ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

Critical Factors for Optimizing Large Multi-modal Models: Mage Resolution and Text Labeling

U. Harita,

Assistant Professor, Koneru Lakshmaiah Education Foundation, Vaddeswaram Guntur,

uharita@gmail.com

Tanaya Ganguly,

Assistant Professor, Koneru Lakshmaiah Education Foundation, Vaddeswaram Guntur,

gangulytanaya11@gmail.com

Abstract

Large Multimodal Models have proven to be remarkably adept at comprehending tasks involving broad vision and language. However, these models frequently face difficulties when handling complex scene understandings and narratives because of the limited supported input resolution (e.g., 448 x 448) and the incomplete description of the training image-text combination. Here, we suggest the Cat as a solution to the issue. We are contributing in two ways: We propose a multi-level description generation method that automatically provides rich information that can guide model to learn contextual association between scenes and objects. 1) Our method can be built upon an existing vision encoder (e.g., vit-BigHuge with 2b parameters) to effectively improve the input resolution capacity up to 896 x 1344 pixels; 2) without pretraining from the beginning. Our comprehensive testing on over 16 different datasets shows that Cat outperforms the current LMMs on basic tasks including General Visual Question Answering (VQA), Document-oriented VQA, and Image captioning.

Figure 1: Example results from the Cat.

The model shows an advanced level of visual comprehension and is good at identifying nuanced

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

details such as logos and contextual objects. It demonstrates a particular ability to understand tiny text and interpret complex layouts within images, inferring context with limited visual cues.

1 Introduction

Large multimodal models (LMMs) have gained momentum in recent years due to their ability to process and analyse a wide range of data kinds, including text and images. Academics have taken notice of their ability in a variety of multimodal tasks, such as labelling images, answering visual questions, and more. State-of-the-art models that illustrate the rapid progress in this field include LLaVA Liu et al. (2023c), MiniGPT4 Zhu et al. (2023), mPLUG-Owl Ye et al. (2023a), and Qwen-VL Bai et al. (2023).

However, there are several obstacles to overcome, especially when dealing with complex scenarios, because to the wide range of image resolutions and the inconsistency in training data quality. Improvements to the image encoder Liu et al. (2023b) and input resolution increases using large datasets Bai et al. (2023); Chen et al. (2023d,c) have been made in an effort to overcome these issues. Furthermore, the originality of LLaVA (Liu et al., 2023c) is in the integration of multimodal instruction-following data, which allows instruction-tuning to be extended into multimodal situations. Despite these developments, these techniques frequently struggle to manage image input sizes sustainably and require substantial training costs. The need for more intricate image descriptions to understand the subtleties of image-text relationships increases as datasets get bigger, and as a result, the succinct, one-sentence captions in current datasets like COYO Byeon et al. (2022) and LAION Schuhmann et al. (2022) cannot satisfy.

Inspired by these constraints, our work, dubbed Cat, presents a resource-efficient way to increase input resolution in the context of the LMM paradigm. By employing pre-existing LMMs, we circumvent the tedious pre-training phase, thanks to the abundance of good open-source work. We suggest a straightforward yet efficient module that uses a sliding window method to divide high-resolution images into more manageable, localised portions. A static visual encoder, multiple LoRA Hu et al. (2021) modifications, and a trainable visual resampler are used to encode each patch independently. The language decoder is then given these patches' encodings along with the global image's encoding for improved image comprehension. Additionally, we have created a technique that combines multi-level cues

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

from several generators, such as BLIP2, to produce abundant and high-quality caption data.Li et al. (2023e), PPOCR Du et al. (2020), GRIT Wu et al. (2022), SAM Kirillov et al. (2023), and ChatGPT OpenAI (2023a).

