

---

# 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>

## 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 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 the 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 (71.3) 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 (Huang et al., 2023; Zhu et al., 2023; Zhang et al., 2023a; Li et al., 2023c) that can describe the visual content and answer visual questions. However, these LMMs are limited to holistic image understanding

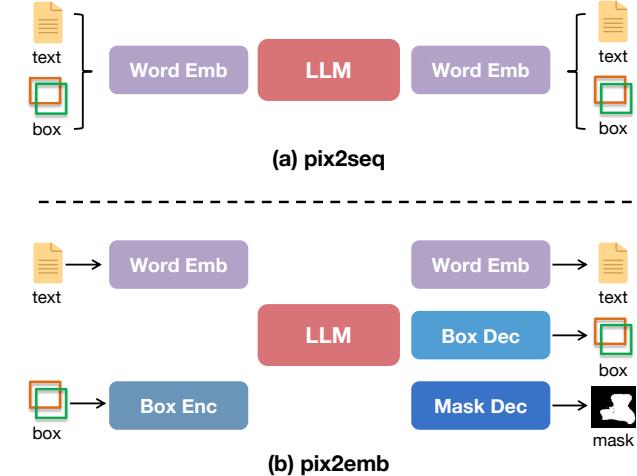


Figure 1. (a) **pix2seq** is the pixel-to-sequence paradigm, which converts bounding boxes into text. (b) **pix2emb** is the pixel-to-embedding paradigm proposed in this paper, which tries to unify all different location formats with embeddings.

without the ability to conduct region-level reasoning, for example, locating the referred objects in the conversation.

Aiming at region-level understanding, some initial works have focused solely on incorporating location data as inputs (Zhang et al., 2023b; Chen et al., 2023a) or outputs (Lai et al., 2023) in LMMs. Nevertheless, these approaches fall short of integrating both location input and output processing within a unified LMM framework. The primary solution (Peng et al., 2023; Wang et al., 2023a; Chen et al., 2023b) for unifying both the location input and output is the pix2seq paradigm (Fig. 1 (a)), where the object coordinates are converted to LLM understandable text tokens (e.g.,  $[x_1, y_1, x_2, y_2]$ ). Consequently, LMMs can generate object coordinates as part of their standard next token prediction process. However, the pix2seq paradigm is limited to discrete coordinate outputs and struggles to provide pixel-level outputs (i.e., segmentation masks), which significantly hinders real-world applications.

To address the limitations, we propose the pix2emb paradigm (Fig. 1), which can accept both locations as inputs and outputs, and accommodate different location formats. The key idea is to model all location information as embeddings, which can be generated from and decoded into different location formats with corresponding encoders

<sup>1</sup>National University of Singapore <sup>2</sup>Tsinghua University. Correspondence to: Ao Zhang <aozhang@u.nus.edu>, Yuan Yao <yaoyuanthu@gmail.com>.

and decoders. Specifically, we introduce a new token, `<trigger>`, which serves as a trigger for localization. During the text generation, the `<trigger>` triggers the location decoding. Then, the hidden states of `<trigger>` can be used for both detection and segmentation, as depicted in Fig. 2. The predicted or provided object location will be encoded back into the embedding for the object referring in the generation. However, the location encoding can only be trained indirectly through the captioning loss, and the location input embedding space is unaligned with the output embedding space. Thus, we propose a cycle consistency loss to align the location input and output embedding space, where the location input ability can also benefit from the location output training.

Building upon the proposed pix2emb method, we introduce a new LMM named NExT-Chat that seamlessly integrates chat, detection, and segmentation capabilities into a single unified model. NExT-Chat is trained in a multi-stage manner, which initially involves training on conversational data with bounding boxes, and then followed by segmentation data. This multi-stage training enables us to take full advantage of the rich and high-quality data (conversation with bounding boxes) accumulated from the pix2seq practices, and then extend to the segmentation with very limited annotations. Benefiting from the multi-stage training, our NExT-Chat can even achieve better performance than baselines using an order of magnitude larger mask annotations (e.g., LISA (Lai et al., 2023)).

To show the effectiveness of our NExT-Chat, 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 cloU of 71.3, 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:

- *Pix2emb 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, detection and segmentation into a single LMM.
- *Comprehensive Experiments.* We provide abundant qualitative and quantitative results to showcase the effectiveness of our proposed method.

## 2. Related Works

**Large multimodal model (LMM).** LMMs are typically built on large language models (LLMs) with visual per-

ception modules, which can generate captions or answer questions based on the given multimodal content. Flamingo (Alayrac et al., 2022) extracts vision information with a pre-trained vision backbone, and incorporate them into the text features with cross-attention layers. Differently, BLIP-2 (Li et al., 2023c) and Kosmos (Huang et al., 2023) directly feed the visual features into the LLMs as soft prompts. Following BLIP-2, MiniGPT-4 (Zhu et al., 2023) and VPGTrans (Zhang et al., 2023a) build LMMs with transfer learning, and significantly reduce the training cost. When considering the training paradigm, researchers find that instruction tuning can better align the LMM with the expected output format. MiniGPT-4 (Zhu et al., 2023) 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 (Liu et al., 2023c) generates the instruction tuning data with the text-only GPT-4 models by feeding the visual information as text sentences. Otter (Li et al., 2023b;a) further propose a MIMIC-IT dataset that can turn the LMM into better in-context learners. However, these LMMs (Alayrac et al., 2022; Liu et al., 2023b;c) can only take the whole image/video as input and output text, and are incapable of handling region understanding tasks.

**LMM for Region Understanding.** GPT4ROI (Zhang et al., 2023b) proposes to encode the regions as features for location input scenarios. LISA (Lai et al., 2023) proposes to combine the LLM with the SAM for object segmentation. However, they are either limited to the location input or single object segmentation. Pix2seq (Chen et al., 2021) first represents object bounding box coordinates as text tokens and thus the LM can take the object locations as both the input and output. Take Kosmos-2 as an example, it discretizes the whole image into  $32 \times 32$  bins, where the bin’s id represents the points lying in it. Shikra (Chen et al., 2023b) points out that introducing too many new tokens will inevitably increase the training difficulties. Thus, Shikra proposes 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 many new tokens, it requires roughly 26 tokens to represent each bounding box, which is ineffective. Different from these works, our NExT-Chat introduces the pix2emb paradigm that can both accept the location inputs and outputs and accomodate different location formats (e.g.bounding boxes and masks). More discussion is in Appendix A.

