Jack of All Tasks, Master of Many: Designing General-purpose Coarse-to-Fine Vision-Language Model

Shraman Pramanick*[1,2†] Guangxing Han*[2] Rui Hou[2] Sayan Nag[3] Ser-Nam Lim[4]
Nicolas Ballas[2] Qifan Wang[2] Rama Chellappa[1] Amjad Almahairi[2]

[1]Johns Hopkins University, [2]Meta, [3]University of Toronto, [4]University of Central Florida

Abstract

The ability of large language models (LLMs) to process visual inputs has given rise to general-purpose vision systems, unifying various vision-language (VL) tasks by instruction tuning. However, due to the enormous diversity in input-output formats in the vision domain, existing general-purpose models fail to successfully integrate segmentation and multi-image inputs with coarse-level tasks into a single framework. In this work, we introduce VistaLLM, a powerful visual system that addresses coarse- and fine-grained VL tasks over single and multiple input images using a unified framework. VistaLLM utilizes an instruction-guided image tokenizer that filters global embeddings using task descriptions to extract compressed and refined features from numerous images. Moreover, VistaLLM employs a gradient-aware adaptive sampling technique to represent binary segmentation masks as sequences, significantly improving over previously used uniform sampling. To bolster the desired capability of VistaLLM, we curate CoinIt, a comprehensive coarse-to-fine instruction tuning dataset with 6.8M samples. We also address the lack of multi-image grounding datasets by introducing a novel task, AttCoSeg (Attribute-level Co-Segmentation), which boosts the model's reasoning and grounding capability over multiple input images. Extensive experiments on a wide range of V- and VL tasks demonstrate the effectiveness of VistaLLM by achieving consistent state-of-the-art performance over strong baselines across many downstream tasks. Our project page can be found at https://shramanpramanick.github.io/VistaLLM/.

1. Introduction

Large language models (LLM) have proven to be the *de-facto* solution to address novel natural language processing (NLP) tasks, thanks to their ability to comprehend user-tailored prompts, instructions, and detailed task descriptions [13, 24, 71, 72, 89, 90]. However, the problem is more challenging the vision domain due to an inherent disparity of input and output formats across different tasks. Though pre-training followed by a fine-tuning strategy is effective for various vision problems [15, 16, 45, 47, 52, 53, 75, 76, 78, 94, 105], with the continuously increasing model parameters, the marginal cost for task-specific tuning comes with significant computational overhead. Hence, it becomes crucial to design general-purpose vision models that can perceive natural-language instructions to solve various vision problems in a zero-shot manner.

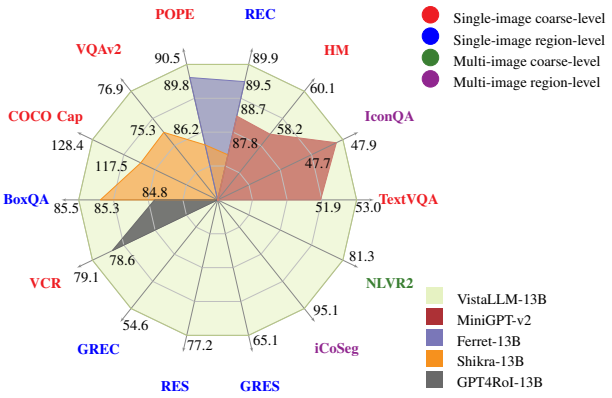


Figure 1. **VistaLLM achieves the state-of-the-art** performance across a broad range of single and multi-image coarse-to-fine grained reasoning and grounding tasks (see Table 1 for details) among general-purpose baselines. Notably, no existing baseline have unified segmentation and multi-image tasks in a single system. We show officially reported numbers for every baseline.

The development of general-purpose vision models faces two significant challenges: first, the unification of diverse input-output formats, and second, an effective representation of visual features for a variety of tasks. Image-level vision tasks such as classification, captioning, and question-answering involve textual outputs and primarily require a broader, coarse-grained image representation, making them relatively straightforward to integrate into a unified framework [14, 21, 57, 121]. In contrast, region-level prediction tasks like object detection and semantic segmentation ne-

*Equal technical contribution.
†Part of this work was done during an internship at Meta.

cessitate fine-grained, pixel-scale visual features and produce dense outputs such as bounding boxes and masks. Converting bounding boxes to natural language sequences is feasible by serializing the coordinates of two corners. However, representing a binary mask as a text sequence poses a more complex challenge, especially when dealing with multiple input images each associated with numerous segmentation masks. Although some recent general-purpose systems have succeeded in unifying coarse-level tasks with object detection [7, 8, 33, 73, 111, 118], they do not incorporate segmentation within the same framework. Furthermore, the capabilities of these existing systems are often limited to processing single-image input, thereby constraining their applicability in broader, more complex scenarios, such as reasoning over multiple images and recognizing and segmenting common objects.

In this work, we present **VistaLLM**, the first general-purpose vision model that addresses coarse- and fine-grained vision-language reasoning and grounding tasks over single and multiple input images. We unify these tasks by converting them into an instruction-following sequence-to-sequence format. We efficiently transform binary masks into a sequence of points by proposing a gradient-aware adaptive contour sampling scheme, which significantly improves over the naive uniform sampling technique previously used for sequence-to-sequence segmentation tasks [9, 10, 58, 120]. Moreover, to preserve global and region-level information from multiple input images, we propose utilizing a QFormer [45] based instruction-guided image tokenizer. Leveraging LLMs' language reasoning ability, we feed our visual features with carefully designed task-specific instructions to LLMs, which generate responses following the instructions. Integrating various tasks with different granularity into such a unified, cohesive, and end-to-end system helps improve the performance of each task by sharing coarse- and fine-grained feature representation.

To train VistaLLM on a versatile form of vision and language tasks, we collect **CoinIt** (**Co**arse-to-f**in**e **I**nstruction-**t**uning Dataset) with 6.8M samples, ranging over four broad categories of tasks - single-image coarse-level, single-image region-level, multi-image coarse-level, and multi-image region-level. We address the lack of publicly-available multi-image region-level datasets by proposing a novel task, AttCoSeg (**Att**ribute-level **Co**-**Seg**mentation), which aims to recognize input images which have objects with common attributes (shape, color, size, position), and segment those objects. AttCoSeg contains 804k training samples, and help VistaLLM to gain significant generalizable reasoning and grounding capability over multiple input images. Other tasks of CoinIt are constructed by converting publicly available benchmarks into instruction-following format, such as COCO [54], Flickr [74], VCR [113], LLaVA [57], VG [37], PASCAL [19] etc. Extensive evaluation on 15 different benchmarks proves the efficacy of VistaLLM, which even surpasses specialist (or fine-tuned) systems in most tasks, including 10.9% CIDEr points gain over Shikra [8] on image captioning, 13.1%, 6.7% precision and gIoU improvements over MDETR [35] on GREC and GRES, 3% $\mathcal{J}$-index gains over CycleSegNet [115] on iCoSeg.

