ENHANCING VISUAL GROUNDING AND GENERALIZATION: A MULTI-TASK CYCLE TRAINING APPROACH FOR VISION-LANGUAGE MODELS

Xiaoyu Yang [1,2], Lijian Xu [2,3] (✉), Hao Sun [3], Hongsheng Li [3,4]& Shaoting Zhang [2]
[1] College of Electronics and Information Engineering, Tongji University, Shanghai
[2] Shanghai Artificial Intelligence Laboratory, Shanghai
[3] Centre for Perceptual and Interactive Intelligence, the Chinese University of Hong Kong, Hong Kong
[4] Department of Electronic Engineering, the Chinese University of Hong Kong, Hong Kong

ABSTRACT

Visual grounding occupies a pivotal position in multi-modality vision-language models. However, current vision-language models concentrate on comprehending images, ignoring the human-computer interaction with multi-tasks instructions, thereby imposing limitations on their versatility and depth of responses. In this study, we propose ViLaM, a large multi-modality model, that supports multi-tasks of visual grounding using the cycle training strategy, with abundant interaction instructions. The cycle training between referring expression generation (REG) and referring expression comprehension (REC) is introduced. It enhances the consistency between visual location and referring expressions, and addresses the need for high-quality, multi-tasks visual grounding datasets. Moreover, multi-tasks of visual grounding are promoted in our model, contributed by the cycle training strategy. The multi-tasks in REC encompass a range of granularities, from region-level to pixel-level, which include referring bbox detection, referring keypoints detection, and referring image segmentation. In REG, referring region classification determines the fine-grained category of the target, while referring region captioning generates a comprehensive description. Meanwhile, all tasks participate in the joint training, synergistically enhancing one another and collectively improving the overall performance of the model. Furthermore, leveraging the capabilities of large language models, ViLaM extends a wide range of instructions, thereby significantly enhancing its generalization and interaction potentials. It is particularly advantageous in domains beyond natural images, such as the medical field. Extensive public datasets corroborate the superior capabilities of our model in visual grounding with muti-tasks. Additionally, validating its robust generalization, ViLaM is validated under open-set and few-shot scenarios. Especially in the medical field, our model demonstrates cross-domain robust generalization capabilities. Furthermore, we contribute a visual grounding dataset, especially with multi-tasks. To support and encourage the community focused on visual grounding, we have made both the dataset and our code public: `https://github.com/AnonymGiant/ViLaM`.

***K*eywords** Vision-Language Models · Visual Grounding · Multi-Task · Cycle Training

1 Introduction

Visual grounding serves as a vital bridge that facilitates communication between AI models and the world, representing a significant milestone in the quest for achieving general intelligence. However, traditional approaches to visual grounding predominantly prioritize image comprehension, striving to extract image features with greater accuracy and align them with textual information. Ignoring the wealth of referential information present in the texts of referring expressions leads to the lack of diverse instructions and limited versatility and generalization.

In contrast to traditional methods, the advent of Large Language Models (LLMs) has significantly mitigated this issue, and spawned rethinking about the vision-language models. Leveraging extensive text data, LLMs have attained remarkable proficiency in generating human-like responses and addressing a wide array of tasks. This breakthrough

ushers in a new paradigm for human-computer interaction. Building upon LLMs, large multi-modality models like BLIP2 [1] and MiniGPT4 [2] utilize a pre-trained image encoder and a text encoder, then align vision-language features with simple linear layers. These models demonstrate impressive joint understanding capabilities of language and images, allowing users to give instructions in natural language to perform specific tasks. Nonetheless, the visual grounding task of aligning visual localization with referring expressions remains a challenge for large multi-modality models, particularly when dealing with localization spanning various granularities, such as bounding boxes, keypoints, and segmentated polygons that range from region-level to pixel-level precision. In particular, pre-trained LLMs mainly developed for specific language tasks often struggle with image-processing tasks, arduous to ensure the consistency between visual localization and referring expression.

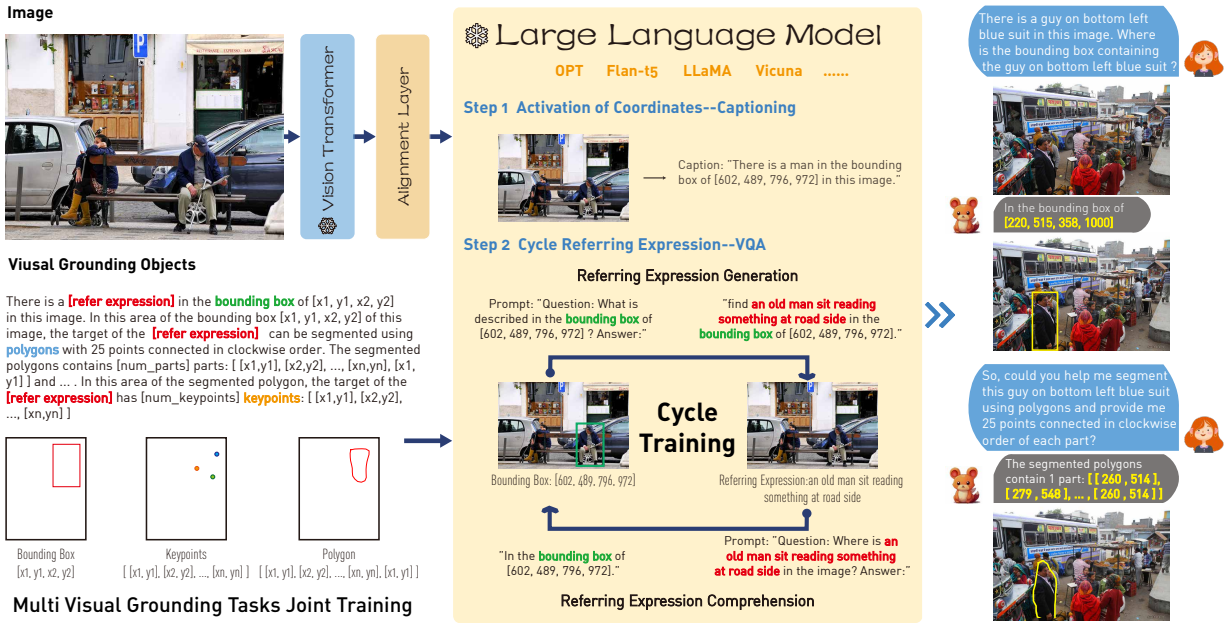


Figure 1: The workflow of our methodology. We design the strategy of cycle training for referring expressions to activate the orientation capability of the vision-language model. The coordinates of the location will cycle through two subtasks: referring expression generation and referring expression comprehension to enhance the consistency between visual location and referring expressions. Furthermore, the joint training of multi-tasks of visual grounding is presented to improve the visual grounding at various levels of granularity, including referring bounding box detection, referring keypoints detection and referring image segmentation. The pre-trained visual model and large language model are frozen in the training.

In this study, we introduce ViLaM, a large multi-modality model, that supports multi-tasks of visual grounding using cycle training strategy, with abundant instructions. First, to enhance the visual grounding capabilities of large multi-modality models, the cycle training strategy is introduced, wherein referring expression generation and referring expression comprehension are cycled to enhance the consistency between visual location and referring expressions. Furthermore, the cycle training approach also contributes to performance improvement through the additional datasets without referring expressions or visual locations. The pairs of referring expressions and visual locations can be generated using the cycle training, thereby harnessing the power of big data to drive the large model. Second, multi-tasks of visual grounding are supported and participate in the joint training in our model. The visual localization capability of the ViLam covers various granularities from the region-level to pixel-level, including bounding box (bbox), keypoints and segmented polygons. Regarding captioning tasks, our proposed model excels in producing fine-grained categorization of the identified targets and generating comprehensive descriptions that encapsulate the attributes of the pointed target. Thereby, the joint training of multiple tasks enhances the consistency between visual localization and referring expressions from various perspectives, leading to improved performance across different tasks and the overall model. Meanwhile, leveraging the power of LLMs, our approach supports a wide array of abundant instruction prompts, significantly enhancing the human-computer interaction and generalization capabilities of our model. Especially in cross-domain tasks, such as the medical field, our model possesses the capability to effectively accomplish novel tasks based on given instructions. Finally, recognizing the demand for higher-quality visual grounding data that encompasses diverse forms of visual location, we developed a dataset called VGCoco. It contains about 240K images with grounding

annotations varying from region-level to pixel-level, including bboxes, keypoints and segmented polygons. In order to foster and support the visual grounding community, we have made our dataset and code publicly available.

To summarize, our contributions are four-fold: (1) We design the cycle training, enhancing the consistency between visual localization and referring expressions, and satisfying the requirements of paired visual grounding datasets for the large model training, both in quantity and quality. (2) Incorporating the LLMs with abundant instructions, multi-tasks of visual grounding are supported in the ViLaM and participate in the joint training, enhancing the capabilities of visual grounding and generalization. (3) We assess the performance of ViLaM in extensive public datasets with multi-tasks, demonstrating its superior capabilities of visual grounding. Besides, the generalization of ViLaM is verified under the open-set or few-shot seniors, especially in cross-domain, such as the medical field. (4) To foster and empower the community, we contribute a visual grounding dataset with multi-tasks annotations. The dataset and our code are public.

