

# NExT-Chat: An LMM for Chat, Detection and Segmentation

Ao Zhang<sup>1</sup> Yuan Yao<sup>1\*</sup> Wei Ji<sup>1</sup> Zhiyuan Liu<sup>2</sup> Tat-Seng Chua<sup>1</sup>

<sup>1</sup>National University of Singapore <sup>2</sup>Tsinghua University

aozhang@u.nus.edu yaoyuanthu@gmail.com

<https://next-chatv.github.io>

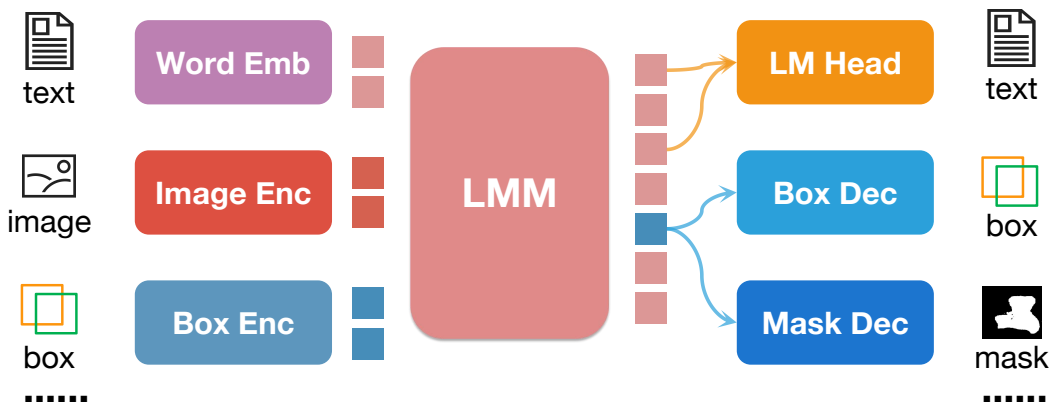


Figure 1. By using the embedding based location modeling method, our NExT-Chat can take bounding boxes as input and output text, bounding boxes and masks in the multimodal conversation.

## Abstract

The development of large language models (LLMs) has greatly advanced the field of multimodal understanding, leading to the emergence of large multimodal models (LMMs). In order to enhance the level of visual comprehension, recent studies have equipped LMMs with region-level understanding capabilities by representing object bounding box coordinates as a series of text sequences (pix2seq). In this paper, we introduce a novel paradigm for object location modeling called pix2emb method, where we ask the LMM to output the location embeddings and then decode them with different decoders. This paradigm allows us to use different location formats (such as bounding boxes and masks) in multimodal conversations. Leveraging the proposed pix2emb method, we train an LMM named NExT-Chat and demonstrate its capability of handling multiple tasks like visual grounding, region captioning, and grounded reasoning. Comprehensive experiments show the effectiveness of our NExT-Chat on various tasks, e.g., NExT-Chat (87.7) vs. Shikra (86.9) on POPE-Random, NExT-

Chat (68.9) vs. LISA (67.9) on referring expression segmentation task, and NExT-Chat (79.6) vs. Kosmos-2 (62.3) on region caption task.

## 1. Introduction

Recently, large language models (LLMs) have shown spreading influence in different areas, among which large multimodal models (LMMs) is one of the most attractive area. Researchers try to equip LLMs with visual perception modules resulting in LMMs [11, 17, 43, 46] that can describe the visual content and answer visual questions. However, these LMMs are limited to holistic image understanding without the ability to conduct region-level reasoning, for example, locating the referred objects in the conversation.

To enable region-level understanding, current solutions [4, 27, 32] utilize the pix2seq [5] paradigm where the object coordinates are converted to LLM understandable text tokens (e.g.,  $[x_1, y_1, x_2, y_2]$ ). Consequently, LMMs can output object coordinates as part of a normal next token prediction problem. However, the pix2seq paradigm is limited to discrete coordinate outputs and struggles to provide other

\*Corresponding author.

fine-grained formats, such as segmentation masks.

To address these limitations, we propose the `pix2emb` paradigm, which can accommodate different location formats. The key idea is to model all location information as embeddings, which can be decoded into the target formats by corresponding decoders. Specifically, we introduce two new tokens, `<trigger>` and `<loc>`, where the `<trigger>` serve as a trigger for localization and `<loc>` act as a placeholder for objects’ location embeddings. During the text generation, the `<trigger>` triggers the location decoding, where the hidden states of `<trigger>` can be used for both detection and segmentation, as depicted in Fig. 2. Then, the predicted or provided object location will be encoded into the embedding of the `<loc>` token for object referring. In addition to supporting flexible output formats, the `pix2emb` modeling also allows for the use of existing localization practices. While the `pix2seq` paradigm can only frame the detection task as a token classification problem, the embedding-based paradigm formulates the localization task as a regression problem, enabling the adoption of established practices such as L1 loss, IoU loss and GIoU loss.