We demonstrate the benefits with a few representative samples from Cat in Fig. 2. First, as can be seen in (a), our model successfully completes the picture captioning assignment by precisely describing nearly every aspect in the image, including the athlete's different accessories and the red flag in the backdrop, all without making any mistakes or omissions. The brown bag in the caption, which might not be immediately seen without close examination of the image, is highlighted by the model in the description in (b). The model is able to draw sensible conclusions from this delicate hint, even though it cannot be verified with certainty. This shows that the model is capable of paying attention to incredibly small items and providing precise and logical descriptions. In (c), the model recognises the many languages and the signs that correspond to them in addition to giving a thorough description of the image. Based on this information, Cat can then reasonably forecast how beneficial the image will be. (d) in the question-answering (QA) challenge demonstrates that, despite the watermark "lifequotes tumblr" in the image being lacking a "e," the model is able to respond to a question regarding it. This shows that our model can read small text in photos after training with higher resolutions. When the model answers the question concerning the date "October 6, 1966" properly in (e), it demonstrates its capacity to interpret data from charts and identify the right response among voluminous text without being distracted by extraneous content. This phenomena shows that the model can accurately represent the alignment of a given text with its matching target.(f) provides more evidence that the model is capable of precisely identifying the response to a query even in complex and hazy language. (g) and (h) highlight the model's world knowledge skills in addition to its relevance to the objective. We also provide comparisons and analysis with other LMMs such asGPT4V OpenAI (2023b) and Qwen-VL-Chat in other sec.

We summarize the advantages of the Cat as follows:

1. **Contextual associations**. We provide a multi-level description generation approach that improves the model's ability to comprehend the relationships between several targets and more effectively explore common knowledge while producing text descriptions, leading to more thorough and insightful findings.
2. **Support resolution up to 1344 x 896 without pretraining**. Above the 448 x 448

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

resolution that is usually used for LMMs, this large resolution boosts the capacity to identify and comprehend small or densely packed objects as well as text.

3. **Performance enhancements on many evaluation datasets**. Our Cat model performed competitively in tasks like Image Captioning, General Visual Question Answering, Scene Text-centric Visual Question Answering, and Document-oriented Visual Question Answering, thanks to our testing over 16 different datasets.

2 Methodology

Our approach involves the use of multi-level text cues to enhance the model's perceptual acuity and streamline the production of superior image-text correspondences. We achieved this by separating the input image into four patches, each of which was processed in parallel by different visual encoders with specialised adapters in order to boost the supported image resolution and better steer the relationship between the image and its description. Concurrently, the entire image is input into an independent visual encoder to support the global feature extraction. In order to better compress the picture data and handle the computational demands of processing long image feature sequences, we use a resampling technique that was inspired by Alayrac et al. (2022). Using trainable vectors as queries, our system, inspired by Bai et al. (2023), makes use of a cross-attention mechanism that interacts with the visual encoder's picture features. Our pipeline diagram illustrates this all-inclusive approach (see Fig. 3).

Figure 3: Overall pipeline of Cat. (a) The pipeline for multi-level description generation of theimage. (b) The overall architecture of our model for high-resolution inputs.

2.1 Amplifying image quality

For reading text and fine-grained picture details, input resolution clarity is essential. Previous studies by Bai et al. (2023); Chen et al. (2023d) recommend beginning with lower resolutions and gradually increasing them through curriculum learning; however, this can need a lot of resources (Qwen-VL can only support resolutions up to 448 x 448). In order to improve resolution more effectively, we choose a less complex method.Given an image $R^{H\times W\times 3}$, we employ a sliding window $WR^{Hv\times Wv}$ (where H_v, W_v denote the resolution of the LMM) to partition the image into smaller, local sections. We also leverage LoRA Hu et al. (2021) within each encoder to address the varied visual elements in different parts of an image. This integration of LoRA helps our encoders to recognize and assimilate detail-sensitive features

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

from each image area effectively, which enhances the understanding of spatial and contextual relationships without a substantial increase in parameter count or computational demand.

