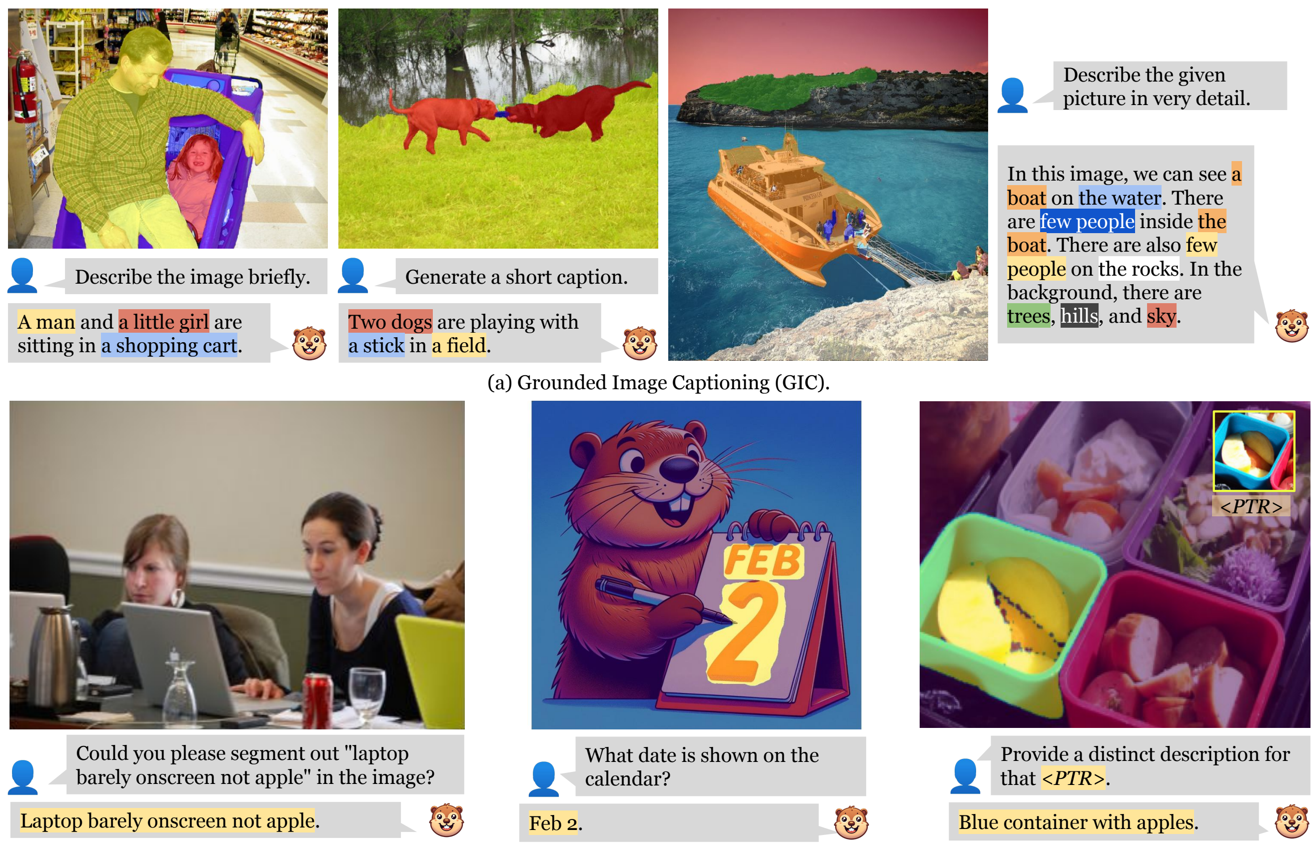
GROUNDHOG 🐹: Grounding Large Language Models to Holistic Segmentation

Yichi Zhang[1†], Ziqiao Ma[1†], Xiaofeng Gao[2], Suhaila Shakiah[2], Qiaozi Gao[2], Joyce Chai[1]
[1]University of Michigan, [2]Amazon AGI
zhangyic@umich.edu
https://groundhog-mllm.github.io/



(a) Grounded Image Captioning (GIC).

(b) Referential Expression Segmentation (RES). (c) Grounded Visual Question Answering (GVQA). (d) Referential Dialogue (RD).

Figure 1. We propose GROUNDHOG, a multimodal large language model that enhances its text output with pixel-level phrase grounding across diverse semantic granularities. The figure demonstrates outputs from our model on the four task types we considered in this work.

Abstract

Most multimodal large language models (MLLMs) learn language-to-object grounding through causal language modeling where grounded objects are captured by bounding boxes as sequences of location tokens. This paradigm lacks pixel-level representations that are important for fine-grained visual understanding and diagnosis. In this work, we introduce GROUNDHOG, *an MLLM developed by grounding Large Language Models to holistic segmentation.* GROUNDHOG *incorporates a masked feature extractor and converts extracted features into visual entity tokens for the MLLM backbone, which then connects groundable phrases to unified grounding masks by retrieving and merging the entity masks. To train* GROUNDHOG, *we carefully curated M3G2, a grounded visual instruction tuning dataset with Multi-Modal Multi-Grained Grounding, by harvesting a collection of segmentation-grounded datasets with rich annotations. Our experimental results show that* GROUNDHOG *achieves superior performance on various language grounding tasks without task-specific fine-tuning, and significantly reduces object hallucination.* GROUNDHOG *also demonstrates better grounding towards complex forms of visual input and provides easy-to-understand diagnosis in failure cases.*

† Work done during internship at Amazon AGI.

1. Introduction

Multimodal large language models (MLLMs) have received an increasing amount of attention to address tasks that necessitate non-linguistic knowledge, e.g., perception and reasoning about the visual world [39, 84]. For fine-grained visual understanding, grounded MLLMs often learn language-to-object grounding by causal language modeling, where grounded objects are captured by bounding boxes as sequences of location tokens. However, bounding boxes are insufficient in indicating amorphous stuff [5], semantic parts of objects [23], finer-grained regions with irregular shapes [26], or groups of instances at the same time. As a result, a single bounding box can often include other irrelevant semantics in order to engulf the target entities, leading to ambiguity in detection. In addition, the generated box coordinate lacks interpretability. When the model hallucinates, such as incorrectly predicting the association between objects and language, it is hard to diagnose whether the problem is due to the model's failure to detect the object, or its incorrect alignment of the object with language.

To address these issues, in this work, we introduce GROUNDHOG, an MLLM developed by grounding Large Language Models to holistic segmentation. Our goal of language grounding is to connect text spans that refer to or can be deduced from visual information, termed as *groundable phrases* [51], to their corresponding regions of visual entities. GROUNDHOG incorporates a masked feature extractor that takes an input image and a set of class-agnostic entity mask proposals, and converts each mask's features into visual entity tokens for an MLLM backbone. This MLLM then connects groundable phrases to unified grounding masks by retrieving and merging the entity masks. Compared to previous grounded MLLMs, GROUNDHOG unlocks unprecedented pixel-level vision-language alignment. It naturally supports visual pointers as input, and can plug-in-and-play with any choice of mask proposal networks, e.g., Segment Anything Model (SAM) [35], domain-specific semantic segmentation models, or user-provided mask candidates. We introduce an enhanced Mask2Former [10] as our default mask proposal network, which detects regions at multiple granularities, e.g., instances (things and stuff), semantic parts, and visual text, leading to a holistic coverage of visual semantics.

To train GROUNDHOG, we curated a Multi-Modal Multi-Grained Grounding (M3G2) dataset consisting of 2.5M text-image pairs for visually grounded instruction tuning, consisting of 36 sub-problems derived and augmented from 27 existing datasets. We present extensive experiments on vision-language tasks that require grounding, including grounded language generation with minimal object hallucination, language-guided segmentation, visual question answering with answer grounding, and referential dialog with spatial pointer inputs (Figure 1). Our empirical results show that GROUNDHOG, without task-specific fine-tuning, can achieve superior or comparable performance with previous models that either require fine-tuning or are specialized only for that dataset. In addition, GROUNDHOG has supports easy-to-understand diagnosis when grounding fails.

2. Our Method: GROUNDHOG

The language grounding task can be succinctly delineated into two fundamental components: *localization* and *recognition*, as established in the literature [51, 69, 92]. Such categorization not only aids in the identification of object presence (objectness) without reliance on specific object classes, but also sets the stage for models to be robust in open-vocabulary settings. Building upon this framework, we formulate the grounding process as an *entity segment selection* problem, which involves (1) proposing entity segmentation masks where the masks encapsulate regions with discernible semantic content, and (2) recognizing the retrieved entities through the understanding of both visual and language context. Concurrently performing both tasks is where MLLMs bring a distinct advantage. This decoupled design of entity mask proposal and language-guided grounding brings several advantages. First, it allows independent improvement of the mask proposal model and MLLM, where specialized data, training, and inference setups can be applied. Second, by decoupling language grounding, it becomes straightforward to determine if a failure is due to the model's inability to propose the entity segment, or its misalignment of the object with the language, thus improving the interpretability of the whole framework. Third, as shown later, when connecting the two parts to work in tandem in a model-independent manner, the MLLM can benefit from multiple different vision specialist models in a plug-and-play fashion. In the remainder of this section, we give details of our model design.

2.1. Building Entity Features from Masks

Our approach assumes the availability of a mask proposal model, which is capable of generating a set of class-agnostic entity masks from an image with high coverage. In contrast to prior studies that relied on low-level features [8, 13, 46, 57], GROUNDHOG interprets the image as a collection of entities. The primary challenge then becomes the derivation of effective visual features to accurately represent these entities. To achieve a complete decoupling of the MLLM from the mask proposal model responsible for providing the masks, we propose to condition the entity features solely on the binary masks without using any embeddings from the mask proposal model. Specifically, the mask corresponding to each entity is employed to extract patch features from pretrained vision foundation models, such as CLIP [63] and DINOv2 [55], through a convolutional mask pooling layer [12]. Given that the feature map dimensions

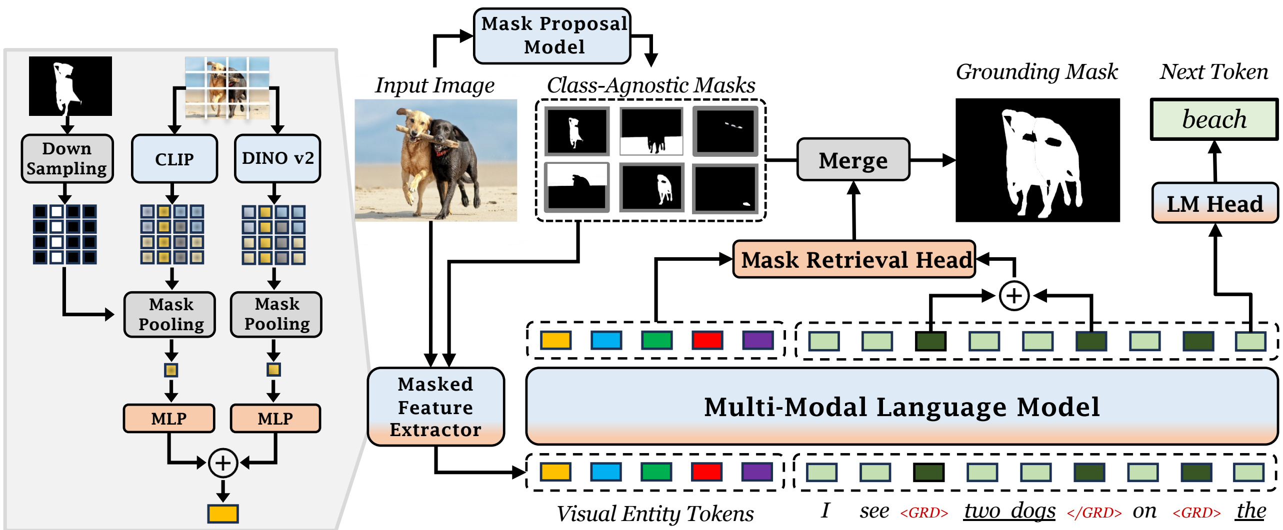


Figure 2. The model architecture of GROUNDHOG model. Given a set of class-agnostic entity mask proposals, the masked feature extractor first extracts the feature of each entity as the visual input of the multi-modal large language model (left). The output hidden states of the grounding tokens are averaged and used to retrieve the entities to ground, which will be merged into a single grounding mask for the phrase. Modules are colored by their trainability: parameter-free operators (grey), frozen (blue), trainable (orange), and partially trainable (mix).

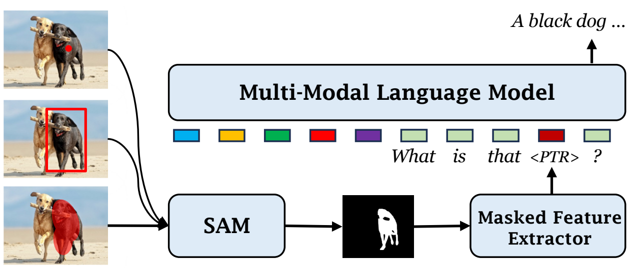


Figure 3. GROUNDHOG can take arbitrary spatial prompts that can be resolved by an interactive segmentation model, such as SAM. The placeholder pointer token <PTR> will be replaced by the extracted entity features and fed as input to the model.

are usually smaller than those of the mask proposals, we resize the masks to match the size of the feature maps prior to pooling. The pooled features are then fed into a Multi-Layer Perceptron (MLP) network to align with the input embeddings of the MLLM. We empirically find the combination of CLIP and DINOv2 features yields the best result, and these features are added to obtain the final input visual entity tokens to the MLLM.

Spatial Prompts Furthermore, for grounded MLLMs to be more broadly applicable, they must be capable of interpreting multi-modal user inputs, including spatial prompts. Thanks to the mask model agnostic design, GROUNDHOGcan seamlessly support such inputs. As demonstrated in Figure 3, by applying an interactive segmentation model such as Segment-Anything (SAM) [35], arbitrary spatial prompts can be translated into binary masks and processed by the same masked feature extractor we just introduced. This extracted feature for the pointed entity will replace the pointer token <PTR> placeholder in the textual input.

