updates to the policy are conservative, preventing drastic changes that could lead to instability. The pseudocode of this approach is shown in Algorithm 2.

Algorithm 2: Proposed online-PPO mode

Input: Dataset $\mathcal{D}$, Number of iterations N for SFT, Number of iterations N_1 for PPO, pretrained language model π_θ, quantity of answers sampled for SFT Q;

Generate dataset for SFT;

SFT:

  for $k = 1$ **to** N_1 **do**

    $(p, \hat{y}) \leftarrow$ *Sample from* $\mathcal{D}_{SFT}$;

    $y \leftarrow \pi_\theta(p)$;

    $\pi_\theta \leftarrow \pi_\theta - \alpha\nabla(-\hat{y}\log(y))$

  end

PPO: ;

  for $j = 1$ **to** N **do**

    $(d, I, bbox) \leftarrow$ *Sample from* $\mathcal{D}$;

    $p \leftarrow$ *"Paraphrase sentence: {d}"* y $\leftarrow \pi_\theta(p)$;

    $(bboxes, masks) \leftarrow$ *Detection model output for* y;

    $r \leftarrow$ *Compute reward* $R(IoU,\ score|y)$;

    $\mathcal{L} \leftarrow \min(r(\theta)\hat{A}, clip(r(\theta), 1-\epsilon, 1+\epsilon)\hat{A})$;

    $\pi_\theta \leftarrow \pi_\theta + \beta\nabla\mathcal{L}$;

  end

4. Experiments

In this section, we describe the experiments conducted using our method, outline the outcomes, and present an ablation study. This study assesses the method's generalizability to other detection models and examines the effects of various reward functions on performance results.

4.0.1. Models and datasets.

Language models For our experiments, encoder-only architectures trained on instructive datasets were selected. Specifically, the Llama2(7b)[24], Llama3 [25](8b),

Qwen [26](7b) and Gemma(7b) [27] models were used. The main motivation for choosing these models was their ability to operate in a zero-shot mode for the task of text paraphrasing. Theoretically, it is possible to use a smaller model that can additionally be trained to generate paraphrased texts.

Image recognition models. We implemented our methodology on two detection systems, GroundingDINO and YOLO-World. For our approach to be effective, it is essential that the detection model can accurately process complex queries that may include contextual dependencies within the object's description, such as "a plush bear lying on the right-hand side table".

In addition to object detection, we also applied our best-performing training strategy to the task of open-vocabulary segmentation by fine-tuning the language model for integration with the HyperSeg [28] segmentation model. HyperSeg was selected due to its status as one of the most advanced and high-performing segmentation models currently available.

Datasets. We conducted training and evaluation on two primary datasets: RefCOCOg and our proposed GOLD-A. The RefCOCOg dataset was selected due to its prominence in the field and its high degree of linguistic variability, which provides a robust testbed for evaluating language grounding capabilities. The GOLD-A dataset was chosen because it closely aligns with the training distribution of the target detection and segmentation models, ensuring relevance and task-specific performance assessment.

In addition to these core datasets, we introduced an auxiliary test set referred to as *External* during the evaluation phase. This dataset comprises labeled real-world photographs, intended to broaden the evaluation scope and assess generalization. For each image, we used Chat GPT-4 to generate detailed textual descriptions of the identified objects, resulting in a total of 50 unique descriptions spanning 15 object classes.

Implementation details. We implement our approach using Transformer Reinforcement Learning(TRL) library [29]. Our experiments are conducted on a single H100 GPU. During the training phase, our configuration includes a batch size of 16. In our online modes, we allocate 30k steps for the RL phase and 20k

steps for the SFT phase, while our offline mode involves 50k training steps.

4.1. Results

Table 2: Comparison of learning outcomes of different LLMs

Detection Model	**Prompt tuning approach**	**External, IoU**	**GOLD-A, IoU**	**RefCOCOg, IoU**
GroundingDINO	*Original prompt*	0.561	0.572	
	Gemma-7b (Pretrained)	0.379	0.436	
	Reframing **offline** (Gemma-7b))	0.398	0.523	
	Reframing **online** (Gemma-7b)	0.542	0.564	
	Llama2-7b (Pretrained)	0.428	0.550	
	Reframing **offline** (Llama2-7b)	0.448	0.562	
	Reframing **online** (Llama2-7b)	**0.572**	**0.585**	
	online-GRPO (Llama2-7b)	**0.572**	**0.593**	
YOLO-World	*Original prompt*	0.501	0.535	
	Gemma-7b (Pretrained)	0.270	0.433	
	Reframing **offline** (Gemma-7b)	0.302	0.508	
	Reframing **online** (Gemma-7b)	0.442	0.521	
	Llama2-7b (Pretrained)	0.411	0.511	
	Reframing **offline** (Llama2-7b)	0.453	0.528	
	Reframing **online** (Llama2-7b)	**0.504**	**0.545**	
HyperRef	*Original prompt*	-	-	
	Llama2-7b (Pretrained)	-	-	
	Reframing **online-GRPO** (Llama2-7b)	-	-	

Through our approach, we trained several versions of Large Language Models (LLMs) and evaluated their performance using the two detection models, calculating metrics such as IOU on original misses, misses obtained from LLMs without tuning, and models trained with online and offline modes.

