Multi-modal Instruction Tuned LLMs with Fine-grained Visual Perception

Junwen He[*1,2], Yifan Wang[1], Lijun Wang[†1], Huchuan Lu[1], Jun-Yan He[2],
Jin-Peng Lan[2], Bin Luo[2], and Xuansong Xie[2]

[1]Dalian University of Technology
[2]DAMO Academy, Alibaba Group

junwen.he@mail.dlut.edu.cn, {wyfan, ljwang, lhchuan}@dlut.edu.cn
{leyuan.hjy, lanjinpeng.ljp, luwu.lb}@alibaba-inc.com, xingtong.xxs@taobao.com

Abstract

Multimodal Large Language Model (MLLMs) leverages Large Language Models as a cognitive framework for diverse visual-language tasks. Recent efforts have been made to equip MLLMs with visual perceiving and grounding capabilities. However, there still remains a gap in providing fine-grained pixel-level perceptions and extending interactions beyond text-specific inputs. In this work, we propose **AnyRef***, a general MLLM model that can generate pixel-wise object perceptions and natural language descriptions from multi-modality references, such as texts, boxes, images, or audio. This innovation empowers users with greater flexibility to engage with the model beyond textual and regional prompts, without modality-specific designs. Through our proposed refocusing mechanism, the generated grounding output is guided to better focus on the referenced object, implicitly incorporating additional pixel-level supervision. This simple modification utilizes attention scores generated during the inference of LLM, eliminating the need for extra computations while exhibiting performance enhancements in both grounding masks and referring expressions. With only publicly available training data, our model achieves state-of-the-art results across multiple benchmarks, including diverse modality referring segmentation and region-level referring expression generation. Code and models are available at* https://github.com/jwh97nn/AnyRef

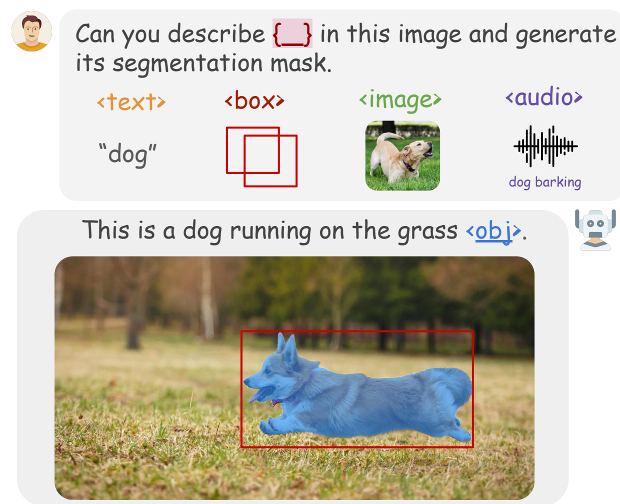

Figure 1. **Multi-modality Referring Segmentation** and **Expression Generation** with **AnyRef**. Our model possesses the capacity to generate natural language descriptions as well as pixel-wise grounding masks for the referred object. It accommodates various referring modalities such as **text**, **bounding boxes**, **images** and **audio**, enabling more flexible user interactions.

1. Introduction

Large language models (LLMs) have garnered widespread influence across various domains, and advancements have been achieved by augmenting LLMs with visual perception modules to bridge the gap between vision and language tasks [6, 18, 23, 61], thereby transforming them into Multimodal Large Language Models (MLLMs). Most recent research aims to further endow MLLMs with finer-grained visual understanding abilities, like visual grounding and referring expression generation, through user-defined formats (*e.g.*, coordinates, bounding boxes, etc.) [4, 31, 57], surpassing the confines of textual responses alone.

Despite the encouraging results demonstrated by existing MLLMs in grounding linguistic expressions to visual scenes, their capacity for precise localization remains restricted to coarse-grained levels (bounding boxes), falling short of pixel-level perceptions (As illustrated in Tab. 1).

*Work done during internship at DAMO Academy.
†Corresponding author

The most recent work, as exemplified by [16], has focused on enhancing MLLMs by integrating segmentation models that generate binary segmentation masks based on textual descriptions. However, this approach is constrained by its reliance solely on textual referring instructions, thereby limiting the versatility of MLLMs in various multimodal interaction scenarios, such as region-based referring or audio comprehension tasks. The interactive segmentation model SEEM [63] attempts to receive audio inputs, but it turns audio into textural prompts with the off-the-shelf speech recognition model Whisper [34], so essentially it is still the textual references.

In light of the above observation, we propose **AnyRef**, a novel multi-modal instruction-tuned LLM with fine-grained visual perception. As shown in Tab. 1, **AnyRef** advances existing MLLMs with the strong capability to perform pixel-level object grounding and generate region-aware expressions derived from references of diverse modalities, including text, bounding boxes, images, and audio inputs, (See Fig. 1 as an example). To this end, we first propose a unified representation for referring across different modalities and map them to the token space of LLMs. We extract features from all the modalities mentioned above to form the *Unified Referring Representation*, which can be processed uniformly by the LLM, utilizing its ability of understanding and reasoning in generating the grounded output. This enables flexible referring beyond textual descriptions, without requiring modality-specific designs or changes to the existing model.