Building upon the proposed `pix2emb` method, we introduce a new LMM named `NExT-Chat`. `NExT-Chat` is designed to handle various conversation scenarios, including visual grounding (Fig. 4), region caption (Fig. 6), and grounded image caption (Fig. 7). Thanks to the incorporation of LLM, `NExT-Chat` is also capable of handling scenarios that requires grounded reasoning. By providing an extensive array of examples, we effectively demonstrate `NExT-Chat`’s remarkable proficiency in understanding various components, including background elements, minute objects, and associating the objects with related knowledge. Moreover, we validate our `NExT-Chat` on various datasets. On the POPE-Random dataset, `NExT-Chat` achieves an impressive accuracy of 87.7, surpassing `Shikra`’s 86.9. In referring expression segmentation (RES), it attains an average cIoU of 68.9, outperforming `LISA`’s 67.9. Moreover, `NExT-Chat` achieves a remarkable 79.6 in CIDEr score for RefCOCOg region captioning, significantly exceeding `Kosmos-2`’s 62.3.

To summarize, our contributions can be listed as follows:

- *Effective Method.* We propose the `pix2emb` method, which can accommodate different output formats such as bounding boxes and segmentation masks.
- *NExT-Chat Model.* Based on the proposed `pix2emb` method, we build `NExT-Chat` that can unify the chat, region input, detection and segmentation in a single LMM.
- *Experiments and Demos.* We provide abundant qualitative and quantitative results to showcase the effectiveness of our proposed method.

## 2. Related Works

### 2.1. LMM

Large multimodal models (LMMs) are typically built on large language models (LLMs) and equipped with visual perception modules to enable the multimodal perception ability, which can generate captions or answer questions based on the given multimodal content. `Flamingo` [1] tries to extract vision information by a pre-trained vision backbone with a resampler, and incorporate them into the text features with a cross-attention mechanism. Instead of using cross-attention layers, `BLIP-2` [17] and `Kosmos` [11] directly feed the visual features into the LLMs as soft prompts. Following `BLIP-2`, `MiniGPT-4` [46] and `VPG-Trans` [43] build LMMs with transfer learning, and significantly reduce the training cost. For example, `VPGTrans` can use only around 10% GPU hours with non-degenerated performances compared with training a new LMM from scratch. When considering the training paradigm, researchers find that a small scale instruction tuning can better align the LMM with the expected output format. `MiniGPT-4` [46] fine-tunes its model with less than 5,000 self-instruct image-text pairs and turns the model into better conversation robot. Different from `MiniGPT-4`’s self-instruct, `LLaVA` [22] generate the instruction tuning data with the text-only GPT-4 models by feeding the visual information as text sentences. `Otter` [15, 16] further propose a MIMIC-IT dataset that can turn the LMM into better in-context learners. `LLaVA-1.5` proposes to further fine-tune the model on human annotated datasets, which can alleviate the image-level hallucination [20]. However, these LMMs [1, 21, 22] can only take the whole image/video as input and output text, and are incapable of handling region understanding tasks.

### 2.2. LMM for Region Reasoning

`GPT4ROI` [44] proposes to encode the regions as features and thus can accept the region as input. `Pix2seq` [5] first propose to represent object bounding box coordinates as text tokens and thus the language model can output the object locations in a token classification manner. However, `pix2seq` only validate its idea on traditional object detection tasks. `UniTab` [36] and `PEVL` [38] further extend the idea to vision&language tasks like visual grounding [26, 40]. Following this line, `Vision-LLM` [32] and `Kosmos-2` [27] recently applies the token classification concept to LMMs. Take `Kosmos-2` as an example, it discretize the whole image into  $32 \times 32$  bins, which can be used to represent the points lying in it. Additional  $32 \times 32$  tokens are introduced to the LLM’s vocabulary for either coordinates input or output. Thus, the LMM can achieve the region-level reasoning. `Shikra` [4] point out that introducing too much new tokens will inevitably increase the

