PixFoundation: Are We Heading in the Right Direction with Pixel-level Vision Foundation Models?

Mennatullah Siam
University of British Columbia, Canada
mennatullah.siam@ubc.ca

Abstract

Multiple works have emerged to push the boundaries on multi-modal large language models (MLLMs) towards pixel-level understanding. Such approaches have shown strong performance on benchmarks for referring expression segmentation and grounded conversation generation. The current trend in pixel-level MLLMs is to train with pixel-level grounding supervision on large-scale labelled data and use specialized segmentation decoders. However, we show that such MLLMs when evaluated on recent challenging benchmarks, exhibit a weak ability in visual question answering. Surprisingly, some of these methods even downgrade the grounding ability of MLLMs that were not trained with such supervision. In this work, we propose two novel challenging benchmarks and show that MLLMs without pixel-level grounding supervision can outperform the state of the art in such tasks when evaluating both the pixel-level grounding and visual question answering. More importantly, we study the research question of "When does grounding emerge in MLLMs that are not trained with pixel-level grounding supervision?" We show that grounding can coincide with object parts or location/appearance.

1. Introduction

There have been numerous advancements in pixel-level image and video understanding, including image/video segmentation [6, 11, 13, 21], visual grounding and reasoning [7, 12], depth estimation [17] and tracking [16]. The majority of these have been transformed with the emergence of foundation models [1], specifically multi-modal large language models (MLLMs) [3, 9]. Recent efforts explored the shortcomings of MLLMs in vision centric benchmarks [14, 15] with challenging visual tasks. Nonetheless, these benchmarks did not evaluate the recent pixel-level MLLMs.

In this work we provide challenging vision centric benchmarks that are dedicated to evaluate these models. Through these, we answer the first research question; *"Are the current pixel-level MLLMs trained with full grounding supervision heading in the right direction to improve both grounding and visual question answering (VQA)?"* Our findings show that the majority of pixel-level MLLMs still fall short in such challenging setting. We also show that MLLMs that were not trained with pixel-level grounding nor have specialized segmentation decoders can have better performance without

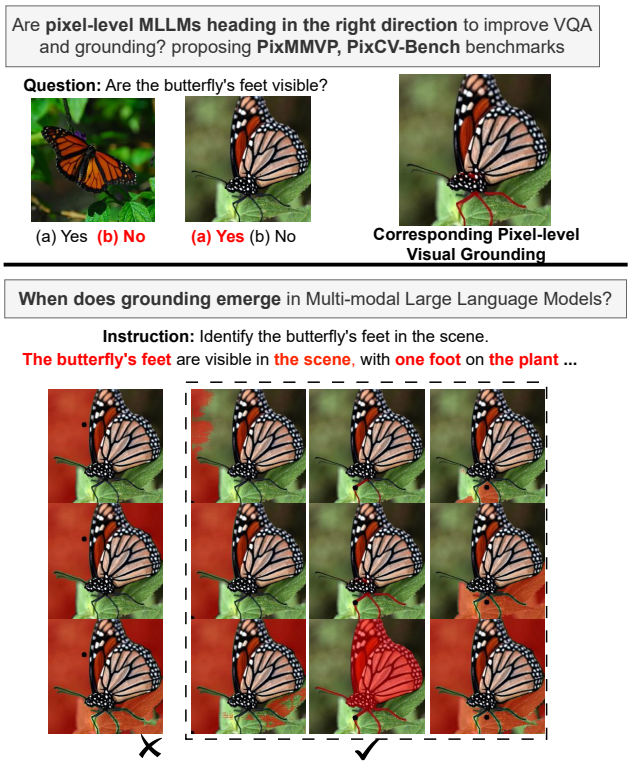


Figure 1. The two major research questions we explore: (i) the grounding and VQA ability of pixel-level MLLMs in challenging scenarios (first row), (ii) the ability of vanilla MLLMs to perform grounding and when does it emerge (second row). Second row shows the noun phrase and its corresponding segmentation mask, highlighted in red, extracted from Llava 1.5 [10] attention maps with three possible output masks to accommodate ambiguity in the point prompt, highlighted as a black point.