To perform pixel-level grounding with LLMs, a possible solution [16] is to trigger the segmentation action by generating a special token <obj>, whose embedding will be subsequently employed as the input to the segmentation model. As opposed to using coordinates sequence of polygons [5, 41] to represent segmentation results, the introduction of the <obj> token effectively simplifies pixel-level visual grounding. Nevertheless, the embedding of the <obj> token is confined in a fixed feature space, due to the nature of next token prediction, leading to limited representational capacity and thus inaccurate segmentation results. To address this constraint, we propose a simple yet effective *refocusing mechanism*, which takes into account the correlation between the grounded expression and the <obj> token. This mechanism utilizes attention scores to weight such correlation, enhancing the mask embedding with additional grounded embeddings, and since the attention scores are intermediate outputs of the self-attention layers, the additional computation introduced by the refocusing mechanism is minimal. Furthermore, the refocusing mechanism also provides a short-cut connection between the generated grounded expression and the segmentation results, allowing pixel-level labels to implicitly supervise the learning process of language expression generation, thereby enhancing the model's regional understanding capability.

To summarize, our contributions are threefold:

- We introduce **AnyRef**, the first general MLLM capable of producing pixel-level object perceptions as well as region-aware referring descriptions. It adeptly accommodates multi-modality references including texts, bounding boxes, images or audio in a general manner, fostering more flexible interactions for users.
- We propose a simple yet effective *refocusing mechanism* to enhance the grounded mask predictions, leveraging the correlations of generated tokens without incurring additional computational overhead, and concurrently yields improvements in regional expression referring.
- Thorough experiments conducted on multiple datasets demonstrate the efficacy of the proposed method, resulting in state-of-the-art performance across a diverse range of multi-modality tasks.

Our model is built upon LLaVA-7B [23], which can be efficiently fine-tuned with 8 NVIDIA 32G V100 GPUs, making our method easily reproducible at a reasonable computational cost.

2. Related Works

2.1. Multi-modal Large Language Model

Multi-modal Large Language Models (MLLMs), built upon large language models (LLMs) as their foundations, extend their capabilities beyond traditional textual understanding to incorporate various modalities such as images, videos, and audio. Building upon the concept of instruction tuning, Flamingo [1] utilizes visual feature inputs as prompts, resulting in impressive performance across diverse visual-language tasks such as image captioning and visual question answering (VQA). Subsequent models, includin BLIP-2 [19], LLaVA [23], InstructBLIP [6], Otter [18] and LLaMa-Adapter [56], utilize additional generated visual instruction-following data for better visual-language alignment, and demonstrate impressive multi-modal chat abilities.

Recent studies expand the capabilities of MLLMs to address localization tasks with region-aware functionalities. KOSMOS-2 [31] and VisionLLM [41] introduce additional location tokens to the vocabulary, enabling the conversion of coordinates into textual representations. These representations are then inputted into LLMs to enhance region understanding. On the other hand, Shikra [4] represents coordinates directly in natural language form. In contrast, GPT4RoI [57] streamlines the process by employing RoI-aligned visual features without incorporating explicit positional information.

Nevertheless, these models lack the capacity to produce fine-grained perceptions (*e.g.*, pixel-level masks), and restrict their referring expressions to textural descriptions and

Method	Image	Referring Format			Pixel-level Grounding	End-to-End Model
		Region	Image*	Audio		
LLaVA (NeurIPS-23) [23]	✓	✗	✗	✗	✗	✓
BuboGPT (arXiv-23) [58]	✓	✗	✗	✓	✗	✗
Vision-LLM (arXiv-23) [41]	✓	✓	✗	✗	✗	✓
DetGPT (arXiv-23) [41]	✓	✓	✗	✗	✗	✓
KOSMOS-2 (arXiv-23) [31]	✓	✓	✗	✗	✗	✓
Shikra (arXiv-23) [4]	✓	✓	✗	✗	✗	✓
GPT4RoI (arXiv-23) [57]	✓	✓	✗	✗	✗	✓
NExT-GPT (arXiv-23) [44]	✓	✗	✗	✓	✗	✓
ASM (arXiv-23) [42]	✓	✓	✗	✗	✗	✓
LISA (arXiv-23) [16]	✓	✗	✗	✗	✓	✓
AnyRef (Ours)	✓	✓	✓	✓	✓	✓

Table 1. **Comparisons of recent Multi-modal Large Language Models.** The term *Referring Format* emphasizes the acceptable modalities used for referencing, whereas *Image** indicates visual references derived from another image.

regions within the image. Our model, leveraging the best of both worlds, not only generates pixel-level grounding masks, but also accommodates a broader range of referring formats (*e.g.*, visual reference from other images or audio) in a unified manner.