## 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

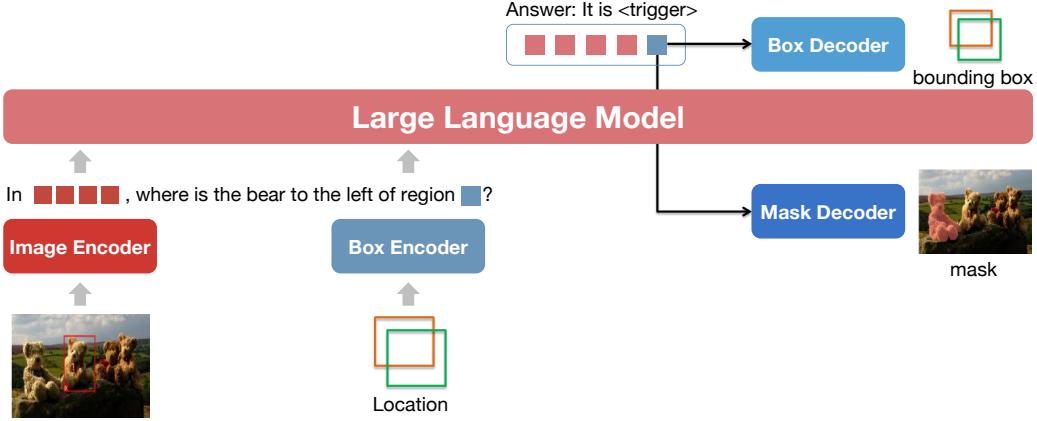


Figure 2. The overall framework of NExT-Chat. The image and given bounding boxes are encoded by image and box encoders respectively. During decoding, the hidden states of the `<trigger>` are fed into box and mask decoders, enabling object detection and segmentation.

(§3.2). Finally, we introduce the multi-stage training (§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 (Radford et al., 2021) 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, which can be considered 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 (Zheng et al., 2023).

### 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 passed to the *Box Decoder*  $\mathcal{F}$  for coordinates regression. Mathematically, the process can be expressed as:

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

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 (Rezatofighi et al., 2019) during training:

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

where  $\mathbf{b}_{gt}$  is the ground truth coordinates, and  $\alpha_d = 2$ ,  $\beta_d = 0.8$  follows the ratio in DETR (Carion et al., 2020).

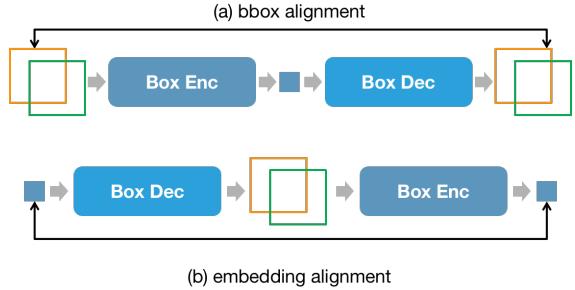


Figure 3. Cycle consistency loss to bind the location encoder and decoder training, which can benefit the location input tasks.

**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 (Lai et al., 2023), we employ SAM (Kirillov et al., 2023) as our mask head, which additionally takes the original image as input. To ensure the compatibility between the LMM’s hidden states and SAM, we project the hidden states to match the dimension of SAM’s prompt embedding using a linear projector, which will be fed into SAM to provide location information. Some astute readers may notice that the detected bounding boxes can also serve as SAM’s prompts. We conduct an ablation study in Table 6 and empirically find that using only the hidden states can achieve the best performance. To train the mask output, we follow the loss in LISA (Lai et al., 2023):

$$\mathcal{L}_{seg} = \alpha_s \text{BCE}(\mathbf{m}, \mathbf{m}_{gt}) + \beta_s \text{D}(\mathbf{m}, \mathbf{m}_{gt}), \quad (3)$$

where BCE and D are BCE Loss and Dice Loss separately.  $\alpha_s$  and  $\beta_s$  are 2 and 0.5 respectively.

