LIRA: Inferring Segmentation in Large Multi-modal Models with Local Interleaved Region Assistance

1. Summary of the Instruction Tuning Data

We provide the detailed composition of our instruction tuning data in Tab. 1, which contains a total of 785K samples. For the comprehension task, we select eleven widely used datasets, comprising a total of 411K samples. These include TextVQA [13], which requires the model to answer questions by reading and reasoning about text within images; DVQA [4], which focuses on processing words and answers related to bar charts; and ChartQA [9], which involves visual and logical reasoning about charts. Additionally, LLaVA150K [6] is a GPT-generated dataset for multimodal instruction-following tasks, while ScienceQA [7] and AI2D [5] are centered around science topics. VQAV2 [1] targets open-ended visual question answering on natural images, and OKVQA [8] extends this by requiring additional world knowledge. AOKVQA [12] further incorporates commonsense reasoning to answer questions about scenes. VizWiz [2] is designed to answer questions posed by blind individuals in real-world scenarios. Following LLaVA , we incorporated the prompt, "When the provided information is insufficient, respond with 'Unanswerable,'" during both training and inference. Finally, GQA [3] is a dataset focused on real-world visual reasoning and compositional question answering. For the segmentation task, we select data from two key tasks: referring expression segmentation (RefSeg) and grounded conversation generation (GCG). In the RefSeg task, which involves object segmentation based on natural language descriptions, we use the RefCOCO [16], RefCOCO+ [16], and RefCOCOg [10] datasets, totaling 168k samples. For the GCG task, which aims to generate detailed image descriptions with corresponding masks for the phrases, we use the GranDf [11] dataset, containing 206K samples.

Task	Dataset	Description	Samples
Comprehension	TextVQA [13]	VQA involving reading and reasoning about text	15k
	LLaVA150k [6]	GPT-generated multimodal instruction-following data	157k
	ScienceQA [7]	Multimodal multiple choice VQA on science topics	15k
	VQAV2 [1]	Open-ended VQA about natural images	60k
	DVQA [4]	Understanding Data Visualizations via Question Answering	10k
	AOKVQA [12]	A Benchmark for Visual Question Answering using World Knowledge	30k
	VizWiz [2]	Answering visual questions from blind people	10k
	AI2D [5]	Multiple choice VQA on science diagrams	30k
	OKVQA [8]	VQA involving world knowledge on natural images	9k
	CharQA [9]	VQA on charts with visual and logical reasoning	15k
	GQA [3]	Real-world visual reasoning and compositional question answering	60k
RefSeg	RefCOCO [16]	Object segmentation based on natural language descriptions	51k
	RefCOCO+ [16]		51k
	RefCOCOg [10]		66k
GCG	GranDf [11]	Generate a detailed image description with corresponding masks for the phrases	206k
Total	-	-	785k

Table 1. Details of the Instruction Tuning Data.

2. More Visualization Results

We present additional visualization results, where Fig. 1 and Fig. 2 showcase LIRA's capabilities across various tasks. As shown in Fig. 1, LIRA demonstrates exceptional scene understanding capabilities, accurately analyzing image content, responding to complex queries, and providing clear and detailed scene descriptions. For instance, LIRA correctly identifies that the man in the first image is walking three dogs and that the car in the center of the second image is brown, showcasing precise recognition of object attributes. Additionally, LIRA not only provides an accurate summary of the scene but also captures the underlying emotions, such as the relaxed atmosphere of the seaside cycling depicted in the first image. As shown in Fig. 2, LIRA is capable of understanding the attributes of objects specified in the instructions, such as "person holding a white dog," "woman on the right," and "girl with slightly curly hair," and accurately segmenting the targets. In the GCG task, LIRA is capable of generating descriptions of the image and accurately segmenting the objects mentioned in the descriptions.

(a)VQA

User: How many dogs is the man wearing a mask leading in the image?
LIRA: man in the image is leading three dogs on a leash.

User: What color is the car in the middle of the picture?
LIRA: The car in the middle of the picture is brown.

User: What is the person in the picture doing?
LIRA: The person in the picture is sitting in a chair, reading a book.