2 Related Work

2.1 LLMs and Multi-modality Pre-training

Large Language Models (LLMs) have recently significantly impacted the field of natural language processing. Through alignment techniques such as supervised learning and reinforcement learning with human feedback, LLMs can effectively generalize to perform a wide range of tasks, even with limited training data. A remarkable application of LLM is ChatGPT, which presents an amazing ability to interact with humans. OpenAI's ChatGPT and GPT4 are prime examples of the impact that AI can have, and there have been extensive open-source efforts to replicate their success, such as OPT [3], BLOOM [4], PALM [5], LLaMA [6].

Multi-modality models have further promoted the development of the vision-language model [7, 8, 9, 1, 2, 10, 11]. GPT-4V [12, 10] has recently shown unprecedented ability in understanding and processing an arbitrary mix of input images and texts. On the other hand, preliminary experiments show that visual grounding accuracy is still limited in the comprehensive scene, like the medical field.

2.2 Visual Grounding

Referring Expression Comprehension Early pioneers typically used a two-stage approach to tackle visual grounding tasks. The initial step involves extracting interest regions, which are subsequently prioritized based on their similarity scores with the language query [13, 14, 15, 16]. Another line of work advocates a one-stage pipeline based on dense anchors [17, 18, 19, 20, 21]. Other Transformer models like SeqTR [22], VGTR [23] and PolyFormer [24] in Vision-Language Tasks are subsequently proposed for the visual grounding task and achieved satisfactory performance.

Generalist Model Recently, the potential of generalist models has been increasingly explored, garnering considerable attention from the research community. Among these, OFA [25] integrates a diverse set of cross-modal and uni-modal tasks within a simple sequence-to-sequence learning framework. It adheres to instruction-based learning in both pre-training and fine-tuning stages, negating the need for additional task-specific layers for downstream tasks. Besides, mPLUG-2 [26] presents a multi-module composition network that utilizes shared universal modules for modality collaboration and separates distinct modality modules to address modality entanglement.

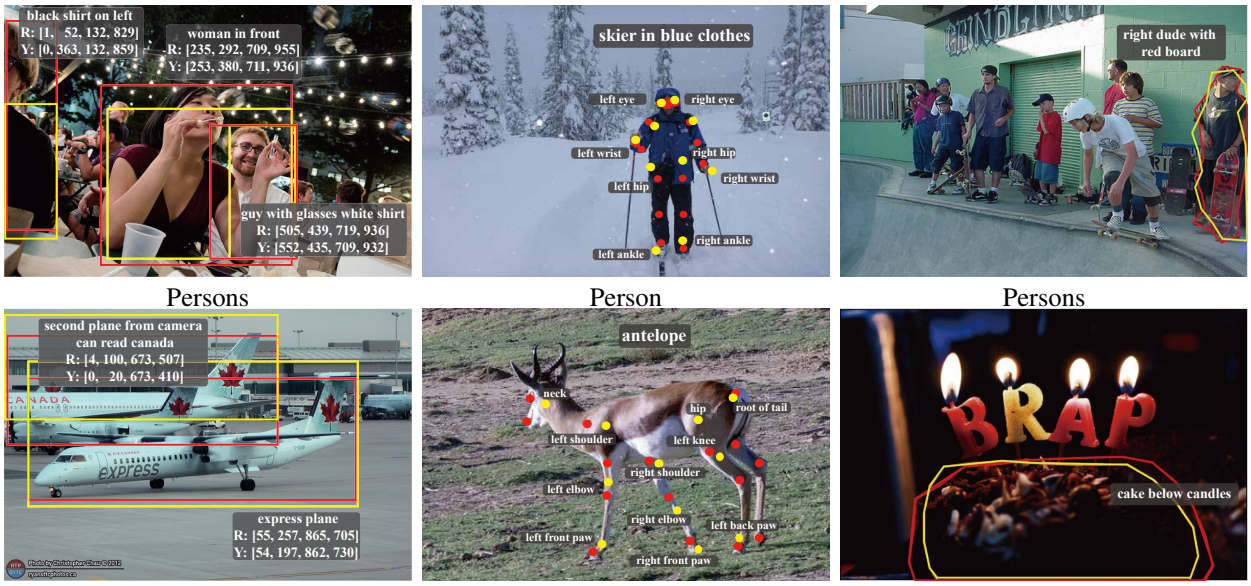
3 Methodology

In this section, we first introduce the architecture of our vision-language model, ViLaM. Then, we explore the activation of object coordinates. Based on robustly outputting these object coordinates, we present the cycle training approach, which serves to reinforce the consistency between visual locations and their corresponding referring expressions. Moreover, the cycle training also allows datasets without referring expressions to participate in the training, thereby facilitating the joint training of multiple tasks. Finally, we introduce our built VGcoco dataset, which includes multi-task annotations for visual grounding.

3.1 Architecture

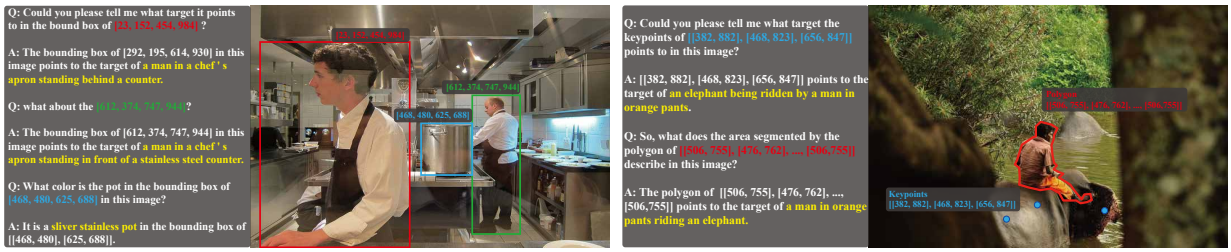
Image encoder: With an input image $x_i \in \mathbb{R}^{H\times W}$, visual features are extracted by image encoder and further projected to feature dimension:

$$v_i = P_{img}(E_{img}(x_i)) \in \mathbb{R}^{(h_f \times w_f)\times d} \tag{1}$$



(a) Referring Bbox Detection (b) Referring Keypoints Detection (c) Referring Image Segmentation

Figure 2: Results of multi-tasks of visual grounding, including (a) Referring bbox detection, (b) referring keypoints detection and (c) referring image segmentation. The red is represented as the ground truth and the yellow denotes the prediction. We choose the person and non-person to display the performance of our model in different tasks.



(a) Referring caption based on the bbox (b) Referring caption based on the keypoints and segmented polygon

Figure 3: Results of referring region caption

where h_f and w_f are the output size of visual features, and d represents the feature dimension. E_{img} can be any common visual backbones and we use Vit-Large in our case. Then by using P_{img}, which is composed of two linear layers, visual features are projected to feature dimension.

Language encoder: With any processed input instruction sequence t_i, text features are extracted by language encoder:

$$l_i = E_{txt}(t_i) \in \mathbb{R}^{n_t \times d} \tag{2}$$

where n_t is the number of input tokens and d represents the feature dimension. In our case, Bert [27] is used as the language encoder.

Multi-modality module: This module follows an encoder-decoder architecture format. Given the input visual features v_i and text features l_i, we first generate fused multi-modality representations by combining the image and text embeddings. These fused features serve as the keys and values in the cross-attention blocks in the decoder. By conditioning on the partial sequence $y_{i<j}$ predicted so far, the decoder recursively makes predictions for the token at position j, effectively generating aligned descriptions across modalities.

$$y_{i,j} = D_{mm}(E_{mm}(concat(v_i, l_i)), y_{i<j}) \in \mathbb{R}^{1 \times d} \tag{3}$$

whereas, D_{mm} and E_{mm} denote the decoder and the encoder of the multi-modality module, respectively.

3.2 Activation of Coordinates

Leveraging its emergent capabilities, the vision-language model exhibits remarkable versatility in scenarios and tasks related to orientation. Initially, the task of activating coordinates is transmuted into conventional object detection within the framework of the vision-language model, devoid of referring expressions. Subsequently, extensive object detection datasets, such as COCO 2017, are integrated into the activation procedure. The considerable quantity of data facilitates the vision-language model in producing coordinates with enhanced robustness and precision.

Subsequently, to reconcile the divergence between semantic and linguistic coordinates, we establish a linguistic representation of coordinates within the large language model: $[x_1, y_1, x_2, y_2]$. Here, x denotes horizontal coordinates, while y signifies longitude. The pair (x_1, y_1) designates the upper-left point, and (x_2, y_2) corresponds to the lower-right point. All coordinates adopt relative positions, normalization to 1000, and rounding.