2.2. Language Grounding to Entity Segmentation

Existing box-grounded MLLMs typically append location tokens after the groundable phrases [8, 9, 57, 85]. However, this method is not readily interpretable. To alleviate this disconnect, we introduce a pair of grounding tokens <GRD> and </GRD> to indicate the start and end of groundable phrases, with the assumption that grounding these phrases requires mapping to certain representations of visual entities irrespective of the visual modality. In Figure 2, a sentence can be represented as `I see <GRD> two dogs </GRD> on <GRD> the beach </GRD>`, with two distinct visual entities grounded. The representation of each groundable phrase, termed as the *grounding query*, is obtained by adding <GRD> and </GRD>'s output embedding from the last transformer layer of the MLLM. The representation is then used to retrieve the entities that the phrase should be grounded to. In particular, we concatenate the grounding query with the last layer output of each visual entity token, and use an MLP to predict a scalar score for each entity. Finally, we merge all the mask proposals into one single mask with pixel-wise maximization:

$$\mathcal{M}_{h,w} = \max_q \left(\mathcal{S}_q \cdot \widehat{\mathcal{M}}_{q,h,w} \right)$$

where $\mathcal{S}_q$ is the normalized score of the q-th mask ranging from 0 to 1, and $\widehat{\mathcal{M}}_{q,h,w}$ denotes the pixel probability at position (h, w) for the q-th mask. Note that a phrase may ground to multiple entities, thus multiple mask proposals may get a high score simultaneously and be selected in conjunction. One of the primary benefits of this decoupled design is its transparency in the selection of entities. Users can easily visualize both the mask proposals and their respective scores, providing a clear understanding of how a grounding mask is predicted. This level of clarity and interpretability is a significant advantage, offering users a tangible insight into the model's grounding process.

2.3. Towards Holistic Entity Mask Proposals

In order to support holistic language grounding to arbitrary segmentations, the entity proposal should have two essential properties. First, the proposals should strike a delicate balance in terms of semantic atomicity. While it is possible to merge multiple proposals later to form multi-entity segmentations, the reverse, i.e., dividing a single proposal into smaller segments, is not feasible. Therefore, instance segmentation is generally preferred over semantic segmentation. However, the segmentation should not be excessively fine-grained to the extent that it compromises basic semantic integrity. Over-segmentation can lead to a loss of the coherent concept of an entity, which is detrimental to the grounding process. Second, the entity proposals should have a high coverage of entities, encompassing a diverse range of granularities. This includes not only tangible objects (things) and amorphous concepts (stuff) but also extends to sub-components of objects (parts of things) and structured regions such as areas containing visual text. The ability to propose entities across this spectrum of granularity is pivotal, as it directly determines the upper bound of the grounding capability of MLLM.

We initiated our study with a Mask2Former model pretrained on the COCO panoptic segmentation dataset, capable of segmenting 134 object categories. However, preliminary experiments revealed its limitations in semantic coverage and adaptability to open-world scenarios. To enhance this, we developed Mask2Former+, an upgraded version designed for multi-grained segmentation. This upgrade involved creating a diverse dataset by merging annotations from various sources, including COCO [5], LVIS [25], Entity-v2 [61], Pascal [16], PACO [64] (Figure 8); MHP-v2 [40] for human part parsing; and TextOCR [68] for text segmentation. Additionally, we expanded the model's capabilities by adding 50 expert queries each for semantic parts and visual text regions, alongside the original 200 entity queries. We assessed Mask2Former+'s performance on 1000 images from validation splits from 4 grounding benchmarks, RefCOCO+ [86], PhraseCut [77], ReasonSeg [37], and TextVQA-X [67]. We use the Any-IoU [30] metric for evaluation, i.e., for each ground truth mask, we extract the most overlapped mask proposals and compute the IoU, then take the average. As Table 1 demonstrates, Mask2Former+ shows consistent improvements across all domains, particularly in those significantly divergent from COCO. This highlights its enhanced adaptability and precision in a broader range of segmentation challenges, providing a good mask proposal model for GROUNDHOG. We refer to Appendix A for more details of the model and data.

Model	RefCOCO+	PhraseCut	ReasonSeg	TextVQA-X
Mask2Former	0.867	0.563	0.602	0.137
Mask2Former+	**0.873**	**0.624**	**0.745**	**0.446**

Table 1. The average Any-IoU of the proposals on each dataset. The vanilla Mask2Former is trained on the COCO-Panoptic dataset and our Mask2Former+ is trained on our combined dataset. Mask2Former+ obtains a consistent improvement in all scenarios, especially in non-COCO domains.

Task	Dataset	Gr. Ann.			Sem. Gran.					# Pairs
		M	B	Po	S	Th	Pa	G	Tx	Train
GCAP	PNG	✓	✓		✓	✓		✓		132k
	Flickr30K-Entity		✓		✓	✓	✓	✓		149k
RES	RefCOCO	✓	✓		✓					113k
	RefCOCO+	✓	✓		✓					112k
	RefCOCOg	✓	✓		✓					80k
	RefCLEF	✓	✓		✓					105k
	gRefCOCO	✓	✓		✓					194k
	PhraseCut	✓	✓		✓	✓	✓	✓		85k
	D-Cube	✓	✓		✓			✓		10k
	ReasonSeg	✓	✓		✓	✓	✓	✓		1k
	RIO	✓	✓		✓			✓		28k
	SK-VG		✓		✓					23k
GVQA	VizWiz-G	✓	✓		✓	✓			✓	6k
	TextVQA-X	✓	✓						✓	15k
	GQA		✓		✓	✓	✓	✓		302k
	VQS		✓		✓					20k
	Shikra-BinaryQA		✓		✓	✓	✓	✓		4k
	EntityCount	✓	✓		✓	✓	✓	✓		11k
	FoodSeg-QA	✓	✓		✓			✓		7k
	LVIS-QA	✓	✓		✓	✓		✓		95k
RD	RefCOCO-REG	✓	✓	✓	✓					17k
	RefCOCO+-REG	✓	✓	✓	✓					17k
	RefCOCOg-REG	✓	✓	✓	✓					22k
	gRefCOCO-REG	✓	✓	✓	✓					20k
	VG-SpotCap		✓	✓	✓	✓	✓	✓		247k
	V7W		✓	✓	✓					23k
	PointQA			✓	✓					64k
	VCR		✓	✓	✓					156k
	ShikraRD		✓	✓	✓	✓	✓	✓		2k
	SVIT-RD		✓	✓	✓	✓	✓	✓		33k
	Guesswhat	✓	✓	✓	✓					193k
	VG-RefMatch		✓	✓	✓	✓	✓	✓		247k
	HierText	✓	✓	✓					✓	6k
M3G2 (Total)										2.5M

Table 2. Summary of datasets included in M3G2. The datasets are grouped by four task types: Grounded Image Captioning, Referring Expression Segmentation, Grounded Visual Question Answering, and Referential Dialogue. We show the availability of Grounding Annotations (Box, Mask, and Pointer inputs), the Semantic Granularity (Stuff, Things, Parts, Groups, and Text), and the number of text-image pairs for training.

3. Our Dataset: M3G2

In this section, we introduce M3G2, a **M**ulti-**M**odal **M**ulti-**G**rained **G**rounding dataset consisting of 2.5M text-image pairs for visually grounded instruction tuning, consisting of 36 sub-problems derived and augmented from 27 existing datasets. We re-organize and augment public datasets of language grounding, visual question answering, referring expression segmentation, and referring expression generation into various forms of visually grounded dialogue for grounded instruction tuning, outlined briefly in Table 2. The dataset is categorized into four main types: (1) Grounded Image Captioning (GIC), (2) Referential Expression Segmentation (RES), (3) Grounded Visual Question Answering (GVQA), and (4) Referential Dialog (RD). We provide illustrated descriptions of our prompt design, accompanied by examples of each task type as depicted in Fi gure 4. We detail the task schema in the following sections and provide the complete sets of templates in Appendix B.

3.1. Grounded Image Captioning (GIC)

The task of *grounded image captioning* requires the model to produce a narrative for the visual scene, and accurately identify and associate the groundable phrases with their respective binary segmentation masks. The objective of this task is to empower the model to articulate the scene while acknowledging various visual elements and their spatial interrelations. We incorporate the Panoptic Narrative Grounding (PNG) dataset [34] for dense and detailed scene descriptions, as well as the Flickr30K-Entity dataset [59] for concise descriptions of the salient contents in the image. We create a collection of task prompt templates that instruct the model to describe the image either in detail or briefly.

3.2. Referring Expression Segmentation (RES)

In contrast to previous tasks, the *referring expression segmentation* task requires that the model generates a segmentation mask based on a given referring expression. Besides the RefCOCO series [44, 53, 86], we have further leveraged existing RES benchmarks [37, 66, 77, 79, 80] for this purpose. To frame a RES task, our prompts clearly direct the model to focus on the segmentation aspect of the given referring expression. An example prompt in Figure 4 could be `Localize the sandwich on the plate at the far end`. The model's expected output would repeat the provided referring expression and accompany it with the correct segmentation mask, as in `Here it is: <GRD> the sandwich on the plate at the far end </GRD>` with a correct mask associated.

3.3. Grounded Visual Question Answering (GVQA)

The task of *grounded visual question answering* requires the model to comprehend a question (with optional pointers) and to produce an answer that is grounded to a binary segmentation mask that justifies the answer. The goal of this task is to enable natural QA-based interaction with users and reasoning in the model with grounded explanations. Specifically, we harvest and adapt a collection of public VQA datasets with grounding annotations [6, 8, 19, 29], QA benchmarks on visual text [65], and create templated QA pairs from segmentation datasets [25, 61]. Our prompt templates instruct the model to respond either as open-ended answers or by selecting from multiple choices, with the response anchored to a segmentation mask (Figure 4).

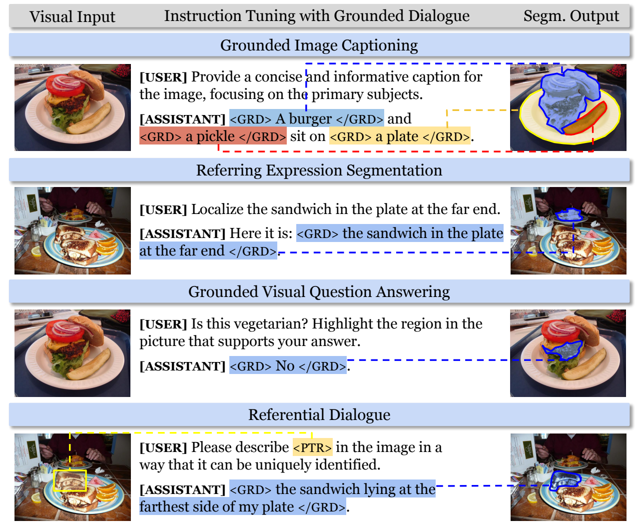


Figure 4. The M3G2 dataset for grounded visual instruction tuning. M3G2 is a diverse dataset of multiple granularities, unifying 4 different task types with visually grounded dialogue.

3.4. Referential Dialogue (RD)

The task of *referential dialogue* requires the model to conduct dialogue communication with users, especially when conditioned on user-provided spatial prompts. This includes existing RD datasets [8, 52, 88, 90, 94], multi-turn augmentations from segmentation datasets [17, 36, 48] as well as the *referring expression generation (REG)* task the RefCOCO series [44, 53, 86]. The REG task differs from the region captioning task in that it demands the description to be a referring expression that distinctly identifies the targeted object. Effective REG calls for the model to engage in dialogue interactions cooperatively, adhering to the Gricean Maxims [24] which dictate that communication should be as informative, truthful, relevant, and clear as necessary.

4. Experiment and Analysis

4.1. Implementation

Learning from Both Box and Mask Supervision. In the M3G2 dataset, not all sub-datasets include mask supervision. We employ different loss functions to effectively benefit from grounded supervision from both mask and box annotations. When the mask annotations are available, we apply the dice loss $\mathcal{L}_{\text{dice}}$ and binary cross-entropy loss $\mathcal{L}_{\text{bce}}$ between the predicted grounding masks and the ground truth masks of each phrase, following Cheng et al. [10]. When the box annotations are present, we apply the projection loss $\mathcal{L}_{\text{proj}}$ as introduced by Tian et al. [70]. The final loss calculation is a linear combination of the language modeling loss $\mathcal{L}_{\text{lm}}$ and these mask-related losses. We refer to Appendix C for more details and explanations of these loss terms.

Model	Single Instance								Multi-/No Instance		Reasoning		
	RefCOCO			RefCOCO+			RefCOCOg		gRefCOCO	PhraseCut	ReasonSeg	RIO	
	val	test-A	test-B	val	test-A	test-B	val-u	test-u	val	test	val	test-c	test-u
Specialist													
MDETR [30]	-	-	-	-	-	-	-	-	-	53.7	-	44.1	22.0
CRIS [76]	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4	55.3	-	-	-	-
LAVT [82]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1	58.4	-	-	-	-
ReLA [44]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0	63.6	-	22.4	-	-
PolyFormer [47]	76.9	78.5	74.8	72.2	75.7	66.7	71.2	71.2	-	-	-	48.8	26.8
UNINEXT-H [43]	82.2	83.4	81.3	72.5	76.4	66.2	74.7	76.4	-	-	-	-	-
Generalist													
LISA$_{7B}$ [37]	74.1	76.5	71.1	62.4	67.4	56.5	66.4	68.5	-	-	44.0	-	-
LISA$_{7B}$ (FT) [37]	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	-	-	52.9	-	-
🐹$_{7B}$	78.5	79.9	75.7	70.5	75.0	64.9	74.1	74.6	66.7	54.5	56.2	57.9	33.9

Table 3. Results on 7 Referring Expression Segmentation (RES) benchmarks with single instance queries [32, 54], multi-/null instance queries [44, 77] and reasoning-based queries [37, 62]. We report cIoU for RefCOCO/+/g and mIoU for other benchmarks, respectively.