In summary, our contributions are threefold: (i)We propose VistaLLM, equipped with a instruction-guided image tokenizer, to seamlessly integrate coarse- and fine-grained vision-language reasoning and grounding tasks over single and multiple input images into a unified general-purpose model. (ii) To efficiently convert segmentation masks into a sequence, we propose a gradient-aware adaptive contour sampling scheme, which improves over previously used uniform sampling by $3-4$ mIoU scores on different segmentation benchmarks. (iii) We construct CoinIt, a large-scale coarse-to-fine instruction-tuning dataset, for model training. Moreover, we introduce a novel task, AttCoSeg, which addresses the lack of publicly available multi-image grounding datasets. We evaluate VistaLLM on a wide-range of vision-language tasks across 15 benchmarks, achieving state-of-the-art performance in all of them, even surpassing specialist systems. We summarize these results in Figure 1.

2. Related Works

General-purpose vision models, also known as multimodal large language models (MLLM), have recently been proven to be an effective way to unify a versatile array of vision and language tasks. These models, which use potent LLMs [4, 13, 17, 24, 29, 71, 72, 88–90, 92, 100, 102, 114, 117] to reason textual instructions, can broadly be categorized into two groups based on their input and output formats:

Coarse-level MLLMs: Early attempts of designing MLLMs focused on image-level vision tasks with textual outputs, such as visual question answering [2, 28, 65, 83] and image captioning [23, 25]. Frozen [91], Flamingo [1], FrozenBiLM [104], MAGMA [18], ClipCap [69], VidIL [98], PICa [106] are among the first few to show the in-context capability of LLMs for few-shot vision tasks. More recent works have focused on using LLMs for visual instruction tuning. To name a few, LLaVA [57], MiniGPT-4 [121], MM-REACT [109], BLIP2 [45], mPLUS-OWL [110], LLaMA-Adapter v2 [21], Otter [41], Instruct-BLIP [14], LLaVA-Med [42] have been proven to be effective. However, these models lack region-specific capabilities and can not perform visual grounding tasks.

Region-level MLLMs: More recently, MLLMs have moved forward to unify region-based referring and grounding tasks into general-purpose vision systems. KOSMOS-2 [73], VisionLLM [95], Shikra [8], GPT4RoI [118], All-Seeing Model [97], CogVLM [96], COMM [32], MiniGPT-v2 [7] and Ferret [111] has shown the capability of MLLMs

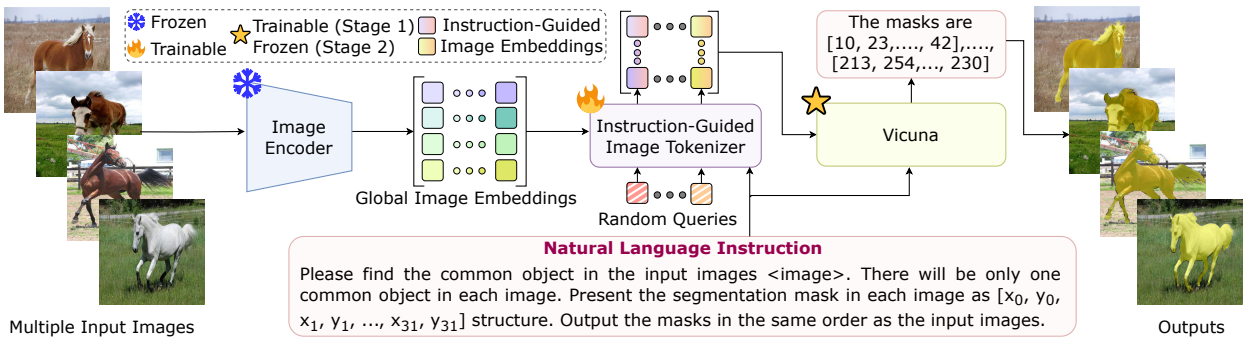


Figure 2. **Overview of the proposed system - VistaLLM**, which integrates single- and multi-image coarse- and fine-grained vision-language tasks into a unified general-purpose framework. VistaLLM contains three key design modules - (i) image encoder to extract the global image embedding, (ii) instruction-guided image tokenizer, which refines and compresses the global image embeddings using task instruction, enabling the model to filter the necessary visual information required for the current task, and (iii) LLM (Vicuna)-based decoder to jointly process image and language features, and generate the desired output. VistaLLM uses a gradient-aware adaptive sampling technique to efficiently represent segmentation masks as a point sequence, described in Section 3.2. All parameters except the image encoder are trained in stage 1, while only the image tokenizer is fine-tuned in stage 2 (See Section 3.1, 5.2 for details).

of fine-grained image comprehension and region-focused conversation. While KOSMOS-2, Shikra, and VisionLLM feed the image coordinates directly into the LLM, GPT4RoI and Ferret use additional feature extractor modules to represent image regions. On a related regime, InternGPT [59], BuboGPT [119], and LISA [38] utilize external vision modules to perform grounding tasks. However, these works are only capable of processing single-input images. In this work, we propose VistaLLM to address all possible reasoning and grounding tasks over single and multiple images. Moreover, we efficiently convert binary masks into sequence by a novel adaptive sampling, which helps to unify segmentation into a general-purpose framework.

3. Method

We start by presenting the model architecture of VistaLLM. Next, we detail the proposed sequence generation approach for segmentation masks and illustrate its efficacy compared to uniform sampling.

3.1. Model Architecture

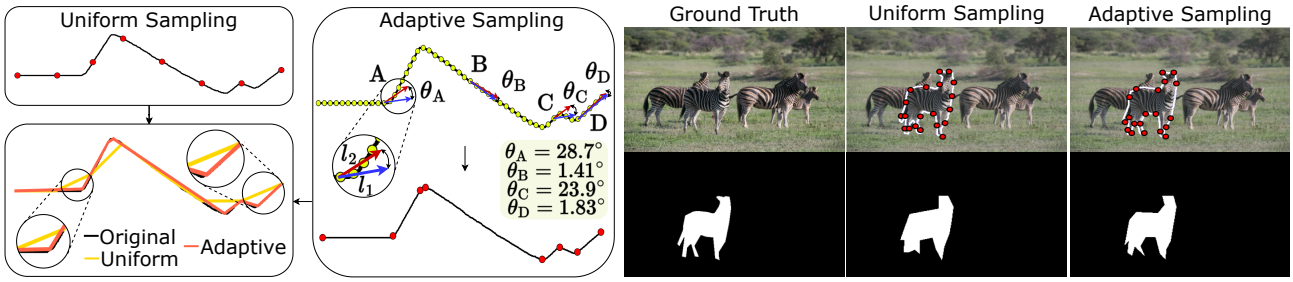
The overall architecture of VistaLLM, shown in Figure 2, consists of three key design modules - (i) image encoder to extract the global image embedding, (ii) instruction-guided image tokenizer, which refines and compresses the global image embeddings using task instruction, enabling the model to filter the necessary visual information required for the current task, and (iii) LLM-based decoder to jointly process image and language features, and generate the desired output.