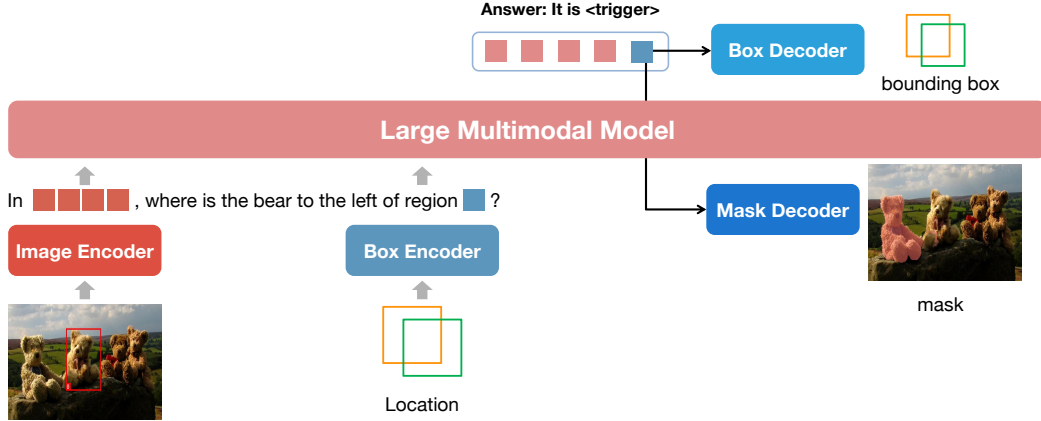


Figure 2. The overall framework of NExT-Chat. The image and given bounding boxes are encoded by image and box encoders respectively. During decoding, the hidden states of the `<trigger>` are fed into box and mask decoders, enabling object detection and segmentation.

training difficulties. Thus, Shikra propose to reuse the LLM’s original vocabulary and turn the box coordinates into normalized numerical values with certain precision like  $[0.111, 0.111, 0.333, 0.333]$ . Although avoiding introducing too much new tokens, it requires roughly 26 tokens to represent each bounding box, which is ineffective. Different from these works, we do not formulate the object localization problem as a token classification problem. Our NExT-Chat introduces an `<trigger>` token as the trigger for location decoding, and then use the hidden states to decode the bounding boxes and the segmentation masks.

### 3. Method

In this section, we present the NExT-Chat framework, starting with an introduction to the overall LMM architecture (§3.1), followed by a description of the pix2emb method (§3.2). Additionally, we provide details on the training process (§3.3).

#### 3.1. LMM Architecture

For the LMM architecture, we adopt a LLaVA-like architecture. Specifically, we employ a CLIP ViT-L/14@336px [29] as the vision encoder. The input image is converted into  $24 \times 24$  patch embeddings and then projected to the same dimension as the word embeddings of the LLM. These patch embeddings serve as visual tokens. Then, the visual tokens will be fed into a decoder-only LLM for conditional text generation. Regarding the selection of LLMs, we opt for the recently released Vicuna-1.5 model [45].

#### 3.2. Pix2Emb Method

**Detection.** To model the object location as output, we introduce a special token, denoted as `<trigger>`, which serves to trigger the localization. As depicted in Fig. 2, the LMM is trained to generate the `<trigger>` token before

predicting the locations. Then, the embedding  $\mathbf{t} \in \mathcal{R}^n$  of `<trigger>` is then passed to the *Box Decoder*  $\mathcal{F}$  for regression. Mathematically, this can be expressed as follows:

$$\mathbf{b} = \mathcal{F}(\mathbf{t}), \quad (1)$$

where  $\mathbf{b} \in \mathcal{R}^4$  represents the predicted bounding box coordinates in the format  $[x_0, y_0, x_1, y_1]$ .

In our NExT-Chat model, the box decoder consists of a 2-layer MLP. To supervise the location output, we employ a joint loss function comprising of the L1 loss and the GIoU loss [30] during training:

$$\mathcal{L}_{det} = \alpha \mathcal{L}_1(\mathbf{b}, \mathbf{b}_{gt}) + \beta \text{GIoU}(\mathbf{b}, \mathbf{b}_{gt}), \quad (2)$$

where  $\mathbf{b}_{gt}$  represents the ground truth coordinates, and  $\alpha = 2$ ,  $\beta = 0.8$  follows the ratio utilized in DETR [3].

**Segmentation.** Similar to the detection process, we utilize the hidden states  $\mathbf{t}$  of the `<trigger>` as input for the mask head. Inspired by LISA [14], we use SAM [12] as our mask head, which also additionally takes the original image as input. To ensure compatibility between the hidden states and SAM, we first project the hidden states to match the dimension of SAM’s prompt embedding using a linear projector. Subsequently, the projected hidden states are fed as the prompt embedding to SAM. For improved performance, we also encode the detected bounding boxes into a prompt embedding with SAM’s prompt encoder and concatenate it with the projected embedding. To train the mask output, we follow the practice outlined in lightning-SAM<sup>1</sup>:

$$\mathcal{L}_{seg} = \text{IoU}(\mathbf{m}, \mathbf{m}_{gt}) + \text{D}(\mathbf{m}, \mathbf{m}_{gt}) + \beta \text{F}(\mathbf{m}, \mathbf{m}_{gt}), \quad (3)$$

where IoU, D, and F are IoU Loss, Dice Loss, and Focal Loss separately.  $\beta$  is set to 20 in our experiments.

<sup>1</sup><https://github.com/luca-medeiros/lightning-sam/tree/main>

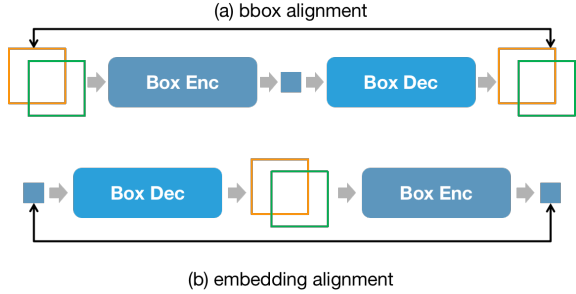


Figure 3. Cycle loss utilized to bind box encoder and decoder training.

