KOSMOS-2: Grounding Multimodal Large Language Models to the World

Zhiliang Peng,* Wenhui Wang,* Li Dong,* Yaru Hao, Shaohan Huang, Shuming Ma, Furu Wei†
Microsoft Research
https://aka.ms/GeneralAI

Abstract

We introduce KOSMOS-2, a Multimodal Large Language Model (MLLM), enabling new capabilities of perceiving object descriptions (*e.g.*, bounding boxes) and grounding text to the visual world. Specifically, we represent refer expressions as links in Markdown, *i.e.*, "`[text span](bounding boxes)`", where object descriptions are sequences of location tokens. Together with multimodal corpora, we construct large-scale data of grounded image-text pairs (called GRIT) to train the model. In addition to the existing capabilities of MLLMs (*e.g.*, perceiving general modalities, following instructions, and performing in-context learning), KOSMOS-2 integrates the grounding capability into downstream applications. We evaluate KOSMOS-2 on a wide range of tasks, including (i) multimodal grounding, such as referring expression comprehension, and phrase grounding, (ii) multimodal referring, such as referring expression generation, (iii) perception-language tasks, and (iv) language understanding and generation. This work lays out the foundation for the development of Embodiment AI and sheds light on the big convergence of language, multimodal perception, action, and world modeling, which is a key step toward artificial general intelligence. Code and pretrained models are available at https://aka.ms/kosmos-2.
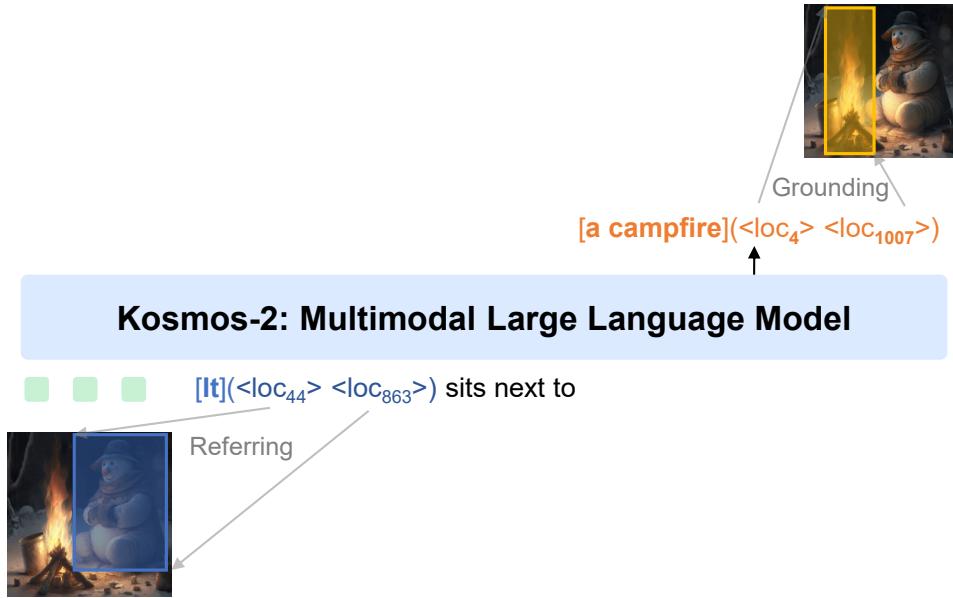
Figure 1: KOSMOS-2 is a multimodal large language model that has new capabilities of multimodal grounding and referring. KOSMOS-2 can understand multimodal input, follow instructions, perceive object descriptions (*e.g.*, bounding boxes), and ground language to the visual world.

---

* Equal contribution. † Corresponding author.
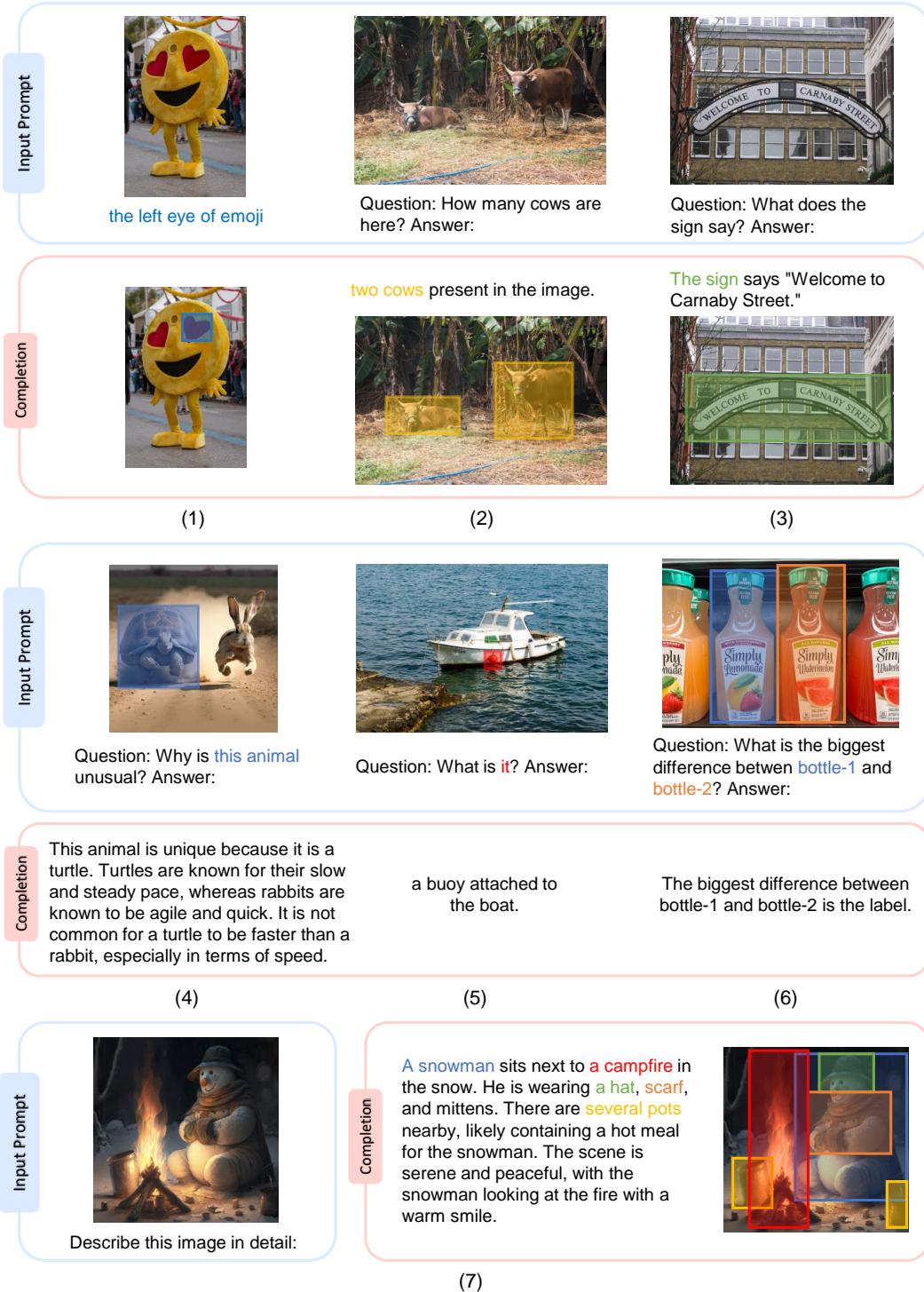
Figure 2: Selected examples generated from KOSMOS-2. The examples include (1) visual grounding, (2)-(3) grounded question answering, (4)-(6) multimodal referring via bounding boxes, and (7) grounded image captioning.