**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.

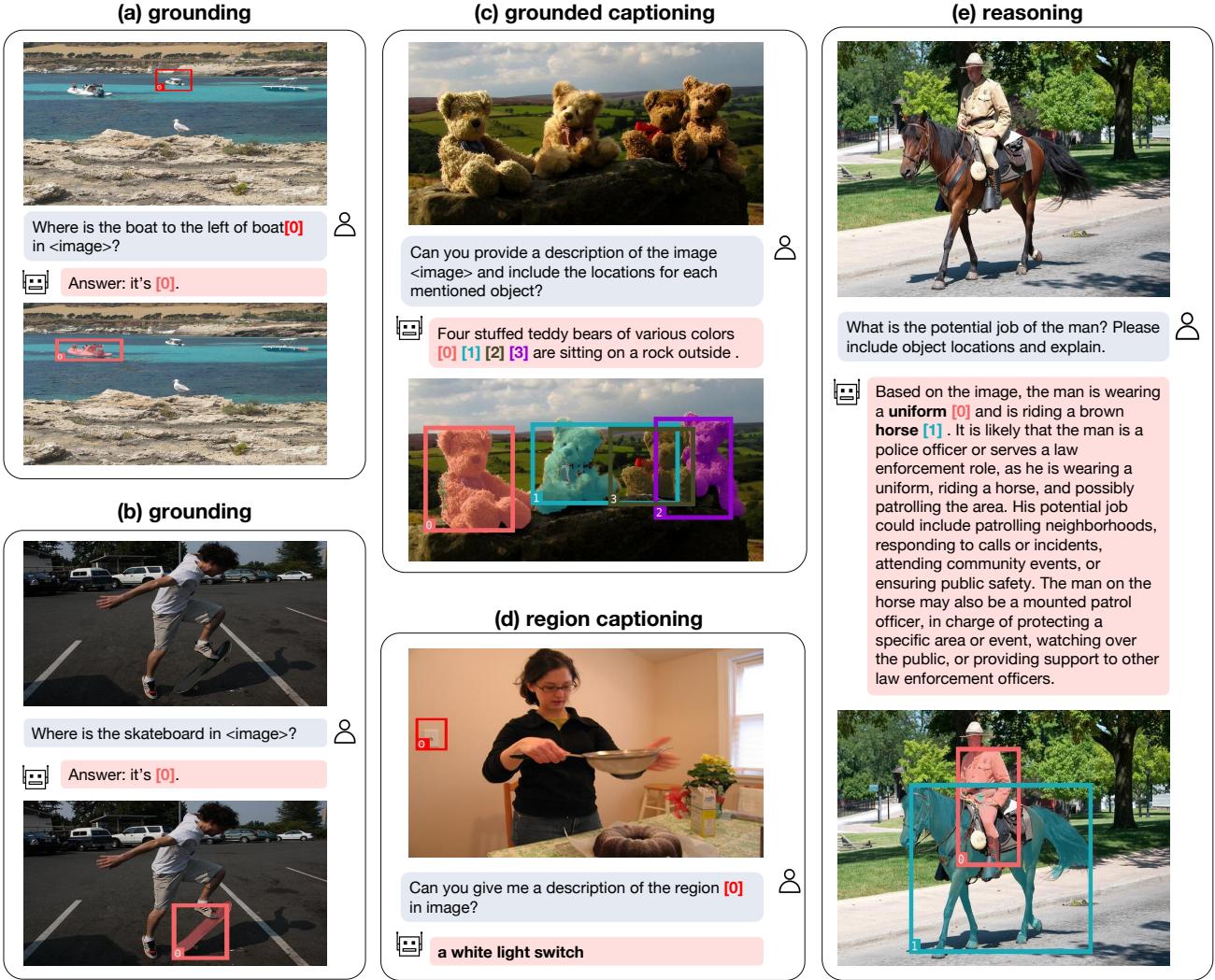


Figure 4. Qualitative examples of NExT-Chat, which shows the model’s capability of (a)&(b) visual grounding, (c) grounded captioning, (d) region captioning, and (e) region-aware reasoning.

Specifically, a 2-layer MLP is introduced as the location encoder  $\mathcal{G}$ . The location encoder takes the bounding box  $b$  as input and converts them into embeddings  $t \in \mathbb{R}^n$ . Due to the inconvenience of drawing masks by humans users, we do not additionally train a mask encoder in our NExT-Chat. If there are special scenarios that require mask input, we convert them into boxes. To train the location encoder, a straightforward way is to ask the LMM to make region descriptions, and thus the location encoder can be supervised through the standard text generation loss  $\mathcal{L}_{text}$ .

However, we observe that the location encoder cannot be effectively trained solely through  $\mathcal{L}_{text}$ . Different from the location decoders that can be directly trained with  $\mathcal{L}_{det}$  or  $\mathcal{L}_{seg}$ , the supervision from the  $\mathcal{L}_{text}$  is indirect and constrained by the amount of the location input data. Inspired by

the weight tying in LLMs’ word embedding layers and token classification layers, we introduce a cycle consistency loss to align the location input and output’s embedding space. Therefore, the location encoder and location decoder can benefit from each other’s training. As illustrated in Fig. 3 (a), a bounding box will be encoded and then decoded. Two bounding boxes are asked to be the same. Similarly, the hidden states of  $\langle trigger \rangle$  will also be used for cycle loss (Fig. 3 (b)). Formally, the  $\mathcal{L}_{cyc}$  can be defined as:

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

where  $\mathbf{b}$  is the provided bounding box and  $\mathbf{t}$  is the predicted embedding.  $\mathcal{L}_1$  and  $\mathcal{L}_2$  correspond to the L1 loss and L2 loss respectively.

**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 ( $\uparrow$ )	<b>87.70</b>	86.90	88.57	79.67	50.37	50.10	53.97
	Precision ( $\uparrow$ )	93.46	<b>94.40</b>	84.09	78.24	50.19	50.05	52.07
	Recall ( $\uparrow$ )	<b>81.87</b>	79.27	95.13	82.20	99.13	100.00	99.60
	F1-Score ( $\uparrow$ )	<b>87.28</b>	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 ( $\uparrow$ )	<b>84.57</b>	83.97	82.77	69.73	49.87	50.00	50.90
	Precision ( $\uparrow$ )	86.54	<b>87.55</b>	76.27	65.86	49.93	50.00	50.46
	Recall ( $\uparrow$ )	<b>81.87</b>	79.20	95.13	81.93	99.27	100.00	99.40
	F1-Score ( $\uparrow$ )	<b>84.14</b>	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 ( $\uparrow$ )	81.93	<b>83.10</b>	72.10	65.17	49.70	50.00	50.67
	Precision ( $\uparrow$ )	82.02	<b>85.60</b>	65.13	61.19	49.85	50.00	50.34
	Recall ( $\uparrow$ )	<b>81.80</b>	79.60	95.13	82.93	99.07	100.00	99.33
	F1-Score ( $\uparrow$ )	81.91	<b>82.49</b>	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

### 3.3. Training Process

We design a three-stage training for the NExT-Chat. Stage-1 is to train the model for simple conversations with bounding boxes. In stage-2, we further enhance the model’s conversation ability. In stage-3, the output format will be extended to segmentation masks with lightweight training.

**Stage-1.** In stage-1, we perform pre-training using a mixture of data from various sources, including Flickr30K Entities (Plummer et al., 2015), Visual Genome (Krishna et al., 2017), RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016), RefCOCOg (Mao et al., 2016), VQAv2 (Antol et al., 2015), PointQA (Mani et al., 2020), Visual7W (Zhu et al., 2016), and VCR (Zellers et al., 2019). 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, box encoder, and decoder are 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 the NExT-Chat 7B model, the stage-1 training takes 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, VCR and Shikra-RD (Chen et al., 2023b). The batch size is set to 64, and the learning rate is set to 2e-5. The loss is the same as stage-1’s loss. For the NExT-Chat 7B model, the stage-2 training uses 8 A100 (80G) GPUs for around 10 hours.

**Stage-3.** After two stages of training, the model has already

been equipped with the conversation ability and the region modeling ability. In stage-3, we only need to extend the bounding box output ability to the mask output. Specifically, we train the linear projector between the LMM and SAM, as well as the decoder of SAM. The loss for the stage-3 is:

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

To prevent catastrophic forgetting, we keep other parameters frozen during the segmentation training. 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 series datasets.

