F-LMM: Grounding Frozen Large Multimodal Models Supplementary Material

Size Wu[1] Sheng Jin[2] Wenwei Zhang[3] Lumin Xu[4] Wentao Liu[2,3] Wei Li[1] Chen Change Loy[1]
[1] S-Lab, Nanyang Technological University [2] SenseTime Research and Tetras.AI
[3] Shanghai AI Laboratory [4] The Chinese University of Hong Kong
size001@e.ntu.edu.sg ccloy@ntu.edu.sg

In Sec S1, we provide more detailed experimental results on both question-answering benchmarks and grounding benchmarks. Then we introduce the architecture of F-LMM's mask decoder in Sec S2. We also summarise the datasets used by F-LMM and existing grounding LMMs in Sec S3. Last, we provide visualisation results in Sec S4, including failure cases of existing grounding LMMs on general question-answering tasks and examples of reasoning segmentation, grounded conversations and visual CoT. The broader impact and limitations of this work are elucidated in Sec S5 and Sec S6, respectively.

S1. Benchmark results

Question-Answering Benchmarks. In addition to the four benchmarks reported in the main text, we test the grounding LMMs on a wider range of question-answering benchmarks as shown in Table S1. Due to corrupted instruction-following abilities, existing grounding LMMs obtain zero or near-zero scores on these benchmarks.

Referring Expression Segmentation. The results reported in the main text only include scores on the `Val` subsets of RefCOCO, RefCOCO+ and RefCOCOg. Here, we provide the grounding LMMs' performances on all their subsets in Table S2. The metric used for Referring Expression Segmentation (RES) is cIoU.

Panoptic Narrative Grounding. In the main text, we only report individual mask recalls on thing and stuff objects as well as the overall average recall. Here, we additionally report the mask recalls on singular and plural object nouns as shown in Table S3. As expected, segmenting plural nouns that refer to multiple object instances is more challenging for all the tested models.

S2. Mask Decoder

The architecture of the mask decoder based on a 3-stage U-Net [19] is shown in Figure S1, in which the feature maps are downsampled and upsampled three times. Downsampling encompasses two convolutional layers with a kernel size of 2 and 1, respectively. Upsampling is achieved using bilinear interpolation followed by two convolutional layers with a kernel size of 1. The number of parameters of the mask decoder is 8M.

S3. Dataset Usage

Training Data Comparison. In Table S4, we show the datasets used by existing grounding LMMs and our F-LMM. Existing methods conduct training on a wide range of standard segmentation datasets for excellent grounding ability and collect grounded conversation datasets to preserve chat ability. In contrast, F-LMM only need the RefCOCO and PNG datasets for segmentation capability, without needing additional grounded instruction-tuning datasets.

Training Data Format. For PNG dataset, an image narrative (*e.g.*, 'A hot air balloon is flying over the river') is formatted as *'User: Describe the image. Model: A hot air balloon is flying over the river.'* The coloured texts indicate the keywords for grounding, which are annotated in PNG's training set. For RES dataset where each image is associated with multiple referring expressions (*e.g.* 'The man in blue T-short', 'The girl who is smiling'). We convert the RES data to PNG format by concatenating the referring expressions into a single sentence: *'User: Describe the image. Model: The man in blue T-short; The girl who is smiling.'*

S4. Visualisation

General Multimodal Question-Answering. In Figure S2, we show some examples of grounding LMMs performing general question-answering tasks. When prompted to answer with single words (*e.g.*, yes or no), existing grounding LMMs (GLaMM [17], LISA [7], and PixelLM [18]) usually fail to follow the user instructions. Besides, we also observe that the grounding LMMs tend to misunderstand the user's questions as segmentation requests and reply mask tokens, *e.g.*, '[SEG]'. Furthermore, these grounding LMMs fail to recognise the celebrity (Musk) and famous natural spot, ex-

Table S1. More evaluation results on question-answering benchmarks.

