Lenna: Language Enhanced Reasoning Detection Assistant

Fei Wei[1], Xinyu Zhang[1], Ailing Zhang[1,2], Bo Zhang[1], Xiangxiang Chu[1*]
[1] Meituan Inc. [2] Beihang University
{weifei06, zhangxinyu35}@meituan.com {annyzhang}@buaa.edu.cn {zhangbo97, chuxiangxiang}@meituan.com

Abstract

With the fast-paced development of multimodal large language models (MLLMs), we can now converse with AI systems in natural languages to understand images. However, the reasoning power and world knowledge embedded in the large language models have been much less investigated and exploited for image perception tasks. In this paper, we propose ***Lenna****, a* ***L****anguage* ***e****nhanced reaso****n****ing detectio****n*** ***a****ssistant, which utilizes the robust multimodal feature representation of MLLMs, while preserving location information for detection. This is achieved by incorporating an additional* `<DET>` *token in the MLLM vocabulary that is free of explicit semantic context but serves as a prompt for the detector to identify the corresponding position. To evaluate the reasoning capability of Lenna, we construct a ReasonDet dataset to measure its performance on reasoning-based detection. Remarkably, Lenna demonstrates outstanding performance on ReasonDet and comes with significantly low training costs. It also incurs minimal transferring overhead when extended to other tasks. Our code and model will be available at* https://github.com/Meituan-AutoML/Lenna.

1. Introduction

Recent accelerating advancement of large language models (LLM) [16, 41, 43] has amplified the model's capacity for natural language comprehension and generation. Bolstered by these large language models, multimodal large language models (MLLM) have achieved significant performance leaps in perception tasks (*e.g.*, detection [37, 46], segmentation [23]), generation tasks (*e.g.*, captioning [25], VQA [29, 55]).

Referring Expression Comprehension (REC), serving as a crucial task for assessing the natural language understanding and positioning capability of multimodal large models, has been the focal point of numerous studies, *e.g.*, InstructDet [11], Shikra [5], miniGPT-v2 [4]. The objective of the REC task is to obtain the position of an object given explicit instructions, such as positional words, colors, shapes, *etc*. Nonetheless, rare attention has been paid to the model's reasoning ability which demands an understanding of implicit intentions. DetGPT [37] represents the inaugural research to propose the reasoning-based object detection task. Yet its weak connection of the reasoning and the localization process, via an intermediate bridge in the form of several

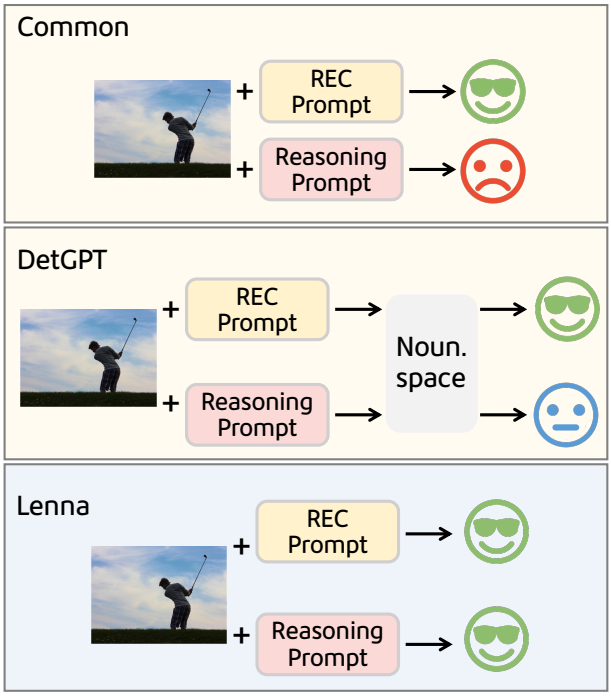


Figure 1. Illustration of common MLLM-based REC methods, DetGPT, and Lenna. A REC prompt comprises explicit instructions of the target, typically describing its position, color, and shape (*e.g.* "*A long, slender golf club*"). A reasoning prompt contains implicit intentions (*e.g.* "*something showing that the man is playing sports*"). Common methods fail to handle the reasoning prompt effectively. DetGPT struggles to translate the input prompts to the target names in the noun space. Lenna excels at handling both.

*Corresponding author.
†This work is done when Ailing Zhang is an intern at Meituan. This project is still under development and the reported results are subject to frequent modifications.

nouns, results in substantial information loss in the multimodal feature space. Figure 1 illustrates the differences among the common MLLM-based REC methods, DetGPT, and our approach.

To facilitate the reseasoning capacity and world knowledge of LLMs in the perception task, we introduce Lenna, a **L**anguage **E**nhanced reaso**N**ing detectio**N** **A**ssistant. Drawing inspiration from LISA [23], we discover that the token embedding of LLMs possesses a robust multimodal feature representation capability, which can transmit localization information to the detector without forfeiting reasoning ability. To achieve this, we incorporate an additional <DET> token as a signal to convey object detection information. Unlike other words in the existing vocabulary, <DET> token is free of explicit semantic context. The hidden embedding of the <DET> token functions as a prompt for the detector, assisting it in pinpointing the relevant target position.