2.2. Referring Segmentation

Referring Expression Segmentation translates explicit textual descriptions into corresponding pixel-level segmentations, requiring a comprehensive understanding of both visual content and linguistic expression. Recent methods including SAM [15], X-Decoder [62] and SEEM [63] unify multiple segmentation tasks within a single model, supporting various human interaction methods. While LISA [16] utilizes the powerful reasoning and comprehension abilities of LLMs to process textural instructions and generate masks through the SAM [15] decoder.

Visual Referring Segmentation can be related to one/few-shot segmentation, where an example of a certain object with its corresponding mask is provided to segment the same object in the query image [12, 30, 43, 44, 55]. Recently, CLIPSeg [28] builds upon the CLIP model to treat the example image as a visual prompt, which can generalize to novel forms of prompts. Painter [43] and SegGPT [44] utilize in-context learning to perform general vision tasks using input task prompts.

Audio-Visual Segmentation aims to generate pixel-level masks for object(s) emitting sound, initially introduced in [60]. AVSegFormer [8] innovatively incorporates learnable audio queries, enabling selective attention to relevant visual features. Additionally, AUSS [21] proposes unmixing self-supervised losses to bridge the gap between audio signals and visual semantics.

While these models have achieved satisfactory results in their respective domains, there is currently a gap in addressing all referring tasks within a single model. Most of the aforementioned methods rely on modality-specific or task-specific designs, which may not generalize well beyond their intended tasks. Our approach leverages the robust comprehension ability of LLMs to concurrently tackle all these tasks while preserving the region-level reasoning capacity. Additionally, the *refocusing mechanism* aids in enhancing region-level referring expression through implicit pixel-level supervisions.

3. Methods

The overall framework of **AnyRef** comprises a vision encoder, multi-modal feature projection layers, a LLM, and a mask decoder, as illustrated in Fig. 2. These initial three components together form a multi-modality LLM, enabling support for various reference formats and generating region-aware grounded textual responses. Additionally, a distinctive <obj> token is introduced to the vocabulary, which provides the input for the mask decoder through a refocusing mechanism, facilitating the generation of pixel-level perceptions.

3.1. Model Architecture

We adopt the pretrained ViT-L/14 from CLIP [33] as the vision encoder, and LLaMA-7B [39] as our LLM. For audio inputs, we choose the pretrained audio encoder from ImageBind [9] to extract audio features. To connect multi-modality information beyond texts to the existing LLM, such as images and audio, we adopt vision-language and audio-language projection layers to project image and audio features to the language space. The input image is converted into a fixed number of 16×16 patch embeddings, while the audio is represented as 3 patch embeddings. Both

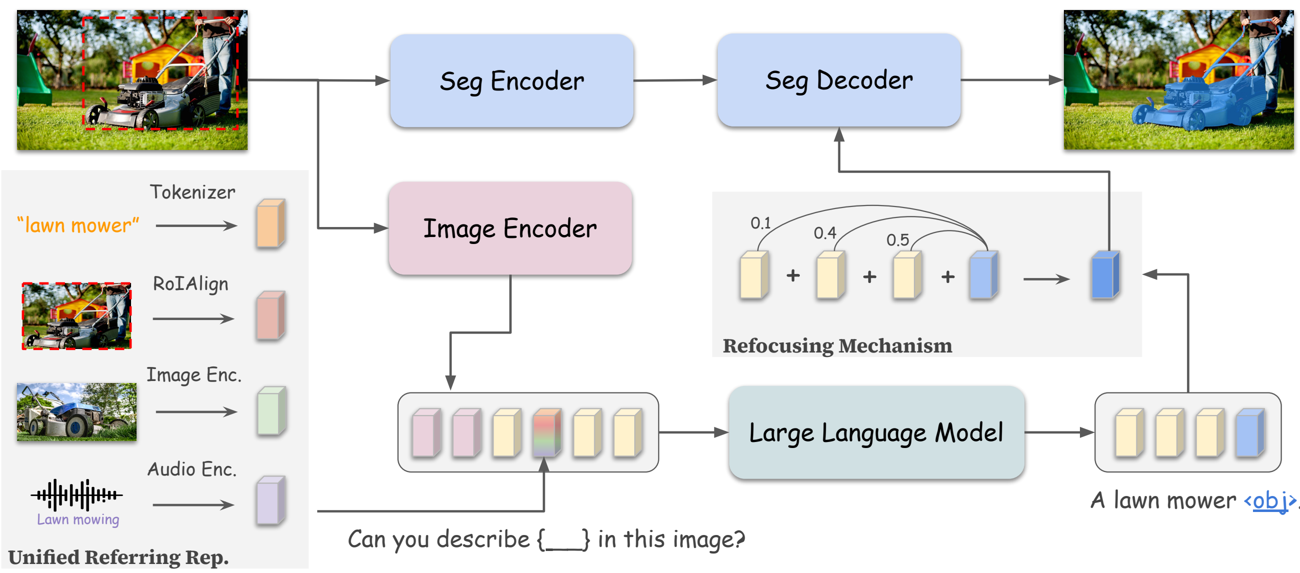

Figure 2. **Overall pipeline of AnyRef.** Vision-language, audio-language projection and MLP layers are omitted for simplicity and clarity. The **Unified Referring Representation** (Sec. 3.1.1) receives references from diverse types of modalities and transforms them into embeddings aligned with the LLM. The **Refocusing Mechanism** (Sec. 3.1.2) enhances the embedding from the single `<obj>` token with grounded textural embeddings, thus providing a broader representational capacity.

the image and audio embeddings are then projected to the same dimension as word embeddings. The LLM takes the interleaved embeddings in the same way as language tokens to generate outputs via an auto-regressive manner.