Model	MME	MMBench	MMVet	LLaVA^{W}	POPE	GQA	VQA^{v2}	AI2D
Existing Grounding LMMs								
PixelLM-7B [18]	309/135	17.4	15.9	46.4	0.0	0.0	0.0	0.0
PixelLM-13B [18]	77/47	18.1	18.1	47.8	0.0	0.0	0.0	0.0
LISA-7B [7]	1/1	0.4	19.1	47.5	0.0	0.0	0.0	0.0
LISA-13B [7]	2/1	0.8	19.8	48.1	0.0	0.0	0.0	0.0
LLaVA-G-7B [24]	-	-	-	55.8	-	-	-	-
GLaMM-7B [17]	14/9	36.8	10.3	32.0	0.94	11.7	24.4	28.2
LaSagnA-7B [22]	0/0	0.0	16.7	34.5	0.0	0.0	0.0	0.0
General-Purpose LMMs								
DeepseekVL-1.3B [14]	1307/225	64.6	34.8	51.1	88.3	59.3	76.2	51.5
MGM-2B [8]	1341/312	59.8	31.1	65.9	83.9	59.9	72.9	62.1
LLaVA-1.5-7B [10]	1511/348	64.3	30.5	69.0	85.9	62.0	76.6	54.8
HPT-Air-6B [21]	1010/ 258	69.8	31.3	59.2	87.8	56.2	74.3	64.8
HPT-Air-1.5-8B [21]	1476/308	75.2	36.3	62.1	90.1	59.4	78.3	69.0
MGM-7B [8]	1523/316	69.3	40.8	75.8	84.2	61.6	76.7	64.3
DeepseekVL-7B [14]	1468/298	73.2	41.5	77.8	88.0	61.3	78.6	65.3
LLaVA-1.6-7B [12]	1519/322	68.1	44.1	72.3	86.4	64.2	80.2	66.6
LLaVA-1.6-Mistral-7B [12]	1501/324	69.5	47.8	71.7	86.8	55.0	80.3	60.8
MGM-HD-7B [8]	1546/319	65.8	41.3	74.0	84.2	61.6	76.7	64.3

Table S2. Detailed comparisons on Referring Expression Segmentation (RES).

Model	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
Specialised Segmentation Models								
MCN [15]	62.4	64.2	59.7	50.6	55.0	44.7	49.2	49.4
LAVT [23]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
GRES [9]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
X-Decoder [26]	-	-	-	-	-	-	64.6	-
SEEM [27]	-	-	-	-	-	-	65.7	-
Existing Grounding LMMs								
PixelLM-7B [18]	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5
LISA-7B [7]	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6
PerceptionGPT-7B [16]	75.1	78.6	71.7	68.5	73.9	61.3	70.3	71.7
LLaVA-G-7B [24]	77.1	-	-	68.8	-	-	71.5	-
GroundHog-7B [25]	78.5	79.9	75.7	70.5	75.0	64.9	74.1	74.6
GLaMM-7B [17]	78.6	81.1	76.1	70.5	74.9	63.0	74.8	74.8
LaSagnA-7B [22]	76.8	78.7	73.8	66.4	70.6	60.1	70.6	71.9
Grounding Frozen General-Purpose LMMs by F-LMM (Ours)								
DeepseekVL-1.3B [14]	75.0	78.1	69.5	62.8	70.8	56.3	68.2	68.5
MGM-2B [8]	75.0	78.6	69.3	63.7	71.4	53.3	67.3	67.4
LLaVA-1.5-7B [10]	75.2	79.1	71.9	63.7	71.8	54.7	67.1	68.1
HPT-Air-6B [21]	74.3	79.4	71.8	64.0	71.7	57.2	67.5	68.3
HPT-Air-1.5-8B [21]	76.3	78.5	70.8	64.5	72.8	55.4	68.5	69.6
MGM-7B [8]	75.7	80.2	70.8	64.8	73.2	55.3	68.3	69.4
DeepseekVL-7B [14]	76.1	78.8	72.0	66.4	73.2	57.6	70.1	70.4
LLaVA-1.6-7B [12]	75.8	79.5	72.4	65.8	75.2	58.5	70.1	71.7
LLaVA-1.6-Mistral-7B [12]	75.7	79.6	71.2	66.5	75.5	58.1	70.1	70.3
MGM-HD-7B [8]	76.1	80.2	72.0	65.2	73.4	55.7	68.5	69.4

hibiting a worse grasp of world knowledge compared with a general-purpose LMM (*e.g.*, LLaVA [11]). In contrast, F-LMM inherits the virtues of general-purpose LMMs in instruction following and world knowledge comprehension, thanks to the 'Frozen' design philosophy.