Owing to the simplistic design of the <DET> token, we broaden the REC-based object detection to reasoning-based object detection without altering their original design. We curate a ReasonDet dataset processed from ReasonSeg [23] to evaluate reasoning detection capacity. Leveraging LLM's world knowledge and intelligence, Lenna also exhibits commendable performance on ReasonDet.

It is noteworthy that in contrast to other existing works, Lenna incurs a significantly lower training cost. We expended merely 20 hours in training on 8 A100 GPUs. Due to the streamlined design, Lenna also holds the potential to be extended to other tasks (*e.g.*, instance segmentation, grounding) with minimal cost implications.

Our main contributions can be summarized as follows.

- We propose Lenna, our language-enhanced reasoning detection assistant, that incorporates REC-based and reasoning-based detection in the same simplistic and extensible framework.
- We curate a benchmark dataset called ReasonDet to quantitatively measure the reasoning detection performance of MLLMs.
- Lenna comes with inexpensive training cost and outperforms previous MLLMs on REC and ReasonDet. Concurrently, the visualization results from ReasonDet affirm Lenna's consistent capability in reasoning object detection.

2. Related Work

2.1. Referring Expression Comprehension

The task of Referring Expression Comprehension (REC) is aimed at locating the objects that are explicitly referred to by free-form guided language expressions. This task has been widely adopted by a plethora of works [5, 11, 47, 50] as a standard for monitoring the performance of their models. The evaluation covers language understanding and object localization, offering a holistic measurement of the model's competencies. GLIP [26] unifies object detection and grounding by reformulating object detection as a phrase grounding task. Grounding-DINO [31] extends an original detector DINO [54] to an open-set detector by performing vision-language modality fusion with BERT [13] in multiple phases. DQ-DETR [30] utilizes dual decoupled queries to alleviate the difficulty of modality alignment between image and text in the DETR [3] framework. Moreover, benefiting from the development of language models, a lot of works [6, 8, 34, 44, 45, 56] cast object detection as a language modeling task and achieve better performance on REC. PEVL [51] reformulates discretized object positions and language in a unified language modeling framework. UniTab [50] considers texts and box predictions as an auto-regressive token generation task and presents a unified encoder-decoder model fully shared for texts, boxes, and alignment predictions. However, in scenarios where the text lacks an explicit referent and necessitates reasoning from the language model, current models are largely constrained by the comprehension capacity, thereby inadequately addressing such situations. Conversely, Lenna, schemed to leverage the benefits of LLM, exhibits proficiency in understanding complex semantics, reasoning in the context, and accurately pinpointing the target content.

2.2. Multimodal Large Language Model

Large Language Models (LLMs) [15, 16, 41, 43, 53] have exceptional performance in diverse natural language processing (NLP) tasks. Expanding large language models to a multimodal version [19, 24, 48, 52] has naturally obtained increasing attention. The exemplary efficacy of incumbent Vision Transformers [9, 10, 14, 32, 39, 42] lays a robust groundwork for the expansion. Flamingo [1] employs adaption layers to pretrained LLM layers to incorporate visual information for the next-token prediction task. LLaVA [29] makes the first attempt to utilize GPT4 [36] to generate visual instruction data and converts image feature into the word embedding space with a simple linear layer. Shikra [5] introduces the task of Referential Dialogue and injects visual grounding capabilities into LLMs. MiniGPT-v2 [4] proposes a unified interface model for handling various vision-language tasks by a task-oriented instruction training scheme. BLIP-2 [25] designs a Q-Former module to feed the most useful visual feature to the LLM to have the desired text output. VisionLLM [46] aligns the definitions of vision-centric tasks with the methodologies of LLMs. Closely related to us, DetGPT[37] represents an initial foray into reasoning object detection, establishing a linkage between MLLM and open-vocabulary detector through the medium of object names. Nonetheless, the reliance on object names substantially curtails the representational prowess of the feature space. In stark contrast, Lenna exploits the rich characterization inherent in MLLM embeddings, which encapsulate and transmit

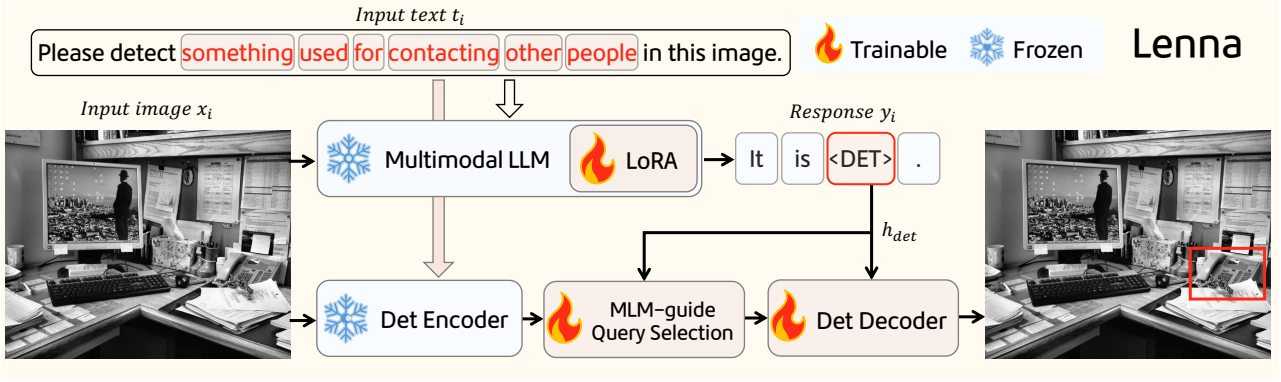


Figure 2. The pipeline of Lenna. Given an input image x_i and text t_i, the multimodal LLM generates a text output. The last-layer embedding for the <DET> token is subsequently incorporated into MLM-guideQuerySelection and DetDecoder to ascertain the object's location. The Det model is architecturally based on Grounding-DINO, which takes text caption t_i' (marked in red) out of t_i and the image x_i as input.

