GSVA: Generalized Segmentation via Multimodal Large Language Models

Zhuofan Xia* Dongchen Han* Yizeng Han Xuran Pan Shiji Song Gao Huang†
Department of Automation, BNRist, Tsinghua University

Abstract

Generalized Referring Expression Segmentation (GRES) extends the scope of classic RES to refer to multiple objects in one expression or identify the empty targets absent in the image. GRES poses challenges in modeling the complex spatial relationships of the instances in the image and identifying non-existing referents. Multimodal Large Language Models (MLLMs) have recently shown tremendous progress in these complicated vision-language tasks. Connecting Large Language Models (LLMs) and vision models, MLLMs are proficient in understanding contexts with visual inputs. Among them, LISA, as a representative, adopts a special [SEG] token to prompt a segmentation mask decoder, e.g., SAM, to enable MLLMs in the RES task. However, existing solutions to GRES remain unsatisfactory since current segmentation MLLMs cannot correctly handle the cases where users might reference multiple subjects in a singular prompt or provide descriptions incongruent with any image target. In this paper, we propose Generalized Segmentation Vision Assistant (GSVA) to address this gap. Specifically, GSVA reuses the [SEG] token to prompt the segmentation model towards supporting multiple mask references simultaneously and innovatively learns to generate a [REJ] token to reject the null targets explicitly. Experiments validate GSVA's efficacy in resolving the GRES issue, marking a notable enhancement and setting a new record on the GRES benchmark gRefCOCO dataset. GSVA also proves effective across various classic referring segmentation and comprehension tasks. Code is available at https://github.com/LeapLabTHU/GSVA.

1. Introduction

Referring Expression Segmentation (RES) [5, 25] is an emerging vision-language (VL) task predicting the masks of the interested objects referred to in the language expression. RES has great potential in many areas, especially embodied AI [17, 28, 54, 58, 66], including VL navigation, human-robot interaction, *etc.* Nevertheless, the simplification in RES formulation that one referring expression must match an individual object in the image [30] has left a gap between current RES algorithms and real-world applications, neglecting multiple-target and empty-target cases.

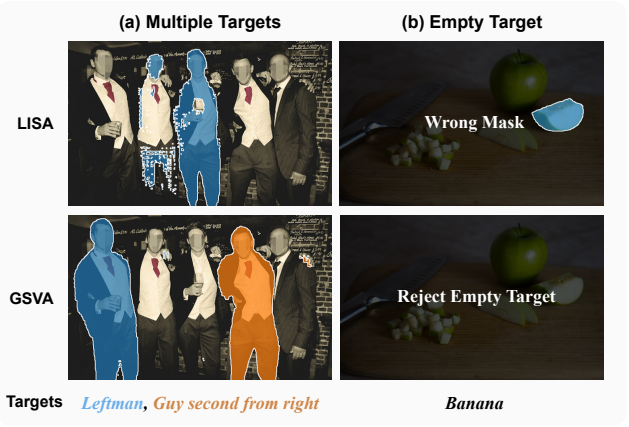


Figure 1. Comparison of the segmentation masks by LISA [32] and GSVA, facing the challenges in Generalized Referring Expression Segmentation (GRES) [38]. (a) LISA fails to segment the correct targets when multiple targets are requested due to the single [SEG] token restriction. GSVA successfully generates all target masks via learning multiple [SEG] tokens. (b) When the referent does not exist in the image, *i.e.*, the empty target is requested, LISA reluctantly produces the wrong mask because of the compulsive [SEG] token output. In contrast, GSVA can reject the empty targets by predicting [REJ] tokens in the output sequence.

To bridge this gap, Generalized Referring Expression Segmentation (GRES) [38] has recently been proposed to support multiple-target and empty-target cases. Practically, users refer to multiple subjects within a single instruction or provide descriptions that do not correspond to any targets in the image. As an example shown in Figure 1 (a), the left man and the second man from the right are targeted simultaneously, while a banana is referred to in a scene of apples in Figure 1 (b). RES takes no account of these cases, which GRES handles. In addition to the multimodal correspondences between images and text prompts in classic RES, GRES poses new challenges in handling more complicated multiple-target and empty-target cases. Therefore, the models must handle complex spatial relationships of the instances in the image [38] to segment the targets at various

*Equal contribution.
†Corresponding author.

locations and reject the empty targets in the wrong places.

The recent blooming Multimodal Large Language Models (MLLMs) [1, 34, 39, 83, 87] meet the requirements of GRES since they show excellency in complex reasoning [8] and instruction following [49] with visual inputs by aligning the LLMs [2, 6, 48, 60–62] and Visual Foundation Models (VFMs) [15, 55, 63, 79] which are typically various Vision Transformers [14, 18, 23, 50, 51, 76, 78] to perceive image or video inputs. To support segmentation output, many works [74, 75, 86] link an MLLM (*e.g.*, LLaVA [39]) and a segmentation model (*e.g.*, SAM [31]) by prompting the decoder with special token embeddings (*e.g.*, [SEG] in LISA [32]) to generate masks of the referents in the user's instructions. Although these models manage to handle RES, GRES is still beyond their reach. As shown in Figure 1, LISA fails to work well in GRES where the multiple-target and empty-target challenges remain uncharted.

To address the above challenges, we propose **Generalized Segmentation Vision Assistant (GSVA)**. We attribute the vulnerability of other segmentation MLLMs in GRES to the single constant [SEG] token that restricts its flexibility. Therefore, we present two pivotal designs in GSVA: (1) learning to predict multiple [SEG] tokens to segment multiple targets; (2) rejecting empty targets in referring expressions by predicting [REJ] tokens. Specifically, when multiple targets are requested in the referring expression, we place multiple weight-sharing [SEG] tokens corresponding to the entities in the expression, encouraging the MLLM to learn to output multiple [SEG] tokens. To distinguish each [SEG] token and avoid ambiguity, we add the expression of each entity in front of the corresponding [SEG] token, hinting each [SEG] token to focus on the specific target, which can be regarded as implicit In-Context Learning, and dynamic neural network [19]. Meanwhile, if the referents are absent in the image, the corresponding [SEG] tokens after the prompts are altered to [REJ] tokens to identify empty targets. The predicted [REJ] tokens are directly assigned with empty masks without decoded, which liberates the segmentation model from seeking non-existing targets in the image. This Benefiting from these novel designs, GSVA takes a big step forward in addressing GRES challenges, as shown in the second row of Figure 1.

Our contributions are summarized as follows: (1) We propose GSVA to solve the GRES problem with MLLM by handling the spatial relationships among targets, and study the GRES problem systematically in the context of LLM for the first time. (2) We propose the non-trivial **shared-weight multiple [SEG] tokens guided by each referent prompt** to address the multiple-target problem. (3) We firstly propose a clean solution, **the [REJ] token**, to reject the empty targets, which can be seamlessly applied to various models. (4) GSVA is intuitive and effective, **achieving state-of-the-art performance on the GRES benchmarks**.