Reasoning Segmentation. We show examples of F-LMM

Table S3. Detailed comparisons on Panoptic Narrative Grounding (PNG).

Model	All	Thing	Stuff	Singular	Plural
Specialist Segmentation Models					
MCN [15]	54.2	48.6	61.4	56.6	38.8
PNG [4]	55.4	56.2	54.3	56.2	48.8
PPMN [2]	59.4	57.2	62.5	60.0	54.0
XPNG [5]	63.3	61.1	66.2	64.0	56.4
Existing Grounding LMMs					
PixelLM-7B [18]	43.1	41.0	47.9	49.1	27.7
GroundHog-7B [25]	66.8	65.0	69.4	70.4	57.7
GLaMM-7B [17]	55.8	52.9	62.3	59.7	45.7
Grounding Frozen General-Purpose LMMs by F-LMM (Ours)					
DeepseekVL-1.3B [14]	64.9	63.4	68.3	68.3	56.1
MGM-2B [8]	65.6	64.4	68.4	69.1	56.9
LLaVA-1.5-7B [10]	64.8	63.4	68.2	68.2	56.1
HPT-Air-6B [21]	65.5	64.0	68.8	68.9	56.6
HPT-Air-1.5-8B [21]	65.4	64.1	68.5	68.9	56.5
MGM-7B [8]	66.3	65.3	68.6	69.8	57.3
DeepseekVL-7B [14]	65.7	64.5	68.5	69.2	56.7
LLaVA-1.6-7B [12]	66.3	65.1	69.0	69.8	57.3
LLaVA-1.6-Mistral-7B [12]	66.5	65.4	69.1	70.0	57.5
MGM-HD-7B [8]	66.7	65.6	69.1	70.1	57.8

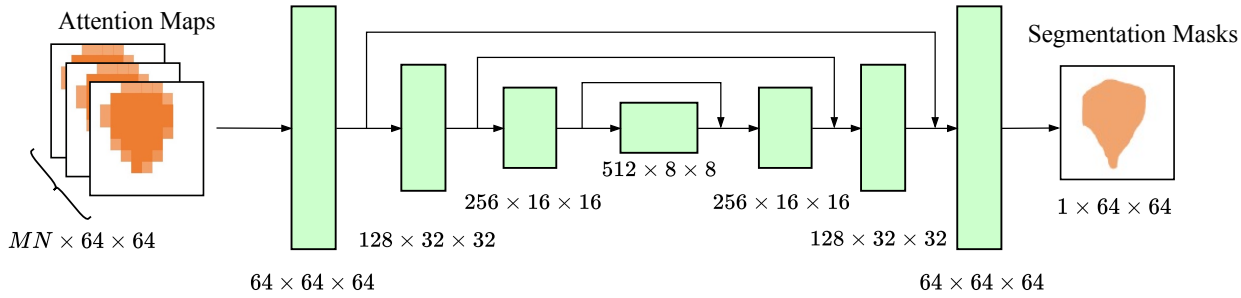

Figure S1. The architecture of the mask decoder is based on a 3-stage U-Net [19] where the feature maps are downsampled and upsampled 3 times.

performing reasoning segmentation in Figure S3. The LMM is first prompted to answer a question relevant to an object in the given image. The content of the answer is regarded as the grounding target. We observe that the LMM (DeepSeekVL-7B) is able to generate the correct answer of the queried object, which is then precisely localised by the mask head of F-LMM.