(b)Caption

User: Provide an in-depth description of this image.
LIRA: The image depicts two people riding bicycles on a boardwalk, enjoying a leisurely day at the beach. One person is riding a bicycle in the foreground, while the other is riding a bicycle in the background. They are both wearing jeans and are positioned on the beach, with the ocean visible in the background. The scene conveys a sense of relaxation and outdoor activity.

User: Provide an in-depth description of this image.
LIRA: The image depicts a man in a forest, kneeling down and looking at a blue backpack. He is standing next to a tent, which is set up in the woods. The man appears to be preparing for a camping trip or organizing his belongings. The backpack is placed on the ground, and the tent is nearby, suggesting that the man is either setting up camp or getting ready to leave.

User: Provide an in-depth description of this image.
LIRA: The image depicts a man sitting on a large cardboard box in a room filled with boxes. He is using a laptop computer, likely working or browsing the internet. The room appears to be in the process of being moved, as there are many boxes scattered around the man. Some of the boxes are placed on the floor, while others are stacked on top of each other. The man is focused on his laptop, possibly organizing or managing tasks related to the move.

Figure 1. The visualization results of the VQA and Image Caption tasks.

(a)RefSeg

User: Please segment the person holding a white dog.

User: Please segment the woman on the right.

User: Please segment the girl with slightly curly hair.

(b)GCG

User: Can you provide a thorough description of the this image? Please output with interleaved segmentation masks for the corresponding phrases.
LIRA: A boy kicks a soccer ball while another boy falls down.

User: Can you provide a thorough description of the this image? Please output with interleaved segmentation masks for the corresponding phrases.
LIRA: A woman, referred to as person , is sitting on rock overlooking sea . The sky is visible over the sea.

User: Can you provide a thorough description of the this image? Please output with interleaved segmentation masks for the corresponding phrases.
LIRA: A man and a woman are doing push-ups on the beach .

Figure 2. The visualization results of the referring expression segmentation (RefSeg) and grounded conversation generation (GCG) tasks.

User: Provide a detailed description of this image.
wo ILVC: The image depicts a man with a beard and a mohawk hairstyle sitting at a desk in a brightly lit room. He is holding a cell phone in his hands, likely engaged in using it. The man is wearing a red and black plaid shirt. The desk is cluttered with various items, including a cup, a pair of glasses, a notebook, and a cell phone. There is also a laptop computer on the desk. The room has a window, which allows natural light to illuminate the space.

User: Provide a detailed description of this image.
w ILVC: The image depicts a man sitting at a desk in a home office, engrossed in his cell phone. He is wearing a red and black plaid shirt and has a beard. On the desk, there is a cup, a notebook, and a pair of glasses. The man appears to be focused on his phone, possibly checking messages or browsing the internet. The office setting includes a window, a chair, and a bookshelf in the background.

User: Provide a detailed description of this image.
wo ILVC: The image depicts a woman, likely a veterinarian, wearing a white lab coat, examining a small white and black dog. The dog is standing on a table, and the woman is holding its paw. She appears to be checking the dog's health or administering a medical procedure. The scene takes place in a medical setting, possibly a veterinary clinic.

User: Provide a detailed description of this image.
w ILVC: The image features a woman, who appears to be a veterinarian, wearing a white lab coat and a stethoscope. She is standing next to a small white dog, possibly a poodle, and is petting the dog. The dog is looking at the woman, and they seem to be interacting in a friendly manner. The scene takes place in a room with a white wall in the background.

User: Provide a detailed description of this image.
wo ILVC: The image depicts three people standing in a field, with two men and a woman. The man on the left is wearing a blue shirt and glasses, while the man on the right is in a white shirt. The woman is positioned in the middle, wearing a pink shirt. They are all standing on a muddy path, with their hands in their pockets.
In the background, there is a small creek or stream running through the field, with water flowing in it. The area is surrounded by trees, and the overall atmosphere appears to be somewhat damp and muddy. The people seem to be observing the surroundings, possibly discussing the condition of the field or the water in the stream.