2. Related Works

RES and GRES. Referring Expression Segmentation (RES) [4, 5, 25, 44] assumes that one expression matches one existing target, and many works explore fusing image and language [3, 16, 35, 37, 81] to segment objects under instructions. Currently, most RES methods adopt the cross-attention module or cross-modal alignment to bridge the modality gap [7, 57, 73, 82, 89]. Another line of research enables the text prompts for segmentation model with a unified decoder [31, 41, 93, 94], offering more flexible outputs. To break the jail for the arbitrary number of targets, DMMI [26] focuses on the one-to-many setting where text expression refers to varying numbers of targets.ReLA [38] proposes the Generalized Referring Expression Segmentation (GRES) task, supporting both the multi-target and empty-target scenarios, which is our main research scope.

MLLM. Multimodal Large Language Models (MLLMs) align the vision and language modalities by various techniques, including cross-attention module [1], prompt tuning tokens [87], Q-Former [9, 34], linear projection layers [39], and unified model architectures [52, 79]. Endowed with unprecedented capabilities in reasoning with world-knowledge and following instructions of users, MLLM shows extraordinary performances in various vision language tasks [43, 65]. Equipped with the segmentation decoders or detection heads, MLLMs can also excel in the vision-centric tasks, such as object detection and segmentation [36, 56, 64, 80, 86]. Among them, LISA [32] makes the most of the reasoning ability with a [SEG] token to address the Reasoning Segmentation problem. However, LISA fails to tackle the challenge in GRES due to the inflexible [SEG] token, which is addressed by our proposed GSVA.

Dynamic Networks. Dynamic neural networks [19] can adapt their architectures [20–22, 24, 27, 67–70] or parameters [11, 12, 53] to different inputs or switch the computation architecture in adjustment to different different time steps [46, 47, 71, 77], in order to achieve better accuracy and efficiency. In GSVA, the weight-sharing [SEG] tokens adapt to multiple targets under the hint of the prepended target prompts and dynamically reject empty targets with an individual prediction of rejection tokens.

3. Generalized Segmentation Vision Assistant

In this section, we initiate with the introduction of the model design of **Generalized Segmentation Vision Assistant** (**GSVA**), which is followed by the analysis of some certain limitations of LISA in Generalized Referring Segmentation (GRES). Subsequently, we delve into the introduction of two pivotal elements of GSVA, segmenting multiple targets and learning the rejection token. These components are fundamental in the conceptualization and construction of GSVA.

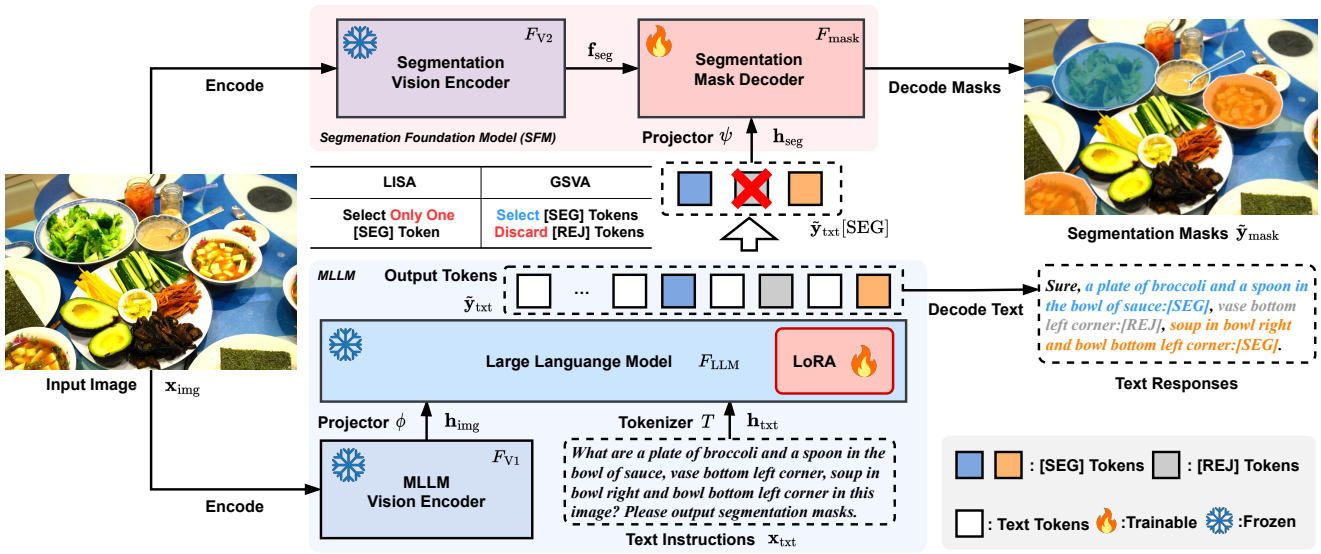


Figure 2. Overview of GSVA. At the bottom of the figure, the MLLM encodes the input image and concatenates the tokenized text tokens to follow instructions. GSVA generates multiple [SEG] tokens to handle multiple referred targets and rejects the objects absent in the image through [REJ] tokens. At the top of the figure, the SFM also encodes the image for segmentation and selects all [SEG] tokens in the output sequence to prompt the mask decoder to segment the target objects referred to in the instructions.

3.1. Model Architecture

The architecture of GSVA is illustrated in Figure 2, resembling LISA [32], which enables high-fidelity segmentation outputs by integrating two types of foundation models: (1) Multimodal Large Language Model (MLLM) as an aligned vision-language cognitive module; (2) Segmentation Foundation Model (SFM) to segment the target out of the input image based on user's instruction. To connect these two modules, LISA proposes a paradigm named *embedding as mask* where an extra [SEG] token is appended to the vocabularies of the MLLM and serves as the prompt of the SFM to segment the target following the intention of the user.