semantic and localization information to the detector.

3. Method

3.1. Architecture Design

We introduce Lenna, an end-to-end language-enhanced reasoning detection assistant. Figure 2 delineates its framework. In essence, Lenna is an amalgamation of a multimodal large language model LLaVA [29] and an open-set detector Grounding-DINO [31]. Analogous to LISA [23], we initially extend the original LLM vocabulary with a special token <DET> to signify the demand for detection output. Upon receiving an image x_i and a text instruction t_i, the multimodal large language model $\mathcal{M}$ engenders a text response y_i, which is formulated as

$$y_i = \mathcal{M}(x_i, t_i). \tag{1}$$

A <DET> token is necessitated to distinguish task types when MLLM is asked to undertake an object detection assignment. Consequently, we obtain an embedding h_{det} that corresponds to the <DET> token, which is laden with both semantic and location information related to the target. Simultaneously, the pre-trained encoder of the detector, denoted as $\mathcal{D}_{enc}$, extracts enhanced image feature f_{img} and text feature f_{txt}, which can be formulated as

$$f_{img}, f_{txt} = \mathcal{D}_{enc}(x_i, t_i'). \tag{2}$$

where t_i' is the object caption (marked in red in Fig. 2) in t_i. Subsequently, h_{det}, f_{img} and f_{txt} are fed into the MLM-guide query selection (MQS) module, which facilitates cross-space alignment between the feature spaces of BERT-based and LLM-based models. For better modality alignment in the implementation, MQS is designed to incorporate both a cross-attention module and a similarity calculation module,

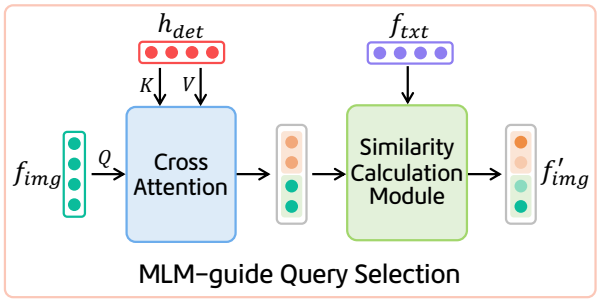


Figure 3. Architecture illustration of MLM-guideQuerySelection module (MQS). The enhanced features f_{img} and f_{txt} are extracted from the encoder of the detector. h_{det} represents the embedding corresponding to the <DET> token. The output of MQS is denoted as f'_{img}, which serves as the input feature of the decoder.

as shown in Figure 3. In the *cross-attention module*, we employ h_{det} as K, V to activate the corresponding features in the enhanced image feature f_{img}. In the *similarity calculation module*, akin to Grounding-DINO, we select features that exhibit greater relevance to the input text feature f_{txt}. The operation of MQS can be formulated as

$$f'_{img} = MQS(h_{det}, f_{img}, f_{txt}) \tag{3}$$

Ultimately, h_{det} is incorporated into each text cross-attention layer of decoder $\mathcal{D}_{dec}$ resulting in the final location $pred$, which can be represented as

$$pred = \mathcal{D}_{dec}(h_{det}, f'_{img}). \tag{4}$$

3.2. Optimization Objectives

We construct an end-to-end training process where Lenna employs a loss function that combines the auto-regressive language modeling loss $\mathcal{L}_{tok}$ and the detection loss $\mathcal{L}_{det}$,

amalgamated according to a specific weight ratio λ_{tok} and λ_{det}, which can be mathematically expressed as

$$\mathcal{L} = \lambda_{tok}\mathcal{L}_{tok} + \lambda_{det}\mathcal{L}_{det} \tag{5}$$

In auto-regressive language modeling loss $\mathcal{L}_{tok}$, we maximize the likelihood of target token y_t conditioned on input image x_i, input text t_i, and previous target tokens $y_{j'}$, which can be formulated as

$$\mathcal{L}_{tok} = -\sum_{j=1}^{L} \log p\left[y_j | y_{j'}, x_i, t_i\right] \quad \text{where} \quad j' < j \tag{6}$$

We follow Grounding-DINO in $\mathcal{L}_{det}$ to use the L1 Loss and the GIOU loss [40] as bounding box regression, and use contrastive loss [39] for classification, which is formulated as

$$\begin{aligned} \mathcal{L}_{det} = & \lambda_{L1} L1(pred, gt) \\ & + \lambda_{GIOU} GIOU(pred, gt) \\ & + \lambda_{Contrast} Contrast(pred, gt) \end{aligned} \tag{7}$$

3.3. Training Data Formulation

As shown in Figure 4, our model is trained on four distinct data types, which serve as a strategy to endow our model with image-text feature matching capabilities and proficiency by aligning two multimodal feature spaces.