Model	Flickr30K-E R@1$_{val}$	Flickr30K-E R@1$_{test}$
Shikra$_{13B}$	77.4	78.4
Ferret$_{13B}$	81.1	84.8
Shikra$_{7B}$	75.8	76.5
Ferret$_{7B}$	80.4	82.2
🐹$_{7B}$	79.2	79.8

Table 4. Top-1 box recall results on Flickr30K-Entity [59].

Model	PNG AR	AR$_{th}$	AR$_{st}$	AR$_{s}$	AR$_{p}$
PiGLET	65.9	64.0	68.6	67.2	54.5
🐹$_{7B}$	66.8	65.0	69.4	70.4	57.7

Table 5. Phrase grounding results on PNG [21].

Model	TextVQA-X [mIoU]
SAB	29.0
🐹$_{7B}$	39.8

Table 6. Visual text QA results on the TextVQA-X [67] validation set.

Model	PointQA$_{Twice}$ Acc	V7W Acc
Shikra$_{13B}$	70.3	85.3
GPT4RoI$_{13B}$	-	84.8
Shikra$_{7B}$	-	-
GPT4RoI$_{7B}$	-	81.8
🐹$_{7B}$	72.4	85.5

Table 7. Results on PointQA$_{Twice}$ [52] and V7W [94] test sets.

Parameter-Efficient Training Details. We adopt the LLaMA2-7B model [71] as our base LLM, and initialized the weight from LLaVA-1.5 [45]. For the vision encoders, we use the OpenAI CLIP@336 [63] model and DINOv2-L/14-reg [15] pretrained checkpoints. We freeze all the parameters of Mask2Former+, CLIP, and DINOv2 during training. We use Low-Rank Adaptation (LoRA) [28] with $r = 16$ and $\alpha = 16$ to tune the LLM, including all the linear layers, input embeddings, and the LM head. We train all the new components introduced for connecting these models, including the MLP projection layer of CLIP and DINOv2, and the mask retrieval head. As a result, less than 2% of the total parameters are trainable in the whole model. We use the AdamW optimizer [49] with an initial learning rate of 2e-4 and a cosine annealing rate. We train our model on the balanced sampled M3G2 dataset for 2 epochs, which takes around 2 days using 8 40G A100 GPUs.

4.2. Generalist in Grounded Vision-Language Tasks

We first demonstrate GROUNDHOG's capabilities as a generalist model for three different types of grounded vision-language tasks. It's worth noting that, unlike previous work that needs dataset-specific fine-tuning on each of the tasks, GROUNDHOG can achieve comparable performance on all the tasks directly after training on M3G2, i.e., all the reported results from our model are from a single set of weights without any dataset-specific fine-tuning.

Language Grounding To Segmentation. We start by evaluating the model on language grounding tasks, which takes text as input and generates segmentation masks as output. We assess GROUNDHOG on Referential Expression Segmentation (RES) [32] and Caption Phrase Grounding (CPG) tasks. While traditional RES benchmarks [32, 54] focus on single-instance referents requiring primarily visual understanding, we expanded our evaluation to include complex scenarios involving multi-instance or negative queries [44, 77], and those necessitating common sense reasoning [37, 62]. For single-instance RES, we report the cIoU; and for the other benchmarks, we report the mIoU. The results, as detailed in Table 3, show GROUNDHOG outperforming the generalist model LISA across all benchmarks and achieving significant improvements over specialist models in multi-instance, null, and reasoning-based RES tasks. It also performs comparably on the competitive RefCOCO series. For CPG tasks, which involve grounding all phrases in a caption and demand a deep understanding of the context for coreference resolution, we first evaluated GROUNDHOG on the Flickr30K-Entity dataset [59]. Since this dataset only has box annotations, we convert the mask predictions of our model to box and compute the top-1 box recall following the merged-box protocol (All-IoU) [30]. Despite not specializing in predicting boxes, GROUNDHOG still outperforms Shikra 7B/13B [8] and is

(a) Grounded short caption generation on Flickr30K-Entity. While only box supervisions are available for this dataset, GROUNDHOG generalize to pixel-level grounding after joint training on M3G2.

(b) Grounded detailed narrative generation on PNG. GROUNDHOG successfully generalize to grounding a novel category *watch* in the generated caption, which is not included in the 80 categories of PNG annotation.

Figure 5. Examples of GROUNDHOG's performance in grounded image captioning.

Model	Bleu-4	METEOR	CIDEr	SPICE	$F1_{all}$
Shikra_{13B}	-	-	73.9	-	-
Ferret_{13B}	37.0	25.5	76.1	18.3	15.1
Ferret_{7B}	35.1	24.6	74.8	18.0	15.0
$_{7B}$	36.7	26.5	91.3	20.4	32.1

Table 8. Grounded Captioning on Flickr30K-Entity [59].

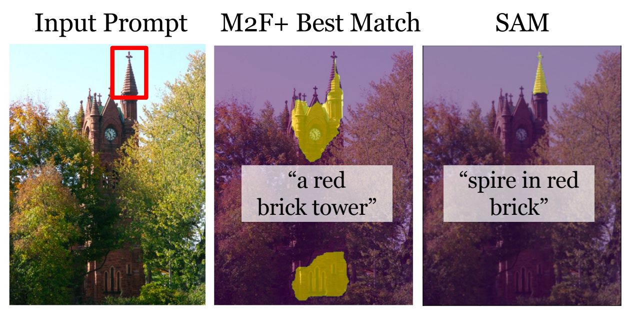


Figure 6. Region caption using the best match proposal from Mask2Former+ versus from SAM. Mask2Former+ fails to propose the exact mask of the spire, leading to a less precise caption.

on par with Ferret-7B [85] in a concurrent work (Table 4). Additionally, on the PNG dataset [34] which tests phrase grounding in longer narratives, GROUNDHOG surpasses the previous state-of-the-art model, PiGLET [22], in all metrics including average recall of grounding masks and detailed scores for things, stuffs, and singular and plural entities (Table 5).

Grounded Language Generation. Our model excels in generating language that accurately grounds to segmentation masks during user conversations. Quantitatively, we assess grounded captioning on the Flickr30K-Entity dataset [59], employing standard text generation metrics such as Bleu-4 [56], METEOR [4], CIDEr [73], and SPICE [2] for language quality; and the $F1_{all}$ score for grounding accuracy following You et al. [85]. As shown in Table 8, GROUNDHOG significantly surpasses existing box-based grounded MLLMs, even their 13B versions, in both language quality and grounding accuracy. This improvement is hypothesized to stem from the diverse task distribution in our M3G2 dataset. We show some generated captions in Figure 5, with a highlight of box-to-pixel generalization (Figure 5a) and novel category grounding (Figure 5b). See the Appendix for more examples. For groundable question answering, we evaluate on the TextVQA-X benchmark [65]. GROUNDHOG outperforms the state-of-the-art specialist model SAB [33] by a significant margin, as measured by the mean IoU of the predicted mask (Table 6).

Spatial Prompt Understanding. For grounded MLLMs, accurately interpreting multimodal instructions is essential, particularly in interactive tasks. We evaluated its performance on two pointer-based QA benchmarks, $\text{PointerQA}_{\text{Twice}}$ [52] and V7W [94], which require the model to answer questions guided by spatial prompts, such as bounding boxes. The model is tasked to generate free-form textual answers in $\text{PointerQA}_{\text{Twice}}$, and selects from multiple-choice options in V7W. GROUNDHOG demonstrates superior performance in these benchmarks, outperforming previous models as shown in Table 7. This highlights its effectiveness in spatial understanding and response accuracy. To further demonstrate the effectiveness of using SAM for the pointer-to-mask conversion, we show the best-matched mask proposal from our Mask2Former+ model in comparison to the mask from SAM in Figure 6. While the best match proposal from the Mask2Former+ model includes a broader area, the SAM-generated mask offers a more precise representation of the specified region, potentially leading to a more accurate caption.

4.3. Trustworthiness and Transparency

Beyond its superior performance as a grounding generalist, we highlight two key improvements for creating a more trustworthy and transparent agent.

Model	Accuracy	Precision	Recall	F1 Score	Yes (%)
Random					
mPLUG-Owl	53.30	51.71	99.53	68.06	96.23
LLaVA	54.43	52.32	99.80	68.65	95.37
MultiModal-GPT	50.03	50.02	100.00	66.68	99.97
MiniGPT-4	77.83	75.38	82.67	78.86	54.83
InstructBLIP	88.73	85.08	93.93	89.29	55.20
Shikra-13B	86.90	94.40	79.26	86.19	43.26
Ferret-13B	90.24	97.72	83.00	89.76	43.26
$_{\text{7B}}$	91.03	85.80	96.40	90.79	45.88
Popular					
mPLUG-Owl	50.63	50.32	99.27	66.79	98.63
LLaVA	52.43	51.25	99.80	67.72	97.37
MultiModal-GPT	50.00	50.00	100.00	66.67	100.00
MiniGPT-4	68.30	64.27	82.40	72.21	64.10
InstructBLIP	81.37	75.07	93.93	83.45	62.57
Shikra-13B	83.97	87.55	79.20	83.16	45.23
Ferret-13B	84.90	88.24	80.53	84.21	45.63
$_{\text{7B}}$	90.13	85.93	93.81	89.70	45.80
Adversarial					
mPLUG-Owl	50.67	50.34	99.33	66.82	98.67
LLaVA	50.77	50.39	99.87	66.98	99.10
MultiModal-GPT	50.00	50.00	100.00	66.67	100.00
MiniGPT-4	66.60	62.45	83.27	71.37	66.67
InstructBLIP	74.37	67.67	93.33	78.45	68.97
Shikra-13B	83.10	85.60	79.60	82.49	46.50
Ferret-13B	82.36	83.60	80.53	82.00	48.18
$_{\text{7B}}$	86.33	85.93	86.63	86.28	49.60

Table 9. Object hallucination results on the POPE [42] benchmark.

Reduced Object Hallucination. Thanks to the varied task distribution and the inclusion of negative question-answering samples in M3G2 dataset, GROUNDHOG significantly reduces object hallucination. We assessed this using the POPE [42] benchmark, which includes binary questions about object existence across three splits, each with a different object distribution (with an order of difficulty *Random* < *Popular* < *Adversarial*). Remarkably, GROUNDHOG consistently outperforms other models in both accuracy and F1 score across all splits, particularly on the more challenging ones. It shows an absolute improvement of 5.2% in accuracy for *Popular* and 4.0% for *Adversarial* over the previously best-performing model. This suggests that our model's enhanced grounding capability plays a significant role in mitigating the object hallucination problem.

Explainability and Diagnosability. Another important highlight of GROUNDHOG is its enhancement of explainability through the decoupled design of entity proposal and selection, as outlined earlier in section 2.2. This is exemplified in the case study illustrated in Figure 7, which illustrates the mask proposal scoring and selective merging process of our model. We show the top-4 masks, where the higher-score masks are labeled in green while the lower-score masks are labeled in red. Users can easily interpret that the failure is due to the incapability of MLLM to recognize the word "KWIK", despite it being successfully localized and proposed as an entity candidate.

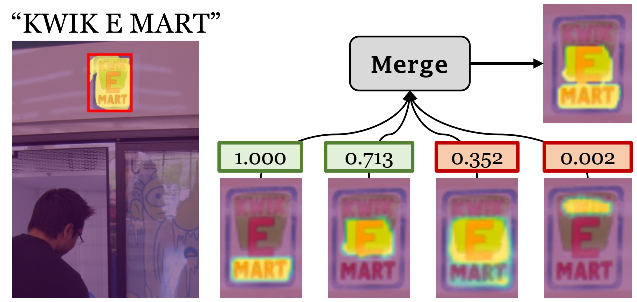


Figure 7. Illustration of a partially correct grounding. The grounding phrase and the ground truth mask are shown on the left. The top-4 mask proposals are presented, with highly-scored masks (green) selected for the merged mask, and low-scored masks (red) excluded. This illustrates the failure to recognize the word "KWIK" by the MLLM, despite its successful proposal.

Setups	RefCOCO+	Flickr30K	TextVQA-X
Mask Proposal Models			
Mask2Former	**67.1**	69.0	9.8
Mask2Former+	66.6	**77.2**	**34.0**
Entity Features			
CLIP	59.8	75.0	32.0
DINOv2	62.3	76.3	28.4
CLIP+DINOv2	**66.6**	**77.2**	**34.0**
Grounding Query			
`<GRD>` only	64.4	67.5	**34.2**
`</GRD>` only	64.4	**77.2**	33.5
Sum	**66.6**	**77.2**	34.0
Eval Input Resolution			
224–480	54.7	67.2	27.6
480–640	65.5	76.7	27.6
800–1024	**66.6**	**77.2**	**34.0**

Table 10. Ablation study on model design choices and evaluation setups. Models are trained on RefCOCO+, Flickr30K, TextVQA-X and tested on corresponding validation sets.

4.4. Ablation Studies

We performed 3 ablation studies to validate our design decisions, training, and evaluating a subset of the M3G2 dataset that includes RefCOCO+, Flickr30K, and TextVQA. These cover a range of visual entities from various image sources and granularities. We start by comparing our Mask2Former+ with the original Mask2Former for mask proposal effectiveness. As indicated in Table 10, the original Mask2Former performs slightly better on RefCOCO, as it is developed specificity on COCO object categories. However, Mask2Former+ significantly surpasses the original in domains with non-COCO entities. Our second set of experiments examined the choice of visual entity features. Although using either CLIP or DINOv2 features alone shows advantages in specific datasets, their combination consistently yields the best results across all datasets. To obtain a robust grounding query representation, we experimented with using the output embedding of the <GRD> token, the </GRD> token, and their sum. We found that the latter approach achieves the best overall results. Finally, we demonstrate that our decoupling design of the mask proposal model and MLLM allows for training at a lower resolution (320px) to expedite grounding training, while scaling up the resolution during evaluation enhances performance.

5. Related Work

5.1. Multimodal Large Language Models

Building on the recent advance of large language models (LLMs), there is an increasing effort in adapting pretrained large language models for multimodal tasks, such as understanding and interpreting visual information [1, 72]. More recently, visual instruction tuning has gained much interest due to its surprising performance with a modest amount of data and computing resources. Various models have been developed, noticeably MiniGPT4[93] , LLaVA [45, 46] and concurrent models [13, 20, 38, 75, 83]. Despite their promising performances, MLLMs often produce objects that are not presented in the given images, a phenomenon referred to as the *object hallucination* problem [14, 31, 42].