Image Encoder. Given a set of k input images $X = \{x_\text{i}\}_1^k$; $x_\text{i} \in \mathbb{R}^{H_i \times W_i \times 3}$, where H_i and W_i denote the height and width of the i^th image, we first feed them into a pre-trained image encoder, EVA-CLIP [87], to extract k image embeddings $Z = \{z_\text{i}\}_1^k$; $z_\text{i} \in \mathbb{R}^{N_\text{i} \times D}$, N_i is number of spatial tokens in the i^th image and D is the hidden dimension. Note that, for larger k, the image feature dimension increases, making it difficult for the LLM decoder to process it as input, which is taken care of in the tokenizer module.

Instruction-guided Image Tokenizer. Unlike many previous general-purpose vision systems [7, 8, 57, 73], which directly feed the global image features into the decoder, we introduce an instruction-guided image tokenizer, which plays three crucial roles: (i) refines the image embeddings in alignment with task description, i.e. for coarse-level tasks, global features are important, whereas for fine-level tasks, only the region features need to be processed. (ii) compresses the image embeddings, which is important when there are many input images, and (iii) flexibly projects multiple input images with different heights and widths into the same feature dimension.

The image tokenizer module takes image embeddings and the language instruction and outputs the refined and compressed visual features. If referring regions (points, boxes, masks) are present in the instruction, they are converted to text-interleaved sequence as described in Section 3.2. Afterwards, we propose to adopt a QFormer [45] network with L ($L < N_\text{i}, \forall i$) randomly-initialized queries, which learns high-level task-specific information using the language instruction. The output from the tokenizer, $F = \{f_\text{i}\}_1^k$; $f_\text{i} \in \mathbb{R}^{L \times D}$, are then flattened to produce the final visual features, $F_\text{v} \in \mathbb{R}^{kL \times D}$ which are fed into the LLM.

LLM. We use Vicuna [12] as our language model, which is a decoder-only LLM [5] with a context length of 2048 build by instruction-tuning LLaMa [89]. The LLM takes the vision features F_v and the language instruction as input, and generates task-specific output. We train the LLM end-to-end by traditional next-token prediction objective calculated over the ground-truth. Since Vicuna only has the digits 0-9 in its vocabulary, we introduce additional tokens 10-999 to represent quantized coordinates. During evaluation, we de-



(a) **Illustration of uniform and adaptive sampling** on a line curve. (b) **Illustration of uniform and adaptive sampling** on a object mask.

Figure 3. **Visualization of uniform and adaptive sampling strategies.** (a) illustration of sampled points and comparison of reassembled curves, (b) illustration of sampled points and comparison of reassembled masks.

quantize the generated number tokens into the image space for metric calculation.

3.2. Sequence Generation for Grounding Tasks

The outputs from grounding tasks typically manifest in one of three formats: points, boxes, and masks. Points and boxes are straightforward to quantify and serialize, as evidenced in [7, 8, 73]. For instance, a point is represented by its coordinates $[x, y]$, while a box is denoted by its diagonal corner points $[x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}}]$, signifying the top-left and bottom-right corners. Conversely, the outline of a mask can assume any free-form shape comprising potentially infinite points. In scenarios where such free-form polygons are referenced in the input instructions, they can be encoded as region features [111, 118]. However, translating segmentation masks into a sequence for output by a general-purpose framework is particularly challenging, and the process necessitates conversion of segmentation masks into a small number of discrete points.

Previously, encoder-decoder-based segmentation approaches [9, 10, 58, 120] uniformly sample N points clockwise from the contour of the mask, and then quantize and serialize them as $[x_1, y_1, x_2, y_2, \ldots, x_{\text{N}}, y_{\text{N}}]$,

$$x_{\text{i}} = \text{round}\left(\frac{\tilde{x}_{\text{i}}}{w} * n_{\text{bins}}\right), \quad y_{\text{i}} = \text{round}\left(\frac{\tilde{y}_{\text{i}}}{h} * n_{\text{bins}}\right) \quad (1)$$

where $(\tilde{x}_{\text{i}},\tilde{y}_{\text{i}})$ are the original floating point image coordinates, w, h are the width and height of the image, n_{bins} is the number of quantization bins, and $(x_{\text{i}},y_{\text{i}})$ are the quantized coordinates. However, as shown in the top-left of Figure 3a, the uniform sampling approach is unaware of the contour curvature and cannot properly represent sharp edges. To alleviate this limitation, we argue that the sampling should preserve more points where the contour has a sharp bend and less where it is almost straight. Based on this observation, we propose a gradient-aware adaptive sampling technique, which we describe in three steps:

- **Contour Discretization.** First, we discretize the continuous contour by uniformly sampling a high number (M) of dense points. Note that these dense points represent the curve well, but such a long sequence is infeasible for training a decoder.
- **Gradient Calculation.** Next, for every point $p_{\text{i}\in\{1,\ldots,M\}}$ on the curve, we draw two lines - l_1 by joining p_{i} with its previous point $p_{\text{i}-1}$, and l_2 by joining $p_{\text{i}-1}$ with the next point $p_{\text{i}+1}$. l_1 and l_2 create an angle θ_{i} $(0° \leq \theta_{\text{i}} < 180°)$ at $p_{\text{i}-1}$. If $\theta_{\text{i}} \simeq 0$, the contour is almost linear at p_{i}, and we can safely discard p_{i} (e.g., points B and D in the right column of Figure 3a). As θ_{i} increases, the curvature at p_{i} becomes sharper, and the importance of keeping p_{i} in the final sampling list increases (e.g., points A and C).
- **Sorting & Quantization:** Finally, we sort $\theta_{\text{i}\in\{1,\ldots,M\}}$ in descending order, and keep the N points $(N \ll M)$ corresponding to the N highest θ_{i}. These N points, which are then quantized (we use 1000 quantization bins, by default) and serialized as in Equation 1, denote the final sampled list.

The right column of Figure 3a depicts the adaptive sampling technique, which produces a better representation of sharp bends of the curve than uniform sampling, shown in the bottom-left of the same figure. We further illustrate the reconstruction from two techniques with a mask from the COCO dataset in Figure 3b, where the uniform sampling loses fine details of the zebra's legs, back, and ears. In contrast, adaptive sampling preserves the mask more precisely.

Both uniform and adaptive sampling techniques inevitably result in a certain amount of information loss from the original ground-truth masks, thereby imposing a constraint on the maximal performance achievable in segmentation tasks. Nonetheless, the extent of this loss is considerably reduced when employing the adaptive sampling approach. For instance, in the RefCOCO validation set for Referring Expression Segmentation (RES), uniform sampling of 32 points from the ground-truth masks yields an mIoU upper bound of 94.70, whereas adaptive sampling achieves 97.26. The superiority of adaptive sampling becomes even more pronounced in the case of complex geometric structures containing numerous sharp bends and intricate details. We delve deeper into the comparative efficacy of these two methods through ablation experiments in Section 5.4.