1 Introduction

Multimodal Large Language Models (MLLMs) [HSD⁺22, ADL⁺22, HDW⁺23, DXS⁺23, Ope23] have successfully played a role as a general-purpose interface across a wide range of tasks, such as language, vision, and vision-language tasks. MLLMs can perceive general modalities, including texts, images, and audio, and generate responses using free-form texts under zero-shot and few-shot settings.

In this work, we unlock the grounding capability for multimodal large language models. Grounding capability can provide a more convenient and efficient human-AI interaction for vision-language tasks. It enables the user to point to the object or region in the image directly rather than input detailed text descriptions to refer to it, the model can understand that image region with its spatial locations. Grounding capability also enables the model to respond with visual answers (*i.e.*, bounding boxes), which can support more vision-language tasks such as referring expression comprehension. Visual answers are more accurate and resolve the coreference ambiguity compared with text-only responses. In addition, grounding capability can link noun phrases and referring expressions in the generated free-form text response to the image regions, providing more accurate, informational, and comprehensive answers.

We introduce KOSMOS-2, a multimodal large language model with grounding capability built upon KOSMOS-1. KOSMOS-2 is a Transformer-based causal language model and is trained using the next-word prediction task. In order to unlock the grounding capability, we construct a web-scale dataset of grounded image-text pairs, and combine it with the multimodal corpora in KOSMOS-1 to train the model. The grounded image-text pairs are built upon a subset of image-text pairs from LAION-2B [SBV⁺22] and COYO-700M [BPK⁺22]. We construct a pipeline to extract and link the text spans (*i.e.*, noun phrases and referring expressions) in the caption to the spatial locations (*e.g.*, bounding boxes) of its corresponding objects or regions in the image. We convert the spatial coordinates of the bounding boxes to a sequence of location tokens, which is then appended after its respective text spans. The data format serves as a "*hyperlink*" to connect the objects or regions of the image to the caption.

Experimental results demonstrate that KOSMOS-2 not only achieves competitive performance on language and vision-language tasks evaluated in KOSMOS-1, but also achieves impressive performance on grounding tasks (phrase grounding and referring expression comprehension) and referring tasks (referring expression generation). As shown in Figure 2, integrating the grounding capability enables KOSMOS-2 to be used for more downstream tasks, such as grounded image captioning, and grounded visual question answering.

2 Construction of Web-Scale Grounded Image-Text Pairs (GRIT)

We introduce GRIT², a large-scale dataset of **Gr**ounded **I**mage-**T**ext pairs, which is created based on image-text pairs from a subset of COYO-700M [BPK⁺22] and LAION-2B [SBV⁺22]). We construct a pipeline to extract and link text spans (*i.e.*, noun phrases and referring expressions) in the caption to their corresponding image regions. The pipeline mainly consists of two steps: generating noun-chunk-bounding-box pairs and producing referring-expression-bounding-box pairs. We describe these steps in detail below:

Step-1: Generating noun-chunk-bounding-box pairs   Given an image-text pair, we first extract noun chunks from the caption and associate them with image regions using a pretrained detector. As illustrated in Figure 3, we use spaCy [HMVLB20] to parse the caption ("*a dog in a field of flowers*") and extract all noun chunks ("*a dog*", "*a field*" and "*flowers*"). We eliminate certain abstract noun phrases that are challenging to recognize in the image, such as "*time*", "*love*", and "*freedom*", to reduce potential noise. Subsequently, we input the image and noun chunks extracted from the caption into a pretrained grounding model (*e.g.*, GLIP [LZZ⁺22]) to obtain the associated bounding boxes. Non-maximum suppression algorithm is applied to remove bounding boxes that have a high overlap with others, even if they are not for the same noun chunk. We keep noun-chunk-bounding-box pairs with predicted confidence scores higher than 0.65. If no bounding boxes are retained, we discard the corresponding image-caption pair.

---

²A subset of GRIT can be downloaded at https://aka.ms/kosmos-2.

3
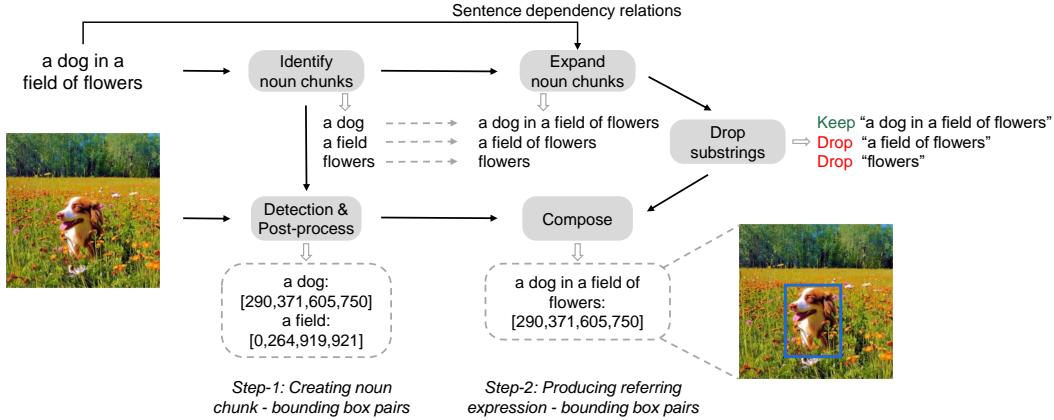
Figure 3: The pipeline of constructing web-scale grounded image-text pairs.

Dataset	Images	Objects	Text Spans	Avg Expression Length
Flickr Entities [PWC+15]	31,783	275,775	513,644	-
RefCOCOg [MHT+15]	26,711	54,822	85,474	8.43
RefCOCO [YPY+16]	19,994	50,000	142,209	3.61
RefCOCO+ [YPY+16]	19,992	49,856	141,564	3.53
Visual Genome [KZG+16]	108,077	4,102,818	-	-
GRIT (Ours)	90,614,680	137,349,210	114,978,233	4.7

Table 1: Comparison GRIT with existing visual grounding datasets.

Step-2: Producing referring-expression-bounding-box pairs In order to endow the model with the ability to ground complex linguistic descriptions, we expand noun chunks to referring expressions. Specifically, we use spaCy to obtain dependency relations of the sentence. We then expand a noun chunk into a referring expression by recursively traversing its children in the dependency tree and concatenating children tokens with the noun chunk. We do not expand noun chunks with conjuncts. For noun chunks without children tokens, we keep them for the next process. In the example shown in Figure 3, the noun chunk '*a dog*' can be expanded to "*a dog in a field of flowers*", and the noun chunk '*a field*' can be expanded to "*a field of flowers*".