We employ the captioning task to prompt our model to output coordinates that express orientation, owing to its proven effectiveness in capturing information in knowledge-intensive scenarios [28]. During training, we utilize the captioning format as follows:

find the <object> in the region of $[x_1, y_1, x_2, y_2]$.

Due to the absence of referring expressions, more than one coordinate may correspond to multiple objects in the image. The captioning form has more feasibility and practicability for the vision-language model to establish links between orientation and linguistic coordinates robustly, without inferring in the prompt.

In the training of captioning, the proposed model is expected to output image captions containing object-related coordinates and compute the loss. For activation of coordinates, we optimize using cross-entropy loss:

$$\mathcal{L}_{\text{ce}} = -\sum_{i=1}^{n}\sum_{j=1}^{|y|} \log P_\theta(y_{i,j}|y_{i,<j}, x_i, t_i) \tag{4}$$

where n is the batch size, θ represents the model parameters, x_i represents the input image, t_i stands for the input instruction, and $y_{i,j}$ denotes the output token at position j for the ith sample at each batch. We follow the training strategy of BLIP2, which only trains the alignment layer and freezes the pre-trained visual model and large language model. To enhance the quality of generation during inference, we employ various decoding techniques, such as beam search.

3.3 Cycle Training

Obtaining the ability of localization in the vision-language model, we design the cycle training to align and enhance the consistency between visual localization and referring expressions, as shown in Fig.1. Inspired by Cycle-GAN [29], the cycle training is expected to learn alignment relationships between two domains X and Y given training samples $\{x_i\}_{i=1}^{N} \in X$ and $\{y_j\}_{j=1}^{M} \in Y$ in the vision-language model, where X denotes the visual grounding features and Y represents the linguistic referring expression.

The cycle training consists of two processes: referring expression generation (REG) represented as $G : X \to Y$ and referring expression comprehension (REC) formulated as $F : Y \to X$. The form of VQA is utilized to organize data, wherein we pose questions involving visual coordinates, and exploit the answers with referring expressions obtained from REG to construct new questions for REC. The visual localization from the answer of cycled REC is expected to be the same as the original. Vice versa, we also perform the cycle from REC to REG.

With the above form of VQA, the visual localization and referring expression are cycled training in the vision-language model. We argue that the learned alignment relationships should be cycle-consistent: for every visual localization x belonging to domain X, the cycle-referring expression should possess the capability to restore x to its nearby coordinates, indicated $x \to G(x) \to F(G(x)) \approx x$. Similarly, for each referring expression y from domain Y, y should be reduced to its original form, i.e. $y \to F(y) \to G(F(y)) \approx y$. The cycle referring expression can be incentivized by the consistency loss:

$$\mathcal{L}_{\text{cyc}}(G, F) = \mathcal{L}_{\text{ce}}(F(G(x)), x) + \mathcal{L}_{\text{ce}}(F(G(y)), y) \tag{5}$$

Benefiting from the cycle training, more normal object detection datasets without referring expressions could be expanded for the visual grounding training, such as COCO 2017. REG generates referring expressions from bounding boxes in COCO 2017, and REC then inferences the bounding boxes from the generated referring expression. Thereby, we can generate big data with visual localization and referring expressions to drive the large vision-language model to achieve visual grounding.

Table 1: Experimental settings of tasks, datasets and metrics. It is worth noting that COCO 2017 does not contain referring expressions. Therefore, we do not utilize reference expressions when activating coordinates. While, in the tasks of referring bbox detection (RBD), referring keypoints detection (RKD) and referring image segmentation (RIS), the cycle training strategy generates the referring expressions for COCO 2017, enabling its participation in the training.

Tasks	Datasets		Metrics
	Training	Valiadation	
Activation of Coordinates	COCO 2017		
RBD	RefCOCO/+/g COCO 2017	RefCOCO/+/g	Acc@0.5
RKD	COCO 2017	COCO 2017 HumanArt AP-10K	AP
RIS	RefCOCO COCO 2017	RefCOCO	IoU
RRCls		COCO 2017	ACC / mAP
RRCap		RefCOCOg Visual Grenome	METEOR CIDEr
Medical Foreign Object Detection		Object-CXR	Acc@0.5
Disease Localization	ChestXray14 (20-shot)	ChestXray14 TBX11K RSNA	Acc@0.5

3.4 Multi-Tasks Joint Learning

We integrate multiple visual grounding tasks for joint training. Firstly, two types of tasks are participated in the training, namely referring expression comprehension (REC) and referring expression generation (REG). They are key factors of the cycle training.

Among REC, various expression forms of visual coordinate are supported in our model, including bounding boxes, keypoints and segmented polygons, with varying granularity from the region-level to the pixel-level. To leverage these different modalities, we conducted joint training of REC by combining tasks of referring bounding box detection (RBD), referring keypoints detection (RKD), and referring image segmentation (RIS). The coarse-grained RBD provides the target position information to guide the more fine-grained RKD and RIS tasks. In turn, the keypoints can effectively assist the polygons in deforming and outlining the boundaries of the targets. Vice versa, the pixel-level polygons also contribute to the refinement of bounding box and keypoint coordinates.

While in REG, referring region classification (RRCls) aims to determine the fine-grained category of the located target, and referring region caption (RRCap) focuses on generating a comprehensive description of the pointed target, including its category, location, color, size, and relationship with the surroundings. Through the cycle training strategy, the consistency between description and visual features is enhanced by the multi-tasks joint training.

3.5 Building VGcoco for Visual Grounding

Due to the inherent challenge of obtaining comprehensive location information, including bounding boxes, keypoints, and segmented polygons, most visual grounding datasets struggle to offer location details at various levels of granularity. Recognizing the demand for higher-quality referring expression data that encompasses diverse forms of visual location, we developed a dataset called VGCoco. It contains about 240K images with grounding varying from region-level to pixel-level and corresponding referring expressions.

We extend the open-source datasets, namely COCO-Pose [31], AP-10K [32] and a part of COCO 2017 [33]. With the aid of the cycle training strategy, the reference expressions required by COCO-Pose are generated by our model. Similarly, AP-10 only contains the keypoints of animal skeletons and bounding boxes, so the segmented polygons and referring expressions are produced by our ViLaM. In addition to persons at COCO-Pose and animals at AP-10K, a part

Table 2: Evaluation results of referring bbox detection on RefCOCO, RefCOCO+ and RefCOCOg datasets. The best-performing multi-tasks models are highlighted in red, and the second-best in blue. Acc@0.5 is applied to evaluate the performance of different methods.

Models	Type	Visual Encoder	Language Model	RefCOCO			RefCOCO+			RefCOCOg	
				val	testA	testB	val	testA	testB	val-u	test-u
SeqTR [22]	Specialized Models	Darknet-53	Bi-GRU	83.72	86.51	81.24	71.45	76.26	64.88	74.86	74.21
MDETR [19]		EfficientNet-B3	RoBERTa-base	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
VGTR [23]		EfficientNet-B3	RoBERTa-base	79.30	82.16	74.38	64.40	70.85	55.84	66.83	67.28
OFA [25]	Generalist Models	ResNet-152	BART-Large	**92.04**	**94.03**	**88.44**	**87.86**	**91.70**	**80.71**	**88.07**	**88.78**
mPLUG-2 [26]		ViT-L/14	BERT-Large	90.33	92.80	86.05	-	-	-	84.70	85.14
FERRET [30]		ViT-L/14	Vicuna-7B	87.49	91.35	82.45	80.78	87.38	73.14	83.93	84.76
Ours		ViT-L/14	Vicuna-7B	**92.99**	**95.90**	**90.39**	**90.96**	**94.78**	**86.93**	**90.05**	**89.51**

of common objects at COCO 2017 are joined in the designed datasets. We use skeletonization and clustering methods to obtain the keypoints of the target, and generate its reference expression through our model. To support and encourage the community focused on visual grounding, we have released this dataset to the public. For further information about VGcoco, please refer to the supplementary materials.

4 Experiments

4.1 Experimental Settings

The experimental settings are exhibited in the Table.1, including tasks, dataset and metrics. Our training dataset primarily contains the COCO 2017 [33], RefCOCO [34], RefCOCO+ [34], and RefCOCOg [35, 36]. Besides, we use the ChestXray14 [37] as the 20-shot training dataset to perform a case study of generalization of disease localization on chest X-rays.

For validation, in addition to assessing the accuracy of our model on the respective closed-set dataset, we also evaluated its generalizability in an open-set scenario, such as HumanArt [38] and AP-10K [32] in referring keypoints detection, Visual Grenome [39] in referring region caption. Besides, the generalization is validated from different domains, such as Object-CXR [40] in medical foreign object detection and RSNA Pneumonia dataset [41], TBX11K dataset [42] and ChestXray14[37] in disease localization on Chest X-ray.