## 4. Qualitative Results

In this section, we present qualitative results that showcase the capabilities of our NExT-Chat model across various scenarios. More results are included in the Appendix E.

**Visual Grounding.** First of all, NExT-Chat is capable of grounding objects with given queries. As shown in Fig. 4 (a), it can accurately locate the skateboard based on the text description. In addition to the text-only prompt, our NExT-Chat can also accept the location input in visual grounding. For example, it can locate the boat to the left of the middle boat marked by a bounding box in Fig. 4 (b).

**Grounded Captioning.** Another compelling application of our NExT-Chat model is its ability to describe the visual content with grounded objects in the images. Fig. 4 (c) demonstrates that our model can identify the major 4 bears in the image, and effectively organize them into a coherent

**Table 2. RES:** comparison between our NExT-Chat and baselines on RES. The evaluation metric is **cIoU**. **ft** refers to a task-specific fine-tuning for LMM based models. **box+SAM** is the result of using NExT-Chat predicted box and SAM without stage-3 training.

Type	Methods	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
non-LMM	MCN (Luo et al., 2020)	62.4	64.2	59.7	50.6	55.0	44.7	49.2	49.4
	VLT (Ding et al., 2021)	67.5	70.5	65.2	56.3	61.0	50.1	55.0	57.7
	CRIS (Wang et al., 2022b)	70.5	73.2	66.1	65.3	68.1	53.7	59.9	60.4
	LAVT (Yang et al., 2022b)	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
	GRES (Liu et al., 2023a)	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
	X-Decoder (Zou et al., 2023a)	-	-	-	-	-	-	64.6	-
	SEEM (Zou et al., 2023b)	-	-	-	-	-	-	65.7	-
	PolyFormer (B) (Liu et al., 2023d)	74.8	76.6	71.1	67.6	72.9	59.3	67.8	69.1
LMM	PolyFormer (L) (Liu et al., 2023d)	76.0	78.3	73.3	69.3	74.6	61.9	69.2	70.2
	LISA-7B (Lai et al., 2023)	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5
	LISA-7B (ft) (Lai et al., 2023)	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6
	GLaMM (Rasheed et al., 2023)	54.7	58.1	52.2	42.5	47.1	39.5	54.8	55.6
	GLaMM (ft) (Rasheed et al., 2023)	78.3	81.5	74.4	68.0	75.7	61.8	72.5	72.0
	NExT-Chat (box+SAM)	69.4	75.9	63.5	61.2	69.0	52.3	62.4	64.0
	NExT-Chat	76.9	80.5	72.4	67.6	73.7	59.4	69.5	70.3
	NExT-Chat (ft)	<b>80.3</b>	<b>82.4</b>	<b>76.1</b>	<b>73.5</b>	<b>78.5</b>	<b>66.0</b>	<b>74.8</b>	<b>75.3</b>

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

**Region Captioning.** NExT-Chat can also accept the location as input and describe the region’s content. Note that, existing image-level LMMs have already been able to generate captions for the main objects in the image. Thus, to evaluate our model’s region captioning ability, we ask the model to describe a small light switch in the background, which is generally ignored by the image-level LMMs. As depicted in Fig. 4 (d), our model can accurately describe the input as a white light switch, which demonstrates the effectiveness of our model in generating region-based captions.

**Reasoning.** Thanks to the incorporation of LLM, our model is also capable of answering questions with region-aware explanations. As shown in Fig. 4 (e), given the question “what is the potential job of the man?”, our model can not only hypothesize the man’s job as a police officer but also ground the visual evidence which is the man’s uniform and horse. The example showcases the model’s capability for nuanced reasoning and the ability to link the mentioned objects with the regions in the images.

## 5. Experiment

In this section, 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, referring expression segmentation (RES), referring expression comprehension (REC), and region captioning.

### 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 (Chen et al., 2023b), InstructBLIP (Dai et al., 2023), MiniGPT-4 (Zhu et al., 2023), LLaVA (Liu et al., 2023c), MM-GPT (Gong et al., 2023) and mPLUG-OWL (Ye et al., 2023) on the POPE dataset (Li et al., 2023d).

**Results.** The results, presented in Table 1, demonstrate that our NExT-Chat exhibits competitive performance compared with existing SOTA models. Notably, our model achieves the best performance for the random and popular splits and achieves the second-best performance for 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 ex-

**Table 3. REC:** comparison between our NExT-Chat and baselines on REC. The evaluation metric is **Acc@0.5**.

Type	Methods	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
non-LMM	MAttNet (Yu et al., 2018)	76.4	80.4	69.3	64.9	70.3	56.0	66.7	67.0
	OFA-L (Wang et al., 2022a)	80.0	83.7	76.4	68.3	76.0	61.8	67.6	67.6
	OFASys	-	80.1	-	-	-	-	-	-
	TransVG (Deng et al., 2021)	81.0	82.7	78.4	64.8	70.7	56.9	68.7	67.7
	UNITER (Chen et al., 2020)	81.4	87.0	74.2	75.9	81.5	66.7	74.0	68.7
	VILLA (Gan et al., 2020)	82.4	87.5	74.8	76.2	81.5	66.8	76.2	76.7
LMM	UniTAB (Yang et al., 2022a)	86.3	88.8	80.6	78.7	83.2	69.5	80.0	80.0
	G-DINO-L (Liu et al., 2023e)	90.6	93.2	88.2	82.8	89.0	75.9	86.1	87.0
	Kosmos-2 (Peng et al., 2023)	52.3	57.4	47.3	45.5	50.7	42.2	60.6	61.7
	VisionLLM-H (Wang et al., 2023a)	-	86.7	-	-	-	-	-	-
	Shikra-7B (Chen et al., 2023b)	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2
	Shikra-13B (Chen et al., 2023b)	87.8	91.1	81.8	82.9	87.8	74.4	82.6	83.2
NExT-Chat-7B ( <b>ours</b> )		85.5	90.0	77.9	77.2	84.5	68.0	80.1	79.8

pression segmentation (RES) splits of RefCOCO, RefCOCO+, and RefCOCOg. As for baselines, we choose both the LMM based methods (LISA (Lai et al., 2023) and GLaMM (Rasheed et al., 2023)) and non-LMM based methods including MCN (Luo et al., 2020), VLT (Ding et al., 2021), CRIS (Wang et al., 2022b), LAVT (Yang et al., 2022b), GRES (Liu et al., 2023a), X-Decoder (Zou et al., 2023a), SEEM (Zou et al., 2023b) and PolyFormer (B/L) (Liu et al., 2023d). 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. Against LMM-based methods, specifically the LISA-7B and GLaMM, NExT-Chat demonstrates enhanced performance in six dataset splits, notably achieving a substantial 6.3 point improvement in the RefCOCO+ testA split than LISA. It is noteworthy that NExT-Chat is trained with a significantly smaller mask annotations, comprising only 127k object masks, in contrast to baselines such as LISA, which utilize mask annotations more than an order of magnitude larger.