Furthermore, we only retain referring expressions or noun chunks that are not contained by others. As shown in Figure 3, we keep the referring expression "*a dog in a field of flowers*" and drop "*a field of flowers*" (as it is entailed by "*a dog in a field of flowers*") and '*flowers*'. We assign the bounding box of the noun chunk ('*a dog*') to the corresponding generated referring expression ("*a dog in a field of flowers*").

In the end, we obtain approximately 91M images, 115M text spans, and 137M associated bounding boxes. We compare GRIT with existing publicly accessible visual grounding datasets in Table 1. Data samples of GRIT are shown in the Appendix.

3 KOSMOS-2: A Grounded Multimodal Large Language Model

KOSMOS-2 is a grounded multimodal large language model, which integrates grounding and referring capabilities compared with KOSMOS-1. The model can accept image regions selected by the user using bounding boxes as input, provide visual answers (*i.e.*, bounding boxes), and ground the text output to the visual world. KOSMOS-2 adopts the same model architecture and training objective as KOSMOS-1. We add grounded image-text pairs into the training data to endow the model with grounding and referring capabilities. For a text span (such as noun phrase and referring expression) and its corresponding bounding boxes in a grounded image-text pair, We discretize continuous coordinates of bounding boxes into a sequence of location tokens to encode with text tokens in a unified way. Then we link the location tokens and their corresponding text span via a "*hyperlink*" data

format. The model is trained to establish a mapping between image regions and their corresponding location tokens and connect the image regions with their associated text spans.

3.1 Grounded Input Representations

Given a text span and its associated bounding boxes in a grounded image-text pair, we first convert the continuous coordinates of bounding boxes into a sequence of discrete location tokens [CSL$^+$21]. For an image with width W and height H, we evenly divide both the width and height into P segments each. $P \times P$ bins are obtained and each bin consists of $(W/P) \times (H/P)$ pixels. For each bin, we use a location token to represent the coordinates within that bin. We use the coordinates of the center pixel of each bin to determine bounding boxes on the image. In total, $P \times P$ location tokens are introduced, and these tokens are added to word vocabulary to enable unified modeling with texts.

The bounding box can be represented using its top-left point (x_1, y_1) and bottom-right point (x_2, y_2). We discretize the top-left and bottom-right corner points to location tokens, respectively. We concatenate the top-left location token `<loc`$_1$`>`, the bottom-right location token `<loc`$_2$`>`, and special boundary tokens `<box>` and `</box>`, to represent a single bounding box: "`<box><loc`$_1$`><loc`$_2$`></box>`". If the text span is associated with multiple bounding boxes, we use a special token `<delim>` to concatenate the location tokens of these bounding boxes: "`<box><loc`$_1^i$`><loc`$_2^i$`><delim>`...`<loc`$_1^j$`><loc`$_2^j$`></box>`".

Then we arrange the text span and its associated location tokens in a format resembling a "*hyperlink*" in markdown. For the text span with a single bounding box, the resulted sequence is "`<p>` *text span* `</p><box><loc`$_1$`><loc`$_2$`></box>`", where `<p>` and `</p>` are special tokens indicating the beginning and end of the text span. The data format tells the model that image regions within the bounding box are associated with the text span.

For the example shown in Figure 1, the input representation is:

> `<s> <image>` Image Embedding `</image> <grounding> <p>` It `</p><box><loc`$_{44}$`><loc`$_{863}$`></box>` seats next to `<p>` a campfire `</p><box><loc`$_4$`><loc`$_{1007}$`></box> </s>`

where `<s>` and `</s>` indicate start- and end-of-sequence, and `<image>` and `</image>` represent the beginning and end of encoded image embeddings. `<grounding>` is a special token to tell the model ground the text output to the visual world. We map input text tokens and location tokens to embeddings via a lookup table. Following KOSMOS-1, a vision encoder and a resampler module are used to obtain image embeddings for input images.

For language-only data, cross-modal paired data (*i.e.*, image-text pairs), and interleaved multimodal data, we use the same input representations as of KOSMOS-1.

3.2 Grounded Multimodal Large Language Models

Based on KOSMOS-1, KOSMOS-2 enhances multimodal large language models by incorporating grounding and referring capabilities. KOSMOS-2 also uses a Transformer-based causal language model as the backbone and is trained with the next-token prediction task.

In addition to multimodal corpora used in KOSMOS-1 (including text corpora, image-caption pairs, and interleaved image-text data), we add grounded image-text pairs into training. The training loss only considers discrete tokens, such as text tokens and location tokens. The model can learn to locate and understand image regions by their location tokens and the whole image, associate text spans to image regions, and output bounding boxes of the image region using location tokens.

KOSMOS-2 shows new capabilities of grounding and referring. The referring capability enables us to point out image regions with bounding boxes. KOSMOS-2 can understand the image regions users refer to by the coordinates of bounding boxes. The referring capability provides a new interaction method. Different from previous MLLMs [ADL$^+$22, HSD$^+$22, HDW$^+$23], which can only provide text output, KOSMOS-2 can provide visual answers (*i.e.*, bounding boxes) and ground text output to the image. The grounding capability enables the model to provide more accurate, informative, and comprehensive responses. In addition to vision, language, and vision-language tasks evaluated in

KOSMOS-1, the model can be used for more downstream tasks, such as grounded image-captioning, grounded VQA, referring expression comprehension and generation.

3.3 Model Training

Training Setup   We train the model on newly added grounded image-text pairs, monomodal text corpora, image-caption pairs, and interleaved image-text data. Our training process involves a batch size of 419K tokens, consisting of 185K tokens from text corpora, 215K tokens from original and grounded image-caption pairs, and 19K tokens from interleaved data. We train KOSMOS-2 for 60k steps, equivalent to around 25 billion tokens. The AdamW optimizer is employed with $\beta = (0.9, 0.98)$. We set the weight decay to 0.01 and the dropout rate to 0.1. The learning rate increases to 2e-4 during the first 375 warm-up steps and linearly decays to zero. We train the model on 256 V100 GPUs and the training takes approximately one day to complete. In order to tell the model when to ground text output to the visual world, we prepend the '`<grounding>`' token to the grounded caption during training.

Following KOSMOS-1, the vision encoder has 24 layers with 1,024 hidden size and 4,096 FFN intermediate size. The multimodal large language model component is a 24-layer MAGNETO Transformer [WMH+22, MWH+22] with 2,048 hidden dimensions, 32 attention heads, and 8,192 FFN intermediate size. The total number of trainable parameters amounts to approximately 1.6B. The image resolution is set to 224×224 and the patch size is 14×14. We divide the width and height of the image into 32 bins, with each bin consisting of 7×7 pixels. A total of 32×32 location tokens are added to the vocabulary. KOSMOS-2 uses the weights of KOSMOS-1 for initialization, the newly added word embeddings of location tokens are initialized randomly. We update all the parameters during training and instruction tuning.