4. Coarse-to-fine Instruction-tuning Dataset

To train VistaLLM on a versatile form of vision and language tasks, we collect CoinIt (**Co**arse-to-f**i**n**e** **I**nstruction-

Dataset	Task	Corpus	Multi img?	Reg. level?	Input format	Output format	Metrics (%)
COCO [54]	Caption	Train, Eval	✗	✗	I	T	SPICE, CIDEr
	VQAv2	Train, Eval	✗	✗	I + Q	T	Accuracy
	REC	Train, Eval	✗	✓	I + R	B	Pr@0.5
	GREC	Train, Eval	✗	✓	I + R	M	Pr@0.5, N-acc
	RES	Train, Eval	✗	✓	I + R	B	mIoU
	GRES	Train, Eval	✗	✓	I + R	M	gIoU, N-acc, T-acc
	REG	Train	✗	✓	I + B	T	–
	AttCoSeg	Train	✓	✓	I	M	–
Flickr [74]	Spot Caption	Train	✗	✓	I	T + B	–
VG [37]	REG	Train	✗	✓	I + B	T	–
VCR [113]	Reasoning	Train, Eval	✗	✓	I + Q + B	T	Accuracy
LLaVa [57]	VQA	Train	✗	✗	I + Q	T	–
LT-QA [68]	BQA	Train, Eval	✗	✓	I + Q + B	B	Accuracy
V7W [122]	PQA	Train, Eval	✗	✓	I + Q + P	T	Accuracy
	BQA	Train, Eval	✗	✓	I + Q + B	T	Accuracy
TextVQA [83]	Reading comp.	Eval	✗	✓	I + Q	T	Accuracy
IconQA [65]	Reasoning	Eval	✓	✓	I + Q	T	Accuracy
HM [36]	Classification	Eval	✗	✗	I	T	Accuracy
POPE [51]	Hallucination	Eval	✗	✗	I + Q	Y/N	Prec., Recall, F1
NLVR [85, 86]	Reasoning	Train, Eval	✓	✗	I + Q	Y/N	Accuracy
PASCAL [19] iCoSeg [3] MSRC [101]	CoSeg	Train, Eval	✓	✓	I	M	Precision ($\mathcal{P}$), Jaccard Index ($\mathcal{J}$)

Table 1. **Training and evaluation datasets, input-output formats, and metrics.** To train VistaLLM on versatile form of vision and language tasks, we collect CoinIt, which is a unified set of 14 benchmarks. We quantitatively evaluate the trained model on 15 tasks without additional fine-tuning, among which TextVQA, IconQA, POPE, and HM contain unseen tasks during training, assessing the system's generalization capability. I: Image, T: General Text, Q: Question, R: Referring Expression, P: Point coordinate, B: Bounding Box, M: Segmentation Mask, Y/N: Yes or No.

tuning Dataset), which is a unified set of 14 benchmarks containing $6.8M$ samples, among which (i) 13 are publicly available which we convert to instruction-tuning format, and (ii) we construct a new benchmark, AttCoSeg (**Att**ribute-level **Co**-**Seg**mentation), to alleviate the lack of multi-image region-level datasets. We quantitatively evaluate the trained model on 15 benchmarks without additional fine-tuning. Notably, 4 of these 15 downstream contain entirely unseen tasks during training, helpful for assessing the system's generalization capability. To ensure data integrity, we confirm that no images from the validation or test sets appear during training, thus eliminating the risk of data leakage. We have grouped these diverse tasks into four main categories based on their input and output formats, summarized in Table 1:

- Single-image coarse-level tasks, such as visual question answering (VQA) and image captioning on COCO [54] and LLaVa [57] require global understanding of a single input image.
- Single-image region-level tasks, like generalized referring expression comprehension (GREC) [22] and segmentation (GRES) [56], spot captioning [8], visual commonsense reasoning (VCR) [113], box question answering (BQA) and point question answering (PQA) [68, 122] require fine-grained dense predictions over one input image. These tasks contain points, bounding boxes and segmentation masks in inputs and outputs.
- Multi-image coarse-level tasks, like natural language for visual reasoning (NLVR) [85, 86] and icon question answering (IconQA) [65] involve comprehending global perception across multiple input images.
- Multi-image region-level tasks, such as object-level co-segmentation (CoSeg) [40, 80] demands fine-grained reasoning and grounding on various input images.

Method	General-purpose?	VQAv2 Val	VQAv2 Dev	VQAv2 Std	COCO Cap. SPICE	COCO Cap. CIDEr
METER [16]	✗	–	76.4	76.4	23.0	128.2
FIBER [15]	✗	–	78.6	78.4	23.1	**128.4**
Unified-IO [64]	✓	–	77.9	–	–	122.3
Flamingo-80B [1]	✓	–	56.3	–	–	84.3
Shikra-13B [8]	✓	75.3	77.4	77.5	–	117.5
VistaLLM-13B	✓	**76.9**	**79.1**	**79.0**	**23.3**	**128.4**
$\Delta_{\text{Ours - Shikra-13B}}$	–	1.6 ↑	1.7 ↑	1.5 ↑	–	10.9 ↑

Table 2. **Performance on VQAv2 and COCO captioning.** VistaLLM yields significant gains over existing general-purpose and fine-tuned baselines. Reported captioning results of METER and FIBER are without CIDEr optimization [79].

AttCoSeg, newly proposed benchmark: Existing multi-image region-level object co-segmentation datasets [3, 19, 101] are small-scale and simple to solve. Hence, we argue that these datasets are insufficient to train VistaLLM to have generalized grounding ability over many input images, and we construct a more challenging larger-scale multi-image region-level dataset. We use Group-wise RES [103] annotations to sample high-quality images containing objects with similar fine-grained attributes (shape, color, size, position). We refer to such images as positives. While training VistaLLM, we input these positive image pairs, ask the model to segment the object with common traits in both of them. We name this task attribute-level co-segmentation (AttCoSeg), which contains over 804k training samples, and help VistaLLM to gain significant generalized reasoning and grounding ability over multiple input images. Notably, we do not collect new images or perform new annotations ourselves when constructing AttCoSeg. Detailed statistics of every dataset are given in the supplementary.

5. Experiments

5.1. Instruction Prompts

Carefully designed language instructions are crucial for general-purpose vision models on diverse tasks with different input-output formats [8, 95]. Since we address closely related tasks like REC, RES, GREC, GRES, we use detailed instructions. Figure 2 illustrates an example instruction for CoSeg. More example instructions are shown in supplementary. We use a special token <image>, which we later replace with the instruction-guided image features to generate interleaved image-text input to the LLM.