We note that the LMM baselines like LISA and GLaMM also report the task-specific fine-tuned results in their paper. Following the practice, we also conduct a task-specific fine-tuning with the whole LLM parameters tuned. As depicted in Table 2, NExT-Chat (ft) can achieve the best performance compared with LISA (ft) and GLaMM (ft) across all of the data splits, which demonstrates the benefits of boosting the

segmentation ability with bounding box data.

### 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): Kosmos-2 (Peng et al., 2023), VisionLLM-H (Wang et al., 2023a), and Shikra (Chen et al., 2023b). We also include the non-LMM based methods: MAttNet (Yu et al., 2018), OFA-L (Wang et al., 2022a), UniTab (Yang et al., 2022a) and etc.

**Results.** First of all, our NExT-Chat can achieve excellent REC results and even beat a series of specifically fine-tuned non-LMM methods like VILLA (Gan et al., 2020), UNITER (Chen et al., 2020) and TranVG (Deng et al., 2021) on all of the splits. When compared with the LMM baselines, our NExT-Chat can outperform the Kosmos-2 and VisionLLM-H. Concretely, we can achieve a 3.3 higher Acc@0.5 than VisionLLM-H on RefCOCO testA. There is an interesting phenomenon that our NExT-Chat is slightly lower than Shikra-7B, which uses a similar data recipe for detection training. We hypothesize the reasons are that: (1) a fixed weight of the detection loss is sub-optimal and requires further exploration for a dynamic balance with the text loss. (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.

**Table 4. Region Captioning:** comparison between our NExT-Chat and baselines on RefCOCOg (google). ft indicates to a task-specific fine-tuning for LMMs.

Methods	RefCOCOg	
	CIDEr	METEOR
GRIT (Wu et al., 2022)	71.6	15.2
Kosmos-2 (Peng et al., 2023) (0-shot)	60.3	12.2
Kosmos-2 (Peng et al., 2023) (2-shot)	62.2	13.8
Kosmos-2 (Peng et al., 2023) (4-shot)	62.3	14.1
ASM (Wang et al., 2023b)	41.9	13.6
GLaMM (Rasheed et al., 2023)	104.0	15.7
GLaMM (ft) (Rasheed et al., 2023)	105.0	16.2
NExT-Chat	79.6	12.0
NExT-Chat (ft)	<b>114.0</b>	<b>17.4</b>

#### 5.4. Region Captioning

**Experiment Setup.** In addition to the location output, we also validate the model’s ability to take locations as input. The RefCOCOg (google) dataset 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 (Wu et al., 2022), Kosmos-2 (Peng et al., 2023), ASM (Wang et al., 2023b) and GLaMM (Rasheed et al., 2023).

**Results.** As shown in Table 4, our model is capable of achieving better CIDEr across all of the baselines except GLaMM, which shows the superiority of our NExT-Chat. It is worth noting that GLaMM includes the RefCOCOg (google) dataset in their model training and thus can achieve higher performance. Therefore, we also conduct a task-specific fine-tuning and report it as NExT-Chat (ft) in Table 4. After the fine-tuning, our NExT-Chat (ft) can achieve the best performance for both CIDEr and METEOR compared with all of the baselines, which shows the effectiveness of our region-modeling strategy.

## 6. Ablation Study

**The influence of  $\mathcal{L}_{cyc}$ .** We find that the cycle consistency loss  $\mathcal{L}_{cyc}$  is important for the location input tasks. To quantitatively evaluate its influence on the region captioning, we train the NExT-Chat on the RefCOCO series data for both referring expression comprehension and region captioning. Then, we compare the results with and without  $\mathcal{L}_{cyc}$ . As shown in Table 5, the model with the  $\mathcal{L}_{cyc}$  can achieve 68.7 for CIDEr and 11.3 for METEOR, while the model without  $\mathcal{L}_{cyc}$  can only achieve 65.1 for CIDEr and 10.9 for METEOR, which indicates the benefits of  $\mathcal{L}_{cyc}$  for the location input tasks.

**The influence of stage-3 training.** The SAM model can directly take the predicted bounding boxes as input and

**Table 5.** The influence of  $\mathcal{L}_{cyc}$  for region captioning.

Methods	RefCOCOg	
	CIDEr	METEOR
w/o $\mathcal{L}_{cyc}$	65.1	10.9
w $\mathcal{L}_{cyc}$	<b>68.7</b>	<b>11.3</b>

**Table 6.** The influence of different factors for RES. **Training** refers to the stage-3 training. **Box** indicates whether to feed the box into SAM. **Emb.** indicates whether to feed the `<trigger>`’s hidden states into SAM.

training	box	emb.	RefCOCO		
			RefCOCO	RefCOCO+	RefCOCOg
	✓		69.6	60.8	63.2
✓	✓		75.3	65.5	68.2
✓		✓	<b>76.6</b>	<b>66.9</b>	<b>69.9</b>
✓	✓	✓	76.1	66.5	69.5

output corresponding segmentation masks without our stage-3 training. To study the effect of stage-3 training, we show the non-training result (line 1) in Table 6. We find that the training can significantly improve the performance on RES task, which shows the necessity of the adaptation training.