2.2 Multi-level Description Generation

For model pretraining, earlier studies like LLaVA Liu et al. (2023c) and Qwen-VL Bai et al. (2023) used large datasets from sources such Laion Schuhmann et al. (2022), COYO Byeon et al. (2022), and CC3M Sharma et al. (2018). Despite their size, these datasets frequently offer image-text combinations that are overly basic and devoid of more detailed image information. Due to such rudimentary or subpar text labels, LMMs are unable to accurately build a link between visual features and basic captions, even when trained with high-resolution images. This undermines the synergy between language understanding and visual processing. We employ an innovative model and tool set to close this gap. To provide more complex picture descriptions, we make use of pretrained systems such as BLIP2 Li et al. (2023e), PPOCR Du et al. (2020), GRIT Wu et al. (2022), SAM Kirillov et al. (2023), and ChatGPT OpenAI (2023a).

2.3 Multi-task Training

Our objective is to train a model that can comprehend diverse image formats for a range of jobs while being economical. In accordance with Bai et al., we pooled many datasets and used the same type of instructions for every task (2023). This improves the model's ability to learn and increases its productivity.

We are working on tasks that need the model to comprehend text and images, such as creating captions for photos and responding to queries regarding images. To create simple captions, we instruct the model to "Generate the caption in English:"; for more intricate ones, we instruct it to "Generate the detailed caption in English:". We established a straightforward structure for questions: "User: question Assistant: answer."

3 Experiments

We assess the visual cognition of our model by putting it to the test on a variety of common vision-

language tasks, such as creating descriptions for images, responding to various visual inquiries, and identifying specific phrases in images.

3.1 Implementation Details

Model Configuration. We do tests using the pre-trained big multimodal model, LLM from Qwen-VL, and a well-trained Vit-BigHuge Ilharco et al. (2021). We skip the instruction-

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

tuning step and go straight to the visual encoder's pretrained state. The maximum sequence length during instruction tuning is 2048 owing to device capacity, while Hv and Wv are set to 448 to match the QwenVL encoder.

Model	Image Caption	VQAv2	General VQA		SciQA	VizWiz
	TextCaps		OKVQA	GQA		
Flamingo-9B	-	51.8	44.7	-	-	28.8
Unified-IO-XI	-	77.9	54.0	-	-	-
Kosmos-1	-	51.0	-	-	-	29.2
Kosmos-2	-	51.1	-	-	-	-
BLIP-2 (Vicuna-13B)	-	65.0	45.9	32.3	61.0	19.6
InstructBLIP	-	-	-	49.5	63.1	33.4
Shikra (Vicuna-13B)	-	77.4	47.2	-	-	-
mPLUG-Owl2	-	79.4	57.7	56.1	68.7	54.5
LLaVA1.5 (Vicuna-7B)	-	78.5	-	**62.0**	66.8	50.0
Qwen-VL(Qwen-7B)	65.1	79.5	58.6	59.3	67.1	35.2
Qwen-VL-Chat	-	78.2	56.6	57.5	68.2	38.9
Cat	**93.2**	**80.3**	**61.3**	60.7	**69.4**	**61.2**

Table 1: Results on Image Captioning and General VQA.

Cat has a big language model with 7.7 billion parameters, a resampling module with 90 million parameters, an encoder with 1.9 billion parameters, and 117 million parameters for LoRA.9.8b is the total parameter for Cat.

Training. We use the cosine learning rate schedule and the AdamW optimizer Loshchilov and Hutter (2017) with a learning rate of 1e-5 during the training phase. Furthermore, we establish β1

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

and β2 values of 0.9 and 0.95, respectively. We use a batch size of 1024 and include a warmup phase of 100 steps. We use a weight decay of 0.1 to prevent overfitting.

3.2 Results

We report the results on Image Caption, General VQA, Scene Text-oriented VQA, and Document- oriented VQA.

