Metadata of the chapter that will be visualized in SpringerLink

Book Title	Hybrid Artificial Intelligent Systems	
Series Title		
Chapter Title	Reframing: Detector-Specific Prompt Tuning for Enhancing Open-Vocabulary Object Detection	
Copyright Year	2025	
Copyright HolderName	The Author(s), under exclusive license to Springer Nature Switzerland AG	
Corresponding Author	Family Name	**Avshalumov**
	Particle	
	Given Name	**Mikhail**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Moscow Institute of Physics and Technology
	Address	141701, Dolgoprudny, Russia
	Email	avshalumov.ms@phystech.edu
	ORCID	http://orcid.org/0009-0006-2424-3260
Author	Family Name	**Volovikova**
	Particle	
	Given Name	**Zoya**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Moscow Institute of Physics and Technology
	Address	141701, Dolgoprudny, Russia
	Division	
	Organization	AIRI
	Address	121170, Moscow, Russia
	Email	
	ORCID	http://orcid.org/0009-0005-1058-7743
Author	Family Name	**Yudin**
	Particle	
	Given Name	**Dmitry**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Moscow Institute of Physics and Technology
	Address	141701, Dolgoprudny, Russia
	Division	

	Organization	AIRI
	Address	121170, Moscow, Russia
	Division	
	Organization	Fedral Research Center “Computer Science and Control” of Russian Academy of Sciences
	Address	Moscow, Russia
	Email	yudin@airi.net
	ORCID	http://orcid.org/0000-0002-1407-2633
Author	Family Name	**Panov**
	Particle	
	Given Name	**Alexandr**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Moscow Institute of Physics and Technology
	Address	141701, Dolgoprudny, Russia
	Division	
	Organization	AIRI
	Address	121170, Moscow, Russia
	Division	
	Organization	Fedral Research Center “Computer Science and Control” of Russian Academy of Sciences
	Address	Moscow, Russia
	Email	panov@airi.net
	ORCID	http://orcid.org/0000-0002-9747-3837

Abstract	Object detection in images using natural language prompts has recently become popular due to the extensive development of open-vocabulary perception tasks. We propose an enhancement strategy for open-vocabulary object detection models through a fine-tuning approach named Reframing. The cornerstone of our method is transforming a user’s query into a prompt optimally aligned with the requirements of a specific detection model. This transformation is facilitated by integrating an additional Large Language Model, which undergoes further training. We propose online and offline modes of Reframing, employing deep learning techniques and feedback from the detection models to refine its query transformation process. Our methodology has been applied to various detection models, showcasing its capability to significantly improve prediction accuracy. Developed code and dataset are available at https://github.com/ZoyaV/reframing.

Keywords (separated by '-')	Open-vocabulary object detection - Image recognition - Prompt tuning - Large language model

Author Proof

Check for updates

Reframing: Detector-Specific Prompt Tuning for Enhancing Open-Vocabulary Object Detection

Mikhail Avshalumov[1(✉)], Zoya Volovikova[1,2], Dmitry Yudin[1,2,3], and Alexandr Panov[1,2,3]

[1] Moscow Institute of Physics and Technology, 141701 Dolgoprudny, Russia
avshalumov.ms@phystech.edu
[2] AIRI, 121170 Moscow, Russia
{yudin,panov}@airi.net
[3] Fedral Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia

Abstract. Object detection in images using natural language prompts has recently become popular due to the extensive development of open-vocabulary perception tasks. We propose an enhancement strategy for open-vocabulary object detection models through a fine-tuning approach named Reframing. The cornerstone of our method is transforming a user's query into a prompt optimally aligned with the requirements of a specific detection model. This transformation is facilitated by integrating an additional Large Language Model, which undergoes further training. We propose online and offline modes of Reframing, employing deep learning techniques and feedback from the detection models to refine its query transformation process. Our methodology has been applied to various detection models, showcasing its capability to significantly improve prediction accuracy. Developed code and dataset are available at https://github.com/ZoyaV/reframing.