Object detection data. Typically, the ground truth data for object detection contains the positional and categorical information of all targets within a predefined list of categories. During the training phase, we formulate questions and answers adhering to a specific template: *"**User:** <image> Please detect the* {category} *in this image. **Assistant:** Sure,* <DET>.", wherein the {category} is randomly chosen from the ground truth categories present in the image, <image> is the placeholder of image tokens. The indicator <DET> denotes that the current input requests the model to compute the detection loss. For this purpose, we adopt COCO dataset [28] for training. In addition, to augment the model's discriminative capacity for diverse categories, we employ all category names from the training dataset as the textual input for Grounding DINO.

Referring expression comprehension data. REC data typically provides images and the descriptive phrase corresponding to the target bounding box. We formulate questions and answers with a subsequent template: *"**User:** <image> What is* {caption} *in this image? Please output object location. **Assistant:** It is* <DET>." where the {caption} is the descriptive phrase supplied in the dataset, generally a noun description augmented with several adjectives. To guarantee the diversity of the training data, we have also devised a variety of similar question templates, which are randomly chosen during the training process. We adopt most widely used REC dataset RefCOCO [21], RefCOCO+ [21], RefCOCOg [35].

Table 1. Comparison of training costs across various models.

Model	GPU days
DQ-DETR [30]	70
Shikra [5]	40
MiniGPT-v2 [4]	35
Lenna	**7**

Reasoning detection data. To enhance the model's applicability in scenarios necessitating comprehension of reasoning questions, such as embodied AI robots, we processed the ReasonSeg dataset [23] at the instance level to yield a reasoning detection benchmark called **ReasonDet**. Similar to the partition of ReasonSeg data from LISA [23], the training set comprises 239 images and 1326 texts, while the validation set includes 200 images and 344 texts. During training, the questions are divided by their lengths. The template for short questions aligns with the REC template. For long questions, we adhere to the following template: *"**User:** <image>* {question} *Please output object location and explain the reason. **Assistant:** Sure, the detection result is* <DET>, {reason}." where {question} is a long question asked in a natural language scenario, and {reason} is the explanation given by the model for the current box prediction.

Visual question answering data. To preserve the inherent visual question answering (VQA) capability of the multimodal large language model, we also incorporated the LLaVA-Instruct-150k [29] dataset generated by GPT-4 during training.

4. Experiment

4.1. Implementation Details

Network architecture. Lenna consists of a multimodal large language model $\mathcal{M}$ and an open-set detector $\mathcal{D}$. Within the multimodal large language model, we employ pre-trained LLaVA-7B [29] for efficient training. Our open-set detector is designed on Grounding-DINO with Swin-Large Transformers [32] as the vision backbone. During the training stage, LoRA [18] is utilized for an efficient finetuning of LLM. We thoroughly train the MQS module and decoder in the detector, while all other parameters are frozen to preserve the original capabilities of the pre-trained model.