Moreover, the instruction must vary for different samples to support flexible user inputs. To generate high-quality instructions with minimal cost, we manually write one example description of each task and resort to GPT-3.5 [5] to create hundreds of variations. Next, we refine and ensure the quality of every instruction with GPT-4 [70]. During

Method	General-purpose?	Ref			Ref+			Refg	
		val	testA	testB	val	testA	testB	val	test
UniTAB [105]	✗	86.3	88.8	80.6	78.7	83.2	69.5	80.0	80.0
MDETR [35]	✗	86.8	89.6	81.4	79.5	84.1	70.6	81.6	80.9
SeqTR [120]	✗	83.7	86.5	81.2	71.5	76.3	64.9	74.9	74.2
OFA-L [93]	✓	80.0	83.7	76.4	68.3	76.0	61.8	67.6	67.6
VisionLLM-H [95]	✓	–	86.7	–	–	–	–	–	–
Shikra-13B [8]	✓	87.8	91.1	81.8	82.9	87.8	74.4	82.6	83.2
MiniGPT-v2 [7]	✓	88.7	91.7	**85.3**	80.0	85.1	74.5	84.4	84.7
Ferret-13B [111]	✓	89.5	92.4	84.4	82.8	88.1	75.2	85.8	86.3
VistaLLM-7B	✓	88.1	91.5	83.0	82.9	89.8	74.8	83.6	84.4
VistaLLM-13B	✓	**89.9**	**92.5**	85.0	**84.1**	**90.3**	**75.8**	**86.0**	**86.4**
$\Delta_{\text{Ours - Ferret-13B}}$	–	0.4 ↑	0.1 ↑	0.6 ↑	1.3 ↑	2.2 ↑	0.6 ↑	0.2 ↑	0.1 ↑

(a) **Performance on referring expression comprehension (REC).** VistaLLM yields better results than existing baselines across all splits.

Method	General-purpose?	Ref			Ref+			Refg	
		val	testA	testB	val	testA	testB	val	test
CGAN [66]	✗	64.9	68.0	62.1	51.0	55.5	44.1	51.0	51.7
VLT [55]	✗	65.7	68.3	62.7	55.5	59.2	49.4	53.0	56.7
LTS [34]	✗	65.4	67.8	63.1	54.2	58.3	48.0	54.4	54.3
CRIS [99]	✗	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4
SeqTR [120]	✗	71.7	73.3	69.8	63.0	66.7	59.0	64.7	65.7
RefTr [48]	✗	74.3	76.8	70.9	66.8	70.6	59.4	66.6	67.4
LAVT [107]	✗	74.5	76.9	70.9	65.8	71.0	59.2	63.3	63.6
PolyFormer [58]	✗	76.0	77.1	73.2	70.7	**74.5**	64.6	69.4	69.9
VistaLLM-7B	✓	74.5	76.0	72.7	69.1	73.7	64.0	69.0	70.9
VistaLLM-13B	✓	**77.2**	**78.7**	**73.9**	**71.8**	74.4	**65.6**	**69.8**	**71.9**
$\Delta_{\text{Ours - PolyFormer}}$	–	1.2 ↑	1.6 ↑	0.7 ↑	1.1 ↑	0.1 ↓	1.0 ↑	0.4 ↑	2.0 ↑

(b) **Performance on referring expression segmentation (RES).** VistaLLM is the first general-purpose model to unify RES.

Table 3. **Performance on (a) REC, and (b) RES.** While none other general-purpose systems can solve RES, VistaLLM sets a new state-of-the-art for both tasks across all splits.

Method	General-purpose?	GREC		Method	General-purpose?	GRES		
		Pr	N-acc.			gIoU	N-acc.	T-acc.
MCN [67]	✗	28.0	30.6	MattNet [112]	✗	48.2	41.2	96.1
VLT [55]	✗	36.6	35.2	VLT [55]	✗	52.0	47.2	95.7
MDETR [35]	✗	41.5	36.1	LAVT [107]	✗	58.4	49.3	96.2
VistaLLM-7B	✓	52.7	69.4	VistaLLM-7B	✓	64.4	68.8	96.6
VistaLLM-13B	✓	**54.6**	**70.8**	VistaLLM-13B	✓	**65.1**	**70.0**	**96.8**
$\Delta_{\text{Ours - MDETR}}$	–	13.1 ↑	34.7 ↑	$\Delta_{\text{Ours - LAVT}}$	–	6.7 ↑	20.7 ↑	0.6 ↑

Table 4. **Performance on generalized referring expression comprehension (GREC) and generalized referring expression segmentation (GRES).** VistaLLM is the first general-purpose system to address both tasks, and gains huge improvements over existing specialist models.

training, we randomly pick one instruction for each sample.

5.2. Implementation Details

We use EVA-CLIP [87] pre-trained on LAION-400M [82] and QFormer [45] pre-trained by InstructBLIP [14] as our visual encoder and instruction-guided image tokenizer. We feed the input images into EVA, which produces 256×1408 dimensional features for 224×224 images. The number of spatial tokens quadratically increases with the input image dimension. The Qformer has 12 encoder layers with 12 heads and outputs 32 queries per image with a hidden size of 768, thus working as an efficient feature compressor. For a fair comparison with existing general-purpose baselines [7, 8, 95, 111, 118], we use Vicuna7B and Vicuna13B [12] as the LLM. All other dense layers are initialized from scratch. For serializing the segmentation masks, we sample 32 points using the proposed adaptive sampling technique.

VistaLLM is trained in two stages. In the first stage, we only use the single-image datasets and do not introduce the instruction-guided image tokenizer. We freeze EVA and train the rest of the model end-to-end for 2 epochs. In the second stage, we only tune the image tokenizer on the multi-image datasets for 5 epochs. VistaLLM is trained using AdamW optimizer [61] and cosine scheduler [60] with linear warmup for the first 3% steps. We use a peak learning rate of $2e{-}5$ and a global batch size of 256. The model from the first stage is used to evaluate single-image datasets, whereas the model from the second stage is used to evaluate multi-image datasets. Training takes $2/3$ days for the first stage and $22/30$ hours for the second stage with $7/13$B models on 32 A100 GPUs, each having 80G memory.

Task	Method	LookTwice-QA			Task	Method	V7W
		Any	Super cls.	Object			
PQA	Mani et al. [68]	56.5	59.1	62.8		V7W [122]	56.1
	Shikra-13B [8]	70.0	70.2	71.9		CMNs [26]	72.5
	VistaLLM-13B	**71.1**	**71.2**	**72.5**		ViLBERT [63]	82.8
	$\Delta_{\text{Ours - Shikra-13B}}$	1.1 ↑	1.0 ↑	0.6 ↑	BQA	ViLBERT$_{\text{FT}}$ [63]	83.4
BQA	Mani et al. [68]	60.2	59.8	61.4		GPT4RoI-13B [118]	84.8
	Shikra-13B [8]	70.3	71.4	72.3		Shikra-13B [8]	85.3
	VistaLLM-13B	**71.4**	**72.5**	**73.0**		VistaLLM-13B	**85.5**
	$\Delta_{\text{Ours - Shikra-13B}}$	1.1 ↑	1.1 ↑	0.7 ↑		$\Delta_{\text{Ours - Shikra-13B}}$	0.2 ↑

Table 5. **Performance of point question answering (PQA) and box question answering (BQA) on LookTwice-QA and Visual-7W.** LookTwice-QA questions based on input point/box on three different level of referential clarity in the question, e.g. "How many of these [items/vehicles/cars] are there?" Visual-7W questions in 'which box' setting, i.e. choose one of the four bounding box options based on given query.

5.3. Main Results

We use **boldface** and underline for the best and second-best performing methods in every table and indicate the performance improvements over the state-of-the-art with Δ.