Multimodal Large Language Model. The MLLM consists of a decoder-based language model F_{LLM} to autoregressively generate text responses following the user's inputs, a vision encoder F_{V1} to extract features from the input image, and a linear projector ϕ to align the representations between image and text modalities. Specifically, the pretrained LLaVA [39] variants with CLIP-ViT-L/14 [55] and Vicuna-7B/13B [6] are employed. Given an input image $\mathbf{x}_{\text{img}}$, the vision encoder F_{V1} first encodes it into image features, and then the projector ϕ maps the features into the visual token embeddings in the LLM input space:

$$\mathbf{h}_{\text{img}} = \phi(F_{\text{V1}}(\mathbf{x}_{\text{img}})), \tag{1}$$

where the input image $\mathbf{x}_{\text{img}}$ is typically resized to $h \times w \times 3$, and the image tokens $\mathbf{h}_{\text{img}} \in \mathbb{R}^{n_{\text{img}} \times d}$ is aligned with the language modality. For CLIP-ViT-L/14, the input image with $h = w = 224$ is encoded with ViT of patch size in 14, therefore the length of tokens $n_{\text{img}} = hw/14^2 = 256$, and the LLM dimensions d are 4096 and 5120 for Vicuna-7B/13B, respectively. Along with the input image, the text instructions describing the targets to segment are tokenized into text tokens by the LLM tokenizer T:

$$\mathbf{h}_{\text{txt}} = T(\mathbf{x}_{\text{txt}}). \tag{2}$$

The image tokens and text tokens are concatenated together and then fed into the LLM after prepending a sequence of fixed prompt tokens $\mathbf{h}_{\text{prompt}}$ (omitted in the figure). The output token embeddings $\tilde{\mathbf{y}}_{\text{txt}}$ are generated autoregressively:

$$\tilde{\mathbf{y}}_{\text{txt}} = F_{\text{LLM}}\left([\mathbf{h}_{\text{prompt}}||\mathbf{h}_{\text{img}}||\mathbf{h}_{\text{txt}}]\right), \tag{3}$$

where $||$ is concatenation operation. The text responses are obtained from $\tilde{\mathbf{y}}_{\text{txt}}$ by applying a linear classifier to predict the next words in the vocabulary.

In LISA, a special token [SEG] is appended in the vocabulary to activate the segmentation ability of MLLM. The model learns to predict [SEG] token in the output sequence to indicate there is a target to segment. LISA then selects the output embedding of the [SEG] token $\tilde{\mathbf{y}}_{\text{txt}}[\text{SEG}]$, and projects it into the prompt space of the SFM by an MLP projector ψ:

$$\mathbf{h}_{\text{seg}} = \psi\left(\tilde{\mathbf{y}}_{\text{txt}}[\text{SEG}]\right). \tag{4}$$

The segmentation model is hence ready to decode the target mask from the query token $\mathbf{h}_{\text{seg}}$.

Segmentation Foundation Model. The SFM is a query-based segmentation model, where a frozen vision encoder F_{V2} takes in images of higher resolution than the vision encoder in MLLM to keep more details, followed by a trainable mask decoder F_{mask} to decode masks from the queries. The pretrained SAM [31] with ViT-H backbone is instantiated as the SFM to produce high-quality masks. The given

input image $\mathbf{x}_{\text{img}}$ is resized to a larger resolution $H \times W \times 3$, with $H = W = 1024$ and encoded with F_{V2} to extract features $\mathbf{f}_{\text{seg}}$ for segmentation:

$$\mathbf{f}_{\text{seg}} = F_{\text{V2}}(\mathbf{x}_{\text{img}}). \tag{5}$$

Condition on the features $\mathbf{f}_{\text{seg}} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C}$ with $C = 256$ for SAM, the mask decoder F_{mask} decodes the segmentation masks from the segmentation queries $\mathbf{h}_{\text{seg}} \in \mathbb{R}^{N_{\text{seg}} \times C}$:

$$\tilde{\mathbf{y}}_{\text{mask}} = F_{\text{mask}}(\mathbf{h}_{\text{seg}} | \mathbf{f}_{\text{seg}}), \tag{6}$$

where each query in $\mathbf{h}_{\text{seg}}$ corresponds to one segmentation mask in $\tilde{\mathbf{y}}_{\text{mask}}$.

LISA assumes that only one target exists to segment in the input image and its corresponding instructions. However, in GSVA, we extend it to a new scenario with multiple targets and empty targets, including multiple [SEG] tokens to invoke segmentation and [REJ] tokens to reject unreasonable instructed targets absent in the image. As shown in Figure 2, GSVA supports multiple [SEG]/[REJ] tokens in the output sequence and selects all the [SEG] tokens and discards every [REJ] token after attaining $\tilde{\mathbf{y}}_{\text{txt}}$ in Eq. (3). Therefore, there is more than one query in $\mathbf{h}_{\text{seg}}$, thus enabling the SFM to segment multiple targets. These designs make GSVA competent in the GRES task, where the awareness of multiple and empty targets is of vital importance.

3.2. GRES: Task and Challenges

Task. Generalized Referring Expression Segmentation (GRES) [38] removes the constraint on the number of referred targets in the expression in the conventional Referring Expression Segmentation (RES) [30, 44, 84]. Different from that one expression only refers to one instance or region in RES, GRES allows arbitrary numbers of referred targets, including multiple instances or no target circumstances. In GRES, the user can refer to many instances simultaneously or include the objects that do not exist in the image. For instance, there are three referring expressions in Figure 2, including *a plate of broccoli and a spoon in the bowl of sauce*, *vase bottom left corner*, and *soup in bowl right and bowl bottom left corner*. In the GRES case, the model ought to segment the masks of the objects referred to in the 1st and 3rd expression, meanwhile producing an empty mask for the 2nd expressions since there are no vases present at the bottom left corner.

Challenges. The challenges in GRES are common in practice, especially in embodied AI [17, 54, 58, 66]. The one challenge is **multiple targets**. Take robot navigation and planning [4, 45, 66] as an example, a robot may be asked to perceive multiple targets in the surrounding environment, e.g., to *bring the two bowls of soup* in Figure 2. The vision system of the robot needs to locate and segment the containers holding the referred food one at a time. The other challenge is **empty target**. Suppose the robot is ordered to cut an apple with a knife in the scene of Figure 2, whereas no apple is in the view of the camera, the vision system has to identify that the referred object is not in the scene. If it relies on some conventional RES methods which assume the expression must match something in the image, the output of the vision system could be undefined and potentially dangerous in some real-world cases.

Differences from ReasonSeg. Reasoning Segmentation (ReasonSeg) proposed by LISA [32] emphasizes the complex text instructions in RES. In ReasonSeg, the instructions are more implicit and sophisticated, forcing models to reason using world knowledge. Besides, the logic chain is usually longer and more challenging in ReasonSeg, which requires the model to deduce the final target object referred to in the image. In contrast, GRES increases complexity in another dimension by involving complicated spatial relationships. Hence, the model has to learn to handle this spatial information and understand the relationships between the instances. To meet these requirements, LISA tunes MLLMs with complex instructions paired with masks, while GSVA arouses the spatial modeling capabilities of MLLMs by learning multiple targets and rejecting empty targets.