5.2. MLLM with Language Grounding

The ability to connect language to their corresponding visual elements in the physical world, known as *grounding* [27], is crucial in everyday human communication about our shared surroundings. Grounding datasets have been shown to benefit vision-language pre-training, both in terms of object-level recognition [41] and language learning [51]. Recent works unify text and grounding regions into token sequences [50, 74, 81] in casual language modeling. Based on such paradigm, researchers have developed a family of grounded MLLM, including GPT4ROI [89], Kosmos-2 [57], Shikra [8], PVIT [7], BuboGPT [91], Qwen-VL [3], and Ferret [85]. Despite their promising performance, these models focus on object grounding to bounding box, which cannot handle pixel-level grounding across various semantic granularities. Furthermore, it lacks the diagnosability and explainability in failure cases. We introduce GROUNDHOG to fill this gap.

5.3. Language-Guided Semantic Localization

The field of language-guided semantic localization has a long history in the vision-language research community, requiring that the model localize a given referring expression with bounding boxes or segmentation masks. This task has evolved from early attempts to understand simple referring expressions within images, such as the well-known RefCOCO series [53, 86] and their generalized variant [44] that takes no-target and multi-target into account. The integration of advanced language reasoning from LLMs has enabled research to tackle even more nuanced reasoning tasks that involve complex language contexts [37, 58, 87]. Notably, LISA [37] formulates a reasoning segmentation task to bring language-informed reasoning into semantic segmentation, and contributes a powerful baseline. Our model builds on these developments, but is designed to be more universally applicable as a grounded MLLM.

6. Conclusion

In this study, we introduce GROUNDHOG, a novel framework designed to enable pixel-level explainable grounding in large language models, leveraging holistic segmentation. The system builds upon a pre-trained mask proposal network to provide pixel-level visual features for the large language models, allowing them to retrieve segmentation mask proposals that can be used for grounding. We also present M3G2, a dataset of 1.9M training text-image pairs with 36 sub-problems derived from 27 existing datasets for visually grounded instruction tuning, facilitating precise vision-language alignment at the pixel level. We show that after training on M3G2, GROUNDHOG achieves superior performance on various grounding tasks. Through extensive case studies, we further show that GROUNDHOG unlocks explainability and diagnosability, and demonstrates better grounding towards occluded objects, groups of multiple instances, amorphous background regions, semantic parts of objects, and objects with irregular shapes.

Limitations And Future Work

This work, while exciting, has several limitations that we acknowledge and aim to address in future research. Firstly, the datasets utilized to develop M3G2 consist of a blend of existing academic datasets. The quality of annotations in these datasets varies significantly, and they often lack comprehensive coverage of concepts. To enhance training efficiency, applying data filtering methods could help reduce the size of the dataset without compromising its effectiveness. Additionally, expanding the vision-language grounding data to a web-scale could significantly improve the comprehensiveness of grounding learning.

Secondly, our current model is limited to processing only single images. Although the entity-centric approach we adopted could theoretically extend to other modalities like 3D or video, this potential has not yet been empirically validated. Testing and validating our model on datasets relevant to these modalities would be a valuable direction for future research. This step is crucial to understanding the model's effectiveness across different types of application scenarios and further improving its usefulness.

Acknowledgements

This work was supported by Amazon and NSF IIS-1949634. We would like to thank the anonymous reviewers for their valuable comments and suggestions.

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. 9

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Proceedings of the 14th European Conference on Computer Vision*, pages 382–398, 2016. 7

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 9

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 7

[5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 2, 4, 1

[6] Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19098–19107, 2022. 5, 2

[7] Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. Position-enhanced visual instruction tuning for multimodal large language models. *arXiv preprint arXiv:2308.13437*, 2023. 9

[8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2, 3, 5, 6, 9

[9] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *International Conference on Learning Representations*, 2021. 3

[10] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2, 5, 1, 4

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

[12] Jifeng Dai, Kaiming He, and Jian Sun. Convolutional feature masking for joint object and stuff segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3992–4000, 2015. 2

[13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 2, 9

[14] Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2128–2140, 2023. 9

[15] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 6, 4

[16] Daan de Geus, Panagiotis Meletis, Chenyang Lu, Xiaoxiao Wen, and Gijs Dubbelman. Part-aware panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5485–5494, 2021. 4, 1

[17] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, 2017. 5, 3

[18] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 1

[19] Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. Vqs: Linking segmentations to questions and answers for supervised attention in vqa and question-focused semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1811–1820, 2017. 5, 2

[20] 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. 9

[21] Cristina González, Nicolás Ayobi, Isabela Hernández, José Hernández, Jordi Pont-Tuset, and Pablo Arbeláez. Panoptic narrative grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1364–1373, 2021. 6

[22] Cristina González, Nicolás Ayobi, Isabela Hernández, Jordi Pont-Tuset, and Pablo Arbeláez. Piglet: Pixel-level grounding of language expressions with transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 7

[23] Abel Gonzalez-Garcia, Davide Modolo, and Vittorio Ferrari. Objects as context for detecting their semantic parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6907–6916, 2018. 2

[24] Herbert P Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975. 5

[25] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 4, 5, 1, 3

[26] Kai Han, Yunhe Wang, Jianyuan Guo, Yehui Tang, and Enhua Wu. Vision gnn: An image is worth graph of nodes. *Advances in Neural Information Processing Systems*, 35:8291–8303, 2022. 2

[27] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990. 9

[28] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 6, 4

[29] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 5, 2

[30] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 4, 6

[31] Osman Semih Kayhan, Bart Vredebregt, and Jan C Van Gemert. Hallucination in object detection—a study in visual part verification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2234–2238. IEEE, 2021. 9

[32] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 6, 2

[33] Seyedalireza Khoshsirat and Chandra Kambhamettu. Sentence attention blocks for answer grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6080–6090, 2023. 7

[34] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 5, 7, 1

[35] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 2, 3

[36] 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. 5, 3

[37] 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. 4, 5, 6, 9, 2

[38] 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. 9

[39] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 2023. 2

[40] Jianshu Li, Jian Zhao, Yunchao Wei, Congyan Lang, Yidong Li, Terence Sim, Shuicheng Yan, and Jiashi Feng. Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206*, 2017. 4, 1

[41] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 9

[42] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 8, 9

[43] Fangjian Lin, Jianlong Yuan, Sitong Wu, Fan Wang, and Zhibin Wang. Uninext: Exploring a unified architecture for vision recognition. *arXiv preprint arXiv:2304.13700*, 2023. 6

[44] 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, 9, 2, 3

[45] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 6, 9

[46] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2, 9

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

[48] Shangbang Long, Siyang Qin, Dmitry Panteleev, Alessandro Bissacco, Yasuhisa Fujii, and Michalis Raptis. Towards end-to-end unified scene text detection and layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1049–1059, 2022. 5, 3

[49] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6, 4

[50] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*, 2022. 9

[51] Ziqiao Ma, Jiayi Pan, and Joyce Chai. World-to-words: Grounded open vocabulary acquisition through fast mapping in vision-language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 524–544, 2023. 2, 9

[52] Arjun Mani, Nobline Yoo, Will Hinthorn, and Olga Russakovsky. Point and ask: Incorporating pointing into visual question answering. *arXiv preprint arXiv:2011.13681*, 2020. 5, 6, 7, 3

[53] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 5, 9, 3

[54] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 6, 2

[55] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[56] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 7

[57] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2, 3, 9

[58] Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, and Lingpeng Kong Tong Zhang. Detgpt: Detect what you need via reasoning. *arXiv preprint arXiv:2305.14167*, 2023. 9

[59] 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. 5, 6, 7, 1

[60] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *Proceedings of the 16th European Conference on Computer Vision*, pages 647–664, 2020. 1

[61] Lu Qi, Jason Kuen, Tiancheng Shen, Jiuxiang Gu, Wenbo Li, Weidong Guo, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High quality entity segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4047–4056, 2023. 4, 5, 1, 3

[62] Mengxue Qu, Yu Wu, Wu Liu, Xiaodan Liang, Jingkuan Song, Yao Zhao, and Yunchao Wei. RIO: A benchmark for reasoning intention-oriented objects in open environments. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 6, 2

[63] 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. 2, 6, 4

[64] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151, 2023. 4, 1

[65] Varun Nagaraj Rao, Xingjian Zhen, Karen Hovsepian, and Mingwei Shen. A first look: Towards explainable textvqa models via visual and textual explanations. *arXiv preprint arXiv:2105.02626*, 2021. 5, 7, 2

[66] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *Proceedings of the 14th European Conference on Computer Vision*, pages 817–834. Springer, 2016. 5, 2

[67] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 4, 6

[68] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. 2021. 4, 1

[69] Bharat Singh, Hengduo Li, Abhishek Sharma, and Larry S Davis. R-fcn-3000 at 30fps: Decoupling detection and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1081–1090, 2018. 2

[70] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5443–5452, 2021. 5, 4

[71] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 6

[72] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 9

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

[74] 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. 9

[75] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu

Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023. 9

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

[77] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10216–10225, 2020. 4, 5, 6, 2

[78] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1

[79] Yixuan Wu, Zhao Zhang, Chi Xie, Feng Zhu, and Rui Zhao. Advancing referring expression segmentation beyond single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2628–2638, 2023. 5, 2

[80] Chi Xie, Zhao Zhang, Yixuan Wu, Feng Zhu, Rui Zhao, and Shuang Liang. Described object detection: Liberating object detection with flexible expressions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 5, 2

[81] 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. 2022. 9

[82] 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. 6

[83] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 9

[84] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 2

[85] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 3, 7, 9

[86] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 4, 5, 9, 3

[87] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. *arXiv preprint arXiv:2305.18279*, 2023. 9

[88] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019. 5, 3

[89] 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. 9

[90] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. 5, 3

[91] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023. 9

[92] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 2

[93] 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. 9

[94] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016. 5, 6, 7, 3

Appendix

A. Mask2Former+ Implementation Details

Our enhancement of the original Mask2Former model focuses on broadening its segmentation capabilities beyond the 134 common object categories it currently handles, which include 80 things and 55 stuffs as defined in the COCO dataset. The primary goal is to enable the model to recognize an expanded range of object categories, as well as segmentation masks of various levels of granularities, such as semantic parts and visual text regions.

Training data. We have compiled a comprehensive dataset by combining multiple existing segmentation datasets. This ensemble encompasses a wide spectrum of entities (things and stuff), their semantic parts, and visual text, drawn from sources such as COCO [5], LVIS [25], Entity-v2 [61], Pascal [16], PACO [64], MHP-v2 [40], and TextOCR [68]. The resulting dataset comprises over 200K images and 4.5M masks, as summarized in Table 11. Notably, the annotations from COCO, LVIS, and PACO are based on a shared set of COCO images. We merged these annotations to ensure comprehensive mask proposal coverage, thereby providing holistic instance coverage within each image, as can be illustrated in Figure 8.

Dataset		Granularity			Dataset Size	
Name	Split	Entity	Part	Text	#Image	#Masks
LVIS [25] & PACO [64]	part	✓	✓		15,089	596,687
	no_part	✓			103,178	2,062,536
Entity-v2 [61]	cls	✓			31,913	579,076
Pascal [18]	train	✓	✓		4,998	93,322
	val	✓	✓		5,105	95,462
MHP [40]	train		✓		15,403	410,113
TextOCR [68]	train			✓	21,749	714,770
Total		✓	✓	✓	197,435	4,551,966

Table 11. Summary of the training datasets for Mask2Former+. Entity includes both thing and stuff categories.

Model. Building on the foundation of the original Mask2Former [10], we developed Mask2Former+, a panoptic segmentation model designed for multi-grained segmentation. We initialize our model from the Mask2Former checkpoint with the Swin-L backbone pre-trained on the COCO panoptic segmentation dataset [34]. Besides the 200 entity queries that are trained for thing and stuff proposals, we added 50 additional expert queries for the segmenting parts and the visual text regions, respectively. Given that not all images have annotations for every type of segmentation (for instance, the TextOCR dataset provides annotations only for visual text regions), our model computes the group-wise matching loss exclusively for the annotations available in each dataset. This approach ensures that the model benefits from partial annotations without compromising its ability to recognize other levels of granularity when certain annotations are unavailable. Although most samples in our dataset also have semantic annotations such as object categories, we do not use them but only train the model for class-agnostic mask proposals. We train the model for 20k iterations on our combined segmentation dataset with a batch size of 16 using the Detectron2 library [78].[1]

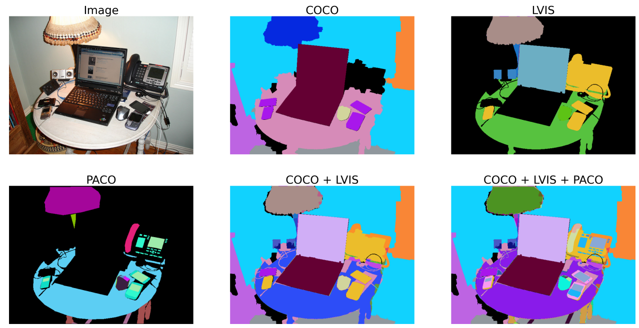


Figure 8. Illustrations of the merged segmentation annotations from COCO Panoptic, LVIS, and PACO datasets.

B. The M3G2 Dataset

In this section, we introduce the M3G2 dataset with Multi-Modal Multi-Grained Grounding. M3G2 is a comprehensive dataset consisting of 36 sub-problems, derived and augmented from 27 existing datasets with grounded vision-language annotations. The dataset is categorized into four main types: (1) Grounded Image Captioning (GIC), (2) Grounded Visual Question Answering (GVQA), (3) Referential Expression Segmentation (RES), and (4) Referential Dialog (RD). Details on the dataset sources, image origins, types of grounding annotations, semantic granularity, and data statistics are summarized in Table 12. All datasets are formatted into the conversation format between a human user and a model assistant, where the user provides task objectives as instructions, and model responses are generated automatically based on the annotations.