User: Provide a detailed description of this image.
w ILVC: The image depicts a group of three people standing in a field near a river. Two men and a woman are standing close to each other, with the man on the left wearing a blue shirt and glasses, the man in the middle wearing a white shirt, and the woman on the right wearing a pink shirt. They appear to be observing the surroundings, possibly discussing something related to the environment or the river. The scene is set in a rural area, with a few trees visible in the background.

Figure 3. Comparison of Hallucinations in Image Caption *wo/w* ILVC. The illusion content is marked in red.

3. Comparing Hallucination *wo/w* ILVC

In Fig. 3, we present additional visualization results to demonstrate the impact of employing ILVC on mitigating hallucination. As shown in the first sub-figure of Fig.3, without ILVC, the model inaccurately generates the description, "There is also a laptop computer on the desk." However, with ILVC applied, the model provides an accurate description without hallucinations. Similarly, in the second sub-figure, the model incorrectly infers the relationship between two objects, stating "the woman is holding its paw" in the absence of ILVC. In the final sub-figure, without ILVC, the model suffers from significant hallucination, describing, "They are all standing on a muddy path, with their hands in their pockets" a scenario that does not appear in the image. In contrast, with ILVC, the model produces a precise description, demonstrating the effectiveness of the ILVC strategy in reducing hallucinations.

4. Details of the AttrEval Dataset

AttrEval is a dataset specifically designed to evaluate the model's ability to understand object attributes. As shown in Fig. 5, it includes two types of tasks: Visual Question Answering (VQA) and Reference Segmentation (RefSeg), with 1436 and 618 samples, respectively. In the VQA task, the model needs to judge the attributes of objects. In the RefSeg task, the model must infer the object's attributes based on the logits corresponding to `<seg>` in the output. We choose the RefCOCO dataset as the basis for constructing AttrEval. The process of building the dataset is as follows: We predefined a set of attribute categories, including category, location, and color. From multiple descriptions of the same object in RefCOCO, we extract unique attributes of color, location, and category. The specific attribute word cloud is shown in Fig. 4. Using these extracted attributes, we construct the VQA and RefSeg tasks based on different descriptions of the same object. For

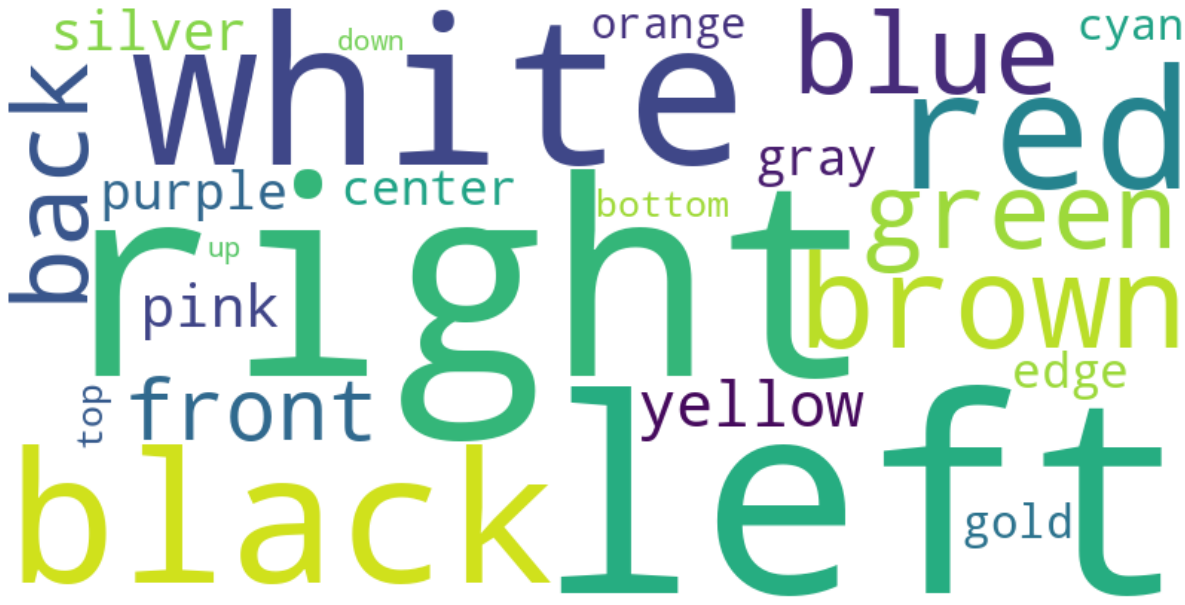

Figure 4. Word Cloud of the AttriEval Dataset.