3.3. Multiple [SEG] Tokens for Multiple Targets

Single [SEG] token. LISA [32] follows the classic RES methods to generate a segmentation mask under given instructions by adding the [SEG] token into the answer. The prompt is formatted as:

> ***User:*** *What is {obj} in this image? Please output segmentation mask.* ***Assistant:*** *Sure, it is [SEG].*

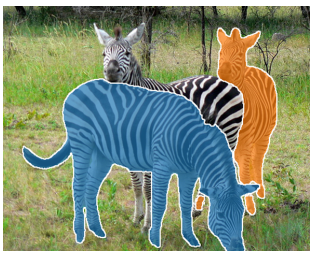
In the above prompt, {*obj*} represents an instance referred to or some semantic area to segment, and the embeddings of the [SEG] token output from the LLM are projected to prompt the SFM. When multiple instances are requested simultaneously, this prompt would confuse the model since only one [SEG] token is forced to match several targets. As shown in Figure 1 (a), LISA coercively predicts the masks of the left man and the guy second from the right, leading to a tattered mask and masking the wrong location.

Multiple [SEG] tokens. To mitigate this issue in GRES, we relax this constraint for GSVA to support multiple target outputs via learning multiple [SEG] tokens. To avoid the ambiguity between the [SEG] tokens and the corresponding objects, we prepend the referring expression before each [SEG] token, *i.e.*, each segmentation prompt in the output text is {*obj_n*}*:[SEG]*, as shown in Figure 3 (a). Streaming all the pairs of object and [SEG] token into one sentence, the question-answer prompt is formatted as (*e.g.*, two targets):

> ***User:*** *What are {obj1}, {obj2} in this image? Please output the segmentation masks.*
> ***Assistant:*** *Sure, {obj1}:[SEG], {obj2}:[SEG].*

This prompt requests the MLLM to identify and distinguish different objects in the image based on the instructions and infuse the corresponding location information

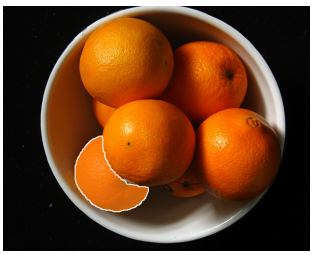
(a) Multiple Targets



User: *What are front zebra, right zebra in this image? Please output segmentation masks.*

Assistant: *Sure, front zebra:[SEG], right zebra:[SEG].*

(b) Empty Target



User: *What are the apple in the bowl, partial orange bottom left in this image? Please output segmentation masks.*

Assistant: *Sure, the apple in the bowl:[REJ], partial orange bottom left:[SEG].*

Figure 3. Example of the prompts and predicted masks of GSVA-Vicuna-7B drawn from gRefCOCO validation set. (a) depicts the multiple-target case, in which two zebras referred to are handled with two separate [SEG] tokens. (b) shows the empty-target case, where no apple is in the bowl. Thus, the null referent is rejected with a [REJ] token, and no segmentation mask will be generated.

of each target into the associated [SEG] token. We regard this ability as an implicit multimodal version of In-Context Learning (ICL), which is demonstrated by many prior works [2, 13, 33, 59, 90]. GSVA takes the target description preceding each [SEG] token as a hint to link this token to the object requested in the image through auto-regressive decoding.

3.4. Rejecting Empty Targets via [REJ] Tokens

Empty targets. In GRES, many expressions match no targets in the image, including absence, incorrect attributes, inaccurate locations, *etc.* These expressions should be treated as empty targets for models to predict all-negative masks. LISA [32] falls short of predicting masks with all-zeros seamlessly since the [SEG] token always calls for a segmentation mask. As shown in Figure 1 (b), LISA incorrectly marks a piece of green apple as the empty target banana.

[REJ] token. We let the MLLM of GSVA predict a [REJ] token for each object that does not exist in the image but is referred to in the instructions, as shown in Figure 3 (b). GSVA predicts the targets marked with [REJ] tokens as empty targets, therefore setting all-zero masks for them. The involvement of [REJ] tokens directly rejects the empty target, liberating the mask decoder of the SFM from learning to segment the inexistent targets. An example prompt with one existing target and one empty target is as follows:

> ***User:*** *What are {obj1} (**absent**), {obj2} (**absent**), {obj3} in this image? Please output the segmentation masks.*
> ***Assistant:*** *Sure, {obj1}:[REJ], {obj2}:[REJ], {obj3}:[SEG].*

The [REJ] token prediction can also be seen as a variant of VQA task, where the specified object and its position in the image need to be considered. Thanks to the unprecedented capabilities of MLLM in understanding the images [9, 34, 39, 92] and reasoning the spatial relationships of the referring objects [52, 88, 91], we make the most of the MLLM in GSVA to unleash the burden of the segmentation model. The proposed empty-target-aware mechanism both improves the quality of masks and ameliorates the errors of identifying nonexistent objects.

4. Experiments

In this section, we conduct comprehensive experiments to validate the efficacy of GSVA. First, we show the results on gRefCOCO [38] dataset to show the superiority of GSVA in GRES tasks. Then we verify GSVA that is also competent with other baselines in classic RES, REC tasks. We move on to ablate some important design choices of GSVA, followed by some qualitative visualization of GRES results.

4.1. GRES

Settings. We adopt gRefCOCO [38] dataset to validate GSVA and LISA [32] on GRES, which contains 278,232 expressions, including 80,022 multi-target and 32,202 empty-target ones, referring to the objects in 19,994 images. The images are split into four subsets: training, validation, test-A, and test-B, following the same UNC partition of RefCOCO [84]. We first add the gRefCOCO training set into the mixed training dataset in LISA to pretrain GSVA and LISA for 50,000 steps and then finetune the models on the gRefCOCO training dataset for another 10 epochs. We evaluate the pretrained and finetuned models on the remaining validation set, test set A, and test set B, respectively. We adopt the gIoU, cIoU metrics for the segmentation mask outputs. Following the implementation in Liu et al. [38], gIoU averages the IoU for each mask, whereas cIoU computes the cumulative intersection area over the cumulative union area across the whole dataset. As for the empty target, we compute the No-target-accuracy (N-acc.), which is the ratio of the correctly classified empty-target expressions over all the empty-target expressions in the dataset. For a correctly classified empty target, the gIoU is set to 1.0, and the cIoU does not take them into account, while for a misclassified empty target, the gIoU is set to 0.0 and the union area is accumulated in the cIoU. Following the criteria in Liu et al. [38], a predicted mask is regarded as empty for LISA if the positive pixels are less than 50, whereas GSVA predicts [REJ] tokens to identify empty targets.