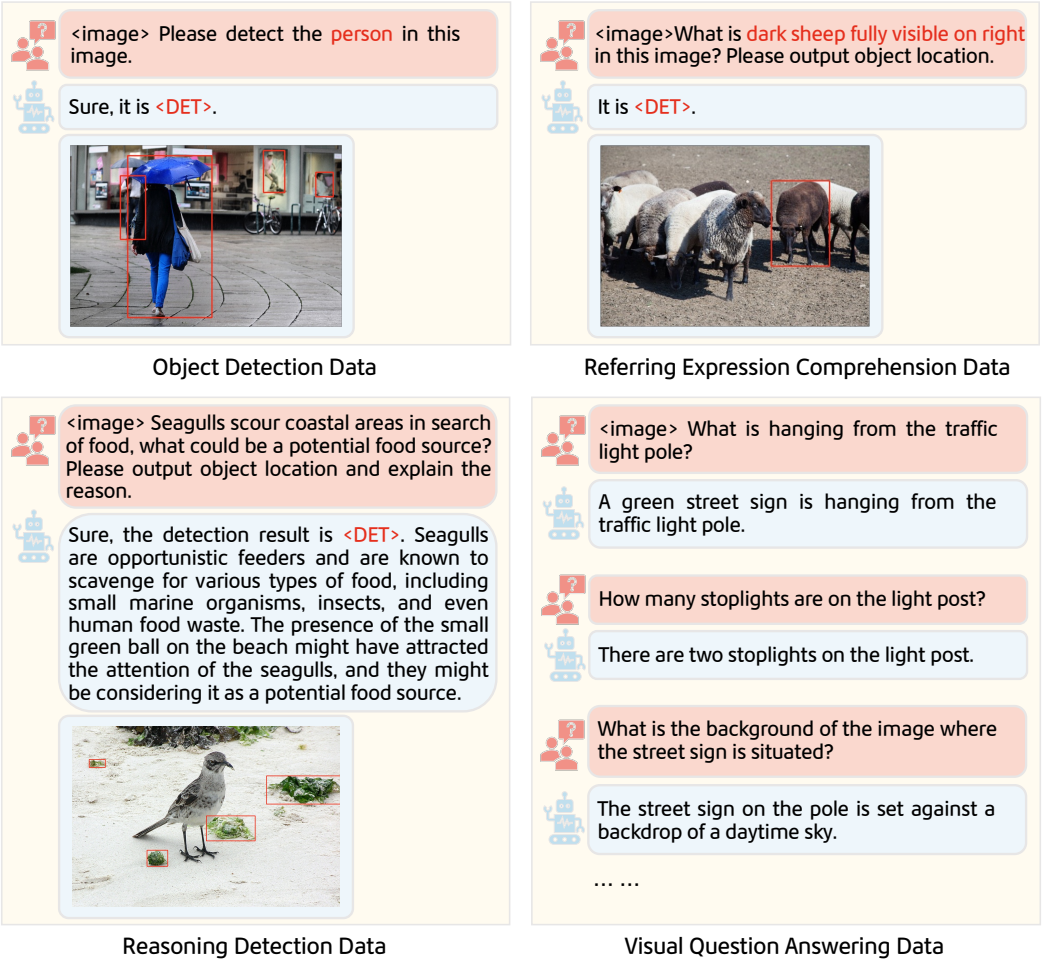


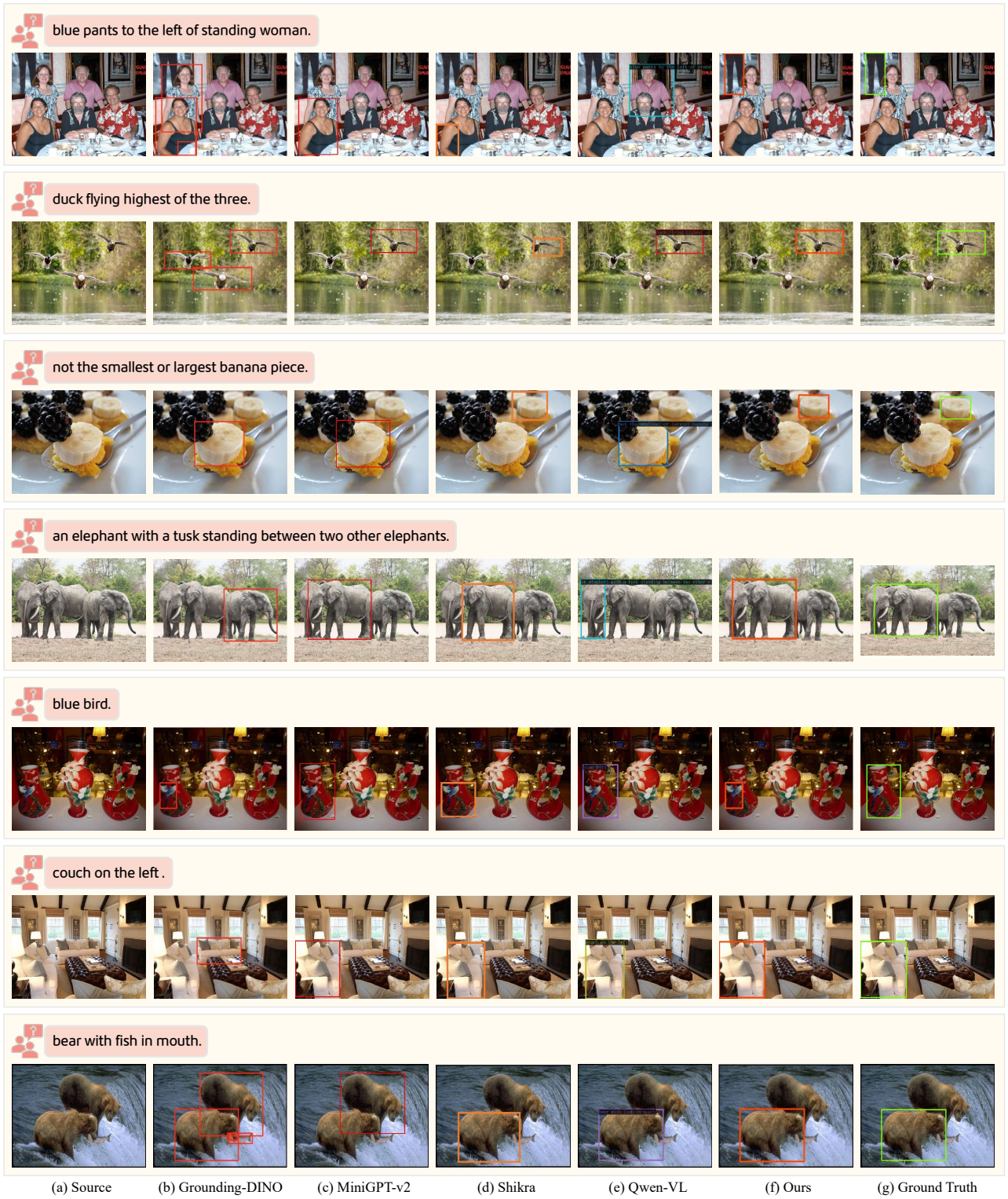
Figure 4. Examples of various types of training data for Lenna, including object detection data (OD), referring expression comprehension data (REC), reasoning detection data (RD), and visual question answering (VQA) data. In addition to the VQA data, the response of the remaining data incorporates the <DET> token, signifying the activation of the detection task. Correspondingly, the ground truth of its associated regression box is also present in red.