When assessing the performance of our model across multiple tasks, the Acc@0.5 evaluates the impact of bbox-related tasks, OKS-based AP validates the performance of referring keypoints detection, and IoU measures the effectiveness of referring image segmentation. Furthermore, in order to evaluate the capability of referring region caption, we select the task of classification to output the category in the bbox with ACC metric and the task of the detailed caption to generate the description with metrics of METEOR and CIDEr. More detailed experimental implementations are given in the supplementary.

4.2 Accurate REC of Multi-Tasks

Qualitatively, Fig.2 exhibits the visual grounding results of multi-tasks, including referring bbox detection, referring keypoints detection and referring image segmentation. For each task, we exhibit the performance of visual grounding in person and non-person. The red denotes the ground truth and the yellow is the prediction.

For referring bbox detection, Fig.2(a) illustrates the superiority of our method. It accurately identifies the object and understands its description words, such as position, color, and text on the object. Significantly, our model demonstrates the ability to recognize objects that are overlapping and occluded. When overlapping targets of the same class are present, our model reveals remarkable capabilities in understanding, discriminating, and locating them. This further corroborates that the cycle training effectively enhances the consistency between visual location and referring expressions.

In the context of referring keypoints detection illustrated in Fig.2(b), our model leverages the power of large language models to proficiently identify the corresponding keypoints by employing simple interactive questions, such as eyes, hips and shoulders. In joint training of multi-tasks, the referring keypoints detection empowers the model to gain a deeper understanding of the relationships among various targets. In addition to the referring expression of keypoints-level, we extend the target-level referring expression through the cycle training strategy, such as changing "person" to "skier in

Table 3: Evaluation results of referring keypoints detection on the close-set scenario with the COCO val2017 dataset, and open-set generalization validation over the HumanArt and AP-10K datasets. The best-performing multi-tasks models are highlighted in red, and the second-best in blue. Average precision is utilized to validate the performance of keypoints detections.

Models	Type	COCO val	HumanArt	AP-10K
PCT [43]	Specialized Models	80.20	63.70	14.60
ViTPose [44]		82.00	64.10	14.70
Unified-IO [45]	Generalist Models	25.00	15.70	7.60
Painter [46]		70.20	12.40	15.30
InstructDiff [47]		**71.20**	**51.40**	**15.90**
Ours		**76.10**	**54.62**	**44.67**

Table 4: Evaluation results of referring image segmentation on RefCOCO dataset. The best-performing generalist models are highlighted in red, and the second-best in blue. IoU is utilized to validate the performance of segmentation.

Methods	Type	RefCOCO		
		val	testA	testB
LAVT [48]	Specialized Models	74.46	76.89	70.94
SeqTR [22]		71.70	73.31	69.82
PolyFormer [24]		74.82	76.64	71.06
Unified-IO [45]	Generalist Models	46.42	46.06	48.05
InstructDiff [47]		61.74	65.20	60.17
LISA [49]		**74.10**	**76.50**	**71.10**
Ours		**74.85**	**76.02**	**74.34**

blue clothes". It enables the integration of target-level referring expressions during the joint training, thereby enhancing the model's comprehension capabilities.

Besides, our model expands the competencies of visual grounding to pixel-level shown in Fig.2(c). Polygon is adopted to accurately delineate the objects denoted by referring expressions, thereby significantly enhancing the alignment between the target shape and the referring expressions through the cycle training strategy.

Quantitatively, Table.12 presents a comparison of referring bbox detection results between our model and various types of visual grounding models, including specialized models and multi-tasks models. It clearly demonstrates that our method achieves SOTA performance across all test datasets. Notably, in the testB split of RefCOCO and RefCOCO+, our model outperforms other methods by a significant margin. This highlights the superiority of our approach in effectively handling the referring expression comprehension with bounding boxes, particularly for non-people objects.

Moreover, Table.3 indicates our model performs excellently in the referring keypoints detection, with the best performance among multi-tasks models. Notably, our model demonstrates significantly stronger generalization performance on the open-set datasets of HumanArt and AP-10K compared to other models. In fact, we surpass specialized models by a significant margin in AP-10K, corroborating our main contribution of the cycle training strategy, which enhances the capability of LLMs in understanding, discriminating, locating and generalization.

Additionally, Table.4 investigates the effectiveness of our model in referring image segmentation. Our performance surpasses other methods, with the exception of being slightly inferior to LISA[49] on the testA split of RefCOCO dataset, which employs ViT-H SAM [50] backbone as the vision backbone.

Table 5: Evaluation results of referring object classification on COCO 2017 val set. The best-performing generalist models are highlighted in red, and the second-best in blue. ACC is utilized to validate the performance of classification.

LLaVA [10]	Shikra[51]	PVIT [52]	Ours		
			keypoints	bbox	polygon
40.04	53.91	64.53	74.52	**80.58**	**81.55**

Table 6: Evaluation results of the referring region caption on RefCOCOg and Visual Genome dataset. The best-performing multi-tasks models are highlighted in red, and the second-best in blue. METEOR and CIDEr is utilized to validate the performance of region captioning.

Methods	RefCOCOg		Visual Genome	
	METEOR	CIDEr	METEOR	CIDEr
GRIT [53]	15.2	71.6	**17.1**	**142.0**
Kosmos-2 [54]	14.1	62.3	-	-
Ours(keypoints)	**26.5**	146.6	17.0	**130.1**
Ours(bbox)	26.3	**167.2**	**19.7**	**131.3**
Ours(polygon)	**26.6**	**165.6**	**18.8**	105.9

4.3 Region Captioning

Referring Object Classification The performance of object classification was evaluated on COCO 2017 dataset using classification accuracy. As shown in Table.5, the referring object classification task achieved excellent results using different visual prompts,i.e., keypoints, bbox, and polygon. Compared to PVIT method, there was an approximately 15% improvement in the ACC values with the bbox prompt of our method. We observed a significant improvement when using the polygon prompt, indicating an inherent relationship between the shape and the object.

Referring Object Caption We further evaluate the region-level captioning ability of our model on the RefCOCOg and Visual Genome datasets. As shown in Table.6, the region-level caption task has also achieved excellent results on the RefCOCOg dataset, with improvements of more than 10% and 90% in METEOR and CIDEr scores, respectively, compared to the results achieved by the Kosmos-2 method. There is no obvious difference among the results obtained by various visual prompts. More ablation experiments for the caption task are provided in the supplementary materials.

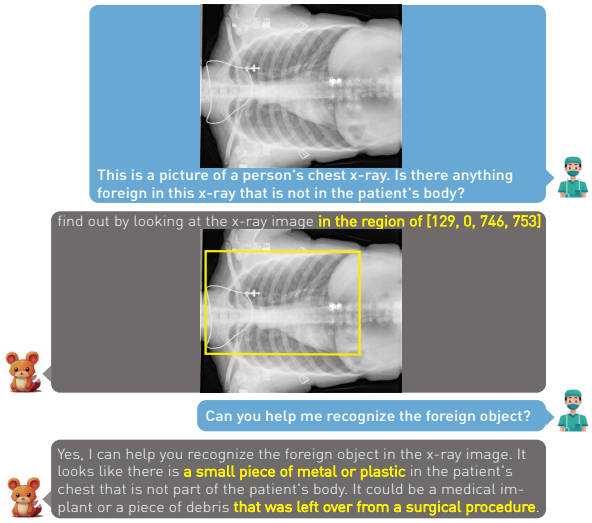


Figure 4: The zero-shot results of visual question answering for foreign objects detection in chest X-ray images. The presence of foreign objects is accurately detected by giving its coordinates. Particularly, the model can deduce that the foreign object is metal or plastic by asking to recognize the foreign object.

4.4 Case Studies of Robust Generalization in the Medical Domain

We conduct several case studies in the medical domain to corroborate the robust generalization of our model, namely medical foreign object detection and disease localization in chest X-rays.

Regarding medical foreign object detection, we employed the model trained on natural images directly, without fine-tuning specifically on chest X-rays. Fig.4 illustrates the results of visual question and answering for medical foreign object detection on the Object-CXR dataset [40]. The necklace around the neck and the clip in the middle retain their inherent shape in X-rays, so our model accurately identifies the presence of these foreign objects outside

the body, attributed to its robust and considerable generalization. Beyond the localization, our model can recognize the detected foreign object upon inquiry, and deduce that the object is likely made of metal or plastic debris. More specially, ViLaM also exhibits an ability to infer the potential origin or source of the debris, leveraging its extensive language understanding capacity. Moreover, on the quantitative of classify the presence or absence of foreign objects, ViLaM achieves an AUC of 93.1%, demonstrating a substantial feasibility and robust generalization.

Table 7: Evaluation results of **disease localization** task with 20-shot setting on four typical disease labels from the chest X-ray datasets. Acc@0.5 is applied to evaluate methods.