**Embedding vs box as inputs for mask decoder.** The mask decoder can take either the bounding box or the `<trigger>` embedding as the input. To compare the influence of different inputs for mask decoder, we show results with the box (line 2), emb (line 3), and box&emb (line 4) inputs in Table 6. When comparing the box and emb., we find that the embedding input has an obvious superiority over the box, where changing from box input to embedding input can result in an over 1 point improvement on all of the 3 splits. Another interesting finding is that further combining the box and embedding will not improve the performance and even cause a slight degeneration. A potential explanation is that the location information has already been encoded in the embedding and the box can not provide any new information.

## 7. Conclusion

In this paper, we present a novel location modeling method called pix2emb, which utilizes embeddings to achieve multiple location output formats, such as bounding boxes and segmentation masks. Based on the pix2emb, we train an 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 region-aware reasoning. The model also achieves SOTA performance on a series of datasets.

## Impact Statement

NExT-Chat is a multimodal conversation model that can conduct region-level understanding. We try to discuss its potential risks and mitigation strategies. First of all, our model may suffer from the hallucination problem. Some generated content may not conform to the visual facts in the given images. Further alignment with high-quality data can be applied to alleviate the hallucination. Secondly, similar to the language model, our model may sometimes generate offensive content for users. If the model is applied in human conversation, filtration algorithms should be employed to avoid exposing offensive content to users. Moreover, our model is mainly trained on the general domain data and thus may be not accurate enough for scenarios like medical image processing. Domain-specific data fine-tuning will be helpful to build conversation robots for the target domain.

## References

- Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.
- Chen, C., Qin, R., Luo, F., Mi, X., Li, P., Sun, M., and Liu, Y. Position-enhanced visual instruction tuning for multimodal large language models. *arXiv preprint arXiv:2308.13437*, 2023a.
- Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., and Zhao, R. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023b.
- Chen, T., Saxena, S., Li, L., Fleet, D. J., and Hinton, G. Pix2seq: A language modeling framework for object detection. *arXiv preprint arXiv:2109.10852*, 2021.
- Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020.
- Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.
- Deng, J., Yang, Z., Chen, T., Zhou, W., and Li, H. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1769–1779, 2021.
- Ding, H., Liu, C., Wang, S., and Jiang, X. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16321–16330, 2021.
- Gan, Z., Chen, Y.-C., Li, L., Zhu, C., Cheng, Y., and Liu, J. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628, 2020.
- Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., Liu, K., Zhang, W., Luo, P., and Chen, K. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.
- Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O. K., Liu, Q., et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment anything. *arXiv:2304.02643*, 2023.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.
- Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., and Jia, J. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.
- Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Yang, J., Li, C., and Liu, Z. Mimic-it: Multi-modal in-context instruction tuning. 2023a.
- Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., and Liu, Z. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023b.
- Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023c.
- Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023d.

- Liu, C., Ding, H., and Jiang, X. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23592–23601, 2023a.
- Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023b.
- Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning, 2023c.
- Liu, J., Ding, H., Cai, Z., Zhang, Y., Satzoda, R. K., Makhaddevan, V., and Manmatha, R. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18653–18663, 2023d.
- Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023e.
- Luo, G., Zhou, Y., Sun, X., Cao, L., Wu, C., Deng, C., and Ji, R. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 10034–10043, 2020.
- Mani, A., Yoo, N., Hinthon, W., and Russakovsky, O. Point and ask: Incorporating pointing into visual question answering. 2020.
- Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., and Murphy, K. Generation and comprehension of unambiguous object descriptions. In *Proceedings of CVPR*, pp. 11–20, 2016.
- Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., and Wei, F. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.
- Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.
- Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R. M., Xing, E., Yang, M.-H., and Khan, F. S. Glamm: Pixel grounding large multimodal model. *ArXiv 2311.03356*, 2023.
- Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S. 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*, pp. 658–666, 2019.
- Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pp. 23318–23340. PMLR, 2022a.
- Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., and Dai, J. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023a.
- Wang, W., Shi, M., Li, Q., Wang, W., Huang, Z., Xing, L., Chen, Z., Li, H., Zhu, X., Cao, Z., et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. *arXiv preprint arXiv:2308.01907*, 2023b.
- Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., and Liu, T. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11686–11695, 2022b.
- Wu, J., Wang, J., Yang, Z., Gan, Z., Liu, Z., Yuan, J., and Wang, L. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022.
- Yang, Z., Gan, Z., Wang, J., Hu, X., Ahmed, F., Liu, Z., Lu, Y., and Wang, L. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *European Conference on Computer Vision*, pp. 521–539. Springer, 2022a.
- Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., and Torr, P. H. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18155–18165, 2022b.
- Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.
- Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. Modeling context in referring expressions. In *Proceedings of ECCV*, pp. 69–85. Springer, 2016.

Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., and Berg, T. L. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1307–1315, 2018.

Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Zhang, A., Fei, H., Yao, Y., Ji, W., Li, L., Liu, Z., and Chua, T.-S. Transfer visual prompt generator across llms. *arXiv preprint arXiv:2305.01278*, 2023a.

Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Chen, K., and Luo, P. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023b.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Zhu, Y., Groth, O., Bernstein, M., and Fei-Fei, L. Visual7w: Grounded question answering in images, 2016.

Zou, X., Dou, Z.-Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15116–15127, 2023a.

Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., and Lee, Y. J. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023b.

## A. Pix2Emb v.s. Pix2Seq

In this section, we compare our pix2emb and different pix2seq variants. Concretely, there are mainly 3 pix2seq variants:

- **4bin** represents bounding boxes using four bins in the format of  $[x_0, y_0, x_1, y_1]$ . Each bin corresponds to a specific location token out of the total 224 tokens available.
- **2bin** employs two bins, with each bin representing a point. The entire image is divided into 1024 discrete bins, and each bin represents the points within it.
- **num** does not introduce any new tokens to the vocabulary and directly uses the textual representation of numerical values with three decimal places. However, the tokens required for each bounding box will be 26. It may also disturb the language model’s ability for number-related generation.

As shown in Table 7, our pix2emb is the only method to model the location output as a regression task, which conforms to the nature of coordinates prediction. When considering the tokens for each box, we only need 2 tokens to represent a single bounding box, which significantly reduces the inference cost especially when compared with the **num**. As for the new vocabulary, our pix2emb only requires 1 additional new token, which reduces the burden of training the extra new parameters. Finally, our NExT-Chat can outperform both the VisionLLM and Kosmos-2 for the REC task. Although our NExT-Chat can not surpass Shikra, our pix2emb pattern can be 169 times more effective than **num** for processing a single bounding box considering the quadratic cost of LM’s self-attention calculation.