Hyperparameter. The training takes approximately 20 hours on 8 NVIDIA A100 GPUs. We adopt AdamW [33] optimizer with a learning rate of 3e-4 and use a learning rate scheduler WarmupDecayLR with the warmup steps of 10. The batch size is set to 2 on each device. The modulating parameters λ_{tok} and λ_{det} in total loss $\mathcal{L}$ are set to 1.0 and 1.0, respectively. In λ_{det}, there are λ_{L1}, λ_{GIOU} and $\lambda_{Contrast}$ which are set to 5.0, 2.0 and 1.0.

4.2. Comparison with State-of-the-art Methods

Training resource consumption. As shown in Table 1, we compare the training resource consumption with DQ-DETR [30], Shikra [5] and MiniGPT-v2 [4] on NVIDIA A100 GPUs. In contrast to existing works, Lenna incurs a significantly lower training cost with its simplistic model architecture and efficient training strategy.

Quantitative comparison. To guarantee the impartiality of the comparison, we evaluate all methods on RefCOCO, RefCOCO+, and RefCOCOg using the accuracy metric with an IoU of 0.5. As shown in Table 2, we first compare our method with specialists that encompass localization models and fine-tuned generalist or foundation models on localization tasks, such as TransVG [12], UNITER [7], VILLA [17], RefTR [27], MDETR [20], UNICORN [49], DQ-DETR [30], InstructDet [11], GroundingDINO [22]. Additionally, our method is also compared with generalist VL models without fine-tuning, including OFA [45], Vision-LLM [46], Shikra [5], MiniGPT-v2 [4], PerceptionGPT [38], Qwen-VL [2]. These generalist VL models are capable of undertaking a variety of vision-language tasks, including image captioning, VQA, REC, *etc*. Among all listed models, Lenna comprehensively demonstrates a promising perfor-



(a) Source (b) Grounding-DINO (c) MiniGPT-v2 (d) Shikra (e) Qwen-VL (f) Ours (g) Ground Truth

Figure 5. Visualization comparison with other VLMs on various referring expressions.

Table 2. Results on referring expression comprehension task. The best results are marked in bold.

Type	Model	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
Specialist SOTAs (Specialist/Finetuned)	TransVG [12]	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73
	UNITER [7]	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77
	VILLA [17]	82.39	87.48	74.84	76.17	81.54	66.84	76.18	76.71
	RefTR [27]	85.65	88.73	81.16	77.55	82.26	68.99	79.25	80.01
	MDETR [20]	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
	UNICORN [49]	88.29	90.42	83.06	80.30	85.05	71.88	83.44	83.93
	DQ-DETR [30]	88.63	91.04	83.51	81.66	86.15	73.21	82.76	83.44
	InstructDet [11]	88.92	90.86	85.57	78.27	83.39	71.04	83.01	82.91
	Grounding-DINO [22]	89.19	91.86	85.00	81.09	87.40	74.71	84.15	84.94
Generalist VL SOTAs (w/o finetuning)	OFA-L [45]	79.96	83.67	76.39	68.29	76.00	61.75	67.57	67.58
	VisionLLM-H [46]	-	86.70	-	-	-	-	-	-
	Shikra-13B [5]	87.83	91.11	81.81	82.89	87.79	74.41	82.64	83.16
	MiniGPT-v2-7B [4]	88.69	91.65	85.33	79.97	85.12	74.45	84.44	84.66
	PerceptionGPT-13B [38]	89.17	93.20	85.96	83.72	89.19	75.31	83.75	84.69
	Qwen-VL-7B [2]	89.36	92.26	85.34	83.12	88.25	77.21	85.58	85.48
ours	Lenna-7B	**90.28**	**93.22**	**86.97**	**88.08**	**90.07**	**83.99**	**90.30**	**90.29**

Table 3. Quantitative results on the validation set of ReasonDet with the metric of accuracy. RD refers to the training set of ReasonDet.

Exp	*Acc.*
Qwen-VL [2]	12.96
Shikra [5]	20.27
MiniGPT-v2 [4]	25.25
Lenna (w/o RD)	37.21
Lenna	**46.84**

Table 4. Comparison on pre-trained parameters. The second column is the model chosen for detector $\mathcal{D}$. We evaluate through the metric of accuracy on the val set of the REC dataset.

Exp	$\mathcal{D}$	RefCOCO	RefCOCO+	RefCOCOg
E1	From Scratch	24.46	23.74	33.48
E2	GDINO-T	89.27	87.47	88.50
E3	GDINO-B	89.36	85.37	**90.57**
E4	GDINO-L	**90.28**	**88.08**	90.30

mance advantage.