Grounded Conversation. In Figure S4, we show some examples of grounding conversation by F-LMM. Our F-LMM maintains the LMMs' original ability to follow the user's instruction and understand unusual scenarios (*e.g.*, the man ironing at the back of a taxi) while being able to localise the keywords and phrases during the conversations precisely. The model used in these examples is DeepSeekVL-7B.

Visual Chain-of-Thought Reasoning. Figure S5 shows examples of visual CoT by F-LMM. The model used in these examples is DeepSeekVL-7B. When the LMM is prompted to answer *'which object is the most relevant to the question'*, the mask head of F-LMM grounds the LMM's answer about the relevant object by generating a segmentation mask, the bounding box of which is used to crop the object region from the original image. Then, the cropped image region is fed to the LMM to obtain the final answer. As shown in Figure S5, the Visual CoT empowered by the LMM's grounding ability is helpful when the LMM needs to focus on question-related regions for visual perception and reasoning. We observe that the LMM can infer the location of text contents (*e.g.*, fax number) in a document and the mask head can localise the relevant region, even though they are not trained on such data (*i.e.* VisCoT [20]).

Attention Maps of Different LMMs. In Figure S6, we compare the attention maps of DeepSeekVL-7B(indicated as 'D') and LLaVA-1.5-7B(indicated as 'L'). (m,n) means attention map at the n-th head of the m-th layer. The two

Table S4. Datasets used by F-LMM and existing grounding LMMs. COCO^s stands for the COCO-Stuff [1] dataset and COCO^p is for the COCO-Panoptic [6] dataset.

Datasets	Language-Based Segmentation					Standard Segmentation							Grounded Conversation			
	RefCOCO	RefCLE	VG	PNG	Flickr	COCO^s	ADE	Mapillary	VLPart	COCO^p	Cityscapes	OpenImage	GranD	MUSE	GVC	M3G2
F-LMM (Ours)	✓			✓												
LISA [7]	✓	✓				✓	✓	✓	✓							
Llava-G [24]	✓		✓		✓	✓				✓					✓	
GLaMM [17]	✓	✓	✓		✓	✓	✓	✓	✓				✓			
GroundHog [25]	✓			✓	✓											✓
PixelLM [18]	✓	✓				✓	✓	✓	✓					✓		
LaSagnA [22]	✓	✓				✓	✓	✓	✓		✓	✓				

Question: *"Barracudas often hunt large fish for food. The 's mouth is adapted to tear through meat. Figure: barracuda. Which fish's mouth is also adapted for tearing through meat?*
A. copperband butterflyfish
B. tiger moray
Answer with the option's letter from the given choices directly."

GLaMM-7B: *"[SEG]."*
LISA-7B: *Sure, the segmentation result is [SEG] .*
PixelLM-7B: *Sure, the segmentation result of the large, predominant fish with a wide open mouth are [SEG0] [SEG1] [SEG2] [SEG3] [SEG4] [SEG5]."*
F-LMM(LLaVA-1.5-7B): *"B"*
GT: *"B"*

Question: *"Does this artwork exist in the form of metalwork? Please answer yes or no."*

GLaMM-7B: *"Sure, it is [SEG]"*
LISA-7B: *"Sure, the segmentation result is [SEG]"*
PixelLM-7B: *"Sure, the segmentation result is [SEG0] [SEG1] [SEG2] [SEG3] [SEG4] [SEG5]."*
F-LMM(LLaVA-1.5-7B): *"No"*
GT: *"No"*

Question: "What is the name of the man in the photo?"

GLaMM-7B: *"Sure, it is [SEG]"*
LISA-7B: *"The man in the photo is named Tesla."*
PixelLM-7B: *"The name of the man in the photo is Sunny Leone."*
F-LMM(LLaVA-1.5-7B): *"The man in the photo is Elon Musk."*
GT: *"The man in the photo is Elon Musk."*

Question: *"What is the name of this famous sight in the photo?"*

GLaMM-7B: *"Sure, it is [SEG]"*
LISA-7B: *"The name of the famous sight in the photo us the Road to Hana."*
PixelLM-7B: *"The famous sight in the photo is known as ""Haleakala," which is a volcanic crater on the island of Hawaii.*
F-LMM(LLaVA-1.5-7B): *"The famous sight in the photo is the Diamond Head, which is a volcanic crater located in Hawaii."*
GT: *"The famous sight in the photo is Diamond Head. "*