VQAv2 & COCO Captioning: Table 2 presents the performance on traditional single-image coarse-level visual question answering and image captioning tasks, which do not necessitate coordinates in the input or output. The input instructions for these tasks are straightforward, such as, "Please generate a simple description of the image <image>." or "Given the image <image>, can you please answer the question <question>", where <question> denotes the input query. On VQAv2, VistaLLM achieves 76.9%, 79.1%, and 79.0% accuracy on the val, dev, and std splits, improving the general-purpose state-of-the-art by over 1.5 points. On image captioning, VistaLLM yields a substantial gain of 10.9 CIDEr points over the best general-purpose baseline [8]. Our model performs on a par with fine-tuned specialist models, signifying the power of LLMs to comprehend and generate strong language descriptions.

REC, RES, GREC & GRES: Next, we evaluate VistaLLM on four single-image grounding tasks. Table 3 shows the

Method	Validation Acc. Q → A	QA → R	Q → AR
ViLBERT [62]	72.4	74.5	54.0
Unicoder-VL [43]	72.6	74.5	54.5
VLBERT [84]	75.5	77.9	58.9
VILLA [20]	78.5	82.6	65.2
GPT4RoI-7B [118]	87.4	89.6	78.6
VistaLLM-13B	**87.8**	**89.9**	**79.1**
$\Delta_{\text{Ours - GPT4RoI-7B}}$	0.4 ↑	0.3 ↑	0.5 ↑

(a) **Performance on visual commonsense reasoning (VCR).**

Method	Acc. TextVQA	IconQA	HM
BLIP-2 [45]	42.5	40.6	53.7
InstructBLIP [14]	50.7	44.8	57.5
MiniGPT-4 [121]	19.9	37.6	–
LLaVA [57]	38.9	43.0	–
MiniGPT-v2 [7]	51.9	47.7	58.2
VistaLLM-13B	**53.0**	**47.9**	**59.1**
$\Delta_{\text{Ours - MiniGPTv2}}$	1.1 ↑	0.2 ↑	0.9 ↑

(b) **Performance on novel tasks - TextVQA, IconQA, and HM.**

Table 6. **Results on (a) VCR, and (b) three novel tasks - TextVQA, IconQA, hateful memes (HM).** VistaLLM achieves consistent gains over existing baselines.

Method	PASCAL Av. $\mathcal{P}$	Av. $\mathcal{J}$	Method	MSRC Av. $\mathcal{P}$	Av. $\mathcal{J}$	Method	iCoSeg Av. $\mathcal{J}$
Quan et al. [77]	89.0	52.0	Rubinstein et al. [81]	92.2	74.7	Rubinstein et al. [81]	70.2
Jerripothula et al. [31]	80.1	40.0	Faktor et al. [19]	92.0	77.0	Faktor et al. [19]	73.8
Li et al. [39]	94.1	63.0	Chen et al. [6]	–	73.9	Jerripothula et al. [30]	70.4
Zhang et al. [116]	94.9	71.0	Li et al. [49]	95.4	82.9	Zhang et al. [116]	89.2
CycleSegNet [115]	96.8	73.6	CycleSegNet [115]	97.9	87.2	CycleSegNet [115]	92.1
VistaLLM-13B	**97.9**	**77.2**	VistaLLM-13B	**98.5**	**90.1**	VistaLLM-13B	**95.1**
$\Delta_{\text{Ours - CycleSegNet}}$	1.1 ↑	3.6 ↑	$\Delta_{\text{Ours - CycleSegNet}}$	0.6 ↑	2.9 ↑	$\Delta_{\text{Ours - CycleSegNet}}$	3.0 ↑

Table 7. **Performance on object co-segmentation (CoSeg) on three datasets - PASCAL, MSRC, and iCoSeg.** VistaLLM is the first general-purpose system to address CoSeg and sets a new set-of-the-art across all datasets, beating previous specialist models.

results of referring expression comprehension (REC) and referring expression segmentation (RES), which aims to ground (detect and segment, respectively) one object in the image described by an input expression. Our model shows promising performance on REC, improving over existing baselines across all evaluation splits. VistaLLM is the first general-purpose system to report results on RES, where we perform as good as fine-tuned specialist models. Such strong results on grounding tasks can be attributed to refined image features, effective sampling techniques, and detailed input instructions. We also evaluate VistaLLM on GREC & GRES, where the output can contain zero, one, or multiple boxes and masks. As shown in Table 4, besides generating high-quality boxes and masks, our model yields an impressive gain of 34.7% and 20.7% N-acc scores over MDETR [35], reflecting the ability of VistaLLM to detect samples without any matching objects in the image.

PQA & BQA: Table 5 shows our performance on point question answering (PQA) and box question answering (BQA), which can have coordinate points and bounding boxes as input and output. LookTwice-QA asks the model to answer a question about a specified region, either mentioning a point or a box. The system needs to comprehend the area in the context of the whole image, e.g., "How many of these [cars] are there in the image?" Visual-7W contains MCQs where the model needs to choose a box from four options. VistaLLM sets new state-of-the-art on both tasks, proving its mighty region-referring ability.

VCR & Novel (Unseen) Tasks: Table 6a shows results on visual commonsense reasoning (VCR) - a single-image fine-grained reasoning task containing questions with referring bounding boxes. VistaLLM produces 0.5% improvement over GPT4RoI [118] in the most challenging $Q \rightarrow AR$ setting. We also access our model's generalization ability by evaluating it on three novel tasks in Table 6b - TextVQA, IconQA, and hateful memes (HM). VistaLLM achieves strong results on all three benchmarks, proving its ability to comprehend novel tasks given well-designed instructions.

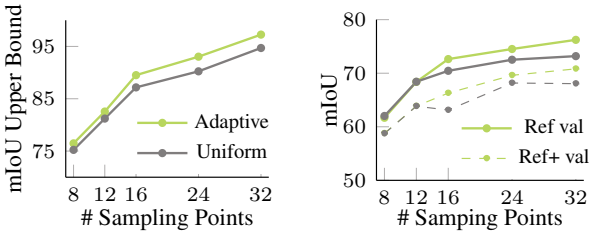
Method	General-purpose?	NLVR dev	test-P
VisualBERT [46]	✗	67.4	67.0
SOHO [27]	✗	76.3	77.3
Oscar [50]	✗	78.1	78.4
Uniter [11]	✗	77.2	77.9
VILLA [20]	✗	78.4	79.3
ALBEF [44]	✗	80.2	80.5
VistaLLM-13B	✓	**80.8**	**81.3**
$\Delta_{\text{Ours - ALBEF}}$	–	0.6 ↑	0.8 ↑

(a) **Results on NLVR.**

Method	R F1	P F1	A F1
mPLUG-Owl	68.4	66.9	66.8
LLaVA [57]	66.6	66.4	66.3
MiniGPT4 [121]	80.2	73.0	70.4
InstructBLIP [14]	89.3	84.7	77.3
Shikra-7B [8]	86.2	83.2	82.5
Ferret-13B [111]	89.8	84.2	82.0
VistaLLM-13B	**90.5**	**84.8**	**82.9**
$\Delta_{\text{Ours - Ferret-13B}}$	0.7 ↑	0.6 ↑	0.9 ↑

(b) **Results on POPE.**

Table 8. **Performance on (a) NLVR, and (b) object hallucination benchmark using POPE evaluation pipeline**. VistaLLM is the first general-purpose model to address NLVR, and beats strong fine-tuned models. VistaLLM demonstrates an intriguing property of alleviating object hallucinations across all three splits. R: Random, P: Popular, A: Adversarial.