Grounded Image Captioning (GIC). GIC focuses on generating image captions that ground to visual entities presented in the image. We incorporate the Panoptic Narrative Grounding (PNG) [34] and Flickr30K-Entity [59] datasets. PNG, derived from Localize Narrative [60] and COCO Segmentation [5], provides long and detailed narratives with an average of 36.5 words per description, exemplified in Figure 15a. These narratives are rich in detail, offering a high coverage of the visual content including the background. Flickr30K-Entity, offering concise captions with box annotations, complements PNG with its larger vocabulary and finer granularity, as shown in 15b. The example instruction templates used to construct the conversation are listed in Ta-

[1] https://github.com/facebookresearch/detectron2

Metadata			Grounding Annotations			Semantic Granularity					Data Size	
Task Type	Dataset Name	Image Source	Mask	Box	Pointer	Thing	Stuff	Part	Multi.	Text	Train	Val / Test
Grd. Captioning (GCAP)	PNG	COCO	✓	✓		✓	✓		✓		132,045	8,435
	Flickr30K-Entity	Flickr30K		✓		✓	✓	✓	✓		148,915	1,000 / 1,000
Referential Expression Segmentation (RES)	RefCOCO	COCO	✓	✓		✓					113,311	-
	RefCOCO+	COCO	✓	✓		✓					112,441	-
	RefCOCOg	COCO	✓	✓		✓					80,322	-
	RefCLEF	ImageCLEF	✓	✓		✓					104,531	-
	gRefCOCO	COCO	✓	✓		✓					194,233	-
	PhraseCut	VG	✓	✓		✓	✓	✓	✓		84,688	-
	D-Cube	GRD	✓	✓		✓			✓		9,499	-
	ReasonSeg	OpenImages & ScanNetV2	✓	✓		✓	✓	✓	✓		1,315	344
	RIO	COCO	✓	✓		✓			✓		27,696	34,170
	SK-VG	VCR		✓		✓					23,404	-
Grounded Visual Question Answering (GVQA)	VizWiz-Grounding	VizWiz	✓	✓		✓	✓			✓	6,494	1,131 / 2,373
	TextVQA-X	OpenImages	✓	✓						✓	14,476	3,620
	GQA	VG		✓		✓	✓	✓	✓		301,623	-
	VQS	COCO		✓		✓					20,380	8,203
	Shikra-BinaryQA	Flickr30K		✓		✓	✓	✓	✓		4,044	1,159
	EntityCount	Entity-v2	✓	✓		✓	✓	✓	✓		11,088	453
	FoodSeg-QA	Recipe1M	✓	✓		✓			✓		7,114	-
	LVIS-QA	COCO	✓	✓		✓	✓		✓		94,860	3,611
Referential Dialog (RD)	RefCOCO-REG	COCO	✓	✓	✓	✓					17,395	-
	RefCOCO+-REG	COCO	✓	✓	✓	✓					17,383	-
	RefCOCOg-REG	COCO	✓	✓	✓	✓					22,057	-
	gRefCOCO-REG	COCO	✓	✓	✓	✓					20,282	-
	VG-SpotCap	VG		✓	✓	✓	✓	✓	✓		247,381	232,935
	V7W	COCO		✓	✓	✓					22,805	10,193 / 57,265
	PointQA-Local	VG			✓	✓					27,426	4,855 / 4,880
	PointQA-Twice	VG			✓	✓					36,762	14,668 / 5,710
	VCR-Open	VCR		✓	✓	✓					58,340	-
	VCR-Multichoice	VCR		✓	✓	✓					97,648	26,534 / 25,263
	ShikraRD	Flickr30K		✓	✓	✓	✓	✓	✓		1,878	-
	SVIT-RD	VG		✓	✓	✓	✓	✓	✓		32,571	-
	Guesswhat-Guesser	COCO	✓	✓	✓	✓					92,136	19,665
	Guesswhat-Oracle	COCO	✓	✓	✓	✓					101,256	21,643
	VG-RefMatch	VG		✓	✓	✓	✓	✓	✓		247,381	-
	HierText	OpenImages	✓	✓	✓					✓	6,058	3,885

Table 12. The full list of datasets used in M3G2.

Instruction Templates for Brief Captioning	Instruction Templates for Detailed Captioning
Describe the image briefly.	Describe the image in detail.
Describe the image in a few words.	Describe the picture's every detail.
Describe the image in a short sentence.	Describe the given picture in very detail.
Describe the image in a clear and concise manner.	Make a fine description of the image.
Generate a short caption for the picture.	Generate a long caption for the given image.
Caption the image in a few words.	Give me a detailed caption of this image.

Table 13. Instruction templates for the GIC task.

ble 13, where we use key words such as "short/briefly" and "in detail" to distinguish between short and long captioning.

Referential Expression Segmentation (RES). RES is a task combining language understanding with precise visual segmentation. Our dataset includes 10 diverse sources. To improve the learning efficiency and enhance contextual understanding, we format queries from the same image into a simulated multi-turn dialog, as illustrated in Figures 16 and 17. We employ the widely used RefCOCO/+/g datasets [32, 54] and RefCLEF [66] for single-object RES. gRefCOCO [44] is employed for multi-object and negative queries. To enhance the visual diversity, we also incorporate PhraseCut [77] and D-Cube [80] that use an image source different than COCO. Additionally, ReasonSeg [37], RIO [62], and SK-VG [79] are included, where a textual context is given and the models need to not only understand that context, but also equips with a certain degree of commonsense knowledge to successfully solve the query, such as shown in Figure 17b, 17c and 17d. The dialogue templates are listed in Table 14.

Instruction Templates For RES	
Highlight "{}" in the image.	Segment "{}" in the image.
Segment: {}.	Help me segment out {}.
Localize "{}" in the image.	Help me localize {}.
Help me highlight the region of {}.	Demonstrate where "{}" is located in this image.
Show me where to find in this photo.	Identify and mark the region of {} for me.
Can you highlight "{}"?	Can you extract the segment: {} for me?
Can you localize "{}" in this image?	Could you please segment out {} in the image?

Table 14. Templates used for the RES task.

Grounded Visual Question Answering (GVQA). The GVQA task extends the visual question answering by additionally requiring visual grounding of the answer. We include 8 datasets for the grounded VQA task in M3G2. First, we collect and organize some existing datasets that can directly fit into our grounded vision-language task framework, including VizWiz-Grounding [6], TextVQA-X [65], GQA [29], VQA [19] and Shikra-BinaryQA [8] (Figure 18). To further improve the data scale and visual concept coverage, we enlarge the GVQA collection by re-purposing existing panoptic segmentation datasets with templated instruc-

Instruction Templates For Short Response VQA.
{}Answer with a single word or a short phrase.
Given the image, answer the question ”{}” with a single word or a short phrase.
Give a short answer to the question ”{}” based on the image.
Instruction Templates For Chain-of-Thought Response VQA.
{}Let’s think step by step.
{}Please include the reasoning process.
{}Before giving the answer, please explain your reasoning.
{}Explain your logic before giving the answer.
Please answer the following question ”{}”, and describe your thought process.
Instruction Templates For Grounding Answer to Masks.
Show where in the image you found your answer.
Mark the part of the image that supports your answer.
Please highlight your evidence in the image.
Point out the evidence from the image.
Indicate the area in the image that justifies your response.
Highlight the section of the image that backs up your answer.
Shade the section of the image that confirms your reply.
Emphasize the part of the image that relates to your answer.
Instruction Templates For Object Presence QA.
Is {} present in the image?
Is there any {} in this image?
Instruction Templates For Object Counting QA.
How many {} can you see in this image?
Count the number of {}.
Instruction Templates For Object Segmentation Request.
Segment {}.
Highlight all the {} in this image.
Show me all the {} presented in the picture.

Table 15. Templates used for the GVQA task.

tions and model responses. Specifically, based on the annotations from LVIS [25] and EntityV2 [61], we design questions about object presence, object counting, and segment query with a possibly negative request (i.e. the target object does not exist in the image), for the model to learn to recognize a diverse set of concepts more faithfully. See Figure 19 for examples of such multi-turn QA, and example question templates used in Table 15.

Referential Dialog (RD) . RD features multi-modal conversations where the user can refer to objects or regions in the image by a spatial prompt (e.g. a bounding box). We include various types of RD in our dataset and the templates used are listed in Table 16. First, we add several existing RD datasets such as V7W [94], PointQA [52], VCR [88], ShikraRD [8] and SVIT [90] without much modifications. We then revisit the RefCOCO series [44, 53, 86] for referential expression generation, where the referred object is given and the goal is to generate a unique description that leads to that object. We use the region caption annotations from the VG dataset [36] for region captioning and a region-matching game. We select a set of region pointers and several descriptions to provide to the model, and the goal is to match the pointed regions with the descriptions (Figure 21b). We repurpose the GuessWhat dataset [17] to make it fit into our RD formulation, as shown in Figure 21a. We also construct a referred text reading task based on the HierText [48] dataset and enhance the model’s capability of text recognition, as shown in Figure 21c.

Instruction Templates For REG.
Provide a distinct description for that <PTR>
Describe the selected area in a unique way. <PTR>
Share a unique description of the region <PTR>
Offer a one-of-a-kind descriptor <PTR>
Describe the selected area <PTR> uniquely.
Point out <PTR> in the picture with a unique description.
Tell me how <PTR> stands out in the photo.
Use your words to highlight just <PTR> in the image.
Please describe <PTR> in the image in a way that it can be uniquely identified.
If you had to describe just <PTR> to someone, how would you do it?
What makes <PTR> different from everything else in the picture?
How can you describe <PTR> in the image in a way that it can be uniquely identified?
Can you provide a referring expression for <PTR> such that it sets it apart from others?
Let’s play a game! Describe <PTR> in the photo so I can find it.
Instruction Templates For Region Captioning.
Describe it <PTR>.
Describe the region <PTR> in a few words.
Describe the region <PTR> in a short phrase.
Describe the selected area <PTR>.
Provide a brief description of this part <PTR>.
Give a short caption for this <PTR>.
Provide a brief description of the area marked <PTR>.
Tell me about the contents in the selected zone <PTR>.
Provide a concise description for this spot <PTR>.
Narrate what you see in the indicated area <PTR>.
What is in the region <PTR>? Describe in a phrase.
What can you see in this area <PTR>?
How would you describe the content at <PTR>?
How would you caption this particular region <PTR>?
What’s depicted in the marked area <PTR>?

Table 16. Templates used for the RD task.

C. GROUNDHOG Implementation Details

Data Balancing. In constructing the M3G2 dataset, we recognized the need to address the varying scales of the multiple constituent datasets to ensure a balanced data distribution during training. To achieve this, we have implemented dataset-specific sampling strategies, adjusting the volume of data from each source dataset through either up-sampling or down-sampling. The ratios we applied are as follows:

- PNG: up-sampled by a factor of 2.
- Flickr30k-Entities: up-sampled by 1.5 times.
- RefCOCO$^+$: up-sampled by 1.5 times.
- RefCOCOg: up-sampled by 1.5 times.
- SK-VG: up-sampled by a factor of 2.
- Dcube (multiturn): up-sampled by a factor of 10.
- ReasonSeg: up-sampled by a factor of 10.
- Shikra-Binary: up-sampled by a factor of 10.
- VCR-Open (multiturn): down-sampled by half.
- VCR-Multiturn: down-sampled to 10%.
- VizWiz: up-sampled by a factor of 3.
- LVIS-QA: down-sampled by half.
- TextVQAX: up-sampled by a factor of 2.
- EntityCount: up-sampled by a factor of 2.
- VG-SpotCap: down-sampled by half.
- Shikra-RD: up-sampled by a factor of 10.
- HierText: up-sampled by a factor of 5.
- GuessWhat-Oracle: down-sampled to 20%.
- GuessWhat-Guesser: down-sampled to 20%.
- SVIT: up-sampled by a factor of 3.

The balanced sampled dataset contains 1.8 million samples in total.

Learning from Both Box and Mask Supervision. In the M3G2 dataset, not all sub-datasets include mask supervision, necessitating a hybrid loss approach to effectively benefit from grounded supervision from both mask and box annotations. We address this by employing different loss functions based on the type of annotation available. When mask annotations are available, we apply the dice loss $\mathcal{L}_{\text{dice}}$ and binary cross-entropy loss $\mathcal{L}_{\text{bce}}$ between the predicted grounding masks and the ground truth masks of each phrase, following Cheng et al. [10]. In cases where only box annotations are present, we apply the projection loss $\mathcal{L}_{\text{proj}}$ as introduced by Tian et al. [70], which selects the mask whose projection on the axis matches the best with the annotated box. Essentially, this can be seen as a 1D dice loss calculated between the projected masks and the edges of the ground truth boxes along both the x and y axes. Given that the primary objective of grounding is to accurately select the correct mask, we assign different weights to these loss components. The mask dice loss and box projection loss are both weighted at 1, while the mask bce loss is given a lower weight of 0.1. The final loss calculation is a summation of the language modeling loss $\mathcal{L}_{\text{lm}}$ and these mask-related losses.

LLM Configuration We adopt the Vicuna-7B model [11] as our base LLM, and use the OpenAI CLIP@336 [63] model and DINOv2-L/14-reg[15] pretrained checkpoints. We use the original conversation template from Vicuna, where all the interactions are formatted as <system_message> <s> USER: <utterance> ASSISTANT: <utterance> </s>.

Parameter-Efficient Training. We freeze all the parameters of the Mask2Former+, the CLIP, and the DINOv2 model during training. We use Low-Rank Adaptation (LoRA) [28] with $r = 16$ and $\alpha = 16$ to tune the LLM, including all the linear layers, input embeddings, and the LM head. We train all the new components introduced for connecting these models, including the MLP projection layer of CLIP and DINOv2, and the mask retrieval head. As a result, less than 2% of the total parameters are trainable in the whole model. We use the AdamW optimizer [49] with an initial learning rate of 2e-4 and a cosine annealing rate. We train our model on the balanced sampled M3G2 dataset for 2 epochs, which takes around 2 days using 8 40G A100 GPUs.