Datasets	TBX11K	RSNA	ChestXray14	
Diseases	Tuberculosis	Pneumonia	Atelectasis	Pneumothorax
VGTR [23]	1.99	4.67	3.70	0
OFA [25]	20.40	14.67	3.90	12.49
Ours	30.84	28.00	11.11	20.83

To further examine the generalization and scalability of our model in the medical field, we conduct preliminary experiments of disease localization on three typical chest X-ray datasets, namely, TBX11K[42], RSNA Pneumonia[41], and ChestXray14[37]. The proposed model is fine-tuned with 20-shot labels for each disease of ChestXray14 datasets. As depicted in Table. 7, ViLaM consistently outperforms other approaches in various disease categories. Particularly noteworthy is the significant improvement of more than 15% in detecting Pneumonia and Pneumothorax compared to alternative methods. This robust performance further validates the exceptional generalization and practicality capability of our model. More experiment results are available in the supplement material.

4.5 Ablation Study

4.5.1 Cycle Training

Table 8: Ablation evaluation results of referring bbox detection on RefCOCO dataset with different modules. ✓denotes the applied module. Acc@0.5 is utilized to evaluate the performance of various conditions.

Modules			RefCOCO		
Coordinates Activation	Cycle Training	Cycle Augment.	val	testA	testB
✓			79.95	79.24	80.99
✓	✓		85.59	87.54	82.60
✓	✓	✓	92.99	95.90	90.39

Firstly, we conduct ablation studies to verify the improvement of the cycle training strategy. The cycle training brings about improvements through two key factors. First, it enhances the consistency between reference expressions and their related locations. Second, additional datasets without referring expressions can participate in the training by generating referring expressions through cycle training.

Consequently, we conduct ablation experiments to evaluate the impact brought by these two key factors of cycle training in Table.8, where "Cycle Training" denotes the performance gain from enhancing consistency using cycle training, and "Cycle Augment." represents the boost from additional datasets, which uses cycle training to generate referring expressions. The referring bbox detection is employed as the task of validation using RefCOCO dataset. Table.8 exhibits that the improvement from the enhanced consistency achieves an Acc@0.5 of 5.64% in val, 8.3% in testA and 1.61%, and from the additional datasets is Acc@0.5 of 7.4% in val, 8.36% in testA and 7.79%.

Upon analysis, we observe that the testA split of RefCOCO exclusively comprises people, thereby leading to a more pronounced improvement resulting from the enhanced consistency achieved through cycle training, as compared to the testB split, which includes non-people. Besides, data augmentation exhibits a significant enhancement effect in different splits.

4.5.2 Joint Training

Furthermore, we implement ablation experiments to validate the effect of multi-tasks joint training of visual grounding in Table.9. The referring keypoints detection and referring image segmentation are selected as the tasks of validation.

Table 9: Ablation evaluation results of multi-tasks joint training of visual grounding. ✓denotes the participated tasks. RIS and RKD are exploited as the validation tasks, and IoU and AP are opted for the metrics, respectively.

Multi-Task Joint Training			RIS(IoU)			RKD(AP)
RBD	RKD	RIS	RefCOCO val	RefCOCO testA	RefCOCO testB	COCO val
✓	✓		-	-	-	70.46
✓		✓	62.83	63.28	61.06	-
✓	✓	✓	74.85	76.02	74.34	76.10

Since the bbox detection is utilized to activate the coordinates of LLMs, the task of referring bbox detection is included in all ablation experiments.

The results of Table.9 showcases the joint learning of multi-tasks of visual grounding substantially boosts the performance of individual tasks. In particular, the referring keypoints detection considerably improves the segmentation task. We analyze that it is mainly because keypoints can well assist polygons in deforming and outlining the boundaries of the targets.

5 Conclusion

We have developed a vision-language model, ViLaM, that enhances visual grounding capabilities and generalization performance based on the foundations of a large language model. Despite being trained solely on the COCO 2017 and RefCOCO/+/g datasets, we are able to generate a considerable amount of additional annotations through the cycle training for multi-tasks, and exhibit competitive performance on multiple referring expression comprehension and generation tasks. We further contribute a multi-task dataset, encompassing referring expression and related annotations of bounding boxes, keypoints, and segmented polygons.

References

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

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

[3] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[4] 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.

[5] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[6] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

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

[8] 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, 2022.

[9] 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.

[10] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[11] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023.

[12] OpenAI. Gpt-4v(ision) system card, 2023.

[13] 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.

[14] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic Graph Attention for Referring Expression Comprehension. pages 4644–4653.

[15] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving Referring Expression Grounding With Cross-Modal Attention-Guided Erasing. pages 1950–1959.

[16] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to Assemble Neural Module Tree Networks for Visual Grounding. pages 4673–4682.

[17] 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.

[18] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 387–404. Springer, 2020.

[19] 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.

[20] 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 *European Conference on Computer Vision*, pages 521–539. Springer, 2022.

[21] 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 (ICCV)*, pages 1769–1779, October 2021.

[22] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. SeqTR: A simple yet universal network for visual grounding. In *Lecture Notes in Computer Science*, pages 598–615. Springer Nature Switzerland, 2022.

[23] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Visual grounding with transformers. In *Proceedings of the International Conference on Multimedia and Expo*, 2022.

[24] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R. Manmatha. PolyFormer: Referring Image Segmentation As Sequential Polygon Generation. pages 18653–18663.

[25] 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, 2022.

[26] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. mPLUG-2: A Modularized Multi-modal Foundation Model Across Text, Image and Video.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[28] Ander Salaberria, Gorka Azkune, Oier Lopez de Lacalle, Aitor Soroa, and Eneko Agirre. Image captioning for effective use of language models in knowledge-based visual question answering. 212:118669.

[29] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232.

[30] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and Ground Anything Anywhere at Any Granularity. In *The Twelfth International Conference on Learning Representations*, 2023.

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

[32] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[33] 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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 740–755. Springer International Publishing.

[34] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling Context in Referring Expressions. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 69–85. Springer International Publishing.

[35] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20.

[36] Varun K. Nagaraja, Vlad I. Morariu, and Larry S. Davis. Modeling Context Between Objects for Referring Expression Understanding. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 792–807. Springer International Publishing.

[37] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[38] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. Human-Art: A Versatile Human-Centric Dataset Bridging Natural and Artificial Scenes. pages 618–629.

[39] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. 123(1):32–73.

[40] JF Healthcare. Object-cxr - automatic detection of foreign objects on chest x-rays.

[41] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

[42] Yun Liu, Yu-Huan Wu, Yunfeng Ban, Huifang Wang, and Ming-Ming Cheng. Rethinking computer-aided tuberculosis diagnosis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2646–2655, 2020.

[43] Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. Human Pose As Compositional Tokens. pages 660–671, 2023.

[44] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. 35:38571–38584, 2022.

[45] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. UNIFIED-IO: A Unified Model for Vision, Language, and Multi-modal Tasks. 2022.

[46] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. pages 6830–6839, 2023.

[47] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Han Hu, Dong Chen, and Baining Guo. InstructDiffusion: A Generalist Modeling Interface for Vision Tasks, 2023.

[48] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. pages 18155–18165.

[49] 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.

[50] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[51] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic, 2023.

[52] Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. Position-enhanced visual instruction tuning for multimodal large language models. *arXiv preprint arXiv:2308.13437*, 2023.

[53] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022.

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

[55] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020.

[56] Cristina González, Laura Bravo-Sánchez, and Pablo Arbelaez. Isinet: An instance-based approach for surgical instrument segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 595–605, 2020.

[57] Yiting Ma, Xuejin Chen, Kai Cheng, Yang Li, and Bin Sun. Ldpolypvideo benchmark: a large-scale colonoscopy video dataset of diverse polyps. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 387–396. Springer, 2021.

[58] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016.

[59] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[60] Haifan Gong, Jiaxin Chen, Guanqi Chen, Haofeng Li, Guanbin Li, and Fei Chen. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Computers in biology and medicine*, 155:106389, 2023.

[61] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. 28:104863.

[62] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017.

[63] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging*, 5(3):036501–036501, 2018.

[64] Susanne G. Mueller, Michael W. Weiner, Leon J. Thal, Ronald C. Petersen, Clifford R. Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga, and Laurel Beckett. Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). 1(1):55–66.

[65] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

[66] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020.

[67] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations*, 2024.

[68] Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. Jack of all tasks, master of many: Designing general-purpose coarse-to-fine vision-language model. *arXiv preprint arXiv:2312.12423*, 2023.

[69] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model. *arXiv preprint arXiv:2403.02330*, 2024.

[70] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[71] Dongsheng Jiang, Yuchen Liu, Songlin Liu, Xiaopeng Zhang, Jin Li, Hongkai Xiong, and Qi Tian. From clip to dino: Visual encoders shout in multi-modal large language models, 2023.

[72] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

[73] 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.

[74] Ao Zhang, Liming Zhao, Chen-Wei Xie, Yun Zheng, Wei Ji, and Tat-Seng Chua. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*, 2023.

[75] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.

[76] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. *arXiv preprint arXiv:2312.02949*, 2023.

[77] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024.

[78] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

[79] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023.

[80] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything.