degrading VQA capabilities. We are inspired by concurrent work that has explored emerging grounding in MLLMs [2] to establish a baseline on our proposed benchmarks. However, unlike their method we focus on the second research question of *"When does grounding emerge in MLLMs?"* Our work documents that it does not necessarily coincide with the language tokens of the object and can coincide with object parts, position, color or context of these objects. Fig. 1 summarizes our research questions.

In summary, our contributions include: (i) Building two challenging benchmarks with referring segmentation corresponding to the visual question answering focused on vision centric tasks [14, 15], which are PixMMVP and PixCV-Bench. (ii) Benchmarking recent efforts in pixel-level MLLMs where we show that they degrade VQA capabilities. More importantly, some of them lag in visual grounding with respect to simple techniques of extracting the segmentation from vanilla MLLMs, i.e., MLLMs that are not trained for pixel-level grounding. (iii) We provide a simple mechanism for extracting segmentation from vanilla MLLMs which we call PixFoundation, with an understanding of when grounding emerges. It uses our observation that grounding can emerge corresponding to different output tokens describing the object's appearance or location.

2. Benchmarks and Method

Benchmarks. We build upon the recently released MMVP [15] which identified clip blind pairs and used them to build a challenging benchmark with the corresponding questions and choices. We augment the aforementioned dataset and manually annotate each question with the corresponding object of interest referring expression, e.g. an elderly person or the butterfly's feet. These expressions correspond to what needs to be grounded in the image to answer the question. Afterwards, we manually label these objects of interest with polygonal annotations using the VGG annotator [4]. We also build upon the 2D component of the recently released CV-Bench [14], to use their source segmentation datasets (ADE20K [20] and COCO [8]). However, the publicly released CV-Bench does not identify the objects in question and their corresponding segmentation. As such we use GPT-4o to parse the questions and identify the objects of interest automatically, followed by manual inspection and correction. Then we use a custom annotation tool to manually filter the segmentation for these identified objects, e.g. selecting only the object mask annotated by the red box when referred to it and filtering out the other instances for that same class. We provide the final PixMMVP and PixCV-Bench with referring expressions and their segmentation annotations that can be used to evaluate the grounding ability in relation to the original VQA task.

A Pixel-level MLLMs Study. We utilize these two benchmarks to evaluate the current trend in pixel-level MLLMs that relies on training with grounding supervision, (i.e., OMG-Llava [19], GLAMM [12], LISA [7] and Llava-G [18]). We inspect the failures of these models through three probing techniques. First, we highlight the degraded performance in VQA from most of these MLLMs, where we use the first prompt *"<IMG><QUESTION>? a.<OPTION1> b.<OPTION2>..."*. Certain pixel-level MLLMs tend to answer the aforementioned question while outputting a corresponding segmentation mask/s for the objects of interest. The second shortcoming we discuss is their degraded ability to visually ground objects, where we prompt these MLLMs to segment the object of interest in the question. Surprisingly, although they were trained with pixel-level grounding supervision, not all of these models show superior grounding performance for fine-grained objects. Third, we highlight another shortcoming, where these MLLMs exhibit degraded ability to follow instructions. In order to probe this, we use the following prompt: *"<IMG><QUESTION>? a.<OPTION1> b.<OPTION2>... Answer with the option's letter from the given."*