(a) **mIoU upper bound** on Ref val set with varying number of points. (b) **mIoU by VistaLLM** on Ref, Ref+ with varying number of points.

Figure 4. **Ablative experiments on RES task.** (a) Comparison of the highest possible mIoU by adaptive and uniform sampling, indicating lesser information loss in adaptive sampling, (b) Effect of number of sampled points on the performance of VistaLLM.

CoSeg & NLVR: Table 7 and Table 8a shows the performance on two multi-image tasks, CoSeg and NLVR. VistaLLM is the first general-purpose model to evaluate both tasks. Given a group of images with a common object, CoSeg aims to recognize and segment the object in every photo. VistaLLM outperforms existing specialist baselines across three different datasets on CoSeg, showing its strong perception and grounding ability. VistaLLM also beats powerful fine-tuned models [20, 44] on NLVR, which aim to reason two input images and answer a query. These results prove the versatility of VistaLLM with more than one input image, which is crucial for real-world use cases.

POPE: We evaluate VistaLLM on POPE object hallucination benchmark in Table 8b, where we perform comparably to strong general-purpose models like Shikra [8], and Ferret [111] across all metrics and splits, and vastly outperform many previous baselines. These results exhibit our model's

Figure 5. **Examples demonstrating VistaLLM's capability for single and multi-image reasoning and grounding tasks.** More visualizations are shown in supplementary. Best viewed when zoomed in and in color.

Method	iCoSeg Av. $\mathcal{J}$	NLVR dev
VistaLLM-13B	**95.1**	**80.8**
w/o Tokenizer	89.7	77.3
w/o Tokenizer PT	94.8	79.5

Table 9. **Ablation on instruction-guided image tokenizer**, which refines global image embeddings.

ability to power against the hallucination problem, essential for its generalized applicability.

5.4. Ablation Study

Adaptive vs. Uniform Sampling: We ablate the quantitative effectiveness of our proposed adaptive sampling method compared to uniform sampling for referring expression segmentation (RES) in Figure 4. With 32 sampled points, the maximum achievable mIoU score on Ref val set by adaptive technique is 97.26, while for uniform sampling, 94.70. However, with fewer sampling points, both methods perform significantly worse. Figure 4b shows that the performance of VistaLLM also improves using adaptive sampling on both Ref and Ref+ val splits, which shows the usefulness of the proposed sampling scheme.

Number of Sampled Points: Figure 4b shows that with a higher number of sampled points, the performance of VistaLLM significantly improves for both Ref and Ref+. When increasing the number of points from 16 to 32, VistaLLM gains 3.6 on Ref and 4.5 on Ref+.

Instruction-guided Image Tokenizer: We ablate the importance of the proposed instruction-guided tokenizer in Table 9. The performance of iCoSeg significantly drops by 5.4 $\mathcal{J}$-index without the tokenizer module. We also see similar effects in captioning, RES, VCR, and NLVR. When using QFormer without pre-trained weights, we observe a substantial drop in all tasks except iCoSeg.

LLM Size: Table 2, 3, 4 shows that larger LLM backbone generally helps improve the performance. We show ablation on the training dataset and image encoder in supplementary.

5.5. Qualitative Results and Error Analysis

Figure 5 visualizes sample results from VistaLLM for single and multi-image reasoning and grounding tasks. As shown in the NLVR and AttCoSeg examples, VistaLLM can successfully parse all input images and comprehend the relation among them. It can also successfully ground all referred objects in foreground and background, as shown in GRES. However, compared to the recently released GPT-4V [108], we perform worse in general and knowledge-based question answering, which can be attributed to the billion scale pre-training of GPT. Nevertheless, VistaLLM's ability to reason over several images and perform precise detection and segmentation makes it unique.

6. Conclusion

We introduce VistaLLM, a powerful general-purpose vision system that integrates coarse- and fine-grained vision-language reasoning and grounding tasks over single and multiple input images into a unified framework. To filter embeddings from various images, VistaLLM uses a language-guided image tokenizer, which provides compressed and refined features following the task description. We also employ a gradient-aware adaptive sampling technique to efficiently represent binary segmentation masks as sequences, significantly improving previously used uniform sampling. We conduct extensive experiments to show the effectiveness of VistaLLM on a wide range of downstream tasks, consistently achieving state-of-the-art performance.

7. Acknowledgement

The codebase for this work is built on the LLaVA [57] and Shikra [8] repository. Shraman and Rama were partially supported by a ONR MURI grant N00014-20-1-2787.

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. In *NeurIPS*, pages 23716–23736, 2022. 2, 5

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *CVPR*, pages 2425–2433, 2015. 2

[3] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *CVPR*, pages 3169–3176, 2010. 5

[4] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *ICML*, pages 2206–2240. PMLR, 2022. 2

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. 3, 5

[6] Hong Chen, Yifei Huang, and Hideki Nakayama. Semantic aware attention based deep object co-segmentation. In *ACCV*, pages 435–450. Springer, 2018. 7

[7] 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. 2, 3, 4, 6, 7

[8] 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. 2, 3, 4, 5, 6, 7, 8

[9] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *ICLR*, 2021. 2, 4

[10] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. *NeurIPS*, 35:31333–31346, 2022. 2, 4

[11] 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 *ECCV*, pages 104–120. Springer, 2020. 7

[12] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023. 3, 6

[13] 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. 1, 2

[14] 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. In *NeurIPS*, 2023. 1, 2, 6, 7

[15] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *NeurIPS*, 35:32942–32956, 2022. 1, 5

[16] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *CVPR*, pages 18166–18176, 2022. 1, 5

[17] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *ICML*, pages 5547–5569. PMLR, 2022. 2

[18] Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. Magma–multimodal augmentation of generative models through adapter-based finetuning. In *Findings of EMNLP*, pages 2416–2428, 2022. 2

[19] Alon Faktor and Michal Irani. Co-segmentation by composition. In *ICCV*, pages 1297–1304, 2013. 2, 5, 7

[20] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, pages 6616–6628, 2020. 7

[21] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 1, 2

[22] Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. Grec: Generalized referring expression comprehension. *arXiv preprint arXiv:2308.16182*, 2023. 5

[23] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. 2019. 2

[24] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An empirical analysis of compute-optimal large language model training. *NeurIPS*, 35:30016–30030, 2022. 1, 2

[25] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019. 2

[26] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *CVPR*, pages 1115–1124, 2017. 6

[27] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*, pages 12976–12985, 2021. 7

[28] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709, 2019. 2

[29] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022. 2

[30] Koteswar Rao Jerripothula, Jianfei Cai, and Junsong Yuan. Image co-segmentation via saliency co-fusion. *IEEE TMM*, 18(9):1896–1909, 2016. 7

[31] Koteswar Rao Jerripothula, Jianfei Cai, Jiangbo Lu, and Junsong Yuan. Object co-skeletonization with co-segmentation. In *CVPR*, pages 3881–3889. IEEE, 2017. 7