Results. We report the GRES segmentation results of GSVA and LISA [32] in Table 1. Three variants of GSVA, including GSVA-Vicuna-7B, GSVA-Vicuna-13B, and GSVA-Llama2-13B show competitive performance without finetuning on gIoU to the strongest non-LLM baseline ReLA [38]. However, LISA models fail to handle the

Method	Generalized Referring Expression Segmentation on gRefCOCO								
	Validation Set			Test Set A			Test Set B		
	gIoU	cIoU	N-acc.	gIoU	cIoU	N-acc.	gIoU	cIoU	N-acc.
MattNet [85]	48.24	47.51	41.15	59.30	58.66	44.04	46.14	45.33	41.32
LTS [29]	52.70	52.30	-	62.64	61.87	-	50.42	49.96	-
VLT [10]	52.00	52.51	47.17	63.20	62.19	48.74	50.88	50.52	47.82
CRIS [72]	56.27	55.34	-	63.42	63.82	-	51.79	51.04	-
LAVT [81]	58.40	57.64	49.32	65.90	65.32	49.25	55.83	55.04	48.46
ReLA [38]	63.60	62.42	56.37	70.03	69.26	59.02	61.02	59.88	58.40
LISA-Vicuna-7B [32]	32.21	38.72	2.71	48.54	52.55	6.37	39.65	44.79	5.00
GSVA-Vicuna-7B	63.32	61.70	56.45	70.11	69.23	63.50	61.34	60.26	58.42
LISA-Vicuna-7B [32] (ft)	61.63	61.76	54.67	66.27	68.50	50.01	58.84	**60.63**	51.91
GSVA-Vicuna-7B (ft)	**66.47**	**63.29**	**62.43**	**71.08**	**69.93**	**65.31**	**62.23**	60.47	**60.56**
LISA-Vicuna-13B [32]	32.73	39.85	3.66	48.76	53.62	4.89	39.49	45.35	4.41
GSVA-Vicuna-13B	62.97	60.18	58.44	67.17	67.59	54.60	58.06	57.28	52.22
LISA-Vicuna-13B [32] (ft)	63.45	62.99	55.25	68.18	69.65	52.16	61.84	**62.24**	56.15
GSVA-Vicuna-13B (ft)	**68.01**	**64.05**	**65.36**	**71.75**	**70.51**	**67.25**	**63.83**	61.28	**63.11**
LISA-Llama2-13B [32]	33.26	39.64	3.27	49.76	53.80	7.28	40.49	45.41	5.73
GSVA-Llama2-13B	63.20	62.38	54.51	69.52	69.86	57.84	62.06	60.77	58.30
LISA-Llama2-13B [32] (ft)	65.24	63.96	57.49	69.99	71.00	55.43	62.11	62.29	56.34
GSVA-Llama2-13B (ft)	**70.04**	**66.38**	**66.02**	**73.29**	**72.79**	**64.72**	**65.45**	**63.20**	**62.47**

Table 1. Generalized referring expression segmentation (GRES) results on gRefCOCO [38] dataset. gIoU and cIoU are IoU metrics averaged by each example and accumulated over whole dataset, respectively. N-acc. is short for the accuracy of correctly classifying null targets. Baselines are copied from Liu et al. [38]. (ft) denotes the model is finetuned on the training set of gRefCOCO.

GRES task without finetuning, showing degradation in both gIoU and cIoU in each model variant. Especially the low N-acc indicates that LISA is unable to correctly reject the empty targets. When finetuned on gRefCOCO training set, GSVA-Vicuna-7B performs better than the finetuned LISA counterpart, with about 4% improvement in gIoU and over 5% in N-acc on all three evaluation splits. GSVA variants with larger LLM incorporated further push the limits, achieving over 70% in gIoU on the validation set, 73% on test set A, and 65% on test set B. The 13B models also consistently outperform LISA by large margins, demonstrating the superiority of GSVA in GRES task.

4.2. Referring Expression Segmentation

Settings. To validate the abilities to handle various tasks, we evaluate GSVA in the classic RES task. Following the common evaluation protocols, we test variants of GSVA and LISA equipped with different LLMs on RefCOCO, RefCOCO+ [30], and RefCOCOg [44]. We follow the UNC split to perform experiments on RefCOCO and RefCOCO+, and UMD split for RefCOCOg. The models are firstly pretrained as in GRES, and then finetuned for 10 epochs with a joint dataset of these three RES training set. The cIoU metric is adopted to measure the model performances.

Results. Table 2 shows the RES results of GSVA. For the pretrained models, all the three variants of GSVA achieve higher cIoU than LISA [32] by clear margins. Our 7B model outperforms LISA-Vicuna about 2% cIoU on almost every data split. After finetuning, the preponderance of GSVA over LISA keeps and even enlarges. Specifically, the cIoU metrics of GSVA-Llama2-13B on 8 sets surpass LISA by at least 2.8%. For the test set of RefCOCOg, the margin even grows to 5.9%, which exhibits GSVA is also competitive in the classic RES task.

4.3. Referring Expression Comprehension

Settings. Since GSVA is capable for RES tasks, it is natural to transfer to Referring Expression Comprehension (REC) tasks, by simply computing the bounding boxes of the masks. The datasets of REC are the same as RES, in which we evaluate GSVA. If the IoU of a predicted bounding box and the ground truth is greater than 0.5, this prediction is marked as correct. We use the same models in RES to evaluate the phrase grounding capability in REC tasks.

Results. As shown in Table 3, we mainly compare our GSVA to LISA in the three variants. Without finetuning, GSVA-Vicuna-7B outperforms LISA with a large margin over 7% in almost all evaluation sets. The similar trends also hold for the Vicuna-13B and Llama2-13B variants. GSVA also benefits a lot from finetuning, *e.g.,* the 7B variant achieves consistently higher Prec@0.5 than uLLaVA-7B [80] with LoRA finetuning, which is another strong baseline that adopts the "mask2bbox" pipelines without direct bounding box supervision. The finetuned GSVA also shows competitive performances to the fully finetuned uLLaVA-7B, suggesting the strong potential of our method. With larger LLMs incorporated, the performances of GSVA continue with over 3% increments over all datasets.

Method	RefCOCO (UNC)			RefCOCO+ (UNC)			RefCOCOg (UMD)	
	Val.	Test-A	Test-B	Val.	Test-A	Test-B	Val.	Test
Referring Expression Segmentation on RefCOCO, RefCOCO+, and RefCOCOg								
MCN [42]	62.4	64.2	59.7	50.6	55.0	44.7	49.2	49.4
VLT [10]	67.5	70.5	65.2	56.3	61.0	50.1	55.0	57.7
CRIS [72]	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4
LAVT [81]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
ReLA [38]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
X-Decoder [93]	-	-	-	-	-	-	64.6	-
SEEM [94]	-	-	-	-	-	-	65.7	-
PolyFormer-L [40]	76.0	78.3	73.3	**69.3**	**74.6**	61.9	69.2	70.2
LISA-Vicuna-7B* [32]	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5
GSVA-Vicuna-7B	76.4	77.4	72.8	64.5	67.7	58.6	71.1	72.0
LISA-Vicuna-7B* [32] (ft)	74.9	**79.1**	72.3	65.1	70.8	58.1	67.9	70.6
GSVA-Vicuna-7B (ft)	**77.2**	78.9	**73.5**	65.9	69.6	**59.8**	**72.7**	**73.3**
LISA-Vicuna-13B [32]	71.7	74.7	68.1	59.4	64.2	52.9	65.2	66.1
GSVA-Vicuna-13B	74.6	77.5	70.5	62.5	66.5	55.5	69.6	71.2
LISA-Vicuna-13B [32] (ft)	76.0	78.8	72.9	65.0	70.2	58.1	69.5	70.5
GSVA-Vicuna-13B (ft)	**78.2**	**80.4**	**74.2**	**67.4**	71.5	**60.9**	**74.2**	**75.6**
LISA-Llama2-13B [32]	73.4	76.2	69.5	62.3	66.6	56.3	68.2	68.5
GSVA-Llama2-13B	77.7	79.9	74.2	68.0	71.5	61.5	73.2	73.9
LISA-Llama2-13B [32] (ft)	76.3	78.7	72.4	66.2	71.0	59.3	70.1	71.1
GSVA-Llama2-13B (ft)	**79.2**	**81.7**	**77.1**	**70.3**	73.8	**63.6**	**75.7**	**77.0**