3.1.1 Unified Referring Representation

To receive multi-modality referring prompts beyond texts, we convert them into fixed-sized tokens and *quote* them between newly introduced special tokens.

For visual prompts including regional bounding boxes or visual examples from another image, we introduce `<img_ref>` and `</img_ref>`, where visual features will be inserted in between. Drawing inspiration from [57], we represent bounding boxes using extracted region-level features from RoIAlign [11] with a fixed size of 4×4. For processing image-level visual examples, we use the same CLIP vision encoder to extract visual features, which are then pooled to 4×4 as well. To refer to them in the same way as textual descriptions, we build prompts such as: "Can you provide a description of `<img_ref><img_feat></img_ref>` in this image?", where `<img_feat>` will be replaced by the extracted visual features.

For audio prompts, we introduce `<aud_ref>` and `</aud_ref>` for LLM to be aware of audio referring inputs, and the extracted audio features will be projected through audio-language projection layer and then inserted in between. And the audio prompted instruction will be built like: "Can you segment the object that makes sound of `<aud_ref><aud_feat></aud_ref>` in this image?". In this way, the referring representation from different modalities is unified, which can be treated the same way as language instructions and easily handled by the LLM.

3.1.2 Refocusing Mechanism

Inspired by [16], we employ another special token `<obj>` to succinctly represent the instance segmentation mask as an embedding. This embedding $\boldsymbol{h}_{obj}$ is derived from the last-layer of LLM associated with the `<obj>` token. It is then projected through an MLP layer γ, before being fed into the segmentation model $\mathcal{S}$. Subsequently, the binary segmentation mask M can be expressed mathematically as,

$$M = \mathcal{S}\Big(\gamma(\boldsymbol{h}_{obj}), \mathcal{V}_{seg}(\boldsymbol{x}_{img})\Big), \tag{1}$$

where $\boldsymbol{x}_{img}$ indicates the input image, and $\mathcal{V}_{seg}$ denotes the vision encoder of the segmentation model.

However, since `<obj>` is a token in the LLM vocabulary, its representation will be limited in a fixed feature range, which will potentially limit its representational capacity and influence the decoded mask quality. Therefore, we propose a *refocusing mechanism* which augments the original mask embedding with grounded text embeddings. The motivation behind is to explicitly force the final mask embedding to focus more on the referring or grounded object with its textural expression. The updated mask embedding can be formulated as

$$\hat{\boldsymbol{h}}_{obj} = \boldsymbol{h}_{obj} + \lambda_f \sum_{i}^{i<obj} \bar{\boldsymbol{a}}_i \cdot \boldsymbol{h}_i, \tag{2}$$

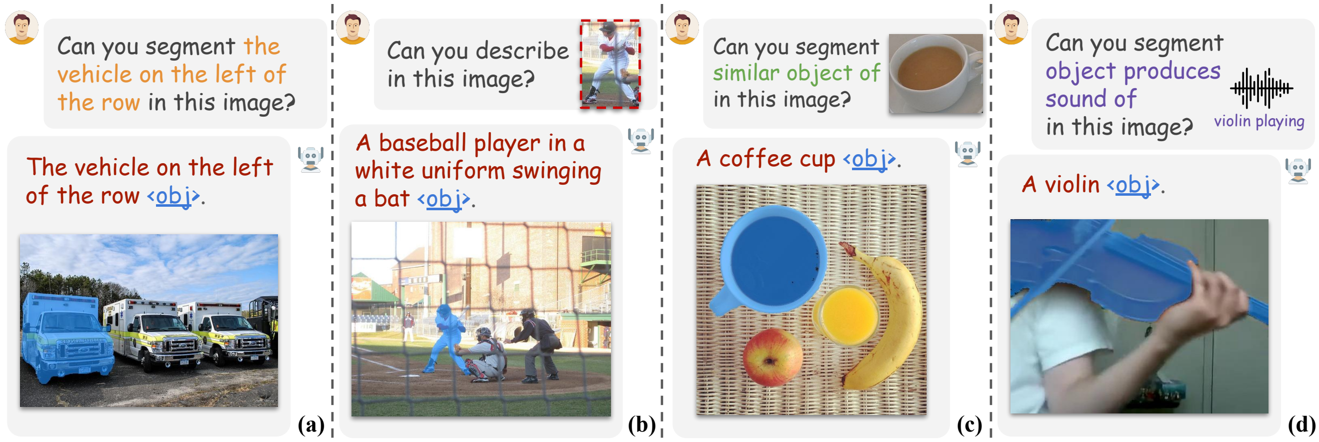

Figure 3. **Qualitative results of AnyRef's applicable capabilities** on multiple tasks, including **(a)** referring expression segmentation, **(b)** region-level captioning and grounding, **(c)** image-level referring segmentation and **(d)** audio-visual segmentation. **AnyRef** demonstrates proficiency in generating both textual responses and pixel-level perceptions across diverse modality instructions.