Appendix

6 Experimental Details

6.1 Datasets

RefCOCO [34], RefCOCO+ [34], and RefCOCOg [35] are three visual grounding datasets that utilize images sourced from MSCOCO [33]. In line with previous approaches, we adopt the train / validation / testA / testB split for both RefCOCO and RefCOCO+ datasets, where testA and testB sets contain only people and only non-people respectively. The split of RefCOCOg-umd [36] on RefCOCOg refers to the splits as the val-u, and test-u. Accuracy@0.5 (Acc@0.5) is used to measure the performance of the visual grounding task, which is right if the IoU between the grounding-truth box and the predicted bounding box is larger than 0.5.

To evaluate the model's generalization capabilities in the medical field, we test its performance on public datasets for disease identification across 6 modalities, namely endoscopy, photography, ultrasound, DR, CT, and MRI, to assess its robustness and adaptability.

The Object-CXR [40] dataset is designed for the automatic detection of foreign objects in chest X-rays. It consists of 5,000 frontal chest X-ray images with foreign objects and 5,000 images without foreign objects. These DR images were captured and collected from approximately 300 township hospitals in China. The ChestXray14 dataset [37] contains 112,120 chest X-ray images with labels for 14 common diseases. Among these, 984 images feature eight key findings with hand-labelled bounding boxes. The RSNA Pneumonia dataset [41] is a binary classification chest X-ray dataset

Table 10: Overview of 12 medical datasets across 6 modalities.

Medical Datasets	Modality	Target	Testset Num
EndoVis18 [55, 56]	Endoscopy	Instrument	1200
LDPolypVideo [57]		Polyp	1040
ISIC16 [58]	Photography	Skin Lesions	379
HAM10000 [59]		Skin Lesions	2000
TN3K [60]	Ultrasound	Thyroid Nodule	614
BUID [61]		Breast Cancer	320
TBX11K [42]	DR	Tuberculosis	1000
RSNA Pneumonia [41]		Pneumonia	1000
Luna16 [62]	CT	Lung Nodule	125
DeepLesion [63]		Lesion	660
ADNI [64]	MR	Hippocampus	1700
LGG [65]		Gliomas	680

Table 11: The training hyperparameters of our method.

Hyperparameters	
Training Steps	70,000
Warmup Steps	1,000
Optimizer	AdamW
Learning Rate	2e-5
Learning Rate Decay	Cosine
Adam β	(0.9, 0.98)
Weight Decay	0.05
Batch Size	12

consisting of 26,683 images. Each radiograph is categorized as either pneumonia or normal. The TBX11K dataset [42] is a large collection comprising 11,000 chest X-ray images, each with corresponding bounding box annotations for tuberculosis areas.

Moreover, we conduct extensive experiments to evaluate the generalization capability of our model on various medical multi-modality datasets, as shown in Table.10. EndoVis18 [55] is a publicly available dataset for endoscopy image analysis. We follow ISINet's annotation and data set division of surgical instrument categories[56]. LDPolypVideo [57] consists of 44 colonoscopy videos for polyp detection, with a total of 18,142 frames, and a resolution of 512×512 pixels. ISIC16 [58] is a collection of dermoscopic images of skin lesions, annotated by dermatologists and skin cancer experts. It consists of 1,267 dermoscopic images of skin lesions, including melanomas and benign lesions, with a resolution of 1024×768 pixels. HAM10000 [59] dataset is a large, publicly available dataset for skin lesion analysis, specifically designed for melanoma detection and skin disease diagnosis. TN3K [60] dataset consists of 2D ultrasound images of thyroid nodules with a resolution of 512×512 for thyroid nodules detection. BUID [61] dataset consists of 780 images with an average image size of 500×500 pixels from 600 female patients for breast cancer detection. Luna16 [62] dataset is a publicly available dataset for lung nodule analysis, specifically designed for lung nodule detection in CT scans. DeepLesion [63] dataset is a large-scale, publicly available dataset for lesion detection and segmentation in CT, with a resolution of 512x512 pixels. ADNI [64] is a large, publicly available dataset for Alzheimer's disease research from magnetic resonance imaging (MRI) scans, specifically designed for the development and evaluation of algorithms for early detection and diagnosis of Alzheimer's disease. LGG [65] (Low-Grade Glioma) dataset is a publicly available dataset for brain tumor segmentation, specifically for detecting low-grade gliomas from MRI scans.

Table 12: Evaluation results of visual grounding on RefCOCO, RefCOCO+ and RefCOCOg datasets. Acc@0.5 is applied to evaluate the performance of two types of visual grounding methods, i.e., specialist and generalist model.

Models	Venue	Visual Encoder	Language Model	RefCOCO			RefCOCO+			RefCOCOg	
				val	testA	testB	val	testA	testB	val-u	test-u
Specialist:											
CM-A-E [15]	CVPR19	ResNet-101	LSTM	87.47	88.12	86.32	73.74	77.58	68.85	80.23	80.37
NMTREE[16]	ICCV19	ResNet-101	Bi-LSTM	85.65	85.63	85.08	72.84	75.74	67.62	78.57	78.21
DGA [14]	ICCV19	ResNet-101	Bi-LSTM	86.34	86.64	84.79	73.56	78.31	68.15	80.21	80.26
MCN [66]	CVPR20	DarkNet-53	Bi-GRU	80.08	82.29	74.98	67.16	72.86	57.31	66.46	66.01
ReSC-Large [18]	ECCV20	DarkNet-53	BERT-base	77.63	80.45	72.30	63.59	68.36	56.81	67.30	67.20
TransVG [21]	ICCV21	ResNet-101	BERT-base	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73
MDETR [19]	ICCV21	EfficientNet-B3	RoBERTa-base	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89
SeqTR [22]	ECCV22	DarkNet-53	Bi-GRU	83.72	86.51	81.24	71.45	76.26	64.88	74.86	74.21
VGTR [23]	ICME22	EfficientNet-B3	RoBERTa-base	79.30	82.16	74.38	64.40	70.85	55.84	66.83	67.28
Generalist:											
OFA [25]	ICML22	ResNet-152	BART-Large	92.04	94.03	88.44	87.86	91.70	80.71	88.07	88.78
mPLUG-2 [26]	ICML23	ViT-L14	BERT-Large	90.33	92.80	86.05	-	-	-	84.70	85.14
Kosmos-2 [54]	ICLR24	ViT-L14	Magneto-1.3B	-	-	-	-	-	-	61.65	86.96
Ferret [67]	ICLR24	ViT-L14	Vicuna-7B	87.49	91.35	82.45	80.78	87.38	73.14	83.93	84.76
VistaLLM [68]	CVPR24	EVA ViT	Vicuna-7B	88.10	91.50	83.00	82.90	89.80	74.80	83.60	84.40
RegionGPT [69]	CVPR24	ViT-L14	Vicuna-7B	-	-	-	-	-	-	60.57	86.96
Shikra [51]	arxiv 23.6	ViT-L14	Vicuna-13B	87.83	91.11	81.81	82.89	87.79	74.41	82.64	83.16
Qwen-VL [70]	arxiv 23.8	ViT-G	QWen-7B	88.55	92.27	84.51	82.82	88.59	76.79	85.96	86.32
COMM [71]	arxiv 23.10	CLIP+DINOv2	Vicuna-7B	91.73	94.06	88.85	87.21	91.74	81.39	87.32	88.33
CogVLM-17B [72]	arxiv 23.11	EVA2-CLIP-E	Vicuna-7B	92.76	94.75	88.99	88.68	92.91	83.39	89.75	90.79
MiniGPT-v2 [73]	arxiv 23.10	EVA ViT	LLaMA2-7B	88.06	91.29	84.30	79.58	85.52	73.32	84.19	84.31
NExT-Chat [74]	arxiv 23.11	ViT-L14	Vicuna-7B	85.5	90.0	77.9	77.2	84.5	68.0	80.1	79.8
SPHINX-2K [75]	arxiv 23.11	ViT-L14	LLaMA2	91.10	92.88	87.07	85.51	90.62	80.45	88.07	88.65
LLaVA-G [76]	arxiv 23.12	Swin Tiny	Vicuna-7B	89.16	-	-	81.68		-	84.82	-
Ferret-v2 [77]	arxiv 24.4	ViT-L14	Vicuna-7B	92.79	94.68	88.69	87.35	92.75	79.3	89.42	89.27
Ours		ViT-L14	Vicuna-7B	92.99	95.90	90.39	90.96	94.78	86.93	90.05	89.51

We strictly follow the official train-test split for EndoVis18, LDPolypVideo, ISIC16, TN3k, Luna16 and ADNI. Due to the absence of an official training/validation/test ratio or a released test set for HAM10000, BUID, TBX11K, RSNA Pneumonia, DeepLesion and LGG, we randomly split each dataset into training/validation/test sets by 7:1:2 for the visual grounding task.