AQ1

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

1 Introduction

Open-Vocabulary Detectors are an approach in computer vision that enables the recognition of a wide array of objects without the need for a predefined, fixed dictionary of classes. These systems exhibit significant flexibility and adaptability, capable of efficiently handling new, previously unseen objects. This makes them particularly valuable for implementation in the dynamically changing conditions of the real world. A key advantage of open-vocabulary detectors is their ability to generalize, allowing the models to successfully apply acquired knowledge

This work was supported by Russian Science Foundation, grant No. 20-71-10116, https://rscf.ru/en/project/20-71-10116/.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
H. Quintián et al. (Eds.): HAIS 2024, LNAI 14858, pp. 1–13, 2025.
https://doi.org/10.1007/978-3-031-74186-9_11

Author Proof

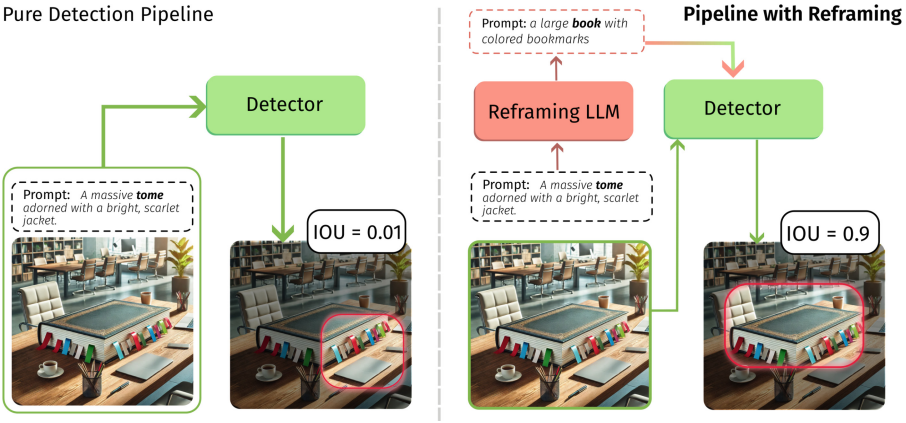


Fig. 1. The proposed idea is reframing. Based on the characteristics of the LLM detector, it paraphrases the input prompt to improve detection quality.

for recognizing objects regardless of the specifics of the domain. For example, GroundingDINO [19] and YOLO-World [1] models exhibit exceptional proficiency in out-of-domain detection, achieving remarkable results in identifying objects that were previously unobserved. Nonetheless, the efficacy of these models is fundamentally contingent upon the quality of the textual prompt inputted, highlighting the importance of well-constructed prompts for optimal model performance.

In our work, we aim to further enhance the efficiency of these detectors during inference by proposing a prompt tuning approach named Reframing. This task can be seen as prompt tuning for various modalities and interest tasks, including image generation, Large Language Model (LLM) queries, and notably, queries for object detection models. There are many techniques for open-set detectors query optimization, such as DetPro [2] or ViLD [5]. However, the success of these methods depends on defining a specific set of identifiable image classes, crucial for prompt generation.

Our concept involves optimizing the user's prompt in such a way that it aligns closely with the specific detector's characteristics. We hypothesize that each detector has a certain textual bias, conditioned by the data on which it was trained. Hence, our objective is to reformulate the user's prompt to better match this bias.

To achieve this goal, we have adapted an additional model based on LLMs, trained to transform the user's prompt into a format most suitable for a specific model. Through metrics such as Intersection over Union (IoU), we can understand which types of requests the detector responds to more effectively. This is achieved by evaluating the quality of work for various prompt requests to a single object and then, based on the assessments, such as IoU, determining the types of requests to which the detector reacts better. The conceptual scheme of our method is presented in the Fig. 1.