where $i{<}obj$ denotes the indices of output tokens before the <obj> token, $\bar{\boldsymbol{a}}_i$ indicates the normalized attention scores between the token i-th token and the <obj> token, and $\lambda_f = 0.1$ controls the focusing weight of augmentation embeddings. This approach enhances the mask embedding, providing a more adaptable feature range compared to the original, thereby expanding its representational capacity.

3.1.3 Training Objectives

The model is trained in the end-to-end manner with a combination of text loss and mask loss. The text loss follows the next word prediction loss [23], and the mask loss includes binary cross-entropy loss and dice loss [29], as

$$\mathcal{L} = \lambda_{text}\mathcal{L}_{text} + \lambda_{bce}\mathcal{L}_{bce} + \lambda_{dice}\mathcal{L}_{dice}, \tag{3}$$

where we choose $\lambda_{text} = 1.0$, $\lambda_{bce} = 2.0$ and $\lambda_{dice} = 0.5$. Due to the *refocusing mechanism*, tokens generated before the <obj> token can receive additional supervisory signals from pixel-level ground truth. This mutual interaction can further benefit the vision-language understanding ability of **AnyRef**, given the interrelated nature of referring expressions and grounding masks.

3.2. Implementation Details.

3.2.1 Training Setup

Unless otherwise specified, we employ the pre-trained CLIP ViT-L/14 as the vision encoder, ImageBind-H [9] as the audio encoder, and LLaMa-7B as the LLM. The vision-language projection layer is initialized from LLaVa [23], while the audio-language projection layer is randomly initialized. The word embeddings of newly introduced special tokens are initialized randomly. Furthermore, the segmentation model utilizes the pre-trained SAM-H [15]. The image resolution is 224×224 for MLLM and 1024×1024 by rescaling and padding for the segmentation model. For audio inputs, we follow settings in [60] to use the 5-second audio clips and convert to 3 fixed-sized embeddings after padding, since the ImageBind [9] audio encoder samples 2-second audio each time.

To ensure training efficiency and preserve generalization ability, we freeze the vision encoders and audio encoder. Fine-tuning of the LLM is conducted using LoRA [13], and the trainable parameters comprise the mask decoder and projection layers, accounting for approximately 7% of the total parameters.

We conduct training using 8 NVIDIA V100 GPUs, each with a batch size of 6, and employ a gradient accumulation step set to 8. The training utilizes mixed precision, converting both the vision and audio encoder to float16 precision. AdamW [26] optimizer with a learning rate of 5e-5 and weight decay of 0.01 is employed, alongside a cosine annealing scheduler incorporating 200 warmup steps. LoRA operates with the rank of 8 and alpha of 16, exclusively applied to query and value projections within the LLM. We employ ZeRO stage-2 [35] with DeepSpeed [37] which completes network training in 10K steps.

3.2.2 Datasets

The training process involves a diverse range of datasets. For general semantic and instance segmentation, COCO-Stuff [3], ADE20K [59], and PACO-LVIS [36] are utilized, with one category chosen per batch. Referring expression segmentation incorporates RefClef, RefCOCO, RefCOCO+ [14], RefCOCOg [52], and PhraseCut [46]. Image-level referring segmentation adopts the method outlined in [27], where samples are chosen from COCO [20], PascalVOC [7], and PhraseCut [46] datasets. Random cropped sam-

Method	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val(U)	test(U)
Specialist Segmentation Models								
CRIS [45]	70.5	73.2	66.1	65.3	68.1	53.7	59.9	60.4
LAVT [49]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
GRES [22]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
PolyFormer [25]	76.0	78.3	73.3	69.3	74.6	61.9	69.2	70.2
UNINEXT [48]	**82.2**	**83.4**	**81.3**	**72.5**	**76.4**	**66.2**	**74.7**	**76.4**
SEEM [63]	-	-	-	-	-	-	65.7	-
Generalist MLLMs								
X-Decoder [62]	-	-	-	-	-	-	64.6	-
LISA-7B [16]	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.4
LISA-7B (ft) [16]	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6
AnyRef	74.1	75.5	70.8	64.1	68.7	57.5	68.1	69.9
AnyRef (ft)	76.9	79.9	74.2	70.3	73.5	61.8	70.0	70.7

Table 2. **Referring expression segmentation** results (cIOU) on RefCOCO(+/g) datasets. (ft) denotes finetuning the model on RefCOCO(+/g) datasets. Our model surpasses all generalist models and most specialist (segmentation-oriented) models.

ples are drawn from images that contain the same category as their corresponding linguistic expressions. Region-level captioning involves RefCOCO(+/g) and Flickr30K Entities [32]. Audio-visual segmentation employs AVSBench [60] with both single and multiple sound sources. To prevent data leakage, samples with images in the validation or test splits are excluded.