Instruction Tuning   After the model is trained, we perform instruct tuning to better align KOSMOS-2 with human instructions. we combine vision-language instruction dataset (*i.e.*, LLaVA-Instruct [LLWL23]) and language-only instruction datasets (*i.e.*, Unnatural Instructions [HSLS22] and FLANv2 [LHV+23]) with the training data to tune the model. In addition, we construct grounded instruction data by utilizing the pairs of bounding boxes and expressions (*i.e.*, noun phrases, and referring expressions) in GRIT. Given an expression-bounding-box pair, we use "`<p>` *expression* `</p>`" as the input instruction, and prompt the model to generate the corresponding location tokens of the bounding boxes. We also use the prompt like "`<p>` *It* `</p><box><loc₁><loc₂></box>` *is*" to ask the model to generate expressions according to its bounding boxes. Table B in Appendix presents more templates.

4   Evaluation

We first evaluate KOSMOS-2 on multimodal grounding and multimodal referring tasks to assess the new capabilities, and then test the model on language and perception-language tasks evaluated in KOSMOS-1.

- Multimodal grounding
  - Phrase grounding
  - Referring expression comprehension
- Multimodal referring
  - Referring expression generation
- Perception-language tasks
  - Image captioning
  - Visual question answering
- Language tasks
 - Language understanding
  - Language generation
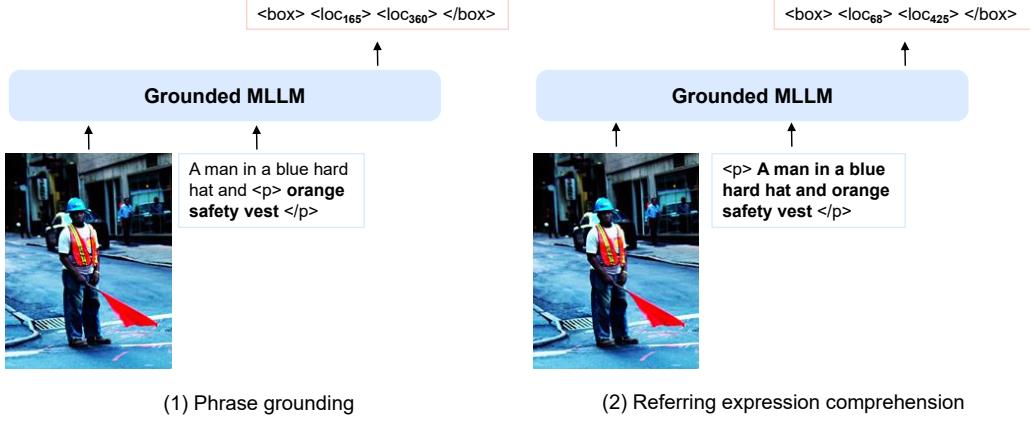
Figure 4: Input format of evaluation on (1) phrase grounding and (2) referring expression comprehension.

4.1 Multimodal Grounding

In order to evaluate the ability of multimodal grounding, we test KOSMOS-2 on widely used phrase grounding and referring expression comprehension tasks in a generation manner. Phrase grounding task requires the model to predict a set of bounding boxes based on one or more given phrases that maybe interrelated within a single caption. Referring expression comprehension task encourages the model to locate the object described in a text referring expression within a given image.

By testing KOSMOS-2 on these two tasks, we can assess how well the model performs in grounding text descriptions to the visual world, which is crucial for developing advanced AI systems capable of handling complex multimodal tasks.

For both phrase grounding and referring expression comprehension tasks, KOSMOS-2 is required to generate location tokens which are then converted to bounding boxes for evaluation. The input format is "`<s><image>` Image Embedding `</image><grounding>`...", where "`<grounding>`" is used to prompt the model to generate locations tokens.

4.1.1 Phrase Grounding

We evaluate phrase grounding task on Flickr30k Entities [PWC+15] val and test splits. In order to reduce ambiguity, we do not prompt the model with individual phrases; instead, we use the current phrase along with the preceding words as input where preceding words serve as context: " ... `<p>` {*phrase*} `</p>`". For the example shown in Figure 4(1), the model needs to predict the locations of phrases "*A man*", "*a blue hard hat*", "*orange safety vest*" and "*an intersection*" in the caption "*A man in a blue hard hat and orange safety vest stands in an intersection.*". To generate the location tokens for the phrase "*A man*" that is the beginning of the caption, the prompt is "`<p>`*A man*`</p>`". For the phrase "*orange safety vest*", the prompt is "*A man in a blue hard hat and* `<p>`*orange safety vest*`</p>`". When multiple men are in the image, the context "*A man in a blue hard hat and*" explicitly helps the model locate the object to reduce ambiguity.

We obtain the location tokens in "`<box>`...`</box>`" from the model response and then covert it into bounding boxes. The generated bounding box is correct if its intersection over union (IoU) with the ground-truth bounding box is greater than 0.5. If KOSMOS-2 generates a location sequence that can not be converted correctly (*e.g.*, "`<box><loc_1></box>`"), we treat it as a negative sample. We use ANY-BOX protocol in MDETR [KSL+21]. We report the R@1, R@5, and R@10 metrics, where R@1/5/10 means calculating the recall using the top 1/5/10 generated bounding boxes. If there are fewer than 5 or 10 bounding boxes generated by KOSMOS-2, we use all available bounding boxes for the calculation.

Results Table 2 presents results on Flickr30k Entities [PWC+15] val and test splits. KOSMOS-2 achieves impressive zero-shot performance and outperforms GRILL [JMC+23], which relies on an attached detector, by a large margin. Moreover, our model outperforms traditional finetuned

Model	Zero-shot	Val Split			Test Split		
		R@1	R@5	R@10	R@1	R@5	R@10
VisualBert [LYY+19]	✗	70.4	84.5	86.3	71.3	85.0	86.5
MDETR [KSL+21]	✗	83.6	93.4	95.1	84.3	93.9	95.8
GLIP [LZZ+22]	✗	86.7	96.4	97.9	87.1	96.9	98.1
FIBER [DKG+22]	✗	87.1	96.1	97.4	87.4	96.4	97.6
GRILL [JMC+23]	✓	-	-	-	18.9	53.4	70.3
KOSMOS-2	✓	77.8	79.2	79.3	78.7	80.1	80.1

Table 2: Phrase grounding results on Flickr30k Entities. We report the R@1, R@5, and R@10 metrics, where R@1/5/10 means calculating the recall using the top 1/5/10 generated bounding boxes.

Model	Zero-shot	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
UNITER [CLY+19]	✗	81.41	87.04	74.17	75.90	81.45	66.70	74.86	75.77
MDETR [KSL+21]	✗	87.51	90.40	82.67	81.13	85.52	72.96	83.35	83.31
OFA [WYM+22]	✗	90.05	92.93	85.26	84.49	90.10	77.77	84.54	85.20
FIBER [DKG+22]	✗	90.68	92.59	87.26	85.74	90.13	79.38	87.11	87.32
VisionLLM [WCC+23]	✗	86.7	-	-	-	-	-	-	-
GRILL [JMC+23]	✓	-	-	-	-	-	-	-	47.5
KOSMOS-2	✓	52.32	57.42	47.26	45.48	50.73	42.24	60.57	61.65

Table 3: Referring expression comprehension results on RefCOCO, RefCOCO+ and RefCOCOg. We report the accuracy metric for all methods.