Figure S2. Examples of grounding LMMs performing general question-answering tasks. The first example is obtained from MM-Bench [13], the second example is extracted from MME [3], and the last two examples are from LLaVA-in-the-Wild [11]. Existing grounding models (GLaMM, LISA, and PixelLM) fail to strictly follow user instructions nor correctly answer questions that necessitate a grasp of general world knowledge. In contrast, F-LMM (built upon LLaVA-1.5 [10] in the above examples), which completely inherits the conversational ability of general-purpose LMMs, performs excellently on these question-answering tasks.

models are similar in showcasing shapes and locations of objects but differ in specific attention map patterns and textures. For example, LLaVA's attention tends to be more concentrated with higher amplitudes.

S5. Broader Impact

This paper addresses an important challenge in large multimodal models—improving the specialised performance while preserving the model's general capabilities. By decoupling the grounding and conversational abilities, building upon the frozen LMMs, the proposed approach allows LMMs to visually ground objects and maintain their broad language capability. Our work is expected to have extensive benefits: (1) It enables the deployment of visually grounding LMMs in real-world applications that require both specialised multimodal capabilities and general language understanding, such as assistive tools and interactive robotics. (2) It paves the way for more flexible and adaptable multimodal AI systems that can be tailored to specific tasks or domains without compromising their core language capabilities. (3) Preserving instruction-following ability and resistance to hallucinations can improve the safety and reliability of the systems, making them suitable for high-stakes applications.

However, similar to many LMM-based systems, there are also potential negative impacts that should be considered: (1) Potential Bias: The pre-trained off-the-shelf LMMs used in the F-LMM approach may already contain

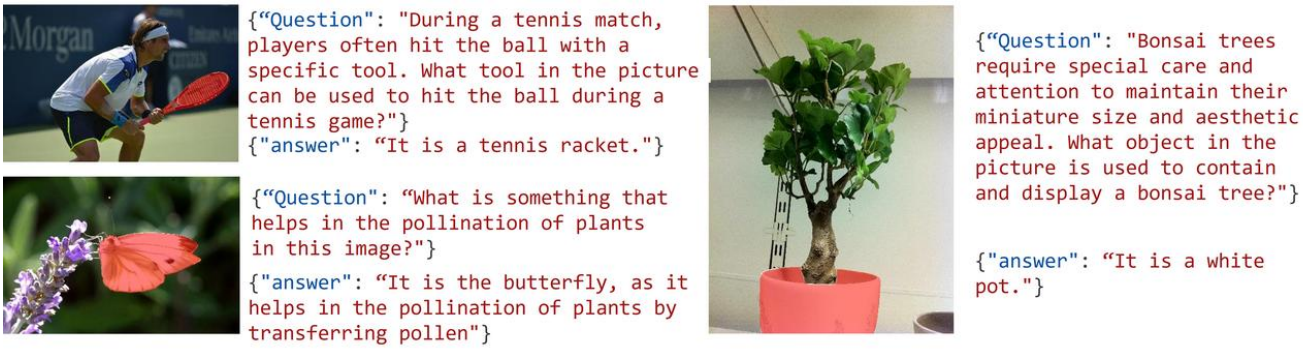

Figure S3. Examples of Reasoning Segmentation. The red masks in the images are segmentation results. The model generating the answers is DeepSeekVL-7B. The sentences in the answers are grounded by the mask head of F-LMM.