**Location as Input.** In addition to the location output, it is essential to incorporate location as input as well. To be consistent with the location output modeling, we also use a single embedding to represent the location information. Therefore, the output location embedding can also serve as the input embedding. Consequently, we introduce another 2-layer MLP, referred to as the location encoder  $\mathcal{G}$ . In order to simplify the problem, we convert all location formats into bounding boxes  $b$  and subsequently transform them into embeddings  $t \in \mathbb{R}^n$  suitable for the LLM. The location encoder can be supervised through the standard text generation loss  $\mathcal{L}_{text}$ . For instance, when inquiring about the relationship between bounding box  $b_1$  and  $b_2$ , the location encoder is compelled to provide precise information.

However, we observe that the location encoder cannot be effectively trained solely through indirect supervision from  $\mathcal{L}_{text}$ . As a result, we introduce an additional cycle loss to facilitate the training of the encoder in conjunction with the decoder. As illustrated in Fig. 3 (a), a bounding box will be encoded and then decoded, where two bounding boxes are asked to be the same. Similarly, the hidden states of `<trigger>` will also be used to calculate the cycle loss (Fig. 3 (b)). Formally, the  $\mathcal{L}_{cyc}$  is defined as:

$$\mathcal{L}_{cyc} = \mathcal{L}_1(b, \mathcal{F}(\mathcal{G}(b))) + \mathcal{L}_2(t, \mathcal{G}(\mathcal{F}(t))), \quad (4)$$

where  $b$  and  $t$  are provided bounding box and predicted embedding respectively. Additionally,  $\mathcal{L}_1$  and  $\mathcal{L}_2$  correspond to the L1 Loss and L2 Loss, respectively.

### 3.3. Training Process

We employ a three-stage training process, consisting of pre-training, instruction tuning, and segmentation training, to train our model. The idea is to train the bounding box decoding ability for the first two stages and then extend to segmentation with a lightweight training.

**Stage 1.** During this stage, we perform pre-training using a mixture of data from various sources, including Flickr30K Entities [28], Visual Genome [13], RefCOCO [40], RefCOCO+ [40], RefCOCOg [26], VQAv2 [2], PointQA [25], Visual7W [47], and VCR [42]. The model is trained with

a batch size of 64 and a learning rate of  $2e-5$  for 65k steps. During this pre-training stage, the entire language model with the box decoder, is trained while keeping the image encoder frozen. The training loss is formulated as:

$$\mathcal{L}_{s1} = \mathcal{L}_{text} + \mathcal{L}_{det} + \mathcal{L}_{cyc}. \quad (5)$$

For NExT-Chat 7B model, the stage-1 training uses 8 A100 (80G) GPUs for around 59 hours.

**Stage 2.** In the second stage, we further fine-tune the model using data from VQAv2, RefCOCO, Flickr30K Entities, LLaVA-instruct, VG grounded captioning, VCR, and Shikra-RD [4]. The batch size is reduced to 64, and the learning rate is set to  $2e-5$ . The loss is the same with stage-1’s loss. For NExT-Chat 7B model, the stage-2 training uses 8 A100 (80G) GPUs for around 10 hours.

**Stage 3.** After the two stages training, the model is equipped with the ability to engage in dialogue and perform image localization. To prevent catastrophic forgetting, we keep most of the parameters frozen during the segmentation training. Specifically, we only train the linear projector between the LMM and SAM, as well as the decoder of SAM. The loss for the stage-3 is:

$$\mathcal{L}_{s3} = \mathcal{L}_{seg}. \quad (6)$$

Thanks to the small amount of training parameters, the training can be done in 3 hours with 8 A100 (80G) GPUs. This training is performed using the referring segmentation splits of RefCOCO, RefCOCO+, and RefCOCOg datasets.

## 4. Experiment

In this section, we begin by conducting a rigorous evaluation to validate the effectiveness of our pix2emb approach in a fair comparison setting. Following that, we demonstrate the potential of our NExT-Chat model by presenting a wide range of qualitative results from different scenarios. Finally, we provide quantitative results to compare the performance of our NExT-Chat model with the current SOTA methods on the image-level hallucination, referring expression segmentation, referring expression detection and region-level caption tasks.

### 4.1. Applications across Different Scenarios

In this section, we present qualitative results that showcase the capabilities of our NExT-Chat model across various scenarios.

**Visual Grounding.** As shown in Fig. 4, we can see that our NExT-Chat accurately detects and segments the queried objects, such as the bears and the sky in the background. To ensure that our model is not biased towards specific objects, we test it with different queries to find all four bears individually. Our model successfully localizes each bear based on

Table 1. **Image Hallucination:** the comparison between our NEXT-Chat with current SOTA models on the POPE benchmark for image hallucination diagnosis.