VisualBert [LYY+19] model by 7.4% R@1 on both val and test splits. In contrast to other models, KOSMOS-2 does not involve prior designs (*e.g.*, object queries or proposals), leading to similar results among R@1, R@5, and R@10. These results demonstrate that KOSMOS-2 can generate high-quality locations without the need for post-processing redundant locations. This capability highlights the effectiveness of our model in handling phrase grounding tasks.

4.1.2 Referring Expression Comprehension

We assess the referring expression comprehension task using three well-established datasets: RefCOCO [YPY+16], RefCOCO+ [YPY+16] and RefCOCOg [MHT+15]. Both RefCOCO and RefCOCO+ were generated through a two-player game, with RefCOCO+ specifically designed to exclude spatial relations, such as "on the left". RefCOCOg incorporates spatial relations and features longer expressions on average. Different from phrase grounding on Flickr30k entities, we measure this task by using referring expression as the input: "<p> *referring expression* </p>". For the example shown in Figure 4(2), the input sequence is "<p>*A man in a blue hard hat and orange safety vest*</p>". Similarly, the predicted bounding box is considered correct only if its IOU with the ground-truth bounding box is greater than 0.5. The failed decoded sequence is also treated as a negative sample. We use the first generated bounding box for the query expression to measure the accuracy.

Results Table 3 reports referring comprehension results on RefCOCO [YPY+16], RefCOCO+ [YPY+16] and RefCOCOg [MHT+15]. KOSMOS-2 also obtains promising zero-shot performance on the comprehension task, significantly outperforming previous zero-shot models on RefCOCOg benchmark. However, compared to previous finetuned works, KOSMOS-2 achieves slightly lower performance on RefCOCO and RefCOCO+ than on RefCOCOg. This discrepancy can be attributed to the data distribution present in RefCOCO and RefCOCO+, where they tend to use a shorter referring expression (*e.g.*, "left bottom") during the two-player game. Hence, one of our future goals is to enhance MLLMs' ability to accurately understand more types of human expressions.
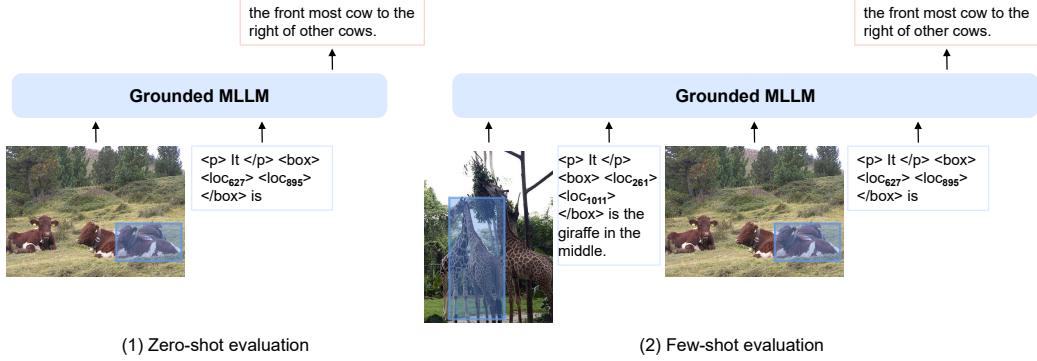
Figure 5: The input format of referring expression generation evaluation under (1) zero-shot and (2) few-shot settings. The bounding boxes shown in the image are for visualization purposes.

4.2 Multimodal Referring

In addition to multimodal grounding tasks, we evaluate the model's ability to understand image regions or objects users refer to via inputting bounding boxes. Compared with previous multimodal LLMs that can only refer image regions or objects to the model via detailed text descriptions, directly referring to image regions using its bounding boxes is more effective and reduces ambiguity.

We evaluate the model on the referring expression generation task, which aims to generate unambiguous text descriptions of specific objects or regions within the bounding box. We employ the widely used RefCOCOg dataset [MHT+15] to evaluate the model's performance under both zero-shot and few-shot settings, showcasing its adaptability in different scenarios.

4.2.1 Evaluation Setup

The model is tasked with generating an associated text description for an object or region given its location tokens of the bounding boxes (*e.g.*, "<box><loc$_1$><loc$_2$></box>"). Benefiting from the unified input format, we use "<p> *It* </p><box><loc$_1$><loc$_2$></box> *is*" as prompt to encourage the model to predict its text description. Figure 5 (1) and (2) demonstrate the input format for zero-shot and few-shot referring expression generation, respectively. Following previous works, we report results using METEOR and CIDEr metrics. The image resolution is 224×224. Greedy search is used for decoding.

4.2.2 Results

Table 4 presents the zero-shot and few-shot results of referring expression generation on RefCOCOg. We compare KOSMOS-2 with a finetuned listener-speaker model, which introduces an added reward-based module (SLR). Our model obtains impressive zero-shot performance on referring expression generation, and even outperforms finetuned SLR by 1.1 CIDEr scores. Moreover, when prompted with fewshot demonstrations, KOSMOS-2 shows further improvements, highlighting its in-context learning ability.

Model	Setting	RefCOCOg	
		Meteor	CIDEr
SLR[YTBB17]	Finetuning	15.4	59.2
SLR+Rerank[YTBB17]	Finetuning	15.9	66.2
KOSMOS-2	Zero-shot	12.2	60.3
	Few-shot ($k = 2$)	13.8	62.2
	Few-shot ($k = 4$)	14.1	62.3

Table 4: Results of referring expression generation on RefCOCOg.

4.3 Perception-Language Tasks

In addition to multimodal grounding and referring tasks, we also evaluate KOSMOS-2 on the vision-language tasks following KOSMOS-1. In particular, we perform zero-shot evaluations on two popular tasks, including image captioning and visual question answering. Image captioning requires the model to generate a text description of the given image, whereas visual question answering seeks to answer a natural language question based on an image. In order to have a fair comparison with KOSMOS-1, we report results without instruction tuning.

4.3.1 Evaluation Setup

For image captioning, we evaluate the model on the widely used Flickr30k *Karpathy split* test set. We employ beam search for caption generation, with a beam size of 5. We report results using CIDEr [VLZP15] metrics evaluated by COCOEvalCap[3]. We use the prompt *"An image of"* to generate the image description.

For visual question-answering, we evaluate zero-shot performance on the test-dev set of VQAv2. Greedy search is used for decoding. We report VQA scores obtained from VQAv2 evaluation server[4]. *"Question: {question} Answer: {answer}"* is used as the prompt for the dataset. The image resolution is 224×224 for both two tasks.

4.3.2 Results

We present the zero-shot performance on Flickr30k and VQAv2 in Table 5. KOSMOS-2 exhibits comparable overall performance to the KOSMOS-1, showing a slight improvement on Flickr30k while experiencing a marginal decrease on VQA. While KOSMOS-2 introduces new capabilities of grounding and referring, the model still achieves competitive performance on perception-language tasks.