Table 2. Referring expression segmentation results on RefCOCO, RefCOCO+ [30] and RefCOCOg [44] dataset. The cIoU metrics of each split are reported. Baselines are excerpted from Lai et al. [32]. (ft) denotes the models are finetuned on the joint training set of the referring expression segmentation datasets. * means the results are excerpted from the original paper.

Method	RefCOCO			RefCOCO+			RefCOCOg	
	Val.	Test-A	Test-B	Val.	Test-A	Test-B	Val.	Test
Referring Expression Comprehension on RefCOCO, RefCOCO+, and RefCOCOg								
u-LLaVA-7B [80] (LoRA)	83.47	87.13	80.21	68.74	76.32	60.98	76.19	78.24
u-LLaVA-7B [80] (full-ft)	86.04	**89.47**	82.26	74.09	**81.16**	66.61	79.87	81.68
LISA-Vicuna-7B [32]	78.68	81.72	75.74	62.92	68.93	56.49	70.10	72.47
GSVA-Vicuna-7B	85.50	88.01	82.49	70.21	75.62	65.11	79.00	79.21
LISA-Vicuna-7B [32] (ft)	85.39	88.84	82.59	**74.23**	79.46	**68.40**	79.34	80.42
GSVA-Vicuna-7B (ft)	**86.27**	89.22	**83.77**	72.81	78.78	68.01	**81.58**	**81.83**
LISA-Vicuna-13B [32]	80.01	83.26	76.26	63.77	70.24	57.42	71.79	73.34
GSVA-Vicuna-13B	83.12	87.01	80.54	68.14	73.90	62.00	77.08	78.89
LISA-Vicuna-13B [32] (ft)	85.92	89.05	83.16	74.86	81.08	68.87	80.09	81.48
GSVA-Vicuna-13B (ft)	**87.71**	**90.49**	**84.57**	**76.52**	**81.69**	**70.35**	**83.90**	**84.85**
LISA-Llama2-13B [32]	82.52	85.56	78.82	67.91	73.77	62.25	75.37	76.83
GSVA-Llama2-13B	86.99	89.54	84.08	73.89	79.10	69.38	80.68	82.07
LISA-Llama2-13B [32] (ft)	85.91	88.84	81.73	74.46	80.56	68.26	80.09	81.27
GSVA-Llama2-13B (ft)	**89.16**	**92.08**	**87.17**	**79.74**	**84.45**	**73.41**	**85.47**	**86.18**

Table 3. Referring expression comprehension results on RefCOCO, RefCOCO+ [30] and RefCOCOg [44] dataset. The metric is precision @ 0.5 IoU threshold. (LoRA) means the LLM in u-LLaVA [80] is finetuned with LoRA adapter, as LISA and GSVA, while (full-ft) represents the LLM in u-LLaVA is fully trained. Results of u-LLaVA-7B with "mask2bbox" strategy are reported for fair comparison.

4.4. Ablation Study

The involvement of the [REJ] token. [REJ] token plays a rather important role in GSVA. We study the effect of the [REJ] token by removing it from the vocabulary in GSVA, yielding a variant unable to reject a target from the text outputs. As shown in the 2nd row of Table 4, after removing [REJ] token, there is a sharp N-acc drop over 25% relatively, followed by the decline of gIoU at about 10% on gRefCOCO validation set. This performance degradation indicates the significance of LLM learning a special token to reject the referred instances absent in the image.

Learning multiple [SEG] tokens. We continue to ablate the multiple [SEG] tokens, which is another core design of GSVA. After removing [REJ] token, we then reduce the number of [SEG] tokens to 1, which is identical to LISA [32] with the referring expression added before the only one [SEG] token: ***Assistant:*** *Sure,* {*obj*}*:[SEG].*. In the early experiments, we have found that stacking multiple expressions before one [SEG] token would result in diver-

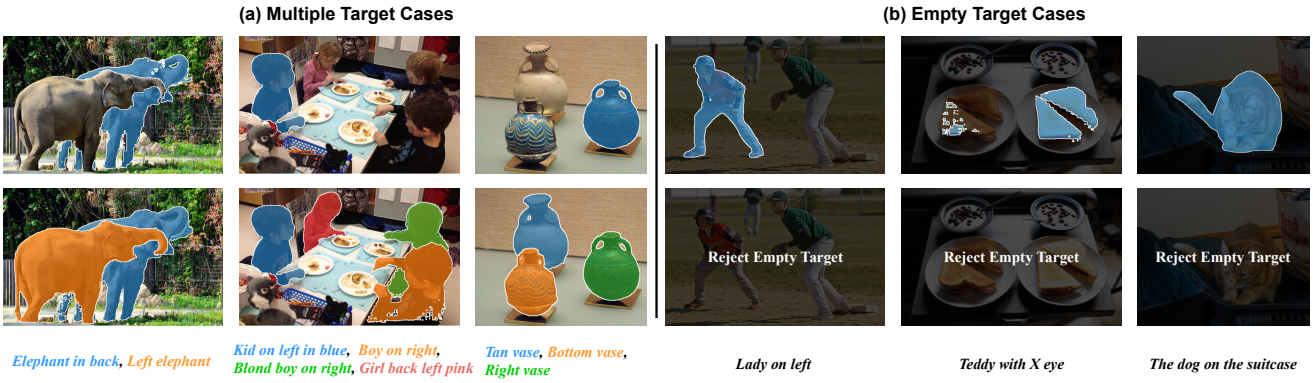


Figure 4. Visualizations of GSVA and LISA [32] in the GRES task. The first row shows LISA's segmentation results, the second row is the masks and rejections of GSVA, and the third row shows the referring expressions in the instructions. In (a) multiple target cases, each target is colored with a specific color. In (b) empty target cases, the images turn darker to highlight the incorrect predictions of LISA. The examples are selected from the gRefCOCO validation set. The masks are generated by the 7B models. Zoom in for the best view.