Datasets	Metrics	NEXT-Chat	Shikra	InstructBLIP	MiniGPT-4	LLaVA	MM-GPT	mPLUG-Owl
Random	Accuracy (↑)	87.70	86.90	88.57	79.67	50.37	50.10	53.97
	Precision (↑)	93.46	94.40	84.09	78.24	50.19	50.05	52.07
	Recall (↑)	81.87	79.27	95.13	82.20	99.13	100.00	99.60
	F1-Score (↑)	87.28	86.19	89.27	80.17	66.64	66.71	68.39
	Yes	45.15	43.26	56.57	52.53	98.77	99.90	95.63
Popular	Accuracy (↑)	84.57	83.97	82.77	69.73	49.87	50.00	50.90
	Precision (↑)	86.54	87.55	76.27	65.86	49.93	50.00	50.46
	Recall (↑)	81.87	79.20	95.13	81.93	99.27	100.00	99.40
	F1-Score (↑)	84.14	83.16	84.66	73.02	66.44	66.67	66.94
	Yes	47.30	45.23	62.37	62.20	99.40	100.00	98.57
Adversarial	Accuracy (↑)	81.93	83.10	72.10	65.17	49.70	50.00	50.67
	Precision (↑)	82.02	85.60	65.13	61.19	49.85	50.00	50.34
	Recall (↑)	81.80	79.60	95.13	82.93	99.07	100.00	99.33
	F1-Score (↑)	81.91	82.49	77.32	70.42	66.32	66.67	66.82
	Yes	49.87	46.50	73.03	67.77	99.37	100.00	98.67

Table 2. **RES:** comparison between our NEXT-Chat and baselines on RES. The evaluation metric is **cIoU**.

Methods	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
MCN [24]	62.4	64.2	59.7	50.6	55.0	44.7	49.2	49.4
VLT [8]	67.5	70.5	65.2	56.3	61.0	50.1	55.0	57.7
CRIS [34]	70.5	73.2	66.1	65.3	68.1	53.7	59.9	60.4
LAVT [37]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
GRES [19]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
X-Decoder [48]	-	-	-	-	-	-	64.6	-
SEEM [49]	-	-	-	-	-	-	65.7	-
LISA-7B [14]	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5
<b>NEXT-Chat (ours)</b>	<b>74.7</b>	<b>78.9</b>	<b>69.5</b>	<b>65.1</b>	<b>71.9</b>	<b>56.7</b>	<b>67.0</b>	<b>67.0</b>

the given queries. Additionally, our model showcases reasoning abilities through challenging grounding problems. For instance, in Fig. 5, our model accurately localizes the remote in response to the query “Where is the object to control the TV in image?” It also localizes the boat based on understanding the given object location input.

**Region Captioning.** To evaluate the effectiveness of our NEXT-Chat model for region input, we conducted experiments where the model generates descriptions based on given bounding boxes. As depicted in Fig. 6, our model consistently produces accurate descriptions specifically tailored to the provided regions, without being influenced by the overall image content or salient regions. We observed this behavior consistently across different examples. Notably, in the second row of Fig. 6, our model demonstrates

the ability to accurately recognize and describe small objects such as flags, as well as background objects like trees. This demonstrates the robustness and effectiveness of our model in generating region-based captions.

**Grounded Captioning.** Another compelling application of our NEXT-Chat model is its ability to describe images by referencing specific objects present within them. Fig. 7 demonstrates that our model can accurately identify and describe the major 2 or 3 objects in an image, effectively organizing them into coherent sentences. By incorporating object references, our model demonstrates a reduced tendency to generate captions containing non-existent objects. This highlights the model’s capability to generate more accurate and contextually grounded image descriptions.

**Reasoning.** In addition to its demonstrated ability in



Table 3. **REC**: comparison between our NExT-Chat and baselines on REC. The evaluation metric is **Acc@0.5**. \* refers to the specialist or fine-tuned methods.

Type	Methods	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
non-LMM	MAttNet* [41]	76.4	80.4	69.3	64.9	70.3	56.0	66.7	67.0
	OFA-L [31]	80.0	83.7	76.4	68.3	76.0	61.8	67.6	67.6
	OFASys	-	80.1	-	-	-	-	-	-
	TransVG* [7]	81.0	82.7	78.4	64.8	70.7	56.9	68.7	67.7
	UNITER* [6]	81.4	87.0	74.2	75.9	81.5	66.7	74.0	68.7
	VILLA* [9]	82.4	87.5	74.8	76.2	81.5	66.8	76.2	76.7
	UniTAB* [36]	86.3	88.8	80.6	78.7	83.2	69.5	80.0	80.0
	G-DINO-L* [23]	90.6	93.2	88.2	82.8	89.0	75.9	86.1	87.0
LMM (pix2seq)	VisionLLM-H [32]	-	86.7	-	-	-	-	-	-
	Shikra-7B [4]	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2
	Shikra-13B [4]	87.8	91.1	81.8	82.9	87.8	74.4	82.6	83.2
LMM (pix2emb)	NExT-Chat-7B (ours)	85.5	90.0	77.9	77.2	84.5	68.0	80.1	79.8

Table 4. **Region Captioning**: comparison between our NExT-Chat and baselines on RefCOCOg.