Author Proof

Our main contributions are as follows:

1. We propose a new pipeline, Reframing, to enhance the capabilities of open-vocabulary detection models. This pipeline utilizes large language models to improve the quality of input queries.
2. We offer two training modes for Reframing: online and offline. These modes imply the use of the detection model either before or during training.
3. We also introduce a new dataset, GOLD-A, which addresses issues such as similarity and monotony in image appearance, as well as object placement in the original dataset.

2 Related Work

Open-Vocabulary Object Detection. Open-vocabulary object detection leverages existing bounding box annotations to recognize a wide range of classes through language generalization. GroundingDINO [10] enhances DINO [19] by integrating vision-language fusion at various stages, underscoring the critical role of prompt selection for each detector's performance.

The You Only Look Once(YOLO) [13] series, known for its efficiency, is constrained to predefined categories, limiting its applicability in open-world scenarios. YOLO-World [1] 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 diversity of methods, the selection of appropriate prompts remains crucial for achieving optimal performance.

Prompt Tuning Algorithms. Prompt tuning strategies exhibit considerable variation. Context Optimization (CoOP) [20] optimizes context for few-shot classification by learning continuous representations from limited data. Vision and Language knowledge Distillation (ViLD) [5] employs a static template such as "a photo of [CATEGORY] in the scene". Detection Prompt (DetPro) [2] builds on the foundations of ViLD and CoOP for Open-Vocabulary Object Detection, introducing automatic prompt learning and a novel approach to background interpretation. However, these methods depend on predefined classes and aim to generalize to novel ones.

Several studies utilize large language models (LLMs) to adapt human prompts. Optimization for text-to-image generative models (OPT2I) [11] modifies input requests based on feedback from a text-to-image model. Promptist [7] employs Proximal Policy Optimization (PPO) to improve text-to-image prompts.

Language Models Fine-Tuning Algorithms. Fine-tuning language models frequently involves reinforcement learning. Reinforcement Learning from Human Feedback (RLHF) [15,21] utilizes PPO and human evaluations to enhance performance. Reinforced Self-Training (ReST) [6] applies offline reinforcement learning

Author Proof

algorithms for further refinement. Despite their successes, these methods often encounter computational complexity and instability. Alternative approaches treat fine-tuning as satisfying constraints within a target probability distribution [4,9], necessitating precise definitions of these constraints. 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 [17].

3 Reframing

Our approach is based on tuning an additional large language model, which we call the Reframing network. This model 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 Reframing 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, the reframing network must be specially trained to work with the specific detector network.

Our framework offers two training modes: **offline** and **online**. 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 mode includes pre-training, where Supervised Fine-Tuning (SFT) is used in the first stage, and then the language model is fine-tuned using the PPO algorithm, utilizing feedback from the detector. Both 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 [8] 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 Reframing.

Author Proof

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.

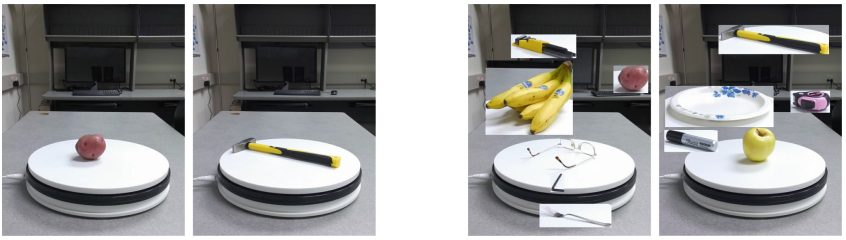
The paraphrasing process is applied to a series of base descriptions linked with specific object IDs and bounding box information. For each entry, the system produces multiple paraphrased versions using the defined prompt, and the results include paraphrased sentences along with a correctness score.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

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

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 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 Fig. 4 (Fig. 2).

Author Proof



(a) GOLD images (b) GOLD-A images

Fig. 2. Example of images from GOLD and GOLD-A datasets.