6.2 Implementation Details

For our language-guided image tokenizer, we leverage the strengths of both BERT[78] and ViT as our text encoder and visual encoder, respectively. We employ ViT-L14 as our visual encoder, which consists of 14 transformer encoder layers and an FFN intermediate size of 4,096. The input image size is set to 224×224, with a patch size of 16×16. The hidden dimensions of the ViT-L14 are 1,024, with 16 attention heads. Meanwhile, we utilize Vicuna-7B, a large language model fine-tuned with instructions, as our text encoder. The Vicuna-7B model boasts 12 transformer layers, with 768 hidden dimensions, 12 attention heads, and an FFN intermediate size of 3,072. The vocabulary size is 30,522, and the maximum input sequence length is 512. To align the text encoder and visual encoder, we employ a Q-former with 12 transformer layers. This Q-former has 768 hidden dimensions, 12 attention heads, and query, key, and value dimensions of 256 each.

In terms of the training progress, the hyperparameters are presented in Table.11. We utilize the AdamW optimizer, which is configured with a cosine annealing schedule as the learning policy. The initial learning rate is set to 2×10^{-5}, and the AdamW optimizer is employed with hyperparameters $\beta = (0.9, 0.98)$. Additionally, we set the weight decay to 0.05 and the dropout rate to 0.1. During the first 1,000 warm-up steps, the learning rate increases to 2×10^{-5}, and subsequently decays to 10^{-7}. Unless otherwise specified, our training protocol consists of 70,000 steps, executed on 4×8 NVIDIA V100 GPUs, which takes approximately two days to complete.

For the annotation, We normalize all coordinates to a uniform range of 0 to 1000, ensuring that all images have a consistent coordinate system. For the polygon representation, we select the point closest to the origin as the starting point and employ a 25-point labelling scheme to describe the polygon sequence in a clockwise direction. To demarcate the beginning and end of the sequence, we utilize <BOS> and <EOS> tags, respectively. For the sampling rule for polygons, we employ isometric sampling, wherein we initially calculate the perimeter of the polygon and subsequently divide it into 25 equal segments to sample the polygon.

7 Additional Experiment Results

Table 13: Evaluation results of referring object classification on LVIS and COCO 2017 val set. ACC is utilized to validate the performance of referring object classification.

Methods	LVIS			COCO 17
	keypoints	bbox	polygon	bbox
LLaVA [53]	50.10	50.30	-	40.04
Shikra[51]	57.82	67.71	-	53.91
Ferret [67]	67.94	79.42	69.77	
Kosmos-2 [54]	-	60.25	-	
GPT4RoI [79]	-	61.76	-	
Ours	58.24	66.42	67.00	80.58

7.1 Referring Object Classification Task

The performance of object classification was evaluated on LVIS and COCO 2017 datasets using classification accuracy. As shown in Table.13, the referring object classification task of our method achieved excellent results using different visual prompts, i.e., keypoints, bbox, and polygon, narrowly surpassed only by the Ferret method on the LVIS dataset. For the COCO 17 dataset, our method yield better performance than LLaVA and Shikra. The comparison results for other methods on the LVIS and COCO 17 datasets in Table.13 are sourced from the Ferret [67]and PVIT [52] papers, respectively.

7.2 VGCoco

In response to the growing need for high-quality referring expression data that captures diverse forms of visual location, we introduce VGCoco, a comprehensive dataset designed to meet this demand. Comprising approximately 240,000 images, VGCoco features a range of grounding annotations, from region-level to pixel-level, accompanied by corresponding referring expressions.

We build upon existing open-source datasets, including COCO-Pose [31], AP-10K [32], and a subset of COCO 2017 [33]. Notably, we leverage the cycle training strategy to generate the reference expressions required by COCO-Pose using our model. COCO-Pose provides a wealth of the keypoints information on the human body with the skeleton for pose estimation, as well as bounding boxes and segmentated polygons, but lacks referring expressions. Therefore, we employ our model to generate the detailed referring expressions of persons in COCO-Pose. Besides, AP-10K is the animal version of COCO-Pose, which consists of keypoints of animal skeletons. With the aid of SAM[80], we get the masks of the target animal and transfer them to polygon by uniform sampling. Then, our model is exploited to produce referring expressions for the target animal. Furthermore, we randomly selected a part of non-people and non-animal targets in COCO 2017 to build the VGCoco dataset. In addition to using our model to generate key points, we use the skeletonization method to obtain the key points of these targets.

7.3 Ablation Experiments

We implement ablation experiments to validate the effect of cycle training on the referring region caption task. As shown in Table.15, regardless of whether keypoints, bounding boxes, or polygons are used as visual prompts, cycle training consistently enhances captioning performance, with improvements of about 5% and 15% in METEOR and CIDEr scores, respectively.

Table 14: Evaluation results of **referring bbox detection** task on 12 typical medical datasets of 6 modalities. 20-shot fine-tuning experiments were performed for non-radiology (Endoscopy, Photography and Ultrasound) and radiology datasets (DR, CT and MRI). Acc@0.5 is applied to evaluate methods.

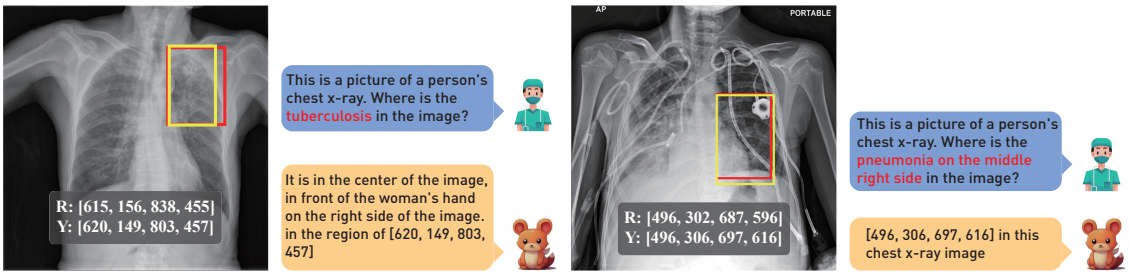
Datasets	Endoscopy		Photography		Ultrasound		DR		CT		MRI	
	EndoVis18	LDPolypVideo	ISIC16	HAM10000	TN3K	BUID	TBX11K	RSNA	Luna16	DeepLesion	ADNI	LGG
VGTR [23]	3.87	7.30	64.12	63.20	12.70	31.46	1.99	4.67	0.00	0.36	2.46	3.67
OFA [25]	7.32	0.30	63.85	61.20	6.81	19.62	20.40	14.67	0.00	2.08	26.26	26.77
Ours	12.53	9.86	67.66	86.00	16.50	38.63	30.84	28.00	0.00	5.23	4.52	18.85

Table 15: Ablation experiment results of referring region caption on RefCOCOg. METEOR and CIDEr are utilized to validate the performance of region captioning.

Visual Prompt	With cycle training		Without cycle training	
	METEOR	CIDEr	METEOR	CIDEr
keyoints	26.5	164.6	19.4	149.3
bbox	26.3	167.2	19.1	152.7
polygon	26.6	165.5	19.8	151.2

Table 16: Evaluation results of **referring bbox detection** task with 20-shot setting on seven labels from the EndoVis18 dataset. Acc@0.5 is applied to evaluate methods. Seven surgical instruments contain Bipolar Forceps (BF), Prograsp Forceps (PF), Large Needle Driver (LND), Monopolar Curved Scissors (MCS), Ultrasound Probe (UP), Suction Instrument (SI) and Clip Applier (CA).

Label	Mean	BF	PF	LND	MCS	UP	SI	CA
VGTR [23]	3.87	12.29	0.00	0.00	14.78	0.00	0.00	0.00
OFA [25]	7.32	22.49	19.73	0.94	4.85	0.00	0.00	3.22
Ours	12.53	16.88	4.17	3.33	13.55	0.00	0.00	16.67



(a) VQA in chest X-ray for the tuberculosis localization in the TBX11K dataset, without referring expressions.

(b) VQA in chest X-ray for the pneumonia localization in the RSNA Pneumonia dataset, with the orientation-related referring expression.

Figure 5: The 20-shot results of disease localization in chest X-ray images. The red box denotes the grounding truth, and the yellow box represents the prediction. (a) Tuberculosis detection in the TBX11K dataset. (b) Pneumonia detection in the RSNA dataset.