D. Additional Qualitative Results

We provide additional qualitative examples generated by GROUNDHOG. See Figure 9 & 10 for GIC, Figure 11 for RES, Figure 12 for GVQA, and Figure 13 & 14 for RD, respectively.

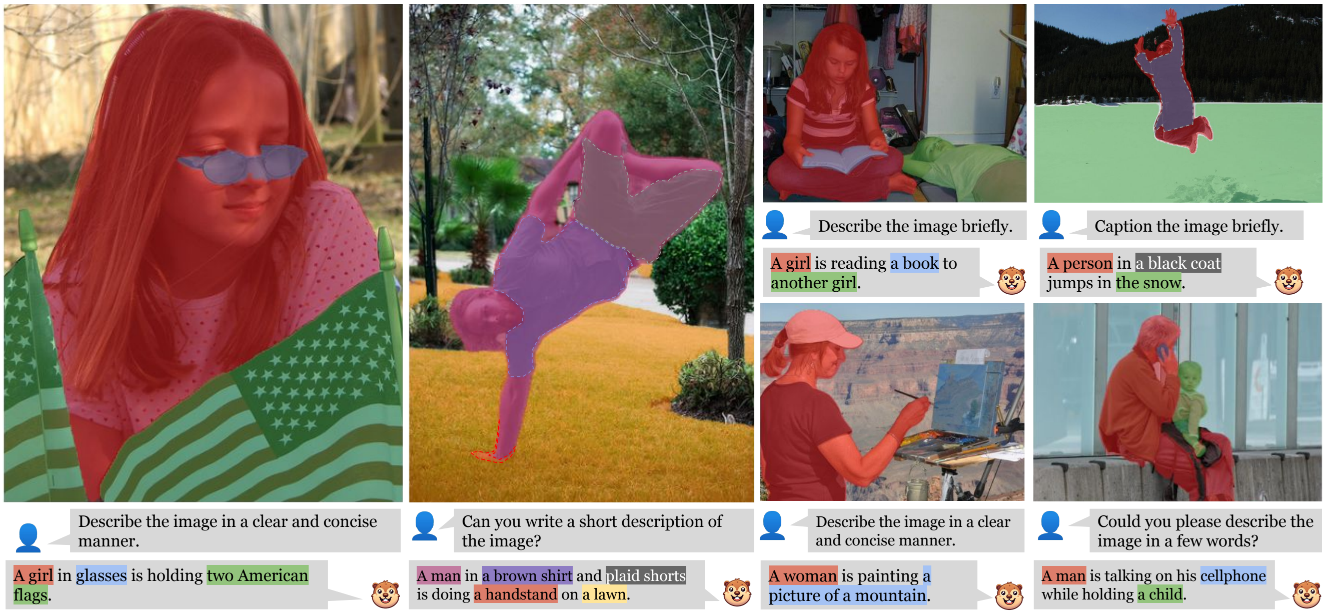


Figure 9. Additional qualitative examples for short grounded captioning.

Figure 10. Additional qualitative examples for detailed grounded captioning.

Figure 11. Additional qualitative examples for referring expression segmentation.

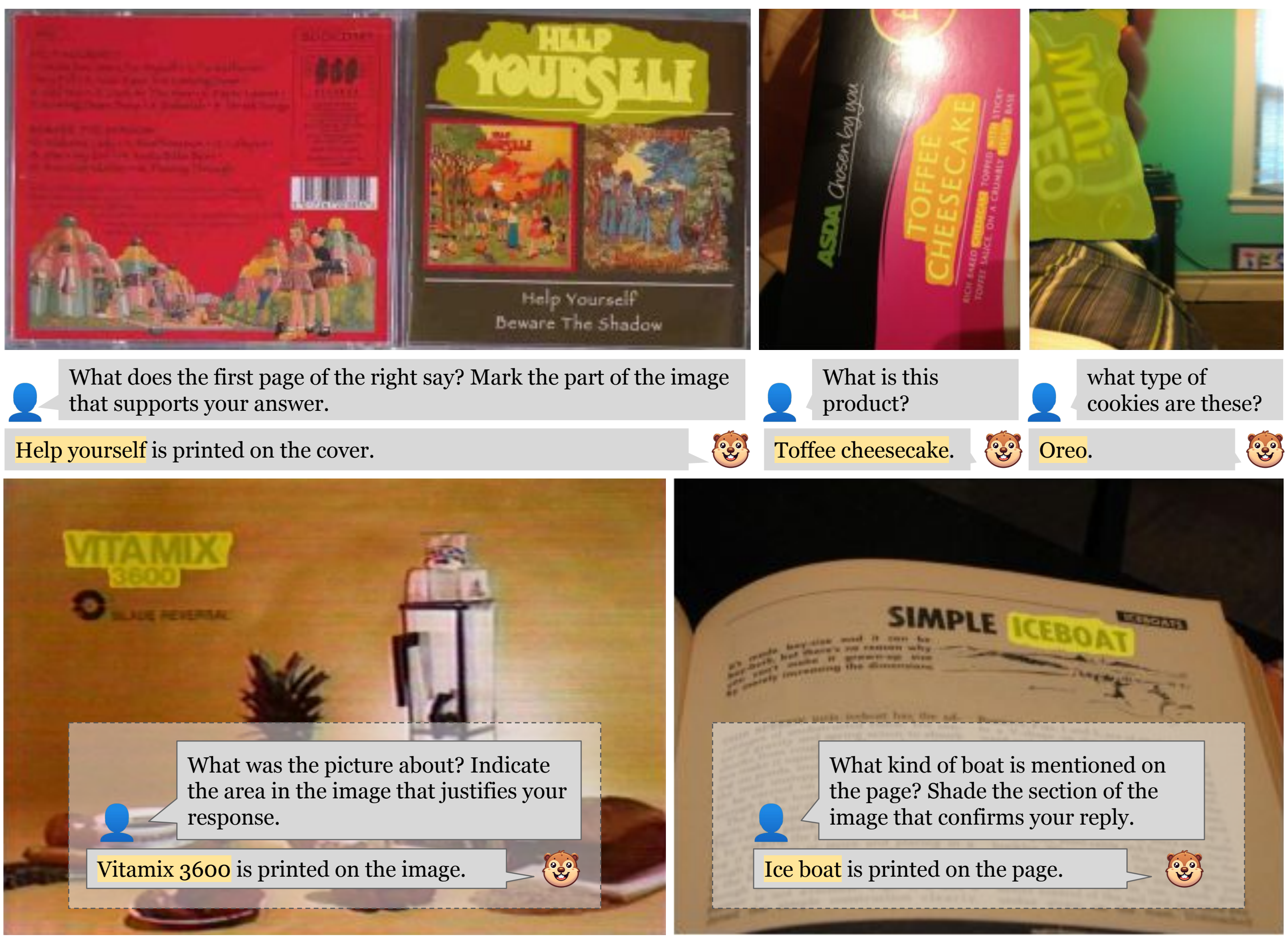


Figure 12. Additional qualitative examples for grounded VQA of visual text.

Figure 13. Additional qualitative examples for referential dialogue with pointer input.

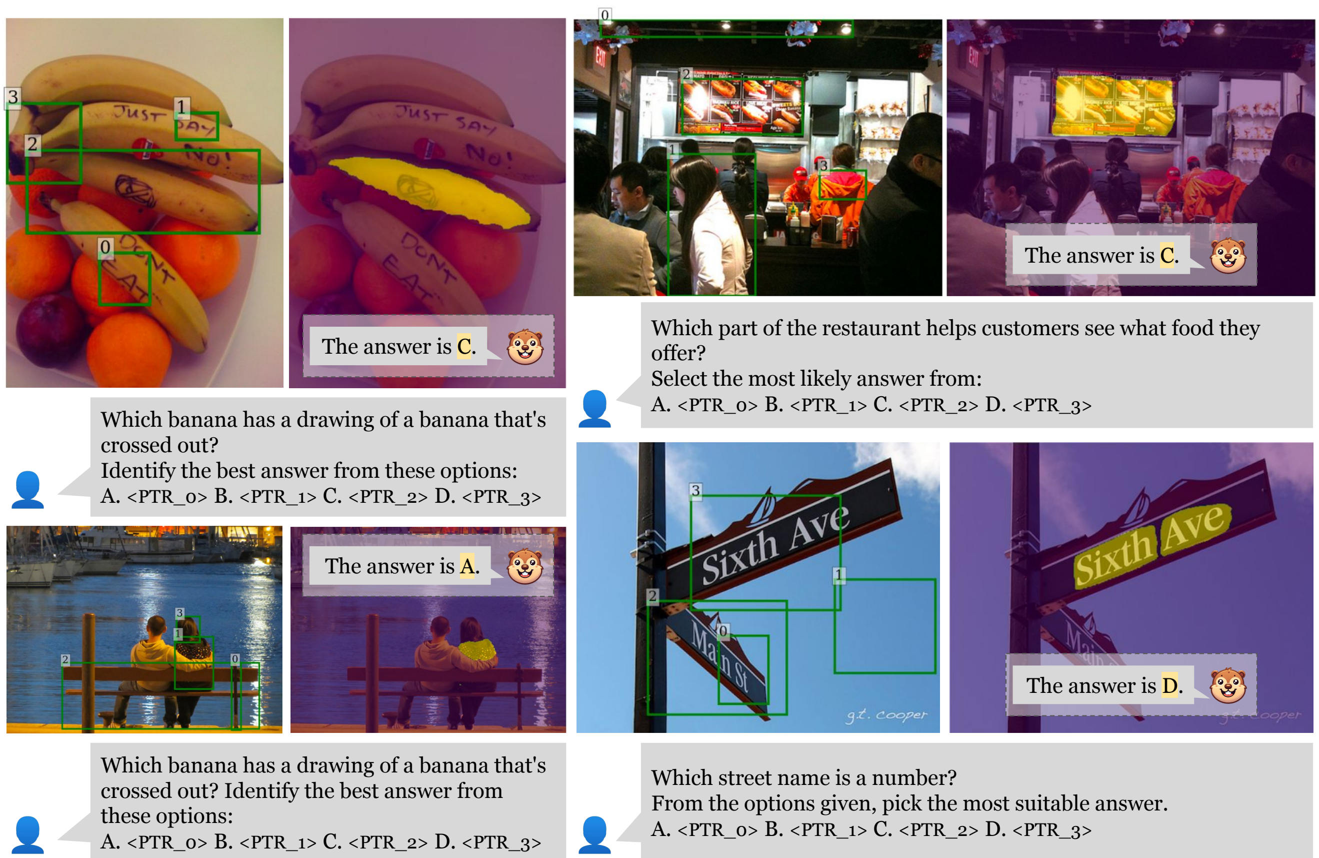
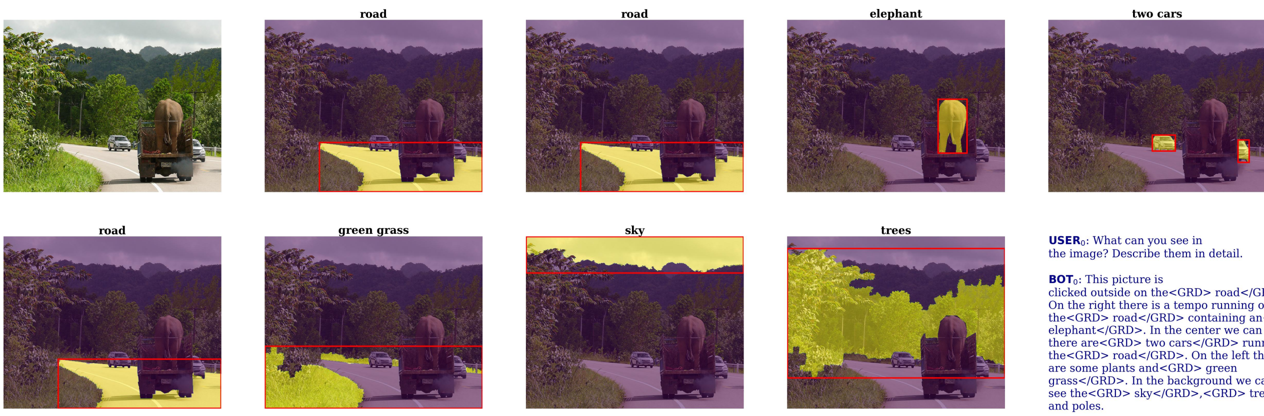


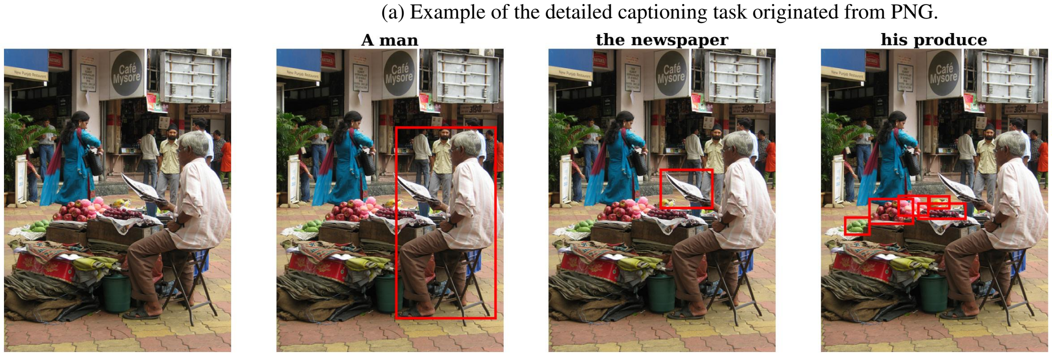
Figure 14. Additional qualitative examples for referential dialogue with pointers as multiple choices input.



USER$_0$: What can you see in the image? Describe them in detail.

BOT$_0$: This picture is clicked outside on the<GRD> road</GRD>. On the right there is a tempo running on the<GRD> road</GRD> containing an<GRD> elephant</GRD>. In the center we can see there are<GRD> two cars</GRD> running on the<GRD> road</GRD>. On the left there are some plants and<GRD> green grass</GRD>. In the background we can see the<GRD> sky</GRD>,<GRD> trees</GRD> and poles.

(a) Example of the detailed captioning task originated from PNG.



USER$_0$: Write a brief, yet comprehensive, description of the image.