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) [12] 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}:

$$\nabla_\theta \mathcal{L}_{DPO}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim D}[\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w)) \\ (\nabla_\theta \log \pi_\theta(y_w|x) - \nabla_\theta \log \pi_\theta(y_l|x))], \tag{1}$$

where $\hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ 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.

Author Proof

Algorithm 1: Reframing 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**
      $\nabla_\theta \mathcal{L}_{DPO}(\pi_\theta; \pi_{ref}) \leftarrow$
      $\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w))\,(\nabla_\theta \log \pi_\theta(y_w|x) - \nabla_\theta \log \pi_\theta(y_l|x))$
    end
  end

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 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) [14], 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:

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

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.

4 Experiments

In this section, we describe the experiments conducted using our Reframing 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.

Models and Datasets. For our experiments, encoder-only architectures trained on instructive datasets were selected. Specifically, the Llama2(7b)

Author Proof

Algorithm 2: Reframing online 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

[16] and Gemma(7b) [3] 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.

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 our proposed GOLD-A dataset, during the testing phase, we also incorporated another dataset referred to as 'External'. This dataset consists of labeled photographs intended to enhance our evaluation process. Utilizing Chat GPT-4, we generated detailed textual descriptions for the objects identified within these photographs, encompassing a total of 50 descriptions across 15 different classes.

Implementation Details. We implement our Reframing approach using Transformer Reinforcement Learning (TRL) library [18]. Our experiments are conducted on a single H100 GPU. During the training phase, our configuration includes a batch size of 16. In our online mode, we allocate 30k steps for the PPO phase and 20k steps for the SFT phase, while our offline mode involves 50k training steps.

Author Proof

4.1 Results

Table 2. Comparison of learning outcomes of different LLMs.

Detection Model	Prompt tuning approach	External, IoU	GOLD-A, 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**
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**

Through our Reframing 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 Reframing tuning, and models trained with online and offline modes of Reframing.

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 Reframing Help Improve the LLM's Ability to Reframe Prompts in Light of the Detection Model's Preferences? Our research explores whether the Reframing 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 of Reframing significantly boost the prompt-tuning capabilities of the language models over their pretrained counterparts. Additionally, models operating in online Reframing mode outperform human-generated prompts. Llama2-7b trained with

Author Proof

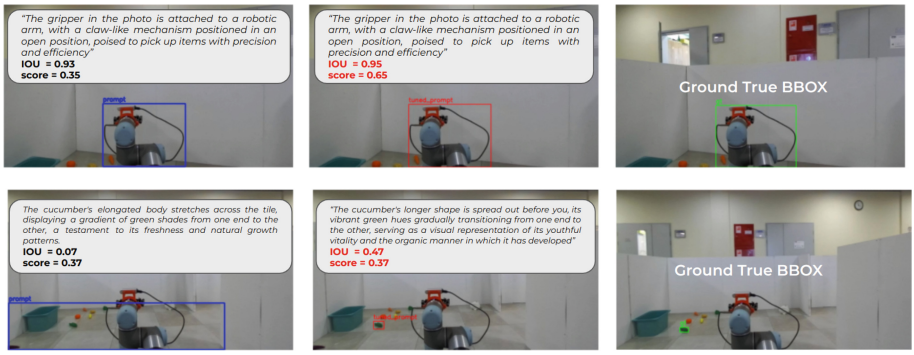


Fig. 3. Detection results: original (left), reframed (center), ground-truth bbox (right). Reframing improves metrics and bbox accuracy.

Table 3. Comparison of pre and post-reframing prompt requests.

Human Prompt	IoU	Prompt after Reframing	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

Reframing online mode shows mean IoU equal to 0.585, while original human prompts show only 0.572 (Fig. 3). AQ2

How Should the Improvements Introduced by the Model Be Interpreted? The evaluation of model performance can be significantly enriched by comparing the detection outcomes on prompt requests, pre and post-reframing for diverse images. Examples of such prompts are shown in Table 3. 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, reframing 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.