Model	Flickr30k	VQAv2
	CIDEr	VQA acc.
FewVLM [JCS+22]	31.0	-
METALM [HSD+22]	43.4	41.1
Flamingo-3B [ADL+22]	60.6	49.2
Flamingo-9B [ADL+22]	61.5	51.8
KOSMOS-1	65.2	46.7
KOSMOS-2	66.7	45.6

Table 5: Zero-shot image captioning results on Flickr30k test set and zero-shot visual question answering results on VQAv2 test-dev set. We report results of KOSMOS-2 and KOSMOS-1 without instruction tuning.

4.4 Language Tasks

We evaluate KOSMOS-2 on eight language tasks, such as cloze and completion tasks (StoryCloze, HellaSwag), Winograd-style tasks (Winograd, Winogrande), commonsense reasoning (PIQA), and three SuperGLUE benchmark [WPN+19] datasets (BoolQ, CB, and COPA). We report the zero-shot results in Table 6. Compared with KOSMOS-1, KOSMOS-2 achieves similar performance on StoryCloze, HellaSwag, Winograd, Winogrande, and PIQA, experiences a decrease in performance on CB, but shows improvement on BoolQ and COPA. In summary, KOSMOS-2 demonstrates the acquisition of new capabilities while experiencing comparable performance on language tasks. This illustrates the potential of the model in balancing and expanding its skills across different domains.

---

[3] https://github.com/salaniz/pycocoevalcap
[4] https://eval.ai/challenge/830/overview

Model	Story Cloze	Hella Swag	Winograd	Winogrande	PIQA	BoolQ	CB	COPA
LLM	72.9	50.4	71.6	56.7	73.2	56.4	39.3	68.0
KOSMOS-1	72.1	50.0	69.8	54.8	72.9	56.4	44.6	63.0
KOSMOS-2	72.0	49.4	69.1	55.6	72.9	62.0	30.4	67.0

Table 6: Zero-shot performance comparisons of language tasks between KOSMOS-2, KOSMOS-1 and LLM. LLM uses the same text data and training setup to reimplement a language model as KOSMOS-1. We report results of KOSMOS-2 and KOSMOS-1 without instruction tuning. Results of KOSMOS-1 and the LLM baseline are from [HDW+23].

5 Conclusion

We present KOSMOS-2, a multimodal large language modal, that can ground to the visual world. Specifically, we pre-train KOSMOS-2 by augmenting the multimodal corpora used in KOSMOS-1 with GRIT, a large-scale dataset of Grounded Image-Text pairs, which is created by extracting and associating noun phrases and referring expressions in the caption to the objects or regions in the scene. KOSMOS-2 enables new capabilities of perceiving image regions and grounding text output to the visual world, which makes grounding as a foundation capability of MLLMs in many downstream applications. Experimental results demonstrate that KOSMOS-2 achieves impressive results on language and vision-language tasks evaluated in KOSMOS-1, grounding tasks including phrase grounding and referring expression comprehension, and referring tasks such as referring expression generation.

Acknowledgement

Some examples (such as Figure 1) are taken from the WHOOPS corpus [BGBH+23].

Ethics Statement

The model presented in this paper is intended for academic and research purposes. The utilization of the model to create unsuitable material is strictly forbidden and not endorsed by this work. The accountability for any improper or unacceptable application of the model rests exclusively with the individuals who generated such content. We also put Microsoft AI Principles[5] into practice when developing the models.

References

[ADL+22] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022.

[AHR+22] Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. CM3: A causal masked multimodal model of the Internet. *ArXiv*, abs/2201.07520, 2022.

[BGBH+23] Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: WHOOPS! a vision-and-language benchmark of synthetic and compositional images. *ArXiv*, abs/2303.07274, 2023.

---

[5] https://www.microsoft.com/ai/responsible-ai

[BPK+22] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset, 2022.

[CLY+19] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, 2019.

[CSL+21] Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geo rey E. Hinton. Pix2seq: A language modeling framework for object detection. *ArXiv*, abs/2109.10852, 2021.

[DKG+22] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, Jianfeng Gao, and Lijuan Wang. Coarse-to-fine vision-language pre-training with fusion in the backbone. *ArXiv*, abs/2206.07643, 2022.

[DXS+23] Danny Driess, F. Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Ho Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Peter R. Florence. Palm-e: An embodied multimodal language model. *ArXiv*, abs/2303.03378, 2023.

[HDW+23] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. *ArXiv*, abs/2302.14045, 2023.

[HMVLB20] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020.

[HSD+22] Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. Language models are general-purpose interfaces. *ArXiv*, abs/2206.06336, 2022.

[HSLS22] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor, 2022.

[JCS+22] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2763–2775, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[JMC+23] Woojeong Jin, Subhabrata Mukherjee, Yu Cheng, Yelong Shen, Weizhu Chen, Ahmed Hassan Awadallah, Damien Jose, and Xiang Ren. Grill: Grounded vision-language pre-training via aligning text and image regions. *ArXiv*, abs/2305.14676, 2023.

[KSL+21] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr - modulated detection for end-to-end multi-modal understanding. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1760–1770, 2021.

[KZG+16] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016.

[LHV+23] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023.

[LLSH23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023.

[LLWL23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[LYY+19] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *ArXiv*, abs/1908.03557, 2019.

[LZZ+22] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022.

[MHT+15] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana-Maria Camburu, Alan Loddon Yuille, and Kevin P. Murphy. Generation and comprehension of unambiguous object descriptions. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20, 2015.

[MWH+22] Shuming Ma, Hongyu Wang, Shaohan Huang, Wenhui Wang, Zewen Chi, Li Dong, Alon Benhaim, Barun Patra, Vishrav Chaudhary, Xia Song, and Furu Wei. TorchScale: Transformers at scale. *CoRR*, abs/2211.13184, 2022.

[Ope23] OpenAI. Gpt-4 technical report. 2023.

[PWC+15] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, J. Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123:74–93, 2015.

[SBV+22] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[VLZP15] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015.

[WCC+23] Wen Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Y. Qiao, and Jifeng Dai. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *ArXiv*, abs/2305.11175, 2023.

[WMH+22] Hongyu Wang, Shuming Ma, Shaohan Huang, Li Dong, Wenhui Wang, Zhiliang Peng, Yu Wu, Payal Bajaj, Saksham Singhal, Alon Benhaim, Barun Patra, Zhun Liu, Vishrav Chaudhary, Xia Song, and Furu Wei. Foundation transformers. *CoRR*, abs/2210.06423, 2022.

[WPN+19] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*, 2019.

[WYM+22] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, 2022.

[YPY+16] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. *ArXiv*, abs/1608.00272, 2016.

[YTBB17] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L. Berg. A joint speaker-listener-reinforcer model for referring expressions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3521–3529. IEEE Computer Society, 2017.

A Hyperparameters

The training hyperparameters of KOSMOS-2 are listed in Table 7.