Methods	RefCOCOg	
	CIDEr	METEOR
GRIT [35]	71.6	<b>15.2</b>
Kosmos-2 [27] (0-shot)	60.3	12.2
Kosmos-2 [27] (2-shot)	62.2	13.8
Kosmos-2 [27] (4-shot)	62.3	14.1
ASM [33]	41.9	13.6
NExT-Chat (ours)	<b>79.6</b>	12.0

single-turn and concise response generation, our NExT-Chat model also possesses the capability for generating detailed explanations in response to given questions. As illustrated in the third example of Fig. 8, our model exhibits the ability to infer the occupation of the man in the image by analyzing contextual cues such as his uniform and the horse he is riding. This inference is supported by the model’s ability to localize relevant regions within the image. Furthermore, for each hypothesis regarding the man’s occupation, our model provides detailed descriptions of the potential duties associated with that occupation. This showcases the model’s capacity for nuanced reasoning and comprehensive explanation generation.

## 5. Comparison with SOTAs

In this study, we evaluate our NExT-Chat model by comparing it with current state-of-the-art (SOTA) models on

various tasks including image-level hallucination diagnose (POPE dataset [18]), referring detection, referring segmentation and region-level captioning (RefCOCOg).

### 5.1. Hallucination

**Experimental Setup.** For a comprehensive evaluation, we benchmarked our NExT-Chat model against current state-of-the-art (SOTA) LMMs including Shikra [4], Instruct-BLIP, MiniGPT-4 [46], LLaVA [22], MM-GPT [10] and mPLUG-OWL [39] on the POPE dataset [18].

**Results.** The results, presented in Table 1, demonstrate that our NExT-Chat model exhibits competitive performance compared with existing SOTA models. Notably, our model achieves the the best performance for the random and popular splits and achieve the second best performance of the adversarial split. These findings indicate that our NExT-Chat model is competent in generating accurate responses, thus positioning it among the top-performing models in the field.

### 5.2. Referring Expression Segmentation

**Experimental Setup.** To rigorously assess our model’s proficiency in generating segmentation masks guided by natural language instructions, we use the referring expression segmentation (RES) splits of RefCOCO, RefCOCO+, and RefCOCOg. As for baselines, we choose both the LMM based method (LISA [14]) and non-LMM based methods including MCN [24], VLT [8], CRIS [34], LAVT [37], GRES [19], X-Decoder [48] and SEEM [49]. cIoU metric is employed to evaluate different methods.

**Results.** As demonstrated in Table 2, NExT-Chat exhibits superior or comparable cIoU scores relative to all baseline models. In comparison with non-LMM based methods, our approach consistently achieves either the highest or second-highest performance across various dataset splits, with the sole exception being the RefCOCO+ val set. Against LMM-based methods, specifically the LISA-7B model, NExT-Chat demonstrates enhanced performance in six dataset splits, notably achieving a substantial 4.5-point improvement in the RefCOCO+ testA split. It is noteworthy that NExT-Chat is trained with a significantly smaller dataset, comprising only 127k object segmentation masks, in stark contrast to baselines such as LISA, which utilize datasets more than an order of magnitude larger. These results underscore the efficiency of our training paradigm in substantially reducing the dependency on extensive and costly segmentation annotation datasets.

### 5.3. Referring Expression Comprehension

**Experimental Setup.** In addition to the segmentation ability, we also validate the detection ability of our method. Concretely, we adopt the REC splits of RefCOCO, RefCOCO+, and RefCOCOg. As for baselines, we first include the LMM method (pix2seq): VisionLLM-H [32], and Shikra [4]. We also include the non-LLM based methods: MAttNet [41], OFA-L [31], UniTab [36], G-DINO-L [23] and etc, where the models with \* mark in the Table 3 refer to the specialist and fine-tuned methods.

**Results.** First of all, our NExT-Chat can achieve excellent REC results and can even beat a series of fine-tuned methods like VILLA [9], UNITER [6] and TranVG [7] on all of the splits. There is also an interesting phenomenon that our NExT-Chat is slightly lower than Shikra-7B even with a similar data recipe for detection training. We hypothesize the reasons are that: (1) it is difficult to seek a perfect balance between the LM loss and localization loss, where the pix2seq methods do not suffer from this problem. (2) LLM is not pre-trained on the regression tasks and will potentially increase the training difficulty. However, we believe that incorporating the regression tasks in the LMM will be necessary, especially for targets like embodied AI.