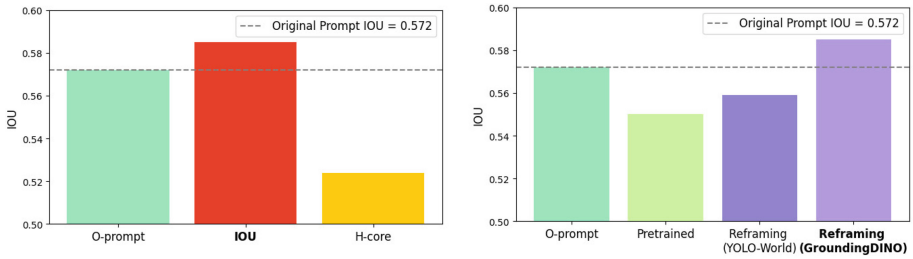
Author Proof

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 Fig. 4a.

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 Reframing for one Detection Model Be Used to Enhance the Performance of Another Detection Model? Our studies investigated whether language models trained using Reframing on one detection model could enhance the performance of another. Results shown in Fig. 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.



(a) (b)

Fig. 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

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.

Author Proof

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.

Acknowledgement. This work was supported by Russian Science Foundation, grant No. 20-71-10116, https://rscf.ru/en/project/20-71-10116/.

References

1. Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: Yolo-world: real-time open-vocabulary object detection. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR) (2024)
2. Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., Li, G.: 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, pp. 14084–14093 (2022)
3. Gemma Team, T.M., et al.: Gemma (2024). https://doi.org/10.34740/KAGGLE/M/3301. https://www.kaggle.com/m/3301
4. Go, D., Korbak, T., Kruszewski, G., Rozen, J., Ryu, N., Dymetman, M.: Aligning language models with preferences through f-divergence minimization. In: International Conference on Machine Learning (2023)
5. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021)
6. Gulcehre, C., et al.: Reinforced self-training (rest) for language modeling. arXiv preprint arXiv:2308.08998 (2023)
7. Hao, Y., Chi, Z., Dong, L., Wei, F.: Optimizing prompts for text-to-image generation. Advances in Neural Information Processing Systems **36** (2024)
8. Kebe, G.Y., et al.: 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). https://openreview.net/forum?id=Yx9jT3fkBaD
9. Khalifa, M., Elsahar, H., Dymetman, M.: A distributional approach to controlled text generation. arXiv preprint arXiv:2012.11635 (2020)
10. Liu, S., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
11. Mañas, O., et al.: Improving text-to-image consistency via automatic prompt optimization. ArXiv **abs/2403.17804** (2024)
12. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems **36** (2024)

Author Proof

13. Redmon, J., Divvala, S.K., Girshick, R.B., Farhadi, A.: 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, pp. 779–788. IEEE Computer Society (2016)
14. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
15. Stiennon, N., et al.: Learning to summarize with human feedback. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 3008–3021. Curran Associates, Inc. (2020)
16. Touvron, H., et al.: Llama 2: open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
17. Volovikova, Z., Skrynnik, A., Kuderov, P., Panov, A.I.: Instruction following with goal-conditioned reinforcement learning in virtual environments (2024). https://arxiv.org/abs/2407.09287
18. von Werra, L., et al.: Trl: Transformer reinforcement learning (2020). https://github.com/huggingface/trl
19. Zhang, H., et al.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)
20. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. Int. J. Comput. Vision **130**(9), 2337–2348 (2022)
21. Ziegler, D.M., et al.: Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593 (2019)

Author Proof

Author Queries

Chapter 11

Query Refs.	Details Required	Author's response
AQ1	This is to inform you that corresponding author has been identified as per the information available in the Copyright form.	
AQ2	Please check and confirm if the inserted citation of Figs. 2 and 3 are correct. If not, please suggest an alternate citation.	