The quantitative results on the ReasonDet dataset are presented in Table 3. We performed a comparative analysis between our proposed method, Lenna, and existing MLLM methods, including MiniGPT-v2 [4], Shikra [5], and Qwen-VL [2]. To ensure a fair comparison, we excluded the ReasonDet data from the training dataset, as indicated in the Lenna (w/o RD) row in Table 3. The results demonstrate that irrespective of whether ReasonDet data is used in training, our method significantly surpasses other techniques. Lenna (w/o RD) achieves a 47.37% improvement over the best-performing MiniGPT-v2 in the SOTAs, and Lenna even exceeds this by 85.50%. This provides solid evidence that Lenna can genuinely comprehend the content within the problem and accomplish precise positioning.

Table 5. Ablation study on training data. We evaluate through the metric of accuracy on the val set of the REC dataset.

Exp	Dataset REC	OD	VQA	RefCOCO	RefCOCO+	RefCOCOg
E1	✓			31.78	30.85	20.53
E2	✓		✓	85.62	85.35	87.01
E3	✓	✓		89.45	85.92	88.77
E4	✓	✓	✓	**90.28**	**88.08**	**90.30**

Qualitative comparison. Figure 5 depicts a qualitative comparison, showcasing an array of scenes and REC results produced by diverse methods. The results suggest that while

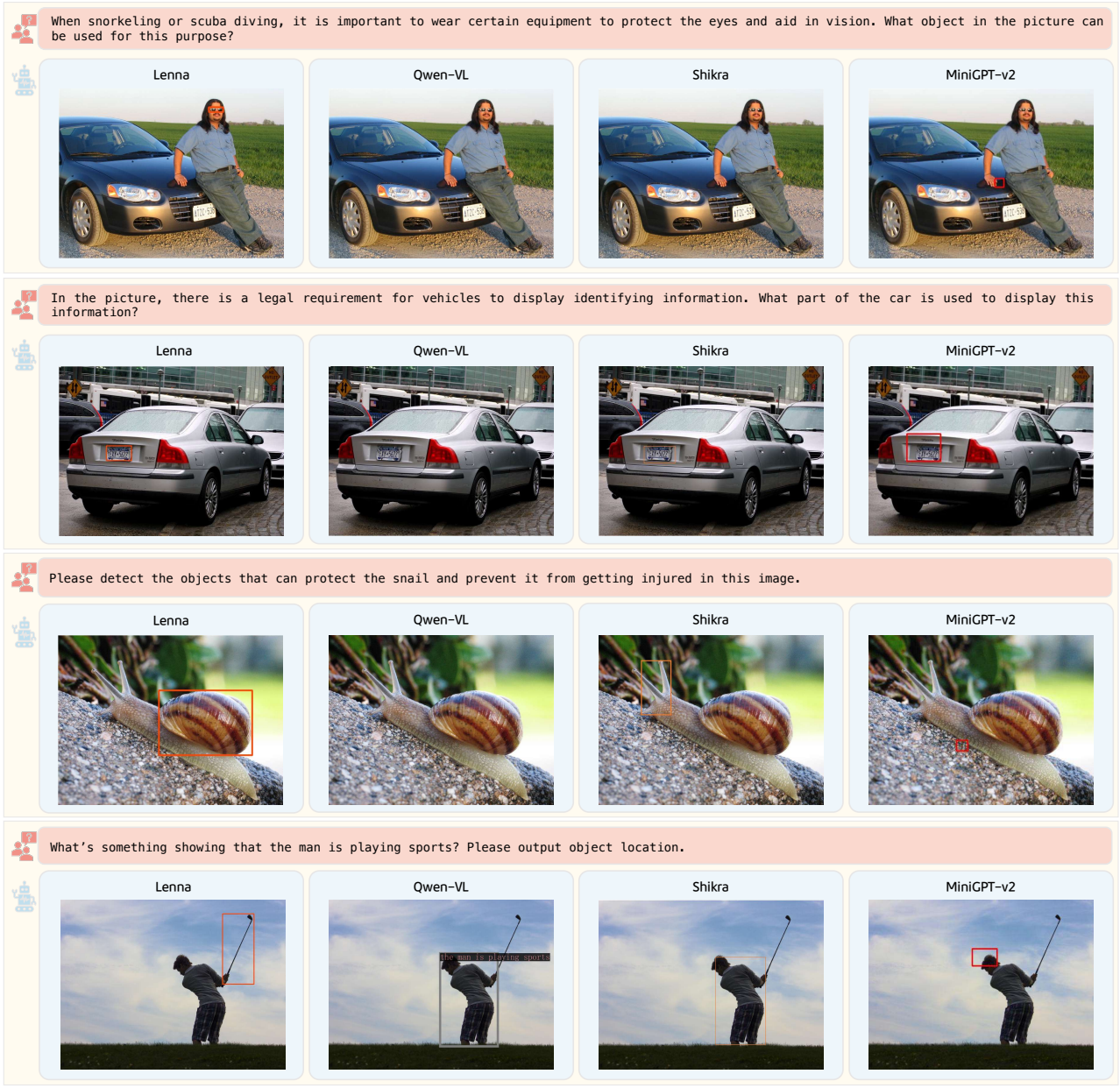


Figure 6. Visualization comparison of reasoning capability with generalist VLMs.