*Table 7.* Comparison between our pix2emb with 3 pix2seq variants including 4bin, 2bin, and num. **Location Output** is the way to predict the coordinates. **Format** is the bounding box format. **Tokes of Box** represents the number of tokens to represent a single bounding box. Note that “[” and “]” also require two tokens. Our pix2emb will use 2 tokens to represent a box consisting of a <trigger> token and the box’s embedding. **New Vocabulary** is the number of new tokens added to vocabulary for location modeling. **Representative Models** are representative models for the given location modeling method.

Method	Location Output	Format	Tokens of Box	New Vocabulary	Representative Model
pix2seq (4bin)	classification	$[x_0, y_0, x_1, y_1]$	6	224	VisionLLM
pix2seq (2bin)	classification	$[p_0, p_1]$	4	1024	Kosmos-2
pix2seq (num)	classification	$[x_0, y_0, x_1, y_1]$	26	0	Shikra
pix2emb (ours)	regression	<trigger> <emb>	2	1	NExT-Chat

## B. Additional Ablation Studies

In this section, we present some additional ablation studies.

**The influence of  $\mathcal{L}_{cyc}$  for the REC.** We find that the cycle consistency loss will not only benefit the location input task but also have a positive effect on the location output tasks. As shown in Table 8, the model trained with the  $\mathcal{L}_{cyc}$  can be consistently better than the non-added one, with over 2 points improvements across all of the three splits.

*Table 8.* The influence of  $\mathcal{L}_{cyc}$  for REC.

Methods	RefCOCO	RefCOCO+	RefCOCOg
w/o $\mathcal{L}_{cyc}$	59.8	45.3	49.6
w $\mathcal{L}_{cyc}$	<b>61.9</b>	<b>48.9</b>	<b>52.2</b>

**The influence of weight of  $\mathcal{L}_{det}$ .** We find that the balance of  $\mathcal{L}_{det}$  and  $\mathcal{L}_{text}$  is essential for the detection performance. We denote the weight of the  $\mathcal{L}_{det}$  is  $k$  in the stage-1 training. We ablate the REC performance with different  $k$  in Table 9 with around 10% data of the stage-1 pre-training and empirically observe that  $k = 1$  is the best.

Table 9. The influence of  $\mathcal{L}_{det}$ 's weight  $k$  for REC.

$k$	RefCOCO	RefCOCO+	RefCOCOg
0.5	74.6	64.6	68.4
1	<b>76.7</b>	<b>66.4</b>	<b>71.1</b>
2	74.6	62.8	67.8
5	71.1	58.0	63.3

## C. Detection v.s. Segmentation

Some astute authors may question why both the detection and segmentation abilities are preserved in the final model. Is the segmentation already enough for the localization? Our answer is no. When working towards multimodal agents, there are some scenarios requiring to indicate only the location rather than concrete objects or stuffs, which is more suitable to use bounding boxes. As shown in Fig. 5, we may ask the model to find a place to put the cake or predict where will be the ball after 1 second.

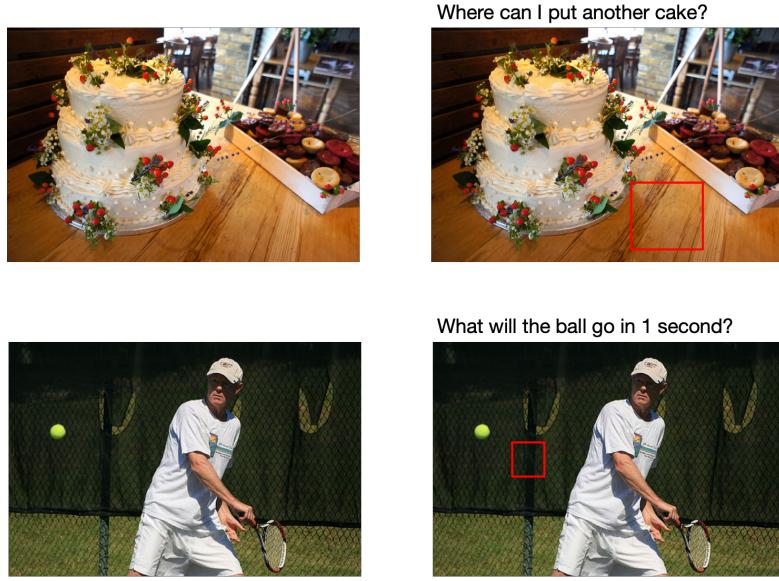


Figure 5. Two examples where the bounding boxes are more suitable than the masks.

## D. Limitations

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. We believe that further training the model with more diverse data and better alignment techniques will make it safer and stronger.

## E. Additional Qualitative Results

In this section, we additionally show some qualitative results of our NExT-Chat.

**Visual Grounding.** As shown in Fig. 6, 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 the given

queries. In addition, our NExT-Chat is also capable of locating the background stuff like the sky in the image.

**Complex Grounding.** Our model also showcases reasoning abilities through challenging grounding problems. For instance, in Fig. 7, our model accurately localizes the remote in response to the query “Where is the object to control the TV in the 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 location input, we conducted experiments where the model generates descriptions based on given bounding boxes. As depicted in Fig. 8, 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. 8, our model demonstrates the ability to accurately recognize and describe small objects such as flags, as well as background objects like trees.

**Grounded Captioning.** Fig. 9 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 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 examples of Fig. 10, our model exhibits the ability to infer the relationship between the man and the woman by analyzing contextual cues. This inference is supported by the model’s ability to localize relevant regions within the image. This showcases the model’s capacity for nuanced reasoning and comprehensive explanation generation.

## Visual Grounding



Figure 6. Visual grounding examples of NExT-Chat.