### 5.4. Region Caption

**Experiment Setup.** In addition to the region output, we also validate the model’s ability of taking regions as input. The RefCOCOg is adopted, where each model is asked to describe the given region. The CIDEr and METEOR are applied as the evaluation metrics. For the baselines, we choose GRIT [35], Kosmos-2 [27] and ASM [33].

**Results.** As shown in Fig. 4, our model is capable of achieving the best performance on CIDEr across all of the baselines, which shows superiority of our NExT-Chat. Espe-

cially for Kosmos-2, we can beat the version with 4-shot examples.

## 6. Conclusion

In this paper, we present a novel location modeling method called pixel2emb, which utilizes embeddings to achieve multiple location output formats, such as bounding boxes and segmentation masks. Through comprehensive exploratory experiments, we demonstrate the effectiveness of the proposed pix2emb method. Additionally, we train a LMM named NExT-Chat, which significantly broadens the range of application scenarios for LMMs. Our NExT-Chat exhibits the ability to handle diverse tasks, including visual grounding, region captioning, grounded captioning and complex question reasoning. In the future, we will continue to enhance the model’s ability on conducting better detection and segmentation. Another promising direction is to extend the NExT-Chat model to multimodal agent which can handle complex tasks that requires region understanding.

## 7. Limitation

In the training procedure, our dataset primarily comprises individual image inputs, resulting in a limitation of our NExT-Chat model when it comes to handling multiple image inputs. Furthermore, the absence of sufficient training data from diverse domains hinders the model’s ability to generate accurate predictions in tasks involving medical and satellite image analysis.

## Author Contributions

Ao Zhang initializes the project, conducts experiments and writes the main part of the paper. Wei Ji and Yuan Yao proof read the paper. Tat-Seng Chua, Zhiyuan Liu and Yuan Yao provides valuable suggestions on the paper structure, experiment design and paper revision.

# Visual Grounding

 <p>Where is the bear wearing the red decoration in &lt;image&gt;?</p> <p>Answer: it's [0].</p> 	 <p>Where is the leftmost bear in &lt;image&gt;?</p> <p>Answer: it's [0].</p> 
 <p>Where is the rightmost bear in &lt;image&gt;?</p> <p>Answer: it's [0].</p> 	 <p>Where is the second bear from left to right in &lt;image&gt;?</p> <p>Answer: it's [0].</p> 
 <p>Where is the skateboard in &lt;image&gt;?</p> <p>Answer: it's [0].</p> 	 <p>Where is the sky in &lt;image&gt;?</p> <p>Answer: it's [0].</p> 

Figure 4. Visual grounding examples of NExT-Chat.



## Complex Grounding

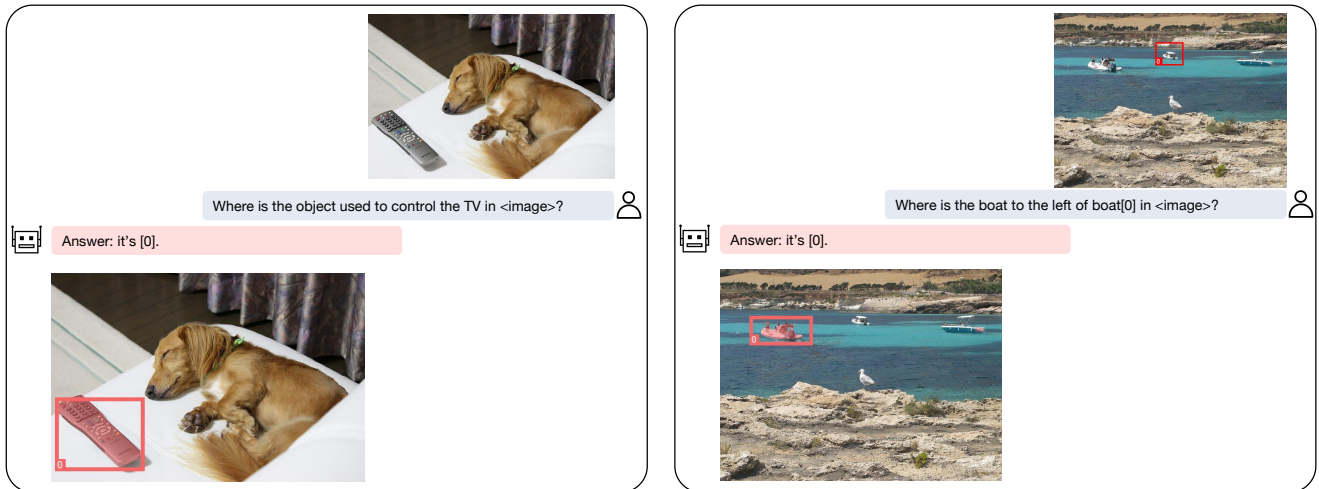



Figure 5. Hard visual grounding examples of NEX-T-Chat.