Image Caption. In order to bridge the gap between visual content and natural language understanding, image captioning is essential. TextCaps Sidorov et al. (2020), an image captioning dataset that demands the model to understand and explain the text in images, is our testing benchmark for the image caption job. In Sec. 4, we also showed that our model can produce comprehensive descriptions of photos. Tab. 1 reports on the TextCaps performance. The outcomes demonstrate that Cat can perform better on the TextCaps dataset.

General VQA. For general visual question answering (VQA), the model must show a thorough comprehension of the relationship between visual and textual information and be able to combine them together successfully. We use five benchmarks for General VQA: ScienceQA Lu et al. (2022b), GQA Hudson and Manning (2019), OKVQA Marino et al. (2019), VQAv2 Goyal et al. (2017), and VizWiz Gurari et al. (2018). The outcomes are displayed in Tab 1. Our models show clear benefits on broad VQA benchmarks. These successes demonstrate how well our strategy works to improve input resolution and detailed data.

Scene Text-oriented VQA. Text is a common aspect of photos in real-world circumstances, hence text-focused question responding is essential. Four datasets—TextVQA Singh et al. (2019), AI2D Kembhavi et al. (2016), STVQA Biten et al. (2019), and ESTVQA Wang et al. (2020)—were used to test our model. The results in Tab. 2 demonstrate the improved performance of our model, especially when processing photos with higher resolution, which makes smaller text details easier to see.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

Model	**TextVQA**	**AI2D**	**STVQA**	**ESTVQA**
Pix2Struct-Large	-	42.1	-	-
BLIP-2	42.4	-	-	-
Instruct BLIP	50.7	-	-	-
mPLUG-DocOwl	52.6	-	-	-
mPLUG-Owl2	54.3	-	-	-
Qwen-VL	63.8	55.9	59.1	77.8
Qwen-VL-Chat	61.5	57.7	-	-
LLaVA-1.5	61.3	-	-	-
Cat	**67.6**	**57.9**	**67.7**	**82.6**

Table 2: Results on Scene Text-oriented VQA.

Model	DocVQA	ChartQA	InfoVQA	DeepForm	KLC	WTQ
Qwen-VL	65.1	**65.7**	35.4	4.1	15.9	21.6
Cat	**66.5**	65.1	**36.1**	**40.6**	**32.8**	**25.3**

Table 3: Results on Doc-oriented VQA.

4 Analysis

4.1 Ablation Study

Input Resolution. Tab. 4 presents the impact of input size for three tasks. As the input size gradually increases, we can observe an improvement in performance, particularly for Deepform, indicating the significant impact of input size on the perception of large multimodal models. Higher-resolution images can provide a clearer and more realistic visual perception. By observing more details and more precise images, the model can better understand visual features such as objects, shapes, and textures within the image, thereby enhancing its visual perception capabilities. When further increasing the input resolution to 1344 x 896 (the largest resolution that Cat can support), our model achieved further improvement on datasets

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

with higher resolution images, such as DeepForm, InfoVQA, and WTQ, as shown in Tab. 5.

Trainable Architectures. As shown in Tab. 4, reducing the LoRA number causes a performance decrease. Using different LoRA for all four encoders compared to not using LoRA provides a better perception of local details, especially with a significant improvement in STVQA. However, to further elaborate, each image block possesses its unique local features, and utilizing four LoRA modules enables a better understanding of contextual information, leading to further performance enhancement. By incorporating LoRA into the four specialized encoders, we can efficiently capture and integrate location-aware details during the encoding process. This approach enables the model to develop a better understanding of the spatial relationships and contextual information within distinct image regions.