Hyperparameters	
Image embedding number	64
Location tokens	1,024
Training steps	60,000
Warmup steps	375
Optimizer	AdamW
Learning rate	2e-4
Learning rate decay	Linear
Adam β	(0.9, 0.98)
Weight decay	0.01
Batch size of text corpora	93
Batch size of original image-caption pairs	1,117
Batch size of grounded image-text pairs	1,117
Batch size of interleaved data	47

Table 7: Training hyperparameters of KOSMOS-2

The instruction tuning hyperparameters are listed in Table 8.

Hyperparameters	
Training steps	10,000
Warmup steps	375
Learning rate	1e-5
Batch size of language instruction data	117
Batch size of vision-language instruction data	351
Batch size of grounded image-text pairs & grounded instruction data	1404
Batch size of text corpora	30
Batch size of interleaved data	15

Table 8: Instruction tuning hyperparameters of KOSMOS-2

B Templates for Grounded Instruction Data

Table 9 presents the instruction templates of expression generation based on its associated bounding boxes during instruction tuning.

- "What is \<p\> it \</p\>\<box\>\<loc$_1$\>\<loc$_2$\>\</box\>? It is {*expression*}."
- "What is \<p\> this \</p\>\<box\>\<loc$_1$\>\<loc$_2$\>\</box\>? This is {*expression*}."
- "Describe \<p\> this object \</p\>\<box\>\<loc$_1$\>\<loc$_2$\>\</box\>. This object is {*expression*}."
- "\<p\> It \</p\>\<box\>\<loc$_1$\>\<loc$_2$\>\</box\> is {*expression*}."
- "\<p\> This \</p\>\<box\>\<loc$_1$\>\<loc$_2$\>\</box\> is {*expression*}."
- "\<p\> The object \</p\>\<box\>\<loc$_1$\>\<loc$_2$\>\</box\> is {*expression*}."

Table 9: Instruction templates used for expression generation.

C Examples of GRIT

We present some examples of the GRIT corpus in Figures 6–9. The grounded image-text pairs span over various domains and contain different numbers of objects.

Figure 6: Example from GRIT. Caption: "*A serving of kale and roasted vegetable salad on an aluminium tray served with a small white bowl filed with creamy light green avocado Caesar dressing*".
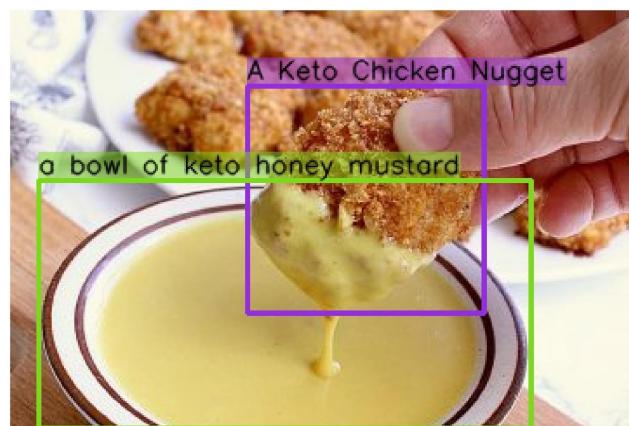
Figure 7: Example from GRIT. Caption: "*A Keto Chicken Nugget being dipped into a bowl of keto honey mustard.*".

Figure 8: Example from GRIT. Caption: "*Solar cells on a red roof are in the foreground. The Sydney skyline is in the background.*".
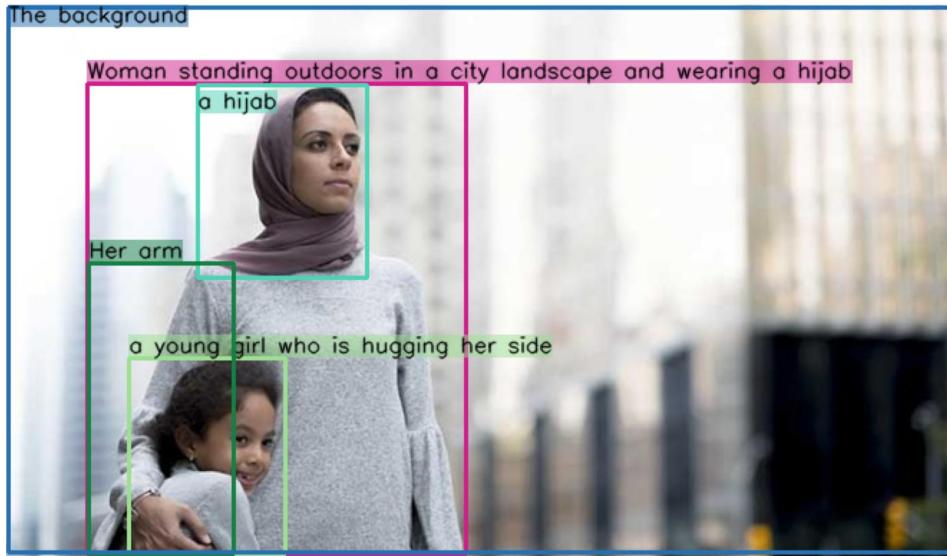
Figure 9: Example from GRIT. Caption: "*Woman standing outdoors in a city landscape and wearing a hijab. Her arm is around a young girl who is hugging her side. The background is blurred.*".

D   More Examples of KOSMOS-2

As illustrated in Figure 10, multimodal referring capability used for visual dialogue can unlock potential in human-AI interaction. In Figure 11, our approach demonstrates its in-context learning ability for fine-grained object detection using both text and image descriptions. Figure 12 and Figure 13 showcase more selected examples, including grounded visual question answering, grounded image captioning, and multimodal referring.