Model w/ Vicuna-7B	Modifications			gRefCOCO Val.		
	RefExp. +[SEG]	Multiple [SEG]	[REJ] Token	gIoU	cIoU	N-acc.
GSVA	✓	✓	✓	63.32	61.70	56.45
	✓	✓	✗	51.57	60.95	30.32
	✓	✗	✗	44.86	59.37	11.96
LISA [32]	✗	✗	✗	32.21	38.72	2.71
	✗	✓	✓	21.83	27.22	0.00

Table 4. Ablation study on the core designs of GSVA. ✓ means the employment of the component while ✗ means not. "RefExp.+[SEG]", "Multiple [SEG]", and "[REJ] Token" are short for adding referring expression before [SEG] in the answer prompt, using multiple [SEG] tokens, involving [REJ] token, respectively.

gence. Therefore we separate multiple targets to prompt the model with one expression at a time. The sharp decrements of gIoU by nearly 7% and N-acc by over 15% in the 3rd row demonstrate the significance of the multiple-[SEG]-token.

Answers without referring expression. To examine the efficacy of the hinting prompts, we remove all the referring expressions before the [SEG] tokens. Based on the removal of multiple [SEG]s and [REJ]s, erasing the added referring expression falls back to the original LISA model, as shown in the 4th row. We further choose only to remove it from GSVA model, keeping other configurations unchanged. Specifically, if there are two referents, the prompts in the answer will turn to ***Assistant:*** *Sure, [SEG], [SEG].*, whose results are in the last row. The zero N-acc shows the model fails to identify any empty target without the help of the expressions, meanwhile the poor gIoU and cIoU indicates the segmentation ability is damaged. This phenomenon also suggests that the added referring expression hint GSVA to associate each [SEG] token to its corresponding target, which is in coherence with our hypothesis.

4.5. Visualization

We visualize some qualitative results of GSVA to verify its effectiveness. As shown in Figure 4, we present two groups of examples from the validation set of gRefCOCO [38] to see how GSVA outperforms LISA in the face of the two main challenges in GRES: multiple targets and empty targets. In part (a), GSVA has managed to segment all the targets referred to, while LISA could only segment one of the requested instances. For example, in the third column, LISA only predicts the mask of the rightmost vase. On the contrary, GSVA separately segments all three targeted vases. In part (b), LISA mistakenly segments the instances in the image that disagree with the referring expression, e.g., in the sixth column, LISA proposes a mask of a cat in response to the request of the dog on the suitcase, whereas GSVA successfully reject all the empty targets.

5. Conclusion

This paper introduces a novel multimodal large language model dubbed **Generalized Segmentation Vision Assistant (GSVA)**. By introducing multiple [SEG] tokens and the new [REJ] token, GSVA effectively achieves multi-objective segmentation and empty target rejection, which addresses the challenging segmentation problems in practical application scenarios, *e.g.*, Generalized Referring Expression Segmentation (GRES). Extensive experiments on GRES, classic RES, and REC tasks fully demonstrate the superior performance of our method, highlighting its significance for future research and applications.

Acknowledgments

This work is supported in part by the National Key R&D Program of China under Grant 2021ZD0140407, the National Natural Science Foundation of China under Grants 62321005 and 62276150. We also thank Dr. Yuan Yao and Prof. Zhiyuan Liu for their insightful and valuable comments on this research project. We also appreciate their generous support of computational resources.

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*, 2022. 2

[2] 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. In *NeurIPS*, 2020. 2, 5

[3] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *IEEE ICCV*, 2019. 2

[4] Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang. Yourefit: Embodied reference understanding with language and gesture. In *IEEE ICCV*, 2021. 2, 4

[5] Ming-Ming Cheng, Shuai Zheng, Wen-Yan Lin, Vibhav Vineet, Paul Sturgess, Nigel Crook, Niloy J Mitra, and Philip Torr. Imagespirit: Verbal guided image parsing. *ACM ToG*, 2014. 1, 2

[6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 2, 3

[7] Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. Mask grounding for referring image segmentation. In *IEEE CVPR*, 2024. 2

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 2

[9] 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. 2, 5

[10] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *IEEE ICCV*, 2021. 6, 7

[11] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *IEEE ICCV*, 2021. 2

[12] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *IEEE CVPR*, 2022. 2

[13] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 5

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[15] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *IEEE CVPR*, 2023. 2

[16] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *IEEE CVPR*, 2021. 2

[17] Chen Gao, Jinyu Chen, Si Liu, Luting Wang, Qiong Zhang, and Qi Wu. Room-and-object aware knowledge reasoning for remote embodied referring expression. In *IEEE CVPR*, 2021. 1, 4

[18] Dongchen Han, Tianzhu Ye, Yizeng Han, Zhuofan Xia, Shiji Song, and Gao Huang. Agent attention: On the integration of softmax and linear attention. *arXiv preprint arXiv:2312.08874*, 2023. 2

[19] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *IEEE TPAMI*, 2021. 2

[20] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Yitian Zhang, and Haojun Jiang. Spatially adaptive feature refinement for efficient inference. *IEEE TIP*, 2021. 2

[21] Yizeng Han, Yifan Pu, Zihang Lai, Chaofei Wang, Shiji Song, Junfeng Cao, Wenhui Huang, Chao Deng, and Gao Huang. Learning to weight samples for dynamic early-exiting networks. In *ECCV*, 2022.

[22] Yizeng Han, Zhihang Yuan, Yifan Pu, Chenhao Xue, Shiji Song, Guangyu Sun, and Gao Huang. Latency-aware spatial-wise dynamic networks. In *NeurIPS*, 2022. 2

[23] Yizeng Han, Dongchen Han, Zeyu Liu, Yulin Wang, Xuran Pan, Yifan Pu, Chao Deng, Junlan Feng, Shiji Song, and Gao Huang. Dynamic perceiver for efficient visual recognition. In *IEEE ICCV*, 2023. 2

[24] Yizeng Han, Zeyu Liu, Zhihang Yuan, Yifan Pu, Chaofei Wang, Shiji Song, and Gao Huang. Latency-aware unified dynamic networks for efficient image recognition. *arXiv preprint arXiv:2308.15949*, 2023. 2

[25] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, 2016. 1, 2

[26] Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. Beyond one-to-one: Rethinking the referring image segmentation. In *IEEE ICCV*, 2023. 2

[27] Gao Huang, Yulin Wang, Kangchen Lv, Haojun Jiang, Wenhui Huang, Pengfei Qi, and Shiji Song. Glance and focus networks for dynamic visual recognition. *IEEE TPAMI*, 2022. 2