Grounding-DINO is adept at comprehending simple and explicit text information such as color and position, it struggles with scenarios where understanding the relationship between multiple target positions within an image is required. A common characteristic of MiniGPT-v2 [4], Shikra [5], and Qwen-VL [2] is that the discrete output form of their language models results in an increased challenge in object location and a certain degree of loss in positioning accuracy. On the other hand, our method consistently outperforms in understanding complex language information and achieving accurate localization.

Figure 6 further illustrates the comparison in reasoning capability. Most existing methods struggle with this task, as the reasoning task diverges from the REC task in that it requires the model to not only understand the question's real meaning but also possess a world knowledge base, which in turn enables reasoning and positioning. Figure 6 showcases some results of MiniGPT-v2 [4], Shikra [5], Qwen-VL [2] and Lenna on the ReasonDet dataset, with Lenna demonstrating superior performance in various reasoning scenarios of differing difficulty levels, such as long-question (first and second rows) and short-question (third and fourth rows).

4.3. Ablation Study

Model scaling. To underscore the significance of the model scale, we have conducted a comprehensive series of comparative experiments on the scale of the detector. As depicted in E2-E4 in Table 4, the results suggest that a larger scale of $\mathcal{D}$ can effectively alleviate the model fitting complexity and confer superior performance enhancement on the model. Furthermore, the pre-trained weights of the detector hold significant importance. As demonstrated in E1 of Table 4, training the detector from scratch results in a decline in model performance.

Training data. In Table 5, we display the contribution of each type of training data to the performance. It is evident that both object detection (OD) data and VQA data exert varying degrees of impact on the model performance. The OD data provides explicit guidance for semantic alignment to the model, while the VQA data contributes to the diversification of the <DET> embedding.

5. Conclusion

We present Lenna, a novel framework that leverages the representational power and world knowledge of large language models (LLM) to enhance reasoning in object detection tasks. Lenna introduces a unique <DET> token embedding to facilitate accurate positioning without losing reasoning information. Lenna stands out due to its efficient training and the ability to extend to various tasks with minimal additional costs. Its design simplicity allows for rapid adaptation and scaling, demonstrating a notable improvement over previous models in terms of training efficiency and versatility. Owing to Lenna's training efficiency and its expansive application potential, we aspire to furnish novel insights for future research and practical deployments in the domain of multimodal large language models.

6. Acknowledgement

This work is supported by National Key R&D Program of China (No. 2022ZD0118700).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 5, 7, 8

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 2

[4] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 1, 2, 4, 5, 7, 8

[5] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1, 2, 4, 5, 7, 8

[6] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021. 2

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 5, 7

[8] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *arXiv*, 2021. 2

[9] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *NeurIPS 2021*, 2021. 2

[10] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. In *ICLR 2023*, 2023. 2

[11] Ronghao Dang, Jiangyan Feng, Haodong Zhang, Chongjian Ge, Lin Song, Lijun Gong, Chengju Liu, Qijun Chen, Feng Zhu, Rui Zhao, et al. Instructdet: Diversifying referring object detection with generalized instructions. *arXiv preprint arXiv:2310.05136*, 2023. 1, 2, 5, 7

[12] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021. 5, 7

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[15] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. PaLM-E: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 2

[16] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022. 1, 2

[17] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020. 5, 7

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 4

[19] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 2

[20] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 5, 7

[21] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. pages 787–798, 2014. 4

[22] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal generation. *arXiv preprint arXiv:2301.13823*, 2023. 5, 7

[23] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 1, 2, 3, 4

[24] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 2

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2

[26] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 2

[27] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34:19652–19664, 2021. 5, 7

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 4

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 2, 3, 4

[30] Shilong Liu, Shijia Huang, Feng Li, Hao Zhang, Yaoyuan Liang, Hang Su, Jun Zhu, and Lei Zhang. Dq-detr: Dual query detection transformer for phrase extraction and grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1728–1736, 2023. 2, 4, 5, 7

[31] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 3

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 4

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[34] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 2

[35] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 4

[36] OpenAI. Gpt-4 technical report, 2023. 2

[37] Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, and Lingpeng Kong Tong Zhang. Detgpt: Detect what you need via reasoning. *arXiv preprint arXiv:2305.14167*, 2023. 1, 2

[38] Renjie Pi, Lewei Yao, Jiahui Gao, Jipeng Zhang, and Tong Zhang. Perceptiongpt: Effectively fusing visual perception into llm. *arXiv preprint arXiv:2311.06612*, 2023. 5, 7

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4

[40] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 4

[41] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. 1, 2

[42] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training

data-efficient image transformers and distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021. 2

[43] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1, 2

[44] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 2

[45] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 2, 5, 7

[46] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023. 1, 2, 5, 7

[47] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. 2023. 2

[48] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *ICLR*, 2022. 2

[49] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *European Conference on Computer Vision*, pages 733–751. Springer, 2022. 5, 7

[50] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 521–539. Springer, 2022. 2

[51] Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Pevl: Position-enhanced pre-training and prompt tuning for vision-language models. *arXiv preprint arXiv:2205.11169*, 2022. 2

[52] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2

[53] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022. 2

[54] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 2

[55] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1

[56] Xueyan Zou*, Zi-Yi Dou*, Jianwei Yang*, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee*, and Jianfeng Gao*. Generalized decoding for pixel, image and language. 2022. 2