Figure 10: Examples of visual dialogue generated from KOSMOS-2.
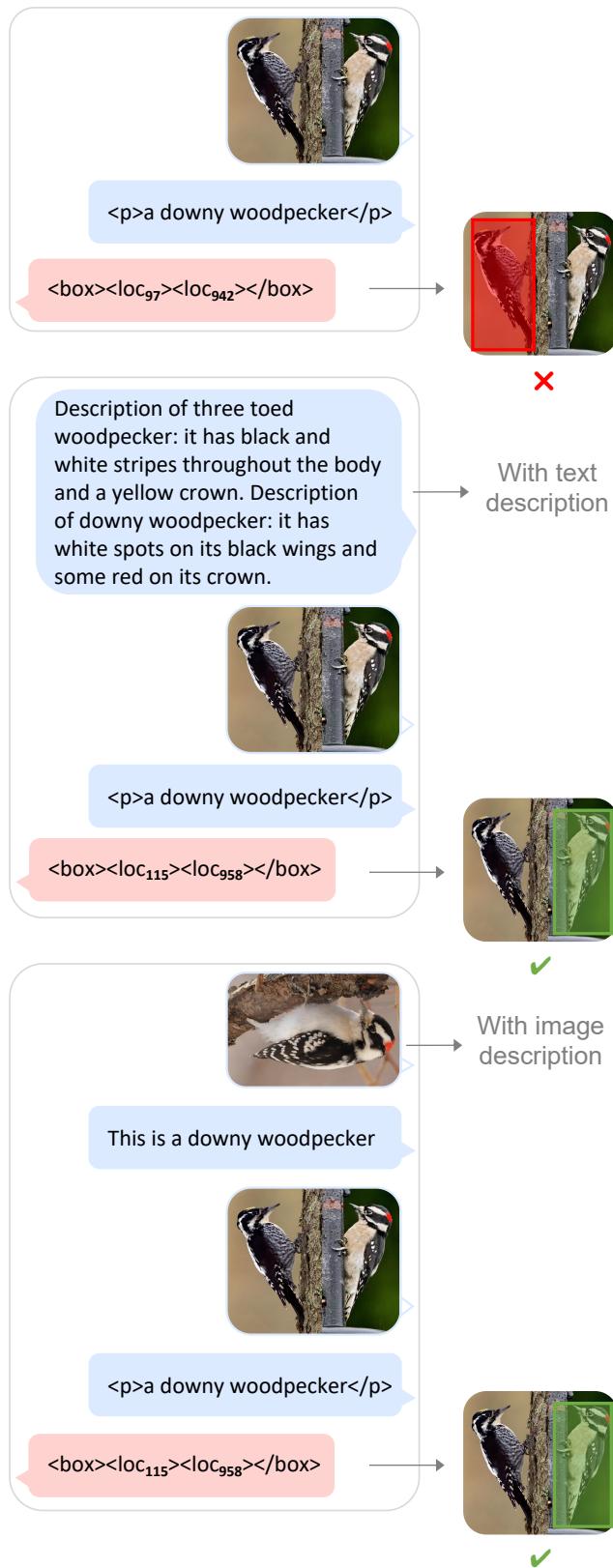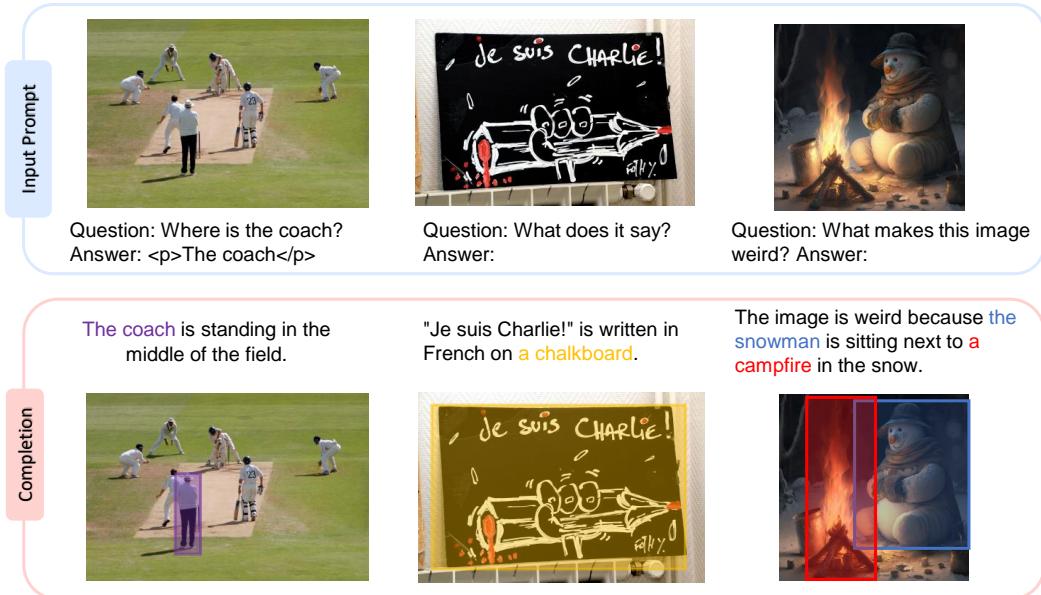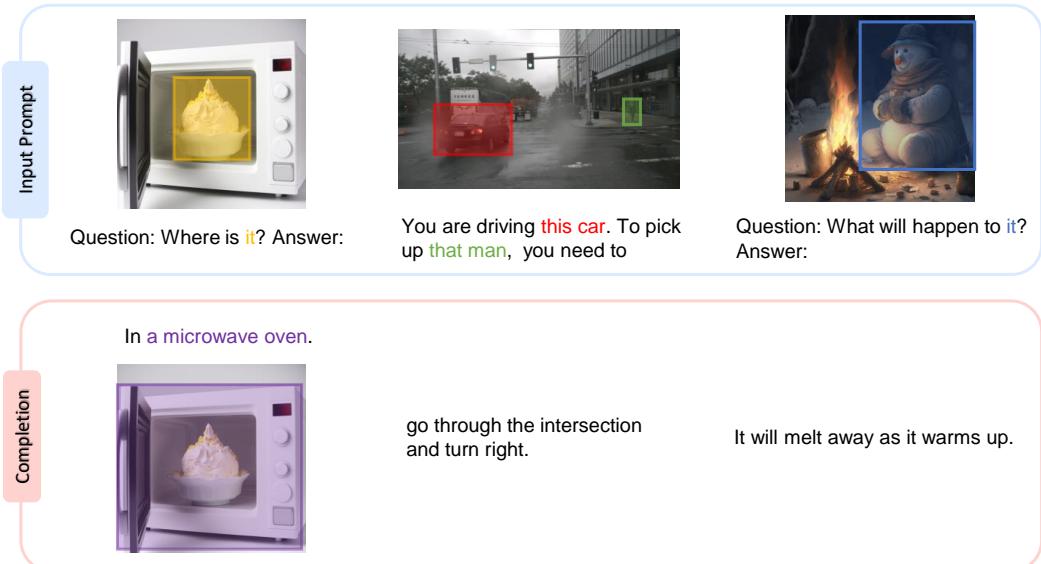
Figure 11: Examples of object detection with multimodal descriptions from KOSMOS-2.

Question: Where is the coach?
Answer: <p>The coach</p>

Question: What does it say?
Answer:

Question: What makes this image weird? Answer:

The coach is standing in the middle of the field.

"Je suis Charlie!" is written in French on a chalkboard.

The image is weird because the snowman is sitting next to a campfire in the snow.

(1) Grounded question answering

Question: Where is it? Answer:

You are driving this car. To pick up that man, you need to

Question: What will happen to it? Answer:

In a microwave oven.

go through the intersection and turn right.

It will melt away as it warms up.

(2) Multimodal referring via bounding boxes

Figure 12: Examples generated from KOSMOS-2.

Input Prompt

Describe this bird in detail:

Completion

The eagle is a large bird with a black and white body and a yellow beak. It has large, dark wings and a pointed tail. It stands on a stump and spreads its wings, displaying its magnificent plumage.

Input Prompt

Describe this image in detail:

Completion

Six fighter jets fly in formation in the sky, creating a beautiful cloud trail.

Input Prompt

Describe this image in detail:

Completion

A man is standing in front of a fire, blowing the flames out of his mouth. He is in the middle of a tropical setting with palm trees surrounding him.

Figure 13: Examples of grounded image captioning generated from KOSMOS-2.