How does each step of our online mode improve image recognition model's performance? We conducted an ablation study to assess the contribution of each stage in our online-GRPO training pipeline to the overall performance of the visual recognition model. The results, summarized in Table 3, highlight the incremental benefits introduced by the supervised fine-tuning and reinforcement learning phases within our framework.

Table 3: Comparison of learning outcomes of different training steps for online-GRPO approach for Llama2 language model

Image recognition Model	**Prompt tuning approach**	**GOLD-A, IoU**
Grounding DINO	*Original prompt*	0.537042
	Pretrained	0.507201
	SFT	0.549440
	GRPO	**0.566567**

Can a language model without any fine-tuning effectively enhance prompts? Without prior training, a language model can refine individual queries by providing more detailed object descriptions. However, as shown in Table 2 it lacks the capability to do this systematically, considering the specific features of the detection model.

Does our methodology help improve the LLM's ability to reframe prompts in light of the detection model's preferences? Our research explores whether the proposed technique improves a language model's ability to modify prompts to better align with the preferences of detection models. The data in Table 2 indicate that both offline and online implementations significantly boost the prompt-tuning capabilities of the language models over their pretrained counterparts. Additionally, models operating in online mode outperform human-generated prompts. Llama2-7b trained with proposed online mode shows mean IoU equal to 0.585, while original human prompts show only 0.572.

Table 4: Comparison of pre and post- paraphrasing prompt requests

Human Prompt	**IoU**	**Prompt after paraphrasing**	**IoU**
A sleek, black **photo frame** with squared-off edges is standing upright on a flat surface, its imposing presence drawing attention.	0.0	This is a black **picture frame**. The frame has a sleek, modern design that looks great	0.97
The **plate** being referred to **is red in color**	0.80	The **item is a red dinner plate** that exudes a vintage aesthetic, characterized by its classic design and timeless appeal.	0.82
A *set* of **blue scissors** with handles.	0.91	A *pair* of scissors with **blue handles**.	0.93

How should the improvements introduced by the model be in-

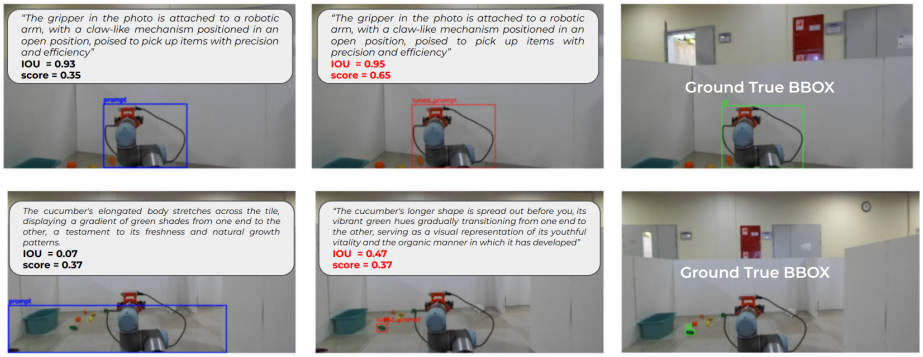
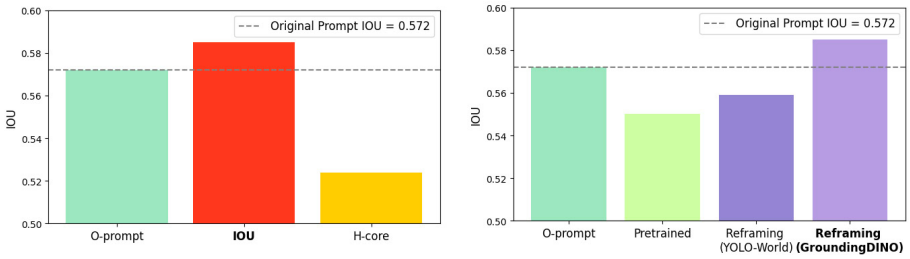


Figure 3: Detection results: original (left), reframed (center), ground-truth bbox (right). Our approach improves metrics and bbox accuracy.

terpreted? The evaluation of model performance can be significantly enriched by comparing the detection outcomes on prompt requests, pre and post-paraphrasing for diverse images. Examples of such prompts are shown in Table 4. Initial assessments with prompts lacking in detail often lead to inaccurate object detection or low confidence in model predictions. This scenario undergoes a marked transformation with the application of reframing. For instance, in one case, our approach leads to a notable boost in the model's performance regarding the predicted bounding box. In another instance, the reframed prompt enables a more precise description, allowing the detection model to accurately delineate the intended object's location. Such observations underscore the importance of detailed prompt formulation in enhancing model prediction accuracy and confidence.