4. Experiments

We assess the capabilities of our model through evaluations on various benchmarks, including different modality referring segmentation (text/image/audio) for pixel-level perception and referring expression generation for regional understanding. Models are categorized as *specialists* or *generalists*, with the former designed exclusively for specific tasks. We provide examples for each task in Fig. 3, and more illustrations can be found in the supplementary material.

4.1. Multi-modality Referring Segmentation

4.1.1 Referring Expression Segmentation

The task involves labeling pixels within an image corresponding to an object instance referred to by a linguistic expression. We instruct our model as: "Can you segment {exp} in this image?", where {exp} is the given explicit description. Evaluation is conducted using Cumulative-IoU (cIoU) as the metric. We make comparisons with state-of-the-art models on validation and test sets of RefCOCO, RefCOCO+ and RefCOCOg [14, 52]. As shown in Tab. 2, our performance surpasses all generalist models and most specialist models except UNINEXT-H [48], which is trained using a considerably larger dataset that includes video samples. Specialist models excel solely at segmentation-related tasks, while generalist models possess additional capabilities for generating textural descriptions and are capable of handling more complex references.

4.1.2 Image Referring Segmentation

Predicting masks using image examples is akin to one- or few-shot segmentation, where regions corresponding to the highlighted object in the example image must be located in a query image. We prompt our model with queries like "Can you find similar object of <img_ref><img_feat></img_ref> in this image?", where <img_feat> denotes pooled features from example images as detailed in Sec. 3.1.1. The evaluation takes place under the in-domain setting on COCO-20^i [20] and PASCAL-5^i [7] for a fair comparison, as most classes are encountered during the training stages. In the few-shot evaluation, the model inferences multiple times using different example images, with the averaged mask serving as the final prediction. In our referring examples, we do not have corresponding mask examples, which is different from the standard setting. we follow [28] to crop out the target object for highlighting, using their segmentation masks. As demonstrated in Tab. 3, our model achieves competitive performance compared to state-of-the-art methods.

4.1.3 Audio-Visual Segmentation

The AVS benchmark comprises single- and multi-sources subsets based on the number of sounding objects. We utilize prompts like, "Can you segment the object(s) that produce sound of <aud_ref><aud_feat></aud_ref> in this image?", to instruct the model for mask predictions. Following [60], evaluation metrics include mean IoU

Method	COCO-20^i		RASCAL-5^i	
	one-shot	few-shot	one-shot	few-shot
Specialist Segmentation Models				
HSNet* [30]	41.7	50.7	68.7	73.8
VAT* [12]	42.9	49.4	72.4	76.3
CLIPSeg [28]	33.2	-	59.5	-
SegGPT [44]	**56.1**	**67.9**	**83.2**	**89.8**
Generalist Multi-task Models				
Painter [43]	32.8	32.6	64.5	64.6
AnyRef	43.5	51.3	74.8	78.6
AnyRef†	46.3	55.2	76.5	80.0

Table 3. Quantitative results of **example-based few-shot segmentation**. * indicates that the categories in training cover that in testing as in [44], and † denotes using mask cropping setting.

Method	Single-source		Multi-source	
	mIOU	F-score	mIOU	F-score
AVS [60]	78.7	0.879	54.0	0.645
BG [10]	81.7	0.904	55.1	0.668
AVSegformer [8]	82.1	0.899	58.4	0.693
AUSS [21]	**89.4**	**0.942**	**63.5**	**0.752**
AnyRef	82.8	0.908	55.6	0.663

Table 4. Quantitative results of **audio-visual segmentation**.

(mIoU) for region similarity and F-score[1] for contour accuracy. The quantitative results in Tab. 4 demonstrate that our model consistently outperforms most methods on single-source split, indicating successful alignment of audio features with the LLM during fine-tuning. However, when confronted with audios containing multiple sound sources, our model encounters challenges in producing masks that cover more than one object. Moreover, owing to the ability of LLM, our model can determine the textural category of the sounding objects, as depicted in Fig. 3 (d).

4.2. Referring Expression Generation

This task involves generating a textual description associated with an object based on its location (bounding box). We evaluate our generated expressions using automatic caption generation metrics, including CIDEr [40] and Meteor [17], on RefCOCO, RefCOCO+ and RefCOCOg. Our model achieves remarkable performance among generalist LLM-based models and demonstrates competitive result to specialist models, as shown in Tab. 5.

Nonetheless, as stated in [2, 24, 54], standard automated evaluation metrics do not authentically capture generation quality due to the constraints of ground-truth expressions. This scenario is particularly pronounced in open-text generation, especially for LLM-based models. These models have the ability to generate rich, natural sentences, while

[1] $F_\beta = \frac{(1+\beta^2)\times \text{precision}\times\text{recall}}{\beta^2\times\text{precision}+\text{recall}}$, where $\beta^2 = 0.3$ following [60]

Figure 4. Comparison of **generated expressions** between ground-truth and LLM-based methods.

the provided ground-truth expressions often tend to be concise, as indicated in Fig. 4.