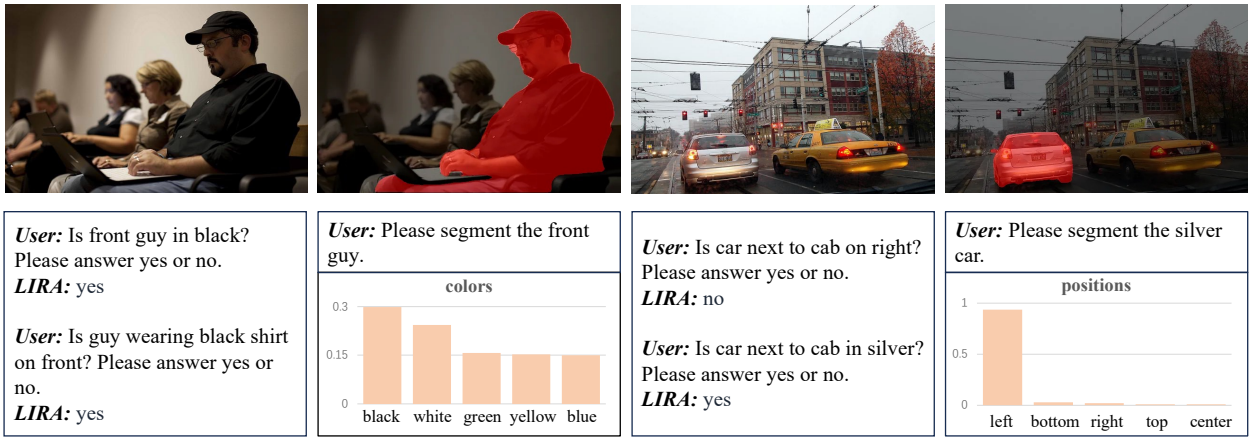

Figure 5. Visualization Results of LIRA on the AttriEval Dataset. The bar chart presents the top five attribute names with the highest probabilities for color or location.

example, if one description of an object does not include color information, we use this description to refer to the object and ask a question about its color, requiring the model to predict the color attribute in the RefSeg task based on the logits corresponding to <seg> in the output. Similarly, if a description lacks location information, we ask a corresponding question about the object's location. In addition, while previous datasets, such as POPE, primarily focus on the existence of objects, our workplaces greater emphasis on the illusion of object attributes.

As shown in Fig. 5, we ask questions about attributes that are not included in the description. For example, in the first image, the description "front guy" does not contain the color attribute, so we ask the question, "Is the front guy in black? Please answer yes or no." with the ground truth (GT) being "yes". Additionally, for the question "Is the front guy in black? Please answer yes or no." we also construct the question "Is the front guy in white? Please answer yes or no." with the GT being "no". Only when both of these questions are answered correctly do we consider the model to have correctly understood the color attribute of the "front guy". In addition, Fig. 5 also shows the visualization of LIRA's answers. LIRA correctly identifies the color and location attributes of the objects. Furthermore, in the RefSeg task, the logits corresponding to the <seg> token correctly contained the color or location attributes of the segmented object. For example, in the second image, LIRA correctly identifies the location of the "silver car" as "left" through the logits.