## Complex Grounding

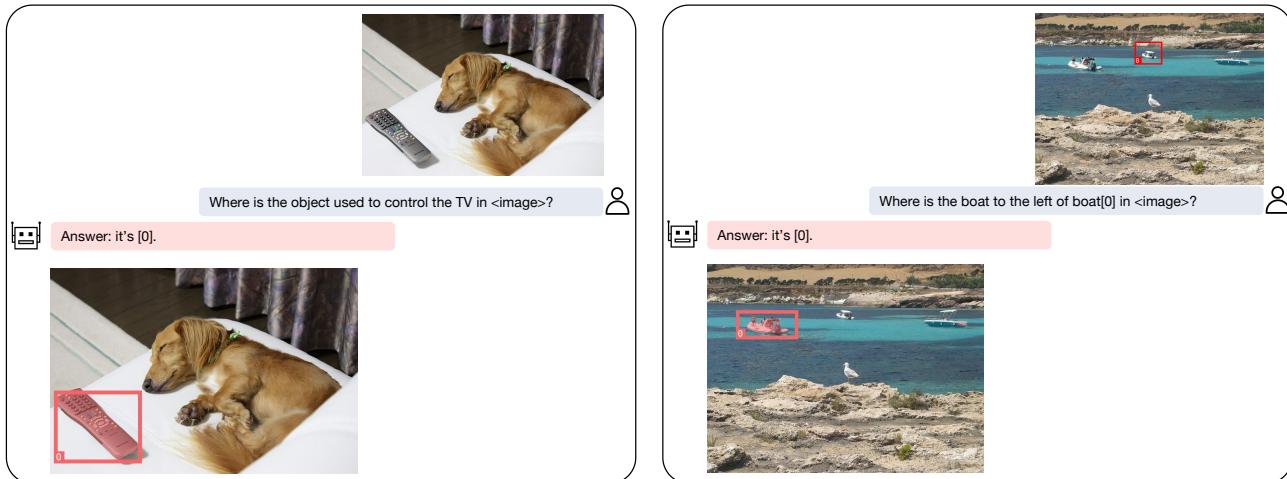


Figure 7. Hard visual grounding examples of NExT-Chat.

## Region Captioning



Figure 8. Region captioning examples of NExT-Chat.

## Grounded Captioning

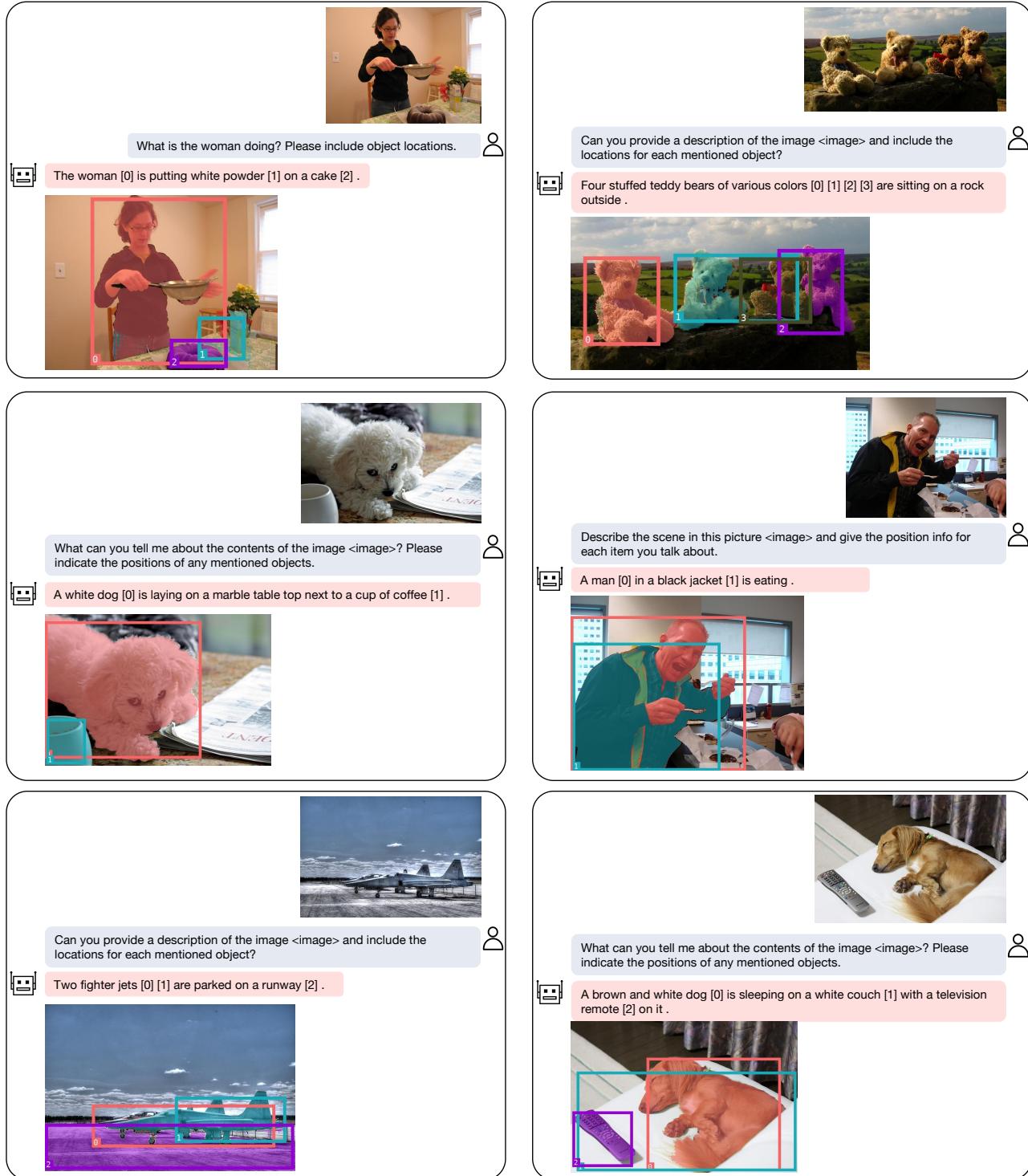


Figure 9. Grounded captioning examples of NExT-Chat.

## Reasoning



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

 Based on the image, it is suggested that the two people, a man [0] and a woman [1], are likely a couple or very close friends. They are both looking at a laptop computer [2], which could be related to work, study, or leisure activities. Their close interaction while looking at the laptop could indicate a shared interest or a shared task, which is typical of a professional or personal relationship, or simply spending time together for entertainment. However, it is important to note that this is only an suggestion based on the visual content, and the actual relationship between the two individuals is not definitively clear from the image alone





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 10. Reasoning examples of NExT-Chat.