To further evaluate the quality of the generated expressions, we conduct human evaluations following [2, 51, 54]. We randomly select 100 images from the validation datasets and ask five human raters to choose the bounding box that best matches the generated expression, and the averaged score is considered the final result. In Tab. 6, we present the results of the human evaluations, including both traditional methods and LLM-based methods. The LLM-based methods produce more detailed descriptions, closely resembling human behavior, which are preferred by the human raters. We provide more examples in supplementary material.

4.3. Ablation Study

We conduct extensive ablation studies to reveal the contribution of each component.

Refocusing Mechanism. We first investigate the effectiveness of enhancing the <obj> token through *refocusing*

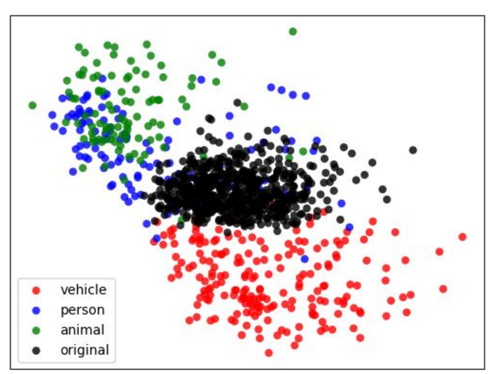

Figure 5. Visualization of mask embeddings before and after the *refocusing mechanism*. **original** denotes original mask embeddings, while **vehicle**, **person**, and **animal** represent the updated mask embeddings corresponding to their respective referring objects contained in the textural expression.

Method	RefCOCO				RefCOCO+				RefCOCOg	
	testA		testB		testA		testB		val	
	Meteor	CIDEr	Meteor	CIDEr	Meteor	CIDEr	Meteor	CIDEr	Meteor	CIDEr
Specialist Models										
Visdif [53]	18.5	-	24.7	-	14.2	-	13.5	-	14.5	-
SLR [54]	29.6	77.5	34.0	132.0	21.3	52.0	21.5	73.5	15.9	66.2
easyREG [38]	<u>31.3</u>	<u>83.7</u>	<u>34.1</u>	132.9	<u>24.2</u>	66.4	22.8	78.7	17.0	77.7
IREG [50]	**34.9**	**105.4**	**37.3**	**154.1**	**30.8**	**89.8**	**26.4**	**97.0**	**19.4**	**101.2**
Generalist MLLMs										
GRIT [47]	-	-	-	-	-	-	-	-	15.2	71.6
KOSMOS-2 [31]	-	-	-	-	-	-	-	-	14.1	62.3
AnyRef	23.9	74.8	26.7	118.6	16.4	59.4	14.3	62.9	16.2	69.0
AnyRef (ft)	30.4	79.5	32.7	<u>138.6</u>	23.2	<u>67.7</u>	20.1	<u>80.1</u>	<u>17.1</u>	<u>79.7</u>

Table 5. Quantitative results on **region-level referring expression generation**. *Generalist models* (LLM-based) perform poorly on automated evaluation metrics due to the limitation of constrained ground-truth expressions, as stated in Sec. 4.2.

Method	RefCOCO		RefCOCO+	
	testA	testB	testA	testB
SLR [54]	66%	62%	43%	38%
SLR+Rerank [54]	73%	77%	49%	46%
KOSMOS-2 [31]	88%	**84%**	63%	65%
Shikra [4]	**91%**	81%	59%	62%
AnyRef	87%	80%	**67%**	**66%**

Table 6. **Human evaluation** on referring expression generation.

λ_f	RefCOCOg	AVSBench	RefCOCOg	
	cIOU	mIOU	Meteor	CIDEr
0.0	68.7	81.4	16.8	71.1
1.0	67.1	80.6	14.3	68.5
0.1	**70.0**	**82.8**	**17.1**	73.7
1.0†	68.0	81.1	15.7	70.0
0.1†	69.3	82.0	17.0	**73.8**

Table 7. Ablation study on **refocusing weight** λ_f. x† indicates trainable λ_f initialized with x.

mechanism, and explore the impact of different refocusing weights λ_f. We evaluate setting different values for λ_f and also try setting it as a learnable parameter along with the model. We conduct evaluations on both referring segmentation and expression generation tasks. Results in Tab. 7 reveal that the refocusing weight significantly affects performance in both tasks. A small weight of 0.1 improves performance, while a larger weight can have detrimental effects, particularly in expression generation. We also experiment with learning λ_f as a parameter along with the model, but we find that the performance varies greatly depending on the initialized value. Thus, for simplicity and stability, we empirically select $\lambda_f = 0.1$ for our experiments.

We further employ PCA to visualize the mask embeddings before and after implementing the *refocusing mechanism* in Fig. 5 We choose three subsets representing different referring objects including vehicles, persons and animals (*e.g.*, the person subset comprises output expressions containing "person," "man," "woman," etc.). The visualization illustrates that the *refocusing mechanism* results in a wider representation range of the mask embedding. Moreover, the updated embeddings demonstrate a clustering pattern aligned with the associated textual expressions, contributing to a more precise decoding of masks.