## Region Captioning





Figure 6. Region captioning examples of NEX-T-Chat.

# Grounded Captioning



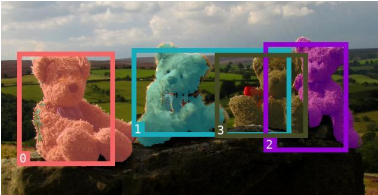

What is the woman doing? Please include object locations.

The woman [0] is putting white powder [1] on a cake [2] .



Can you provide a description of the image <image> and include the locations for each mentioned object?

Four stuffed teddy bears of various colors [0] [1] [2] [3] are sitting on a rock outside .



What can you tell me about the contents of the image <image>? Please indicate the positions of any mentioned objects.

A white dog [0] is laying on a marble table top next to a cup of coffee [1] .

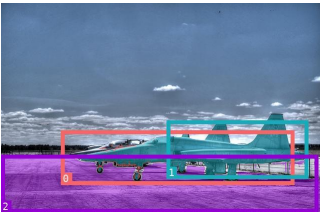

Describe the scene in this picture <image> and give the position info for each item you talk about.

A man [0] in a black jacket [1] is eating .

Can you provide a description of the image <image> and include the locations for each mentioned object?

Two fighter jets [0] [1] are parked on a runway [2] .

What can you tell me about the contents of the image <image>? Please indicate the positions of any mentioned objects.

A brown and white dog [0] is sleeping on a white couch [1] with a television remote [2] on it .

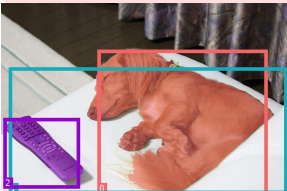


Figure 7. Grounded captioning examples of NEXt-Chat.

## Reasoning

What will the boy do in reaction to the frisbee? Please include object locations and explain.

The boy [0] will be excited and enthusiastic about the frisbee as he is a young boy who is about to catch it. He will try to catch it mid-air and will be happy about his successful catch. Once he catches the frisbee, he will throw it back to the person from whom he threw it in the first place. This will create a fun and engaging recurring play between the boy and the person. The boy will enjoy the game and the healthy competition that comes with playing frisbee.

What is the potential job of the man? Please include object locations and explain.

Based on the image, the man is wearing a uniform [0] and is riding a brown horse [1]. It is likely that the man is a police officer or serves a law enforcement role, as he is wearing a uniform, riding a horse, and possibly patrolling the area. His potential job could include patrolling neighborhoods, responding to calls or incidents, attending community events, or ensuring public safety. The man on the horse may also be a mounted patrol officer, in charge of protecting a specific area or event, watching over the public, or providing support to other law enforcement officers.

Figure 8. Reasoning examples of NEXt-Chat.

## References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 4
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3
- [4] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multi-modal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 1, 2, 4, 6, 7
- [5] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021. 1, 2
- [6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 6, 7
- [7] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021. 6, 7
- [8] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021. 5, 6
- [9] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020. 6, 7
- [10] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023. 6
- [11] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 1, 2
- [12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3
- [13] 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. *International journal of computer vision*, 123:32–73, 2017. 4
- [14] 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, 5, 6
- [15] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyu Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. 2023. 2
- [16] 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
- [17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2
- [18] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 6
- [19] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23592–23601, 2023. 5, 6
- [20] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusion-bench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2023. 2
- [21] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 2
- [22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2, 6
- [23] 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. 6, 7
- [24] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020. 5, 6
- [25] Arjun Mani, Nobline Yoo, Will Hinthorn, and Olga Russakovsky. Point and ask: Incorporating pointing into visual question answering. 2020. 4



- [26] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of CVPR*, pages 11–20, 2016. 2, 4
- [27] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 1, 2, 6, 7
- [28] 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 *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 4
- [29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3
- [30] Hamid Rezaatoughi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 3
- [31] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 6, 7
- [32] Wenhui Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, and Jifeng Dai. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023. 1, 2, 6, 7
- [33] Weiyun Wang, Min Shi, Qingyun Li, Wenhui 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. 6, 7
- [34] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 5, 6
- [35] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022. 6, 7
- [36] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pages 521–539. Springer, 2022. 2, 6, 7
- [37] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022. 5, 6
- [38] Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Pevl: Position-enhanced pre-training and prompt tuning for vision-language models. *arXiv preprint arXiv:2205.11169*, 2022. 2
- [39] 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. 6
- [40] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Proceedings of ECCV*, pages 69–85. Springer, 2016. 2, 4
- [41] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mtnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018. 6, 7
- [42] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4
- [43] Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. Transfer visual prompt generator across llms. *arXiv preprint arXiv:2305.01278*, 2023. 1, 2
- [44] 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
- [45] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. 3
- [46] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2, 6
- [47] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images, 2016. 4
- [48] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023. 5, 6
- [49] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023. 5, 6