LLM	VizWiz	GQA	VQAv2	OKVQA	SciQA	POPE	MMB-en	MMB-cn	RefCOCO			RefCOCO+			RefCOCOg	
									Val	TestA	TestB	Val	TestA	TestB	Val	Test
InternLM2-1.8B	67.8	61.1	77.2	53.7	95.0	89.1	74.0	71.7	79.5	81.9	76.0	72.6	77.4	66.1	75.4	75.1
InternLM2.5-7B	71.5	63.5	80.4	62.9	97.3	88.1	81.1	80.5	81.8	83.4	78.1	76.3	81.1	70.5	78.4	78.2
Qwen2-1.5B	73.1	62.5	79.7	58.5	91.7	87.3	76.8	74.2	80.8	82.3	77.1	74.6	78.2	68.3	76.7	75.9

Table 2. Performance with different LLMs.

5. LIRA with Different Backbones

To further validate the effectiveness of LIRA, we conduct experiments using Qwen2-1.5B [15] from Qwen2VL [14]. As shown in Table 2, on the RefSeg task, LIRA-Qwen2-1.5B achieves an average score of 76.7%, outperforming LIRA-InternLM2-1.8B by 1.2%. On the comprehension task, it attains an average accuracy of 75.5%. LIRA demonstrates strong performance with various backbones, achieving promising results on both the comprehension and segmentation tasks, thereby confirming its effectiveness and generalizability across different backbones.

6. Risks of Error Accumulation and Instance Segmentation

ILVC may introduce error accumulation in multi-object segmentation, as inaccuracies in the initial masks can negatively impact the accuracy of subsequent segmentation results. However, we can use two different prompts to control whether to use ILVC to mitigate this. To investigate this, we follow PSALM and train on the COCO instance segmentation dataset, which features multi-object segmentation. When not using ILVC during training, the baseline mIou is 60.0. When using two prompts and 50% of the data with ILVC and 50% without during training, mIou is 60.6 when inference without ILVC, and mIou is 58.9 when inference with ILVC. The results demonstrate that using two prompts to control whether to use ILVC effectively reduces error accumulation, while incorporating ILVC during training improves instance segmentation performance by 0.6.

7. Computational Overhead

Although our method improves performance, it inevitably introduces some computational overhead, which remains within an acceptable range. The overall training time for LIRA-2B is approximately 22 hours, and the inference speed on referring segmentation tasks is around 21.6 tokens per second. Specifically, SEFE added 4 hours to the training time and reduced inference speed by 1.8 tokens per second. The ILVC module added 3 hours to training time, with no inference overhead for VQA tasks—since segmentation is not required—but resulted in a 1.3 tokens per second reduction for segmentation tasks.

8. Overall Pipeline

To present our method's workflow more clearly, we provide the following pseudocode.

Algorithm 1 Overall Pipeline

Require: Global image $\mathbf{I}$, Text instruction T_{ins}
Ensure: Set of predicted masks $\mathcal{M}$, Generated output sequence S_{out}

$f_g, F_{pixel} \leftarrow \text{SEFE}(\mathbf{I})$
$S \leftarrow \{f_g, T_{ins}\}$
$\mathcal{M} \leftarrow \emptyset$; $M_{current} \leftarrow$ null
while not end-of-generation **do**
  $token \leftarrow \text{LLM.generate}(S)$
  if $token$ is <eos> **then**
    break
  else if $token$ is <seg> **then**
    $M_{current} \leftarrow \text{PixelDecoder}(token, F_{pixel})$
    $\mathcal{M} \leftarrow \mathcal{M} \cup \{M_{current}\}$
    $S \leftarrow S \oplus token$
  else if $token$ is <image_id> **then**
    $I_l \leftarrow \text{CropRegion}(\mathbf{I}, M_{current})$
    $f_l \leftarrow \text{SemanticEncoder}(I_l)$
    $S \leftarrow S \oplus token$
    $S \leftarrow S \oplus f_l$
  else
    $S \leftarrow S \oplus token$
  end if
end while
$S_{out} \leftarrow S$
return $\mathcal{M}, S_{out}$

References

[1] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[2] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 1

[3] 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. 1

[4] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018. 1

[5] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016. 1

[6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[7] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 1

[8] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 1

[9] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 1

[10] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer, 2016. 1

[11] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 1

[12] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022. 1

[13] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 1

[14] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6

[15] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. 6

[16] 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. 1