Baselines and upper bounds. In addition to evaluating state-of-the-art pixel-level MLLMs, we propose two baselines and one upper bound. The first of which is inspired by a concurrent work [2] that identified the emergent grounding in MLLMs without the need for pixel-level grounding supervision. Specifically, we use their attend and segment meta architecture, yet we are the first to discuss when does such grounding emerge in these models. The attend and segment meta-architecture extracts the raw attention map for the i^{th} output token, A_i. Only the attention corresponding to the visual tokens are used and averaged across the layers and heads then normalized across all the output, $\tilde{A}_i = \bar{A}_i - \frac{1}{N}\sum_{j=1}^{N} \bar{A}_j$ for N output tokens. The attend and segment depends on using spaCy natural language processing tool [5] to identify the noun phrases and associate them to the ground-truth referring expressions. Thus, the spaCy embeddings closest to the ground-truth expression are used in the mask selection. This is followed by extracting the maximum attention point from $\tilde{A}_i$ to prompt SAM [6].

For our baseline and upper bound, we build upon the previous pipeline and propose an *oracle* upper bound and an *automatic* baseline. We introduce two main modifications to account for our observation that the correct grounding can occur with different output tokens describing the object not necessarily aligning with the exact ground-truth expression. The first modification is to inspect all the potential output tokens without relying on spaCy embeddings. In the *oracle* we rely on the ground-truth mask to select the correct token and its corresponding attention map with highest intersection over union as an upper bound. The *automatic* baseline uses a simple but powerful mechanism where we use the predicted masks fused with the original image to highlight the potential object of interest. This is followed by feeding these images

Method	PixGr.	PixMMVP					PixCV-Bench				
		$\mathcal{A}\dagger$	$\mathcal{A}$	$\mathcal{M}\dagger$	$\mathcal{M}$	$\mathcal{S}$	$\mathcal{A}\dagger$	$\mathcal{A}$	$\mathcal{M}\dagger$	$\mathcal{M}$	$\mathcal{S}$
Llava 1.5 (7B) [10]	✗	27.3	28.0	-	-	-	17.4	60.3	-	-	-
Llava 1.5 (13B) [10]	✗	39.3	30	-	-	-	14.5	61.4	-	-	-
Cambrian (8B)* [14]	✗	52.0	52.0	-	-	-	62.2	72.2	-	-	-
OMG Llava (7B)** [19]	✓	12.0	12.0	17.8	38.0	18.2	12.0	42.1	-	50.5	**45.9**
GLAMM (7B) [12]	✓	1.3	2.7	31.5	47.4	5.1	-	-	30.2	51.9	-
GLAMM - RegCap (7B) [12]	✓	12.7	6.7	14.5	18.6	15.1	27.8	54.4	3.6	7.4	13.0
LISA (7B) [7]	✓	7.3	-	18.1	42.9	12.5	3.7	-	16.8	48.1	6.9
Llava-G (7B) [18]	✓	9.3	-	17.8	13.5	12.2	14.1	4.4	1.7	17.6	15.7
Llava 1.5 (7B) + (a+s) [2]	✗	27.3	28.0	11.1	11.2	16.0	17.4	60.3	5.2	15.7	24.9
Llava 1.5 (13B) + (a+s) [2]	✗	39.3	30	9.8	11.4	17.7	14.5	61.4	4.7	14.9	24.0
Cambrian (8B)* + (a+s) [2]	✗	52.0	52.0	14.3	15.1	23.4	62.2	72.2	18.6	15.9	29.6
PixFoundation (7B) (Ours)	✗	27.3	28.0	18.8	25.9	26.9	17.4	60.3	5.4	28.5	38.7
PixFoundation (13B) (Ours)	✗	39.3	30	16.9	25.0	<u>30.6</u>	14.5	61.4	4.8	27.6	38.1
PixFoundation (8B)* (Ours)	✗	52.0	52.0	29.6	30.3	**38.3**	62.2	72.2	23.9	33.1	<u>45.4</u>
Upper Bound - Oracle Selection											
PixFoundation† (7B) (Ours)	✗	27.3	28.0	**26.1**	**38.0**	**32.2**	17.4	60.3	**6.3**	**49.7**	**54.5**
PixFoundation† (13B) (Ours)	✗	39.3	30	**23.6**	**38.2**	**38.7**	14.5	61.4	**5.3**	**50.6**	**55.5**
PixFoundation† (8B)* (Ours)	✗	52.0	52.0	**52.0**	**56.1**	**54.0**	62.2	72.2	**54.3**	**64.4**	**68.1**