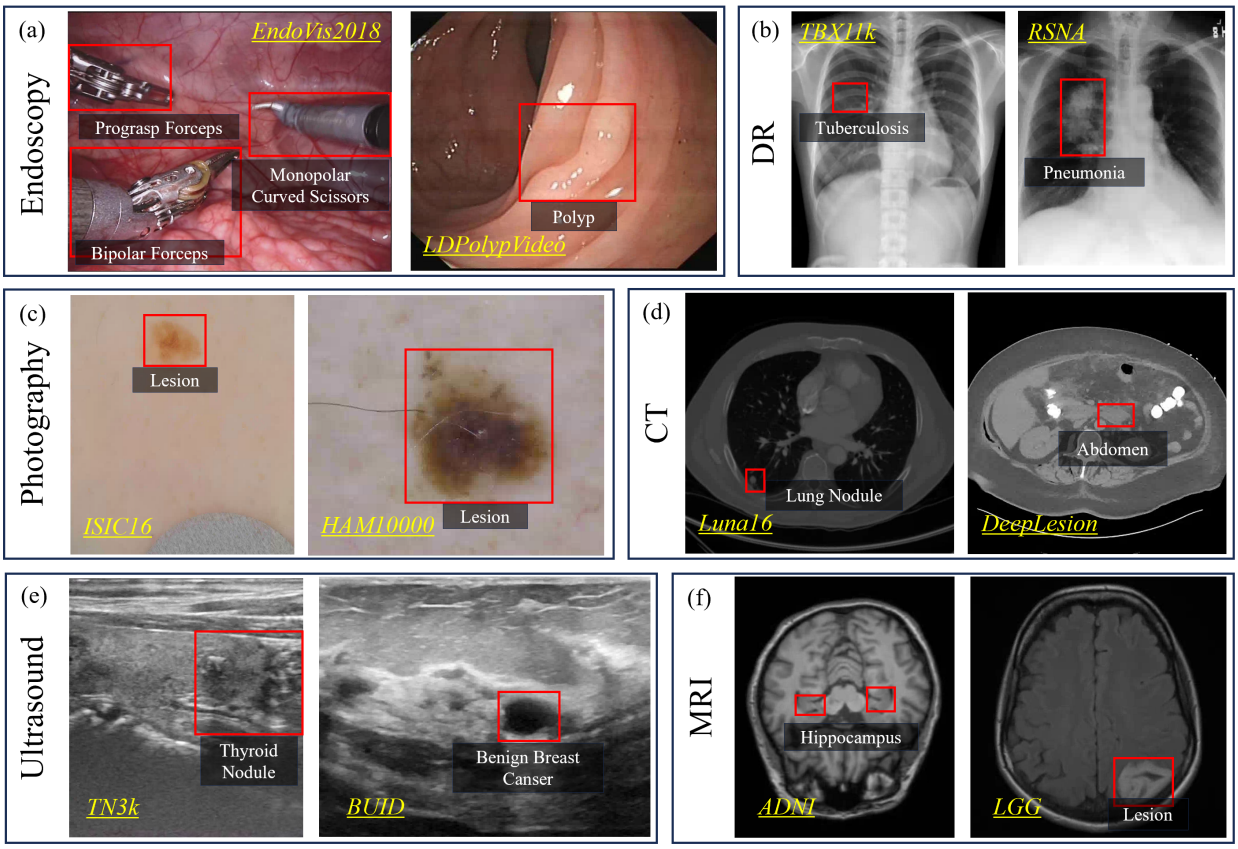


Figure 6: Typical medical datasets for referring bbox detection task, containing 6 modalities: (a) Endoscopy: EndoVis18, LDPolypVideo; (b) DR: TBX11k, RSNA Pneumonia; (c) Photography: ISIC16, HAM10000; (d) CT: Luna16, Deeplesion; (f) MRI: ADNI, LGG

7.4 Generalization Performance in the Medical Domain

To further examine the generalization and scalability of our model, we conduct preliminary experiments of the referring bbox detection task on 12 typical datasets across 6 modalities in the medical domain. The proposed model is fine-tuned with 20-shot labels for each disease of non-radiology and radiology datasets, respectively. We conduct comparative experiments with VGTR and OFA(Large) as representatives of specialist and generalist models, respectively, to evaluate their performance and versatility in the referring bbox detection task.

7.4.1 Non-radiology Images

Endoscopy We evaluated the generalization performance of instrument and disease localization on two typical endoscopy datasets, namely, EndoVis18 and LDPolyVideo. As depicted in Table. 14, ViLaM consistently outperforms other approaches. Table. 16 further demonstrates that the proposed method achieves superior performance in multiple surgical instrument categories. However, there is still room for improvement in some categories, which may be attributed to the issue of data imbalance.

Photography We evaluated the visual grounding performance of three methods on two datasets, ISIC16 and HAM10000, and found that all three methods achieved an accuracy of over 60% on both datasets. This is likely due to the fact that skin disease images and their corresponding features share similar characteristics.

Ultrasound We compared the visual grounding performance of three methods on two ultrasound datasets, TN3K and BUID, and our method achieved the best results. Specifically, our method achieved an accuracy of 16.50% on the TN3K validation set and 38.63% on the BUID breast cancer dataset.

7.4.2 Radiology Images

Chest X-ray Fig.5 illustrates the application of VQA in chest X-ray analysis for the localization of tuberculosis and pneumonia. This demonstrates that our generalist model effectively scales to the medical field, and it can adapt to medical disease localization tasks, with or without the use of referring expressions. Quantitatively, our approach demonstrated a significant advantage over VGTR and OFA, with an improvement of over 10% on the TBX11K and RSNA Pneumonia datasets.

CT We compared the visual grounding performance of three methods on two CT datasets, Luna16 and DeepLesion, and found that all three methods achieved nearly 0% accuracy in the 20-shot finetuning experiment on both datasets. This is likely due to the fact that the features of CT images and general images are quite different, and the lesions, such as lung nodules, are too small, as shown in Fig.6 (d).

MRI Two MRI datasets, the ADNI dataset and the LGG dataset, verify the visual grounding performance of three methods. For Gliomas with larger contrast in the LGG dataset, we have better performance than VGTR. Our result of hippocampus detection is poor due to the low contrast of ADNI, as illustrated in Fig.6 (f).

7.5 Qualitative Examples

This section provides more qualitative examples of multiple tasks, including referring bbox detection, and referring image segmentation as illustrated in Fig.7 and Fig.8, respectively.

Figure 7: More Results of referring bbox detection in RefCOCO. The images of the first row are from testA split containing only people, while the second row images from testB consisting of only non-people. We display typical cases of referring expressions, especially with common indications of orientation, size, color, attachment and markings. The referring expressions of the object are presented in the text box with two coordinates, where R (red) denotes grounding truth and Y (yellow) symbolizes the prediction. The red and yellow bounding boxes are also depicted in the image, respectively.

8 Limitation and Future Work

Our model, leveraging cycle training and multi-task design based on the large language model, exhibited outstanding performance and generalization abilities on a large-scale test set. Nevertheless, there is still room for improvement, particularly in the referring image segmentation task, and the model's generalizability demands further enhancement.

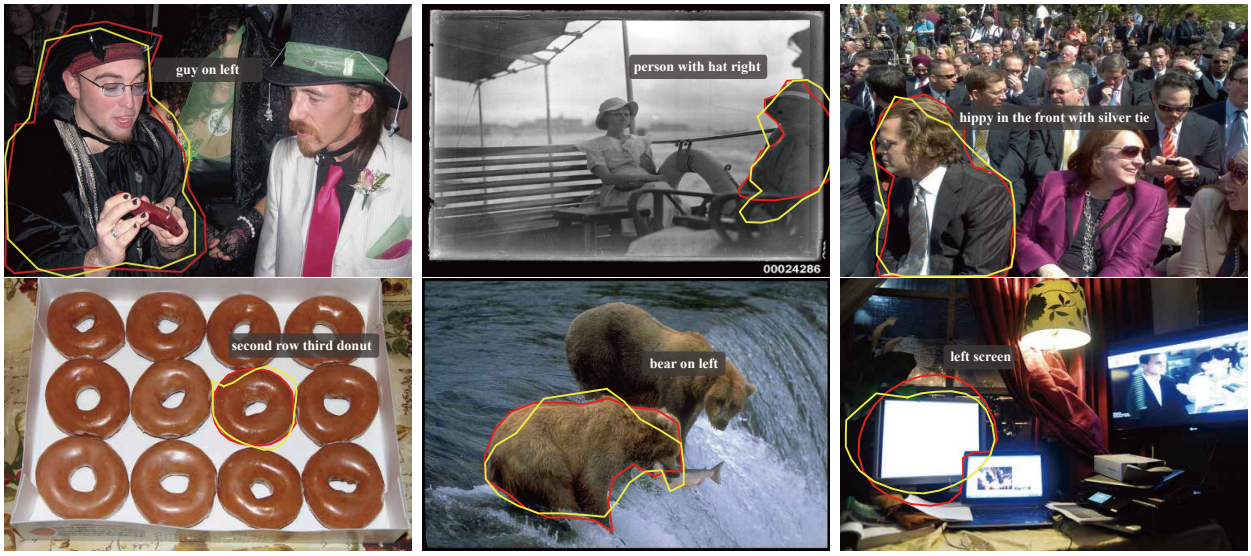


Figure 8: More Results of referring image segmentation in RefCOCO. The images of the first row are from testA split containing only people, while the second row images from testB consisting of only non-people. The referring expressions of the object are presented in the text box. In the image, the red denotes grounding truth and the yellow symbolizes the prediction.