4.2. Ablation study

Impact of the choice of reward function on the quality of results. This study examines how different reward functions affect the quality of query processing in the Llama2 model in our proposed online mode. Specifically, two strategies were analyzed: the use of IoU and evaluation by harmonic mean between IoU and model's confidence score. The comprehensive results of the study are presented in Figure 4a.



(a) (b)

Figure 4: a)Comparison of outcomes of Llama2-7b trained with different detection models and inferenced on GroundingDINO b)Comparison of learning outcomes of Llama2 for the GroundingDINO detector with different reward functions

Further analysis had shown that the most effective approach is to use IoU without any enhancements. Such an approach allows correlating the query's evaluation with its quality: giving a high score for queries that adequately reflect the target object, and a near-zero score for queries that did not find the corresponding object.

Can models trained with our approach for one detection model be used to enhance the performance of another detection model? Our studies investigated whether language models trained using our approach on one detection model could enhance the performance of another. Results shown in 4b demonstrate that while these models show improved outcomes over untrained models on different detection models, direct training on the target detection model yields the most substantial improvements.

5. Conclusion and Discussion

We introduce a novel approach to enhancing open-vocabulary object detection through detector-specific prompt tuning. By integrating an additional LLM trained with deep learning methods, we successfully align user queries with the textual biases inherent to specific detection models. Our methodology boosts detection model performance by refining prompt queries to better match these biases, resulting in notable improvements in prediction accuracy.

Our experiments, conducted across various detection models, demonstrate the effectiveness of our approach. By employing our offline and online approaches and dataset augmentation techniques, we observed improvements in model performance metrics.

Despite the success of our method, there are some limitations. Firstly, our data augmentation technique aligns language models with detection models using deep learning. However, augmenting data with the same model used for fine-tuning might integrate data more effectively. Secondly, the computationally intensive DPO and PPO algorithms we use for fine-tuning our language model require substantial data and resources to produce quality results.

References

[1] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, H.-Y. Shum, Dino: Detr with improved denoising anchor boxes for end-to-end object detection, arXiv preprint arXiv:2203.03605.

[2] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, Y. Shan, Yolo-world: Real-time open-vocabulary object detection, in: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2024.

[3] Y. Zang, W. Li, K. Zhou, C. Huang, C. C. Loy, Open-vocabulary detr with conditional matching, in: European Conference on Computer Vision, Springer, 2022, pp. 106–122.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[5] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: European conference on computer vision, Springer, 2020, pp. 213–229.

[6] X. Gu, T.-Y. Lin, W. Kuo, Y. Cui, Open-vocabulary object detection via vision and language knowledge distillation, arXiv preprint arXiv:2104.13921.

[7] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, et al., Grounded language-image pre-training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10965–10975.

[8] L. Yao, J. Han, Y. Wen, X. Liang, D. Xu, W. Zhang, Z. Li, C. Xu, H. Xu, Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection, Advances in Neural Information Processing Systems 35 (2022) 9125–9138.

[9] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al., Grounding dino: Marrying dino with grounded pre-training for open-set object detection, arXiv preprint arXiv:2303.05499.

[10] P. Wang, S. Wang, J. Lin, S. Bai, X. Zhou, J. Zhou, X. Wang, C. Zhou, One-peace: Exploring one general representation model toward unlimited modalities, arXiv preprint arXiv:2305.11172.

[11] J. Redmon, S. K. Divvala, R. B. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 779–788.

[12] K. Zhou, J. Yang, C. C. Loy, Z. Liu, Learning to prompt for vision-language models, International Journal of Computer Vision 130 (9) (2022) 2337–2348.

[13] Y. Du, F. Wei, Z. Zhang, M. Shi, Y. Gao, G. Li, Learning to prompt for open-vocabulary object detection with vision-language model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 14084–14093.

[14] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, G. Irving, Fine-tuning language models from human preferences, arXiv preprint arXiv:1909.08593.

[15] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, P. F. Christiano, Learning to summarize with human feedback, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 3008–3021.