[28] Rui Huang, Xuran Pan, Henry Zheng, Haojun Jiang, Zhifeng Xie, Cheng Wu, Shiji Song, and Gao Huang. Joint representation learning for text and 3d point cloud. *Pattern Recognition*, 2024. 1

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

[30] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 1, 4, 6, 7

[31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *IEEE ICCV*, 2023. 2, 3

[32] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[33] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 5

[34] 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, 5

[35] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *IEEE CVPR*, 2018. 2

[36] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 2

[37] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *IEEE CVPR*, 2017. 2

[38] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *IEEE CVPR*, 2023. 1, 2, 4, 5, 6, 7, 8

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

[40] 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 *IEEE CVPR*, 2023. 7

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

[42] 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 *IEEE CVPR*, 2020. 7

[43] Tengchao Lv, Yupan Huang, Jingye Chen, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, et al. Kosmos-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419*, 2023. 2

[44] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *IEEE CVPR*, 2016. 2, 4, 6, 7

[45] Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *IEEE CVPR*, 2019. 4

[46] Zanlin Ni, Yulin Wang, Jiangwei Yu, Haojun Jiang, Yue Cao, and Gao Huang. Deep incubation: Training large models by divide-and-conquering. In *IEEE CVPR*, 2023. 2

[47] Zanlin Ni, Yulin Wang, Renping Zhou, Jiayi Guo, Jinyi Hu, Zhiyuan Liu, Shiji Song, Yuan Yao, and Gao Huang. Revisiting non-autoregressive transformers for efficient image synthesis. In *CVPR*, 2024. 2

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

[49] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. 2

[50] Xuran Pan, Chunjiang Ge, Rui Lu, Shiji Song, Guanfu Chen, Zeyi Huang, and Gao Huang. On the integration of self-attention and convolution. In *IEEE CVPR*, 2022. 2

[51] Xuran Pan, Tianzhu Ye, Zhuofan Xia, Shiji Song, and Gao Huang. Slide-transformer: Hierarchical vision transformer with local self-attention. In *IEEE CVPR*, 2023. 2

[52] 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, 5

[53] Yifan Pu, Yiru Wang, Zhuofan Xia, Yizeng Han, Yulin Wang, Weihao Gan, Zidong Wang, Shiji Song, and Gao Huang. Adaptive rotated convolution for rotated object detection. In *IEEE ICCV*, 2023. 2

[54] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *IEEE CVPR*, 2020. 1, 4

[55] 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*, 2021. 2, 3

[56] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023. 2

[57] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *ECCV*, 2018. 2

[58] Qie Sima, Sinan Tan, Huaping Liu, Fuchun Sun, Weifeng Xu, and Ling Fu. Embodied referring expression for manipulation question answering in interactive environment. In *IEEE ICRA*, 2023. 1, 4

[59] Yan Tai, Weichen Fan, Zhao Zhang, Feng Zhu, Rui Zhao, and Ziwei Liu. Link-context learning for multimodal llms. *arXiv preprint arXiv:2308.07891*, 2023. 5

[60] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model, 2023. 2

[61] 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.

[62] 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. 2

[63] 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 *IEEE CVPR*, 2023. 2

[64] 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

[65] 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

[66] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *IEEE CVPR*, 2019. 1, 4

[67] Yulin Wang, Zhaoxi Chen, Haojun Jiang, Shiji Song, Yizeng Han, and Gao Huang. Adaptive focus for efficient video recognition. In *IEEE ICCV*, 2021. 2

[68] Yulin Wang, Rui Huang, Shiji Song, Zeyi Huang, and Gao Huang. Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. In *NeurIPS*, 2021.

[69] Yulin Wang, Yang Yue, Yuanze Lin, Haojun Jiang, Zihang Lai, Victor Kulikov, Nikita Orlov, Humphrey Shi, and Gao Huang. Adafocus v2: End-to-end training of spatial dynamic networks for video recognition. In *IEEE CVPR*, 2022.

[70] Yulin Wang, Yang Yue, Xinhong Xu, Ali Hassani, Victor Kulikov, Nikita Orlov, Shiji Song, Humphrey Shi, and Gao Huang. Adafocusv3: On unified spatial-temporal dynamic video recognition. In *ECCV*, 2022. 2

[71] Yulin Wang, Yang Yue, Rui Lu, Tianjiao Liu, Zhao Zhong, Shiji Song, and Gao Huang. Efficienttrain: Exploring generalized curriculum learning for training visual backbones. In *IEEE ICCV*, 2023. 2

[72] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *IEEE CVPR*, 2022. 6, 7

[73] Zhichao Wei, Xiaohao Chen, Mingqiang Chen, and Siyu Zhu. Learning aligned cross-modal representations for referring image segmentation. *arXiv preprint arXiv:2301.06429*, 2023. 2

[74] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*, 2023. 2

[75] Tsung-Han Wu, Giscard Biamby, David Chan, Lisa Dunlap, Ritwik Gupta, Xudong Wang, Joseph E Gonzalez, and Trevor Darrell. See, say, and segment: Teaching lmms to overcome false premises. *arXiv preprint arXiv:2312.08366*, 2023. 2

[76] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *IEEE CVPR*, 2022. 2

[77] Zhuofan Xia, Xuran Pan, Xuan Jin, Yuan He, Hui Xue, Shiji Song, and Gao Huang. Budgeted training for vision transformer. In *ICLR*, 2023. 2

[78] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Dat++: Spatially dynamic vision transformer with deformable attention. *arXiv preprint arXiv:2309.01430*, 2023. 2

[79] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. *arXiv preprint arXiv:2311.06242*, 2023. 2

[80] Jinjin Xu, Liwu Xu, Yuzhe Yang, Xiang Li, Yanchun Xie, Yi-Jie Huang, and Yaqian Li. u-llava: Unifying multi-modal tasks via large language model. *arXiv preprint arXiv:2311.05348*, 2023. 2, 6, 7

[81] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *IEEE CVPR*, 2022. 2, 6, 7

[82] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *IEEE CVPR*, 2019. 2

[83] 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

[84] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, 2016. 4, 5

[85] 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 *IEEE CVPR*, 2018. 6

[86] Ao Zhang, Liming Zhao, Chen-Wei Xie, Yun Zheng, Wei Ji, and Tat-Seng Chua. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*, 2023. 2

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

[88] 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. 5

[89] Zicheng Zhang, Yi Zhu, Jianzhuang Liu, Xiaodan Liang, and Wei Ke. Coupalign: Coupling word-pixel with sentence-mask alignments for referring image segmentation. In *NeurIPS*, 2022. 2

[90] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*, 2023. 5

[91] 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. 5

[92] 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. 5

[93] 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 *IEEE CVPR*, 2023. 2, 7

[94] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *NeurIPS*, 2023. 2, 7