Exp.	Referring	General	Region Ref.	Image Ref.	cIOU
1	✓				66.2
2	✓	✓			67.0
3	✓	✓	✓		67.7
4	✓	✓		✓	67.4
5	✓	✓	✓	✓	**68.1**

Table 8. Ablation study on **training datasets**.

Training Datasets. The impact of different types of datasets is validated in Tab. 8, and evaluation is carried out on RefCOCOg validation split. Region/Image Ref. refers to region-level and image-level referring data, as explained in Sec. 3.2.2. It becomes apparent that the model's generalization improves as the type of datasets increases.

5. Conclusion

We present **AnyRef**, a pioneering MLLM model capable of generating pixel-level object perceptions and language descriptions from various modality references, including texts, regions, images, and audio. This is made possible by the unified referring representation, which connects different types of inputs to the LLM. We further propose a refocusing mechanism that uses attention scores to improve the segmentation embedding and enhance pixel-level vision perception. Across various downstream tasks, our model exhibits remarkable performance while providing users with enhanced interacting flexibility.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (U23A20386, 62276045, 62293540, 62293542), Dalian Science and Technology Talent Innovation Support Plan (2022RY17).

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] Lior Bracha, Eitan Shaar, Aviv Shamsian, Ethan Fetaya, and Gal Chechik. Disclip: Open-vocabulary referring expression generation. *arXiv preprint arXiv:2305.19108*, 2023. 7

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 5

[4] 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, 3, 8

[5] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. *Advances in Neural Information Processing Systems*, 35:31333–31346, 2022. 2

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1, 2

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 5, 6

[8] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. *arXiv preprint arXiv:2307.01146*, 2023. 3, 7

[9] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 3, 5

[10] Dawei Hao, Yuxin Mao, Bowen He, Xiaodong Han, Yuchao Dai, and Yiran Zhong. Improving audio-visual segmentation with bidirectional generation. *arXiv preprint arXiv:2308.08288*, 2023. 7

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4

[12] Sunghwan Hong, Seokju Cho, Jisu Nam, Stephen Lin, and Seungryong Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2022. 3, 7

[13] 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. 5

[14] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 5, 6

[15] 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. *arXiv:2304.02643*, 2023. 3, 5

[16] 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. 2, 3, 4, 6

[17] Alon Lavie and Abhaya Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, 2007. Association for Computational Linguistics. 7

[18] 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. 1, 2

[19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2

[20] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5, 6

[21] Yuhang Ling, Yuxi Li, Zhenye Gan, Jiangning Zhang, Mingmin Chi, and Yabiao Wang. Hear to segment: Unmixing the audio to guide the semantic segmentation. *arXiv preprint arXiv:2305.07223*, 2023. 3, 7

[22] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23592–23601, 2023. 6

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2, 3, 5

[24] Jingyu Liu, Liang Wang, and Ming-Hsuan Yang. Referring expression generation and comprehension via attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4856–4864, 2017. 7

[25] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18663, 2023. 6

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

[27] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022. 5

[28] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7086–7096, 2022. 3, 6, 7

[29] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 5

[30] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3, 7

[31] 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. 1, 2, 3, 8

[32] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123 (1):74–93, 2017. 6

[33] 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. 3

[34] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 2

[35] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020. 5

[36] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, Amir Mousavi, Yiwen Song, Abhimanyu Dubey, and Dhruv Mahajan. PACO: Parts and attributes of common objects. In *arXiv preprint arXiv:2301.01795*, 2023. 5

[37] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020. 5

[38] Mikihiro Tanaka, Takayuki Itamochi, Kenichi Narioka, Ikuro Sato, Yoshitaka Ushiku, and Tatsuya Harada. Generating easy-to-understand referring expressions for target identifications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5794–5803, 2019. 8

[39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[40] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 7

[41] 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. 2, 3

[42] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023. 3

[43] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 3, 7

[44] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 3, 7

[45] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 6

[46] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020. 5

[47] 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. 8

[48] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, 2023. 6

[49] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, 2022. 6

[50] Fulong Ye, Yuxing Long, Fangxiang Feng, and Xiaojie Wang. Whether you can locate or not? interactive referring expression generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4697–4706, 2023. 8

[51] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 7

[52] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 5, 6

[53] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 8

[54] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7282–7290, 2017. 7, 8

[55] Jian-Wei Zhang, Yifan Sun, Yi Yang, and Wei Chen. Feature-proxy transformer for few-shot segmentation. *Advances in Neural Information Processing Systems*, 35:6575–6588, 2022. 3

[56] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 2

[57] 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, 2023. 1, 2, 3, 4

[58] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023. 3

[59] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 5

[60] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *European Conference on Computer Vision*, pages 386–403. Springer, 2022. 3, 5, 6, 7

[61] 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

[62] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023. 3, 6

[63] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023. 2, 3, 6