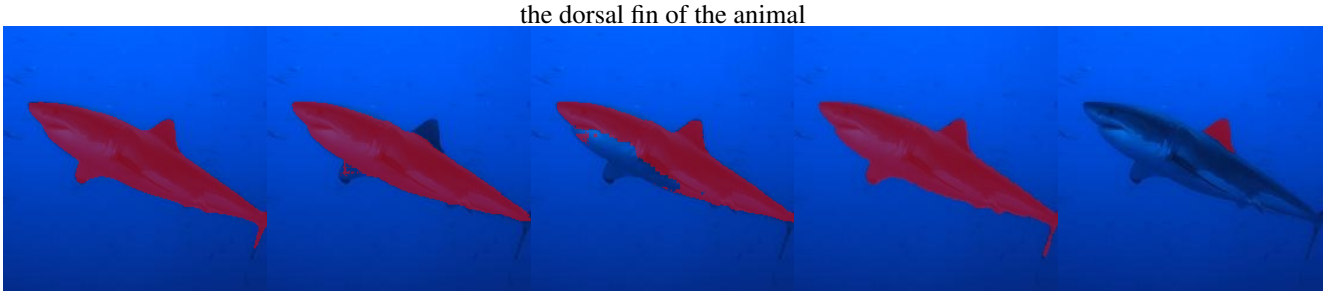
Table 1. **PixMMVP** and **PixCV-Bench** benchmark evaluation of pixel-level MLLMs and baselines. We evaluate the VQA accuracy in the first and third probing (i.e., $\mathcal{A}\dagger$ and $\mathcal{A}$ resp.). Additionally, we evaluate pixel-level visual grounding in the first two probing (i.e., $\mathcal{M}\dagger$ and $\mathcal{M}$ resp.). *, **: models using Llama-3-Ins (8B) and InternLM2 (7B) respectively, unlike the rest that are relying on Vicuna variants (7B and 13B) for the base LLM. - : indicates either the model can not be evaluated in that setting, or has low results below 1% showing complete failure in that setting. $\mathcal{S}$: denotes the score of the MLLM that is the harmonic mean of $\max(\mathcal{A}, \mathcal{A}\dagger)$ and $\max(\mathcal{M}, \mathcal{M}\dagger)$. PixGr.: pixel-level grounding training. The oracle is highlighted in red, the best and second best in $\mathcal{S}$ are bolded and underlined resp.

to GPT-4o inquiring on which is best in highlighting this object. The second modification since SAM has a good understanding of point prompting ambiguity, we process three output masks for each prompt instead of one.