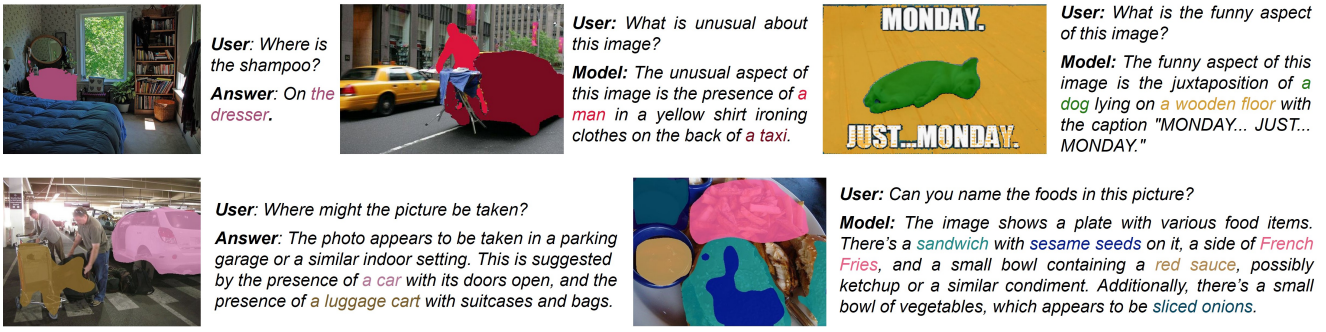

Figure S4. Visualisations of grounded human-AI conversations. The key phrases or words in the conversations can be precisely localised by the mask head of F-LMM. The LMM used is DeepSeekVL-7B.

biases, which could be propagated through the grounding process. (2) Potential for Displacement of Human Labor: The increased capabilities of visually grounding LMMs could lead to the displacement of human labor in certain domains, such as customer service, content creation, or image analysis. (3) Privacy and Ethical Concerns: Integrating visual grounding capabilities with language models raises privacy concerns, as the models could potentially be used to identify individuals or extract sensitive information from images.

To avoid misuse of the model, we will adopt the following safeguards: 1) Access Controls: Strict authentication and authorisation mechanisms will be implemented to ensure that only authorised and responsible individuals or organisations can access and use the models. 2) Usage Policies and Agreements: Clear usage policies and agreements will be established to define the intended purpose of the models. These policies will explicitly prohibit any malicious or harmful activities. Users will be required to agree to these policies and may face legal consequences if they violate them. 3) Transparency: We are committed to promoting transparency by providing comprehensive descriptions of the model's capabilities, limitations, the training pipeline, and the datasets used.

S6. Limitations

While the proposed F-LMM approach demonstrates promising results in preserving conversational abilities while enhancing visual grounding, there are several key limitations that warrant consideration.

- Inherited Biases and Limitations: As the F-LMM method is built upon frozen pre-trained LMMs, it inherits any biases or limitations present in the underlying models. These could include demographic biases, skewed knowledge representations, or other undesirable properties.
- Limited Modality Scope: This work primarily focuses on vision-language multimodal interactions, without exploring other important modalities such as video, audio, and 3D point clouds. Expanding the scope to these additional modalities is a great direction to explore in the future.
- Model Size Constraints: The experiments were restricted to LMMs up to 8 billion in parameter counts due to limited computing resources. Larger and more powerful models beyond this scale were not included. To address these limitations, future research could focus on mitigating biases, expanding the modality scope, and exploring larger-scale models.