[32] 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. *arXiv preprint arXiv:2310.08825*, 2023. 2

[33] 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. *arXiv preprint arXiv:2310.08825*, 2023. 2

[34] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *CVPR*, pages 9858–9867, 2021. 6

[35] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, pages 1780–1790, 2021. 2, 6, 7

[36] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*, pages 2611–2624, 2020. 5

[37] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017. 2, 5

[38] 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. 3

[39] Bo Li, Zhengxing Sun, Qian Li, Yunjie Wu, and Anqi Hu. Group-wise deep object co-segmentation with co-attention recurrent neural network. In *CVPR*, pages 8519–8528, 2019. 7

[40] Bo Li, Lv Tang, Senyun Kuang, Mofei Song, and Shouhong Ding. Toward stable co-saliency detection and object co-segmentation. *IEEE TIP*, 31:6532–6547, 2022. 5

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

[42] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *NeurIPS*, 2023. 2

[43] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020. 7

[44] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, pages 9694–9705, 2021. 7

[45] 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. 1, 2, 3, 6, 7

[46] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 7

[47] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, pages 10965–10975, 2022. 1

[48] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *NeurIPS*, 34:19652–19664, 2021. 6

[49] Weihao Li, Omid Hosseini Jafari, and Carsten Rother. Deep object co-segmentation. In *ACCV*, pages 638–653. Springer, 2019. 7

[50] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, pages 121–137. Springer, 2020. 7

[51] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023. 5

[52] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *CVPR*, pages 23390–23400, 2023. 1

[53] Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. Univtg: Towards unified video-language temporal grounding. In *ICCV*, pages 2794–2804, 2023. 1

[54] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 2, 5

[55] Chang Liu, Xudong Jiang, and Henghui Ding. Instance-specific feature propagation for referring segmentation. *IEEE TMM*, 2022. 6

[56] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *CVPR*, pages 23592–23601, 2023. 5

[57] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2, 3, 5, 7, 8

[58] 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 *CVPR*, pages 18653–18663, 2023. 2, 4, 6

[59] Zhaoyang Liu, Yinan He, Wenhai Wang, Weiyun Wang, Yi Wang, Shoufa Chen, Qinglong Zhang, Yang Yang, Qingyun Li, Jiashuo Yu, et al. Internchat: Solving vision-centric tasks by interacting with chatbots beyond language. *arXiv preprint arXiv:2305.05662*, 2023. 3

[60] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 6

[61] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[62] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 32, 2019. 7

[63] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, pages 10437–10446, 2020. 6

[64] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *ICLR*, 2022. 5

[65] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *NeurIPS Datasets and Benchmarks Track*, 2021. 2, 5

[66] Gen Luo, Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Chia-Wen Lin, and Qi Tian. Cascade grouped attention network for referring expression segmentation. In *ACM MM*, pages 1274–1282, 2020. 6

[67] 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 *CVPR*, pages 10034–10043, 2020. 6

[68] Arjun Mani, Nobline Yoo, Will Hinthorn, and Olga Russakovsky. Point and ask: Incorporating pointing into visual question answering. *arXiv preprint arXiv:2011.13681*, 2020. 5, 6

[69] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2

[70] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[71] TB OpenAI. Chatgpt: Optimizing language models for dialogue. openai, 2022. 1, 2

[72] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023. 1, 2

[73] 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. 2, 3, 4

[74] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *CVPR*, pages 2641–2649, 2015. 2, 5

[75] Shraman Pramanick, Li Jing, Sayan Nag, Jiachen Zhu, Hardik Shah, Yann LeCun, and Rama Chellappa. Volta: Vision-language transformer with weakly-supervised local-feature alignment. In *TMLR*, 2023. 1

[76] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *ICCV*, pages 5285–5297, 2023. 1

[77] Rong Quan, Junwei Han, Dingwen Zhang, and Feiping Nie. Object co-segmentation via graph optimized-flexible manifold ranking. In *CVPR*, pages 687–695, 2016. 7

[78] 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 *ICML*, pages 8748–8763. PMLR, 2021. 1

[79] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, pages 7008–7024, 2017. 5

[80] Carsten Rother, Tom Minka, Andrew Blake, and Vladimir Kolmogorov. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In *CVPR*, pages 993–1000, 2006. 5

[81] Michael Rubinstein, Armand Joulin, Johannes Kopf, and Ce Liu. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*. 7

[82] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 6

[83] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. 2, 5

[84] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2019. 7

[85] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *ACL*, pages 217–223, 2017. 5

[86] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, pages 6418–6428, 2019. 5

[87] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 3, 6

[88] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022. 2

[89] 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. 1, 3

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

[91] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *NeurIPS*, 34:200–212, 2021. 2

[92] Sanh Victor, Webson Albert, Raffel Colin, Bach Stephen, Sutawika Lintang, Alyafeai Zaid, Chaffin Antoine, Stiegler Arnaud, Raja Arun, Dey Manan, et al. Multitask prompted training enables zero-shot task generalization. In *ICLR*, 2022. 2

[93] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, pages 23318–23340. PMLR, 2022. 6

[94] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *CVPR*, pages 19175–19186, 2023. 1

[95] 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. In *NeurIPS*, 2023. 2, 5, 6

[96] 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. 2

[97] 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. 2

[98] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. *NeurIPS*, 35:8483–8497, 2022. 2

[99] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, pages 11686–11695, 2022. 6

[100] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *ICLR*, 2021. 2

[101] John Winn, Antonio Criminisi, and Thomas Minka. Object categorization by learned universal visual dictionary. In *ICCV*, pages 1800–1807, 2005. 5

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

[103] Yixuan Wu, Zhao Zhang, Chi Xie, Feng Zhu, and Rui Zhao. Advancing referring expression segmentation beyond single image. In *ICCV*, pages 2628–2638, 2023. 5

[104] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *NeurIPS*, 35:124–141, 2022. 2

[105] 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 *ECCV*, pages 521–539. Springer, 2022. 1, 6

[106] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI*, pages 3081–3089, 2022. 2

[107] 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*, pages 18155–18165, 2022. 6

[108] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9, 2023. 8

[109] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 2

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

[111] 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 *ICLR*, 2024. 2, 4, 6, 7

[112] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, pages 1307–1315, 2018. 6

[113] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731, 2019. 2, 5

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

[115] Chi Zhang, Guankai Li, Guosheng Lin, Qingyao Wu, and Rui Yao. Cyclesegnet: Object co-segmentation with cycle refinement and region correspondence. *IEEE TIP*, 30: 5652–5664, 2021. 2, 7

[116] Kaihua Zhang, Jin Chen, Bo Liu, and Qingshan Liu. Deep object co-segmentation via spatial-semantic network modulation. In *AAAI*, pages 12813–12820, 2020. 7

[117] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 2

[118] 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. 2, 4, 6, 7

[119] 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

[120] 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 *ECCV*, pages 598–615. Springer, 2022. 2, 4, 6

[121] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2024. 1, 2, 7

[122] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, pages 4995–5004, 2016. 5, 6