3. Experimental results

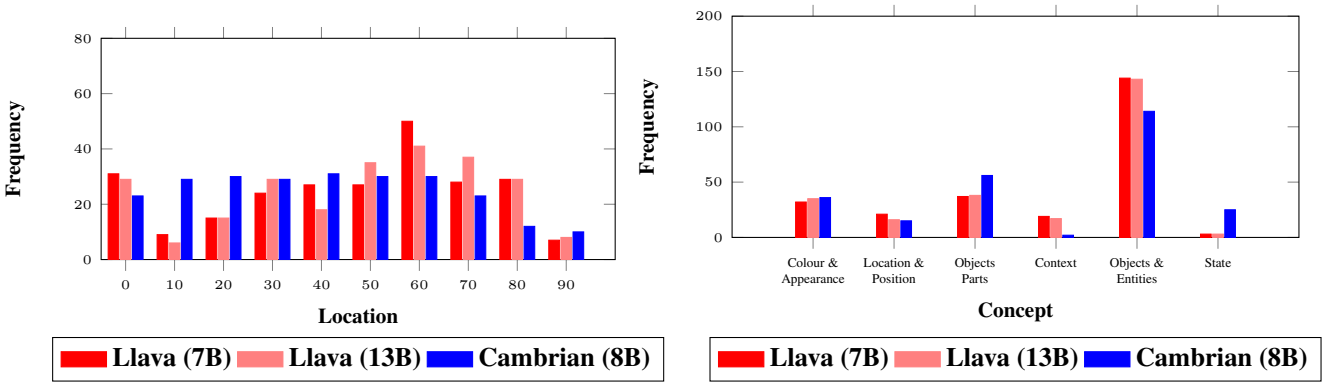
Implementation and evaluation details. We evaluate the VQA and visual grounding capabilities following three probing techniques and reporting their metrics. The first is to evaluate the VQA ability, where the accuracy is computed using GPT-4o following [15] as, $\mathcal{A}\dagger$. If the model generates a segmentation without explicitly asking it to, it is evaluated in terms of mIoU as, $\mathcal{M}\dagger$. The second probing prompts the model to segment the referred expression then evaluates the mIoU reported as, $\mathcal{M}$. The third probing following [14] instructs the model to generate a single option letter and evaluate the accuracy directly without GPT-4o, reported as, $\mathcal{A}$. There is a need for the first probing since some of the recent pixel-level MLLMs face challenges in following instructions. We evaluate the score of each model, $\mathcal{S}$, as the harmonic mean across both tasks, $\mathcal{S} = \frac{2}{\frac{1}{\max(\mathcal{A},\mathcal{A}\dagger)} + \frac{1}{\max(\mathcal{M},\mathcal{M}\dagger)}}$. For GLAMM [12] we use two variants; the original model (GLAMM) and the one fine-tuned for region-level captioning, (GLAMM-RegCap). We evaluate: (i) the attend and segment (a+s), (ii) the *oracle* selection relying on the highest IoU in selecting the correctly predicted masks (PixFoundation†), and (iii) the *automatic* selection, (PixFoundation). These are implemented on top of three base MLLMs which are, Llava 1.5 (7B, 13B) [10] and Cambrian-1(8B) [14].

Are the current pixel-level MLLMs heading in the right direction? We evaluate each of these pixel-level MLLMs capability in VQA and their ability to visually ground the objects of interest in challenging tasks. Table 1 shows the results on PixMMVP and PixCV-Bench benchmarks. From the accuracy of VQA, MLLMs that are not trained with pixel-level grounding surpass their pixel-level counterpart with up to 16%. The best in pixel-level MLLMs score in this aspect is GLAMM-RegCap [19] yet it has degraded ability to generate segmentation. On the other hand, when looking at pixel-level visual grounding we find the best model, GLAMM [12], has a weak ability in VQA or following instructions. Looking at the bottom three rows, the *oracle* confirms that MLLMs that were never trained with pixel-level grounding have the correct grounding within their learned attention maps, refer to Fig. 2. Looking at the final score, $\mathcal{S}$, the oracle variant, PixFoundation† (7B), outperforms the corresponding best pixel-level MLLM, OMG-Llava (7B), by a considerable margin. The automatic baseline outperforms OMG-Llava in PixMMVP with up to 8% and is on-par to it on PixCV-Bench, and the attend and segment baseline [2] still lags behind



(a) OMG-Llava (7B) (b) LISA (7B) (c) GLAMM (7B) (d) Llava-G (7B) (e) PixFoundation† (7B)

Figure 2. **PixMMVP** qualitative comparison in pixel-level visual grounding following the second probing technique. The referred expression is shown on top. It shows that mining for the grounding within the attention maps of vanilla MLLMs using their upper bound is better than MLLMs trained with pixel-level supervision, without degrading their VQA abilities. Thus, questioning whether the current training paradigm of pixel-level MLLMs is in the right direction.