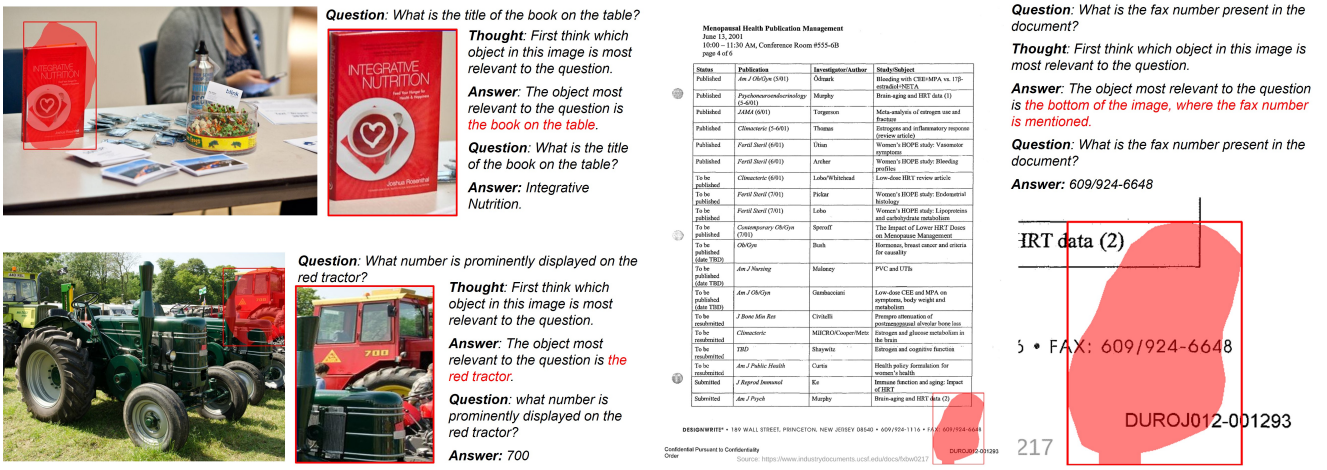

Figure S5. Visual Chain-of-Thought Reasoning. The model used is DeepSeekVL-7B, and the samples are taken from the test set of VisCoT dataset [20]. The LMM is first prompted to think about the question-related object, which is then grounded by the mask head of F-LMM. The region of the question-related object is cropped and fed to the LMM to help answer the question.

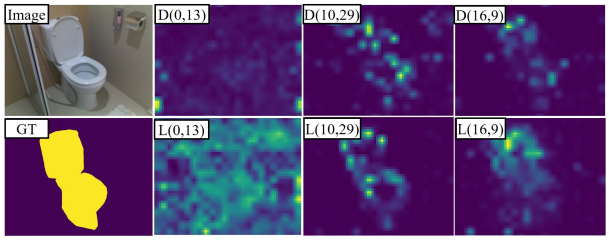

Figure S6. Comparison between the attention maps of DeepSeekVL-7B(indicated as 'D') and LLaVA-1.5-7B(indicated as 'L'). (m,n) means attention map at the n-th head of the m-th layer.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018. 4

[2] Zihan Ding, Zi-han Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Xiaolin Wei, and Si Liu. Ppmn: Pixel-phrase matching network for one-stage panoptic narrative grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5537–5546, 2022. 3

[3] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 4

[4] 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. 3

[5] Tianyu Guo, Haowei Wang, Yiwei Ma, Jiayi Ji, and Xiaoshuai Sun. Improving panoptic narrative grounding by harnessing semantic relationships and visual confirmation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1985–1993, 2024. 3

[6] Alexander Kirillov, Kaiming He, Ross B. Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. *CVPR*, 2019. 4

[7] 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. 1, 2, 4

[8] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv:2403.18814*, 2023. 2, 3

[9] 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. 2

[10] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2, 3, 4

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2, 4

[12] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2, 3

[13] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. 4

[14] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. 2, 3

[15] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020. 2, 3

[16] Renjie Pi, Lewei Yao, Jiahui Gao, Jipeng Zhang, and Tong Zhang. Perceptiongpt: Effectively fusing visual perception into llm. *arXiv preprint arXiv:2311.06612*, 2023. 2

[17] 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. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 2, 3, 4

[18] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. *arXiv preprint arXiv:2312.02228*, 2023. 1, 2, 3, 4

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1, 3

[20] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*, 2024. 3, 6

[21] HyperGAI Team. Hpt 1.5 air: Best open-sourced 8b multimodal llm with llama 3, 2024. 2, 3

[22] Cong Wei, Haoxian Tan, Yujie Zhong, Yujiu Yang, and Lin Ma. Lasagna: Language-based segmentation assistant for complex queries. *arXiv preprint arXiv:2404.08506*, 2024. 2, 4

[23] 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. 2

[24] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, and Jianwei Yang. Llava-grounding: Grounded visual chat with large multimodal models, 2023. 2, 4

[25] Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. *arXiv preprint arXiv:2402.16846*, 2024. 2, 3, 4

[26] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023. 2

[27] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024. 2