[16] C. Gulcehre, T. L. Paine, S. Srinivasan, K. Konyushkova, L. Weerts, A. Sharma, A. Siddhant, A. Ahern, M. Wang, C. Gu, et al., Reinforced self-training (rest) for language modeling, arXiv preprint arXiv:2308.08998.

[17] D. Go, T. Korbak, G. Kruszewski, J. Rozen, N. Ryu, M. Dymetman, Aligning language models with preferences through f-divergence minimization, in: International Conference on Machine Learning, 2023.

[18] M. Khalifa, H. Elsahar, M. Dymetman, A distributional approach to controlled text generation, arXiv preprint arXiv:2012.11635.

[19] Z. Volovikova, A. Skrynnik, P. Kuderov, A. I. Panov, Instruction following with goal-conditioned reinforcement learning in virtual environments`arXiv:2407.09287`.
URL `https://arxiv.org/abs/2407.09287`

[20] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu, et al., Deepseekmath: Pushing the limits of mathematical reasoning in open language models, arXiv preprint arXiv:2402.03300.

[21] G. Y. Kebe, P. Higgins, P. Jenkins, K. Darvish, R. Sachdeva, R. Barron, J. Winder, D. Engel, E. Raff, F. Ferraro, C. Matuszek, A spoken language dataset of descriptions for speech-based grounded language learning, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets

and Benchmarks Track (Round 1), 2021.
URL https://openreview.net/forum?id=Yx9jT3fkBaD

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

[23] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347.

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

[25] Meta AI, Introducing meta llama 3, https://ai.meta.com/blog/meta-llama-3/, accessed: 2025-04-21 (2024).

[26] Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, Z. Qiu, Qwen2.5 technical report (2025). arXiv:2412.15115.
URL https://arxiv.org/abs/2412.15115

[27] T. M. Gemma Team, C. Hardin, R. Dadashi, S. Bhupatiraju, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, et al., Gemmadoi:10.34740/KAGGLE/M/3301.
URL https://www.kaggle.com/m/3301

[28] C. Wei, Y. Zhong, H. Tan, Y. Liu, Z. Zhao, J. Hu, Y. Yang, Hyperseg: Towards universal visual segmentation with large language model, arXiv preprint arXiv:2411.17606.

[29] L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, N. Lambert, S. Huang, Trl: Transformer reinforcement learning, `https://github.com/huggingface/trl` (2020).

Algorithm 3: Proposed online-GRPO mode

Input: Dataset $\mathcal{D}$, Number of iterations N for SFT, Number of iterations N_1 for PPO, pretrained language model π_θ, quantity of answers sampled for SFT Q;

Generate dataset for SFT;

SFT:

  for $k = 1$ **to** N_1 **do**

    $(p, \hat{y}) \leftarrow$ *Sample from* $\mathcal{D}_{SFT}$;

    $y \leftarrow \pi_\theta(p)$;

    $\pi_\theta \leftarrow \pi_\theta - \alpha \nabla(-\hat{y} \log(y))$

  end

GRPO: ;

  for $j = 1$ **to** N **do**

    $(d, I, bbox) \leftarrow$ *Sample from* $\mathcal{D}$;

    $p \leftarrow$ `"Paraphrase sentence: {d}"`;

    $\{y_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(p)$;

    for $i = 1$ **to** G **do**

      $(bboxes_i, masks_i) \leftarrow$ *Detection model output for* y_i;

      $R_i \leftarrow R\big(IoU_i, score_i \mid y_i\big)$;

    end

    $\overline{R} \leftarrow 1G \sum_{i=1}^{G} R_i, \quad \sigma_R \leftarrow \sqrt{1G \sum_{i=1}^{G} (R_i - \overline{R})^2}$;

    for $i = 1$ **to** G **do**

      $\hat{A}_i \leftarrow R_i - \overline{R}\sigma_R + \delta$; `// group-relative advantage`

      $r_i \leftarrow \pi_\theta(y_i \mid p)\pi_{\theta_{old}}(y_i \mid p)$;

    end

    $\mathcal{L}_{\text{clip}} \leftarrow 1G \sum_{i=1}^{G} \min\big(r_i \hat{A}_i,\ \text{clip}(r_i, 1-\epsilon, 1+\epsilon)\, \hat{A}_i\big)$;

    $\text{KL} \leftarrow D_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big)$;

    $\mathcal{L}_{\text{GRPO}} \leftarrow -\mathcal{L}_{\text{clip}} \ + \ \lambda \,\text{KL}$;

    $\theta \leftarrow \theta + \beta \, \nabla_\theta \mathcal{L}_{\text{GRPO}}$;

  end