(a) (b)

Figure 3. Analysis on when grounding emerges on PixMMVP benchmark using the three base MLLMs, Llava 1.5 (7, 13B) and Cambrian-1 (8B), that were not trained with pixel-level grounding supervision. We follow the second probing then report the oracle selection. Analysis on: (a) the output location and (b) the output concept category, that coincides with the best segmentation.

our automatic method. In summary, pixel-level grounding supervision degrades MLLMs ability in VQA and sometimes even their generalization in grounding and vanilla MLLMs using simple mechanisms to extract grounding surpass that.

When does grounding emerge in MLLMs? Taking into account the powerful performance of the *oracle* upper bound, it begs the question of when grounding emerges. We start by looking at when it emerges in terms of the location within the output text. We analyze the word/phrase location with respect to the full output text in terms of a percentage from its total length, (i.e., 0% means the beginning of the text). Fig. 3a shows the location percentages histogram, binned at 10%, for the three base MLLMs reporting the oracle selection and evaluating on PixMMVP benchmark using the second probing. In the Llava 1.5 variants, the highest grounding emerges at the last 40% of the output. For the second analysis we look into the concept category that the correct output word/phrase corresponds to. The previous assumption in other works is that grounding emerges with the exact noun or noun phrase of the object of interest. Except our analysis confirms that this is not necessarily the case. We consider six main concept groups: (i) color and appearance, (ii) location, (iii) object parts, (iv) context, (v) objects and entities, and (vi) state. We then prompt for each of the noun/noun phrase, GPT-4o, to categorize it within these six categories. The histogram of the occurrences of these concepts is shown in Fig. 3b. It clearly conveys that in certain scenarios the correct output when grounding emerges can be describing the position or the color of the object not necessarily the exact referring expression. Finally, we compare our PixMMVP mIoU using the oracle selection and Llava 1.5, at 38.0%, with respect to a random baseline that selects among SAM masks output from random point prompts. We keep the same number of masks/input prompts the oracle is selecting among similar in both cases. The random baseline results in a degradation to 26.4%. It confirms the need for the maximum attention mined from these MLLMs.

4. Conclusion

We proposed benchmarks showing that pixel-level MLLMs degraded VQA and even fine-grained grounding, questioning the current direction of these models. We also provide powerful baselines without pixel-level grounding training.

References

[1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1

[2] Shengcao Cao, Liang-Yan Gui, and Yu-Xiong Wang. Emerging pixel grounding in large multimodal models without grounding supervision. *arXiv preprint arXiv:2410.08209*, 2024. 2, 3

[3] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. 1

[4] A. Dutta, A. Gupta, and A. Zissermann. VGG image annotator (VIA). http://www.robots.ox.ac.uk/ vgg/software/via/, 2016. 2

[5] M Honnibal, I. Montani, S. Van Landeghem, and A. Boyd. spacy: Industrial- strength natural language processing in python. https://spacy.io/, 2020. 2

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 2

[7] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 1, 2, 3

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision, Part V 13*, pages 740–755. Springer, 2014. 2

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2023/. 1

[10] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1, 3

[11] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2021. 1

[12] 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, 2, 3

[13] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[14] Shengbang Tong, Ellis L Brown II, Penghao Wu, Sanghyun Woo, ADITHYA JAIRAM IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems*, 2024. 1, 2, 3

[15] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 1, 2, 3

[16] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19795–19806, 2023. 1

[17] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 1

[18] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Chunyuan Li, Jainwei Yang, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *Proceedings of the European Conference on Computer Vision*, pages 19–35. Springer, 2024. 2, 3

[19] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *arXiv preprint arXiv:2406.19389*, 2024. 2, 3

[20] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 633–641, 2017. 2

[21] Tianfei Zhou, Fatih Porikli, David J Crandall, Luc Van Gool, and Wenguan Wang. A survey on deep learning technique for video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7099–7122, 2022. 1