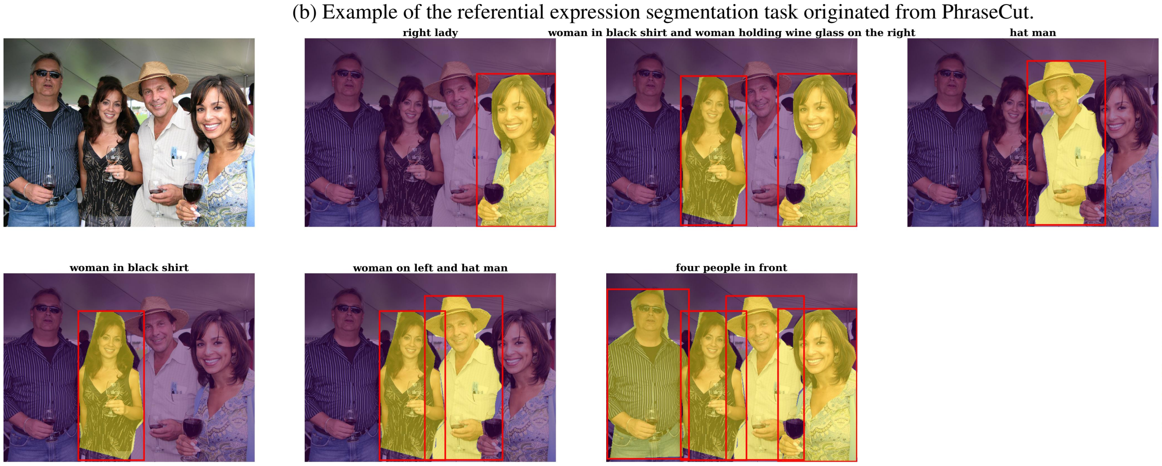
BOT$_0$: <GRD> A man</GRD> in reads<GRD> the newspaper</GRD> and sells<GRD> his produce</GRD>.

(b) Example of the short captioning task originated from Flickr30K-Entity.

Figure 15. Data samples for the GCAP task.

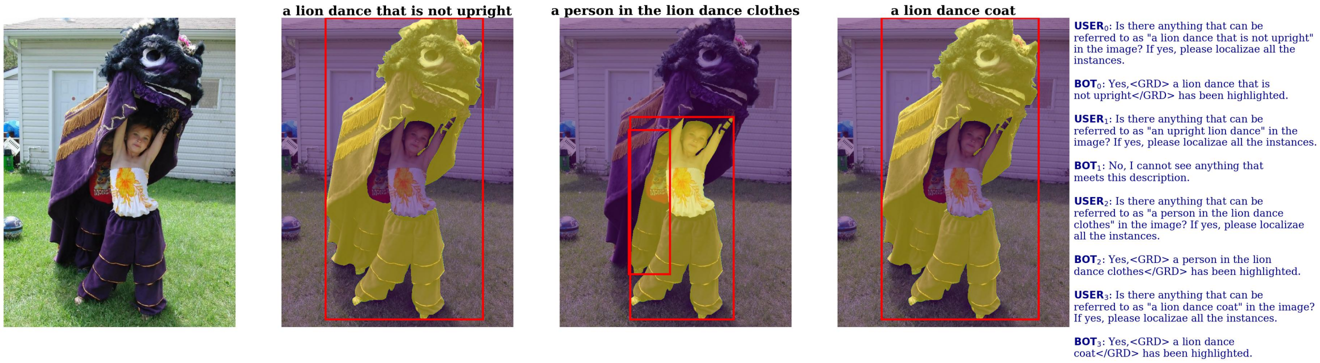
(a) Example of the referential expression segmentation task originated from RefCOCO+.

(b) Example of the referential expression segmentation task originated from PhraseCut.



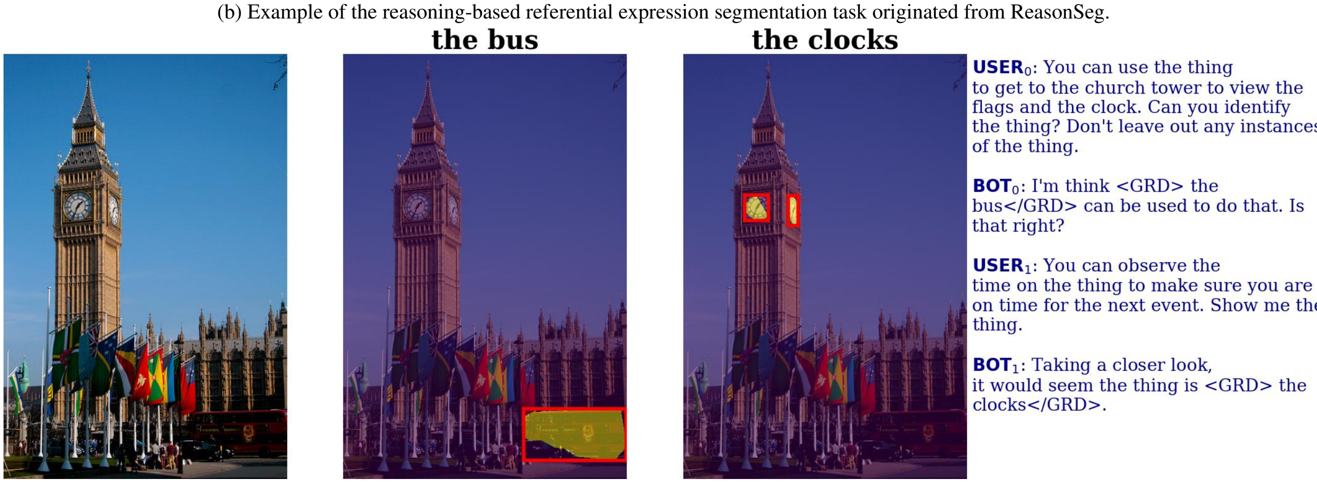
(c) Example of the generalized referential expression segmentation task originated from gRefCOCO.

Figure 16. Data samples for the RES task (part 1).



(a) Example of the generalized referential expression segmentation task originated from D-Cube.

(b) Example of the reasoning-based referential expression segmentation task originated from ReasonSeg.



(c) Example of the reasoning-based referential expression segmentation task originated from RIO.

(d) Example of the reasoning-based referential expression segmentation task originated from SK-VG.

Figure 17. Data samples for the RES task (part 2).

USER$_0$: Give a short answer to the question "Can you tell me the content of this notice?" based on the image. Show where in the image you found your answer.

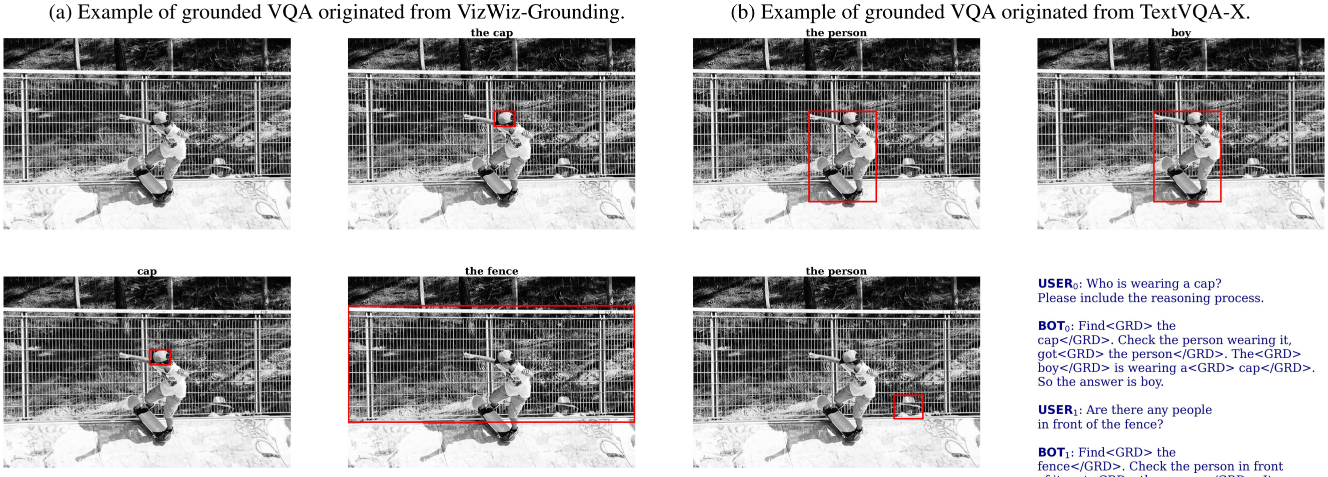
BOT$_0$: <GRD>No smoking</GRD>.

(a) Example of grounded VQA originated from VizWiz-Grounding.

USER$_0$: What are the two words on the bottom of the screen? Please highlight your evidence in the image.

BOT$_0$: <GRD>Livecash midland</GRD> are words at the bottom of the screen.

(b) Example of grounded VQA originated from TextVQA-X.



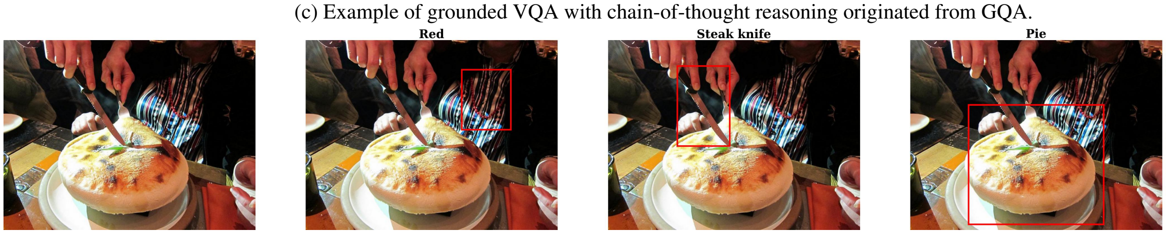
USER$_0$: Who is wearing a cap? Please include the reasoning process.

BOT$_0$: Find<GRD> the cap</GRD>. Check the person wearing it, got<GRD> the person</GRD>. The<GRD> boy</GRD> is wearing a<GRD> cap</GRD>. So the answer is boy.

USER$_1$: Are there any people in front of the fence?

BOT$_1$: Find<GRD> the fence</GRD>. Check the person in front of it, got<GRD> the person</GRD>. It exists. The answer is no.

(c) Example of grounded VQA with chain-of-thought reasoning originated from GQA.



USER$_0$: Give a concise answer to the question "What color is the lady's necklace?" Emphasize the part of the image that relates to your answer.

BOT$_0$: <GRD>Red</GRD>.

USER$_1$: What kind of knife is being used here?

BOT$_1$: <GRD>Steak knife</GRD>.

USER$_2$: What kind of food is this?

BOT$_2$: <GRD>Pie</GRD>.

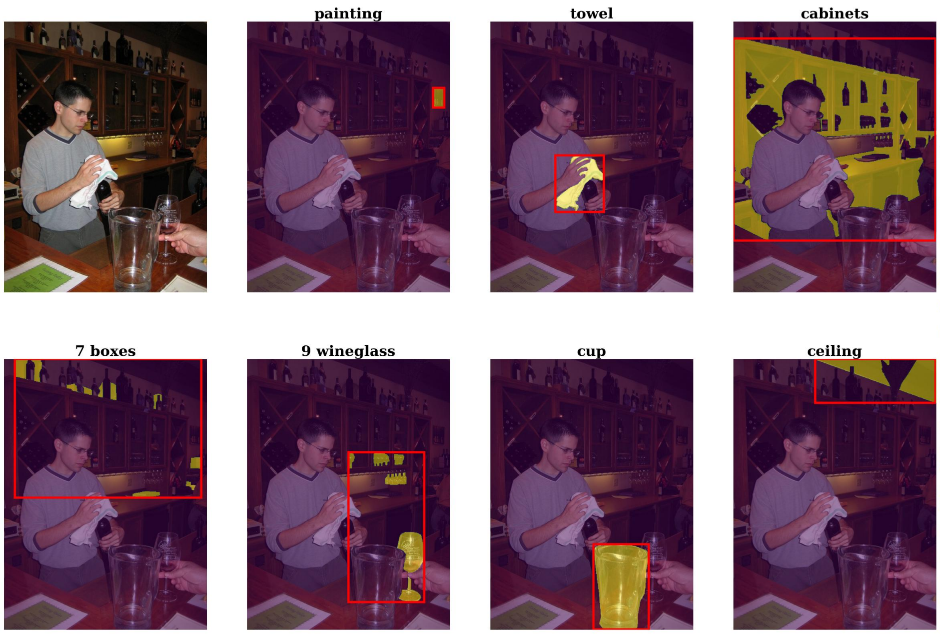
(d) Example of grounded VQA task originated from VQS.

USER$_0$: Is there an event "a girl with a number seven jersey throws a softball to her teammate" happening in the picture? Let's think step by step.

BOT$_0$: In the picture,<GRD> a girl</GRD> is wearing<GRD> a number seven jersey</GRD>. She is indeed throwing<GRD> a softball</GRD> to her<GRD> teammate</GRD>. Based on these details, the event "a girl with a number seven jersey throws a softball to her teammate" is happening in the picture. Therefore, the answer is yes.

(e) Example of the grounded VQA with chain-of-thought reasoning originated from Shikra-BinaryQA.

Figure 18. Data samples of the GVQA task (part 1).



USER$_0$: Have you spotted any sky in this picture?

BOT$_0$: After taking a close look at the image, I couldn't find any trace of sky.

USER$_1$: Can you give me a count of the women present in the image?

BOT$_1$: I've searched the image high and low, but no woman can be found.

USER$_2$: Have you spotted any painting in this picture?

BOT$_2$: Yes, the only <GRD> painting</GRD> is highlighted in the image.

USER$_3$: Can you see any floor in this image?

BOT$_3$: After taking a close look at the image, I couldn't find any trace of floor.

USER$_4$: Can you see any towel in this image?

BOT$_4$: Certainly, there's a single <GRD> towel</GRD>.

USER$_5$: Does the image contain any cabinets?

BOT$_5$: Certainly, there's a single <GRD> cabinets</GRD>.

USER$_6$: How many boxes are there?

BOT$_6$: Let me count! There are <GRD> 7 boxes</GRD> in total.

USER$_7$: How many wineglass can you identified in this image?

BOT$_7$: The count comes up to <GRD> 9 wineglass</GRD>.

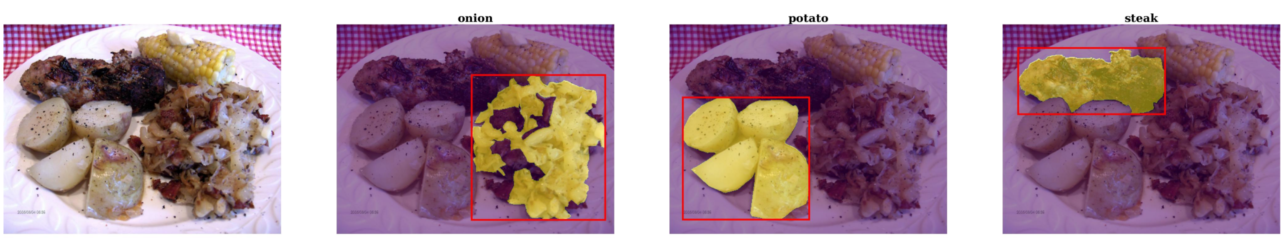
USER$_8$: How many cups are in the image?

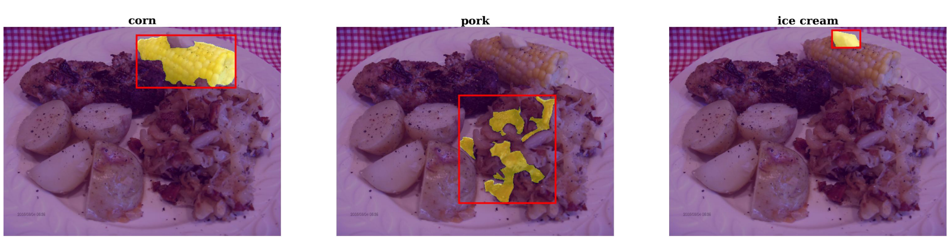
BOT$_8$: The only <GRD> cup</GRD> is highlighted in the image.

USER$_9$: Can you see any ceiling in this image?

BOT$_9$: Yes, here it is: <GRD> ceiling</GRD>.

(a) Example of the grounded VQA task originated from EntityCount.

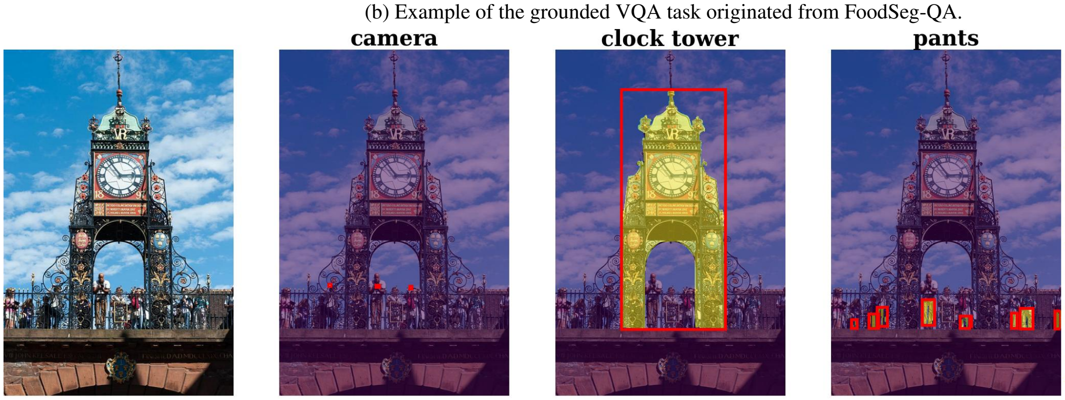




USER$_0$: Detail the ingredients for this dish for me.

BOT$_0$: Here's what goes into this dish: <GRD> onion</GRD>, <GRD> potato</GRD>, <GRD> steak</GRD>, <GRD> corn</GRD>, <GRD> pork</GRD>, and<GRD> ice cream</GRD>.

(b) Example of the grounded VQA task originated from FoodSeg-QA.



USER$_0$: Can you show me all the hats in the image?

BOT$_0$: Emm... It seems that there is no hat in the image. Do you want to try another object?

USER$_1$: How many cameras are there?

BOT$_1$: There are <GRD> 3 cameras</GRD> in the image.

USER$_2$: Segment clock tower.

BOT$_2$: Sure. The <GRD> clock tower</GRD> is highlighted.

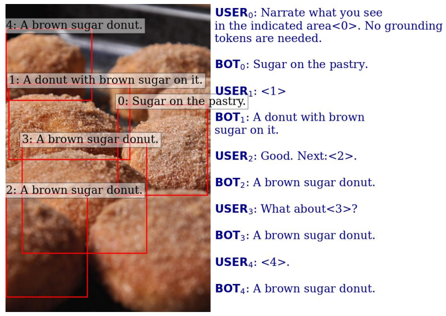
USER$_3$: How many pants can you see?

BOT$_3$: I can see <GRD> 9 pants</GRD> in this photo.

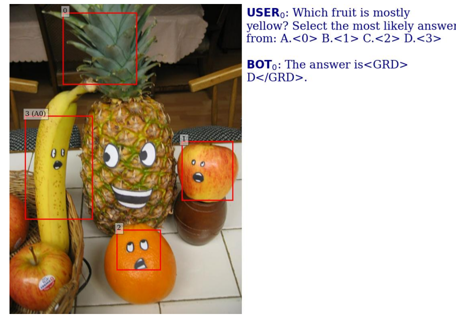
(c) Example of the grounded VQA task originated from LVIS-QA.

Figure 19. Data samples of the GVQA task (part 2).

(a) Refer expression generation (RefCOCO+).



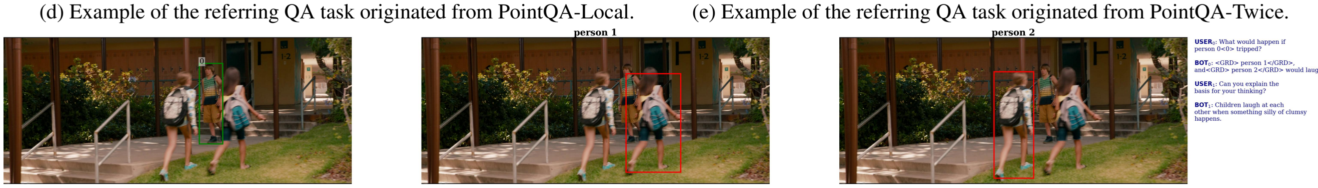
(b) Region captioning (VG).



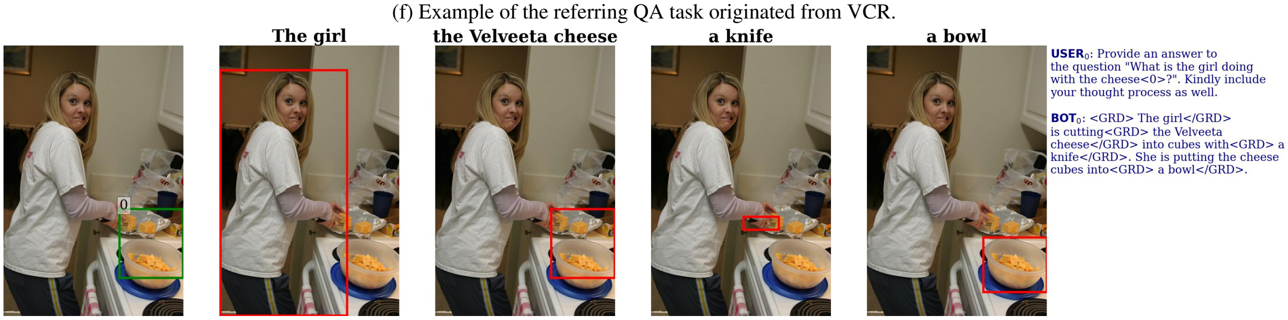
(c) Example of the referring QA task from V7W.

(d) Example of the referring QA task originated from PointQA-Local.

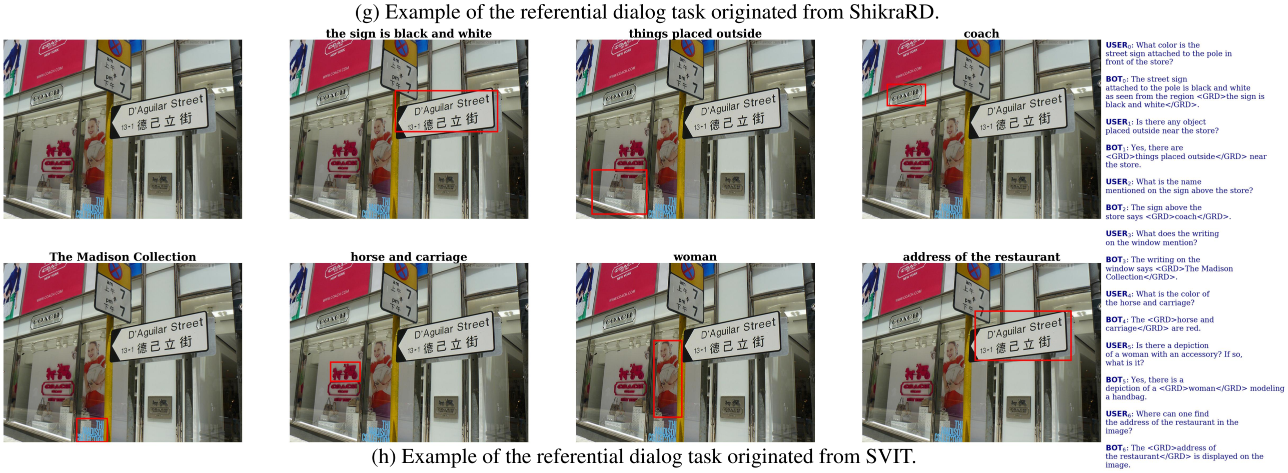
(e) Example of the referring QA task originated from PointQA-Twice.



(f) Example of the referring QA task originated from VCR.

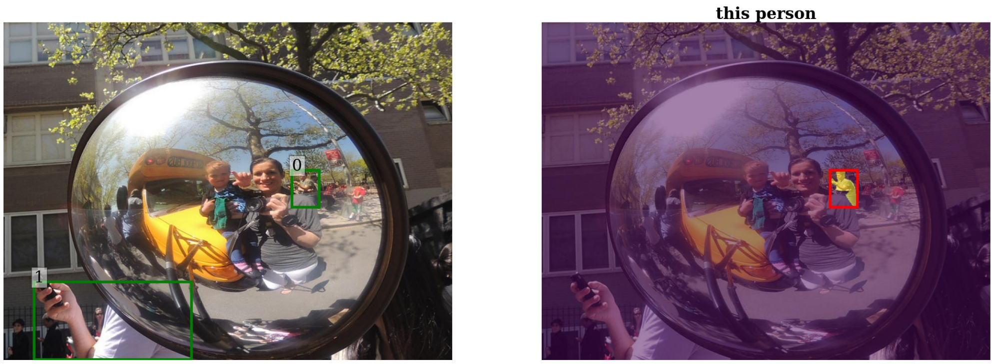


(g) Example of the referential dialog task originated from ShikraRD.



(h) Example of the referential dialog task originated from SVIT.

Figure 20. Data samples of the RD task (part 1).



USER$_0$: Given an object <0> in your mind, answer my yes-no questions to help me identify it. First question: is it a person?

BOT$_0$: Yes

USER$_1$: Is it a kid?

BOT$_1$: No

USER$_2$: Does it have a camera on their hand?

BOT$_2$: No

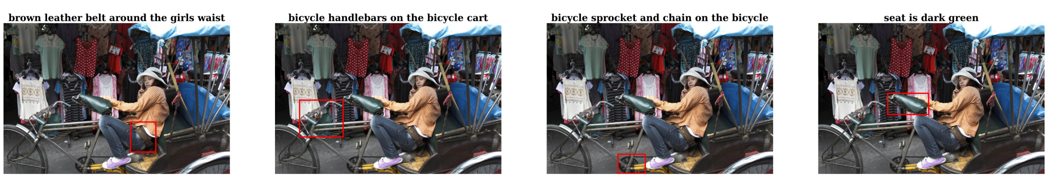
USER$_3$: Is it female?

BOT$_3$: Yes

USER$_4$: Here's my final answer: this person<1>. Did I nail it?

BOT$_4$: Good try, but the answer is <GRD> that person</GRD>.

(a) Example of the guesswhat game originated from GuessWhat.



USER$_0$: Match the descriptions with the regions. Descriptions: red and white polka dot dress, shirt is grey and black striped, light green shirt with white bodice, hanging on a hanger, front wheel of the cart and a silver fender, brown leather belt around the girls waist, bicycle handlebars on the bicycle cart, bicycle sprocket and chain on the bicycle, seat is dark green. Regions:<0> <1> <2> <3> <4> <5> <6> <7>

BOT$_0$: Based on the descriptions, the corresponding regions have been identified:<GRD> red and white polka dot dress</GRD>, <GRD> shirt is grey and black striped</GRD>, <GRD> light green shirt with white bodice, hanging on a hanger</GRD>, <GRD> front wheel of the cart and a silver fender</GRD>, <GRD> brown leather belt around the girls waist</GRD>, <GRD> bicycle handlebars on the bicycle cart</GRD>, <GRD> bicycle sprocket and chain on the bicycle</GRD>, <GRD> seat is dark green</GRD>.

(b) Example of the referential region matching game originated from VG.

USER$_0$: Read the text. <0>

BOT$_0$: <GRD>WATERSPORTS Tidal Wave</GRD>.

USER$_1$: Another one <1>

BOT$_1$: <GRD>843.886.8456</GRD>.

USER$_2$: This? <2>

BOT$_2$: <GRD>www.TidalWaveWatersports.com</GRD>.

(c) Example of the referred text reading task originated from HierText.

Figure 21. Data samples of the RD task (part 2).