Multi-level Description. To further validate the effectiveness of our generated data, we conducted ablation experiments on LLaVA1.5. We utilized a 336-resolution ViT-L as our vision encoder and Vicuna13B Chiang et al. (2023) as our language model. In our experiments, we conducted a comparative analysis by pretraining LLaVA1.5 using the pretraining data from LLaVA, with the modification of replacing 400k instances from the original pre-train data with our generated

Resolution	LoRA	VQAv2	GQA	TextVQA	STVQA	DocVQA	DeepForm
672	4	80	59.6	67.3	67.2	66.4	31.3
784	4	79.9	59.8	67.5	**67.7**	66.5	38.9
896	4	**80.3**	**60.7**	**67.6**	**67.7**	66.5	**40.6**
896	0	80.1	60.4	67.5	65.1	66.1	36.8
896	1	80	60.3	**67.6**	67	**66.7**	36.9

Table 4: Ablation study on input resolution and trainable vision encoders.

Resolution	DeepForm	InfoVQA	WTQ
896×896	40.6	36.1	25.3
896×1344	**42.3** (+1.7)	**39.6** (+3.5)	**26.6** (+1.3)

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

Table 5: Higher input resolution can further enhance performance on documents and other high-resolution images.

Model	Pretrain Data	GQA	VizWiz	TextVQA	POPE	MMBench	MMVet
LLaVA 1.5	Original Caption from CC3M	63.4	56.9	59.8	85.9	68.3	33.5
LLaVA 1.5	Multi-level Description Generation	**63.7 (+0.3)**	**57.7 (+0.8)**	**60.4 (+0.6)**	**86.2 (+0.3)**	**69.3 (+1.0)**	**36.1 (+2.6)**

Table 6: Ablation study on our multi-level generated descriptions.

annotations. The instruction tuning data are all from LLaVA1.5. The results indicate that replacing the 400k pretrained instances with our generated data leads to consistent performance improvements across three VQA tasks: GQA, VizWiz and TextVQA, as well as three widely used evaluation benchmarks: POPE Li et al. (2023g), MMBench Liu et al. (2023d), and MMVet Yu et al. (2023). The results are shown in Tab. 6. We believe that this is due to the detailed captions during the pre-training phase, which help the model focus on more objects and their attributes in the images, resulting in better alignment between vision module and LLM.

Related Work

Large Multimodal Models (LMMs) have attracted widespread attention from researchers in recent years. The exploration of leveraging the robust comprehension and conversational capabilities

of LLMs to enhance performance in visual language tasks has emerged as a focal point of research. Flamingo Alayrac et al. (2022) and OpenFlamingo Awadalla et al. (2023) involve freezing the Vision Encoder and LLM, and adding a Perceiver Resampler module after the Visual Encoder to enhance visual representation. Unified-IO Lu et al. (2022a) trains a comprehensive architecture on over 80 diverse datasets spanning multiple domains.

Enhancing Image Resolution

At present, the input images for the majority of LMMs Liu et al. (2023c); Zhu et al. (2023) are predominantly constrained to 224 x 224, in alignment with the customary input

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

dimensions of the CLIP Radford et al. (2021) utilized in their architecture. However, higher-resolution inputs can capture more details from images and further enhance the model's performance by increasing the sequence length. One approach to improve input resolution is to use a larger visual encoder, such as LLaVA1.5 Liu et al. (2023b), which relies on pre-trained CLIP models like CLIP-ViT-L-336px. Other models Chen et al. (2023c,d); Bai et al. (2023) obtain better performance by continuously increasing the resolution of input images during the training process, which also leads to increased computational complexity, further resulting in larger training costs. OtterHD Li et al. (2023a) use 370K instruction/response pairs to fine-tune Fuyu-8B Bavishi et al. (2023), allowing the original image size to be used during the inference process without scaling. While sharing the same north star in improving image resolution, we focus on developing an approach to continue expanding the image resolution with lower training resources.Optimizing Data Quality

Conclusion

This paper proposes a training-efficient approach to effectively improve the input resolution capacity up to 896 x 1344 pixels without pretraining from the start. To bridge the gap between simple text labels and high input resolution, we propose a multi-level description generation method, which automatically provides rich information that can guide the model to learn the contextual association between scenes and objects. With the synergy of these two designs, our model achieved excellent results on multiple benchmarks. By comparing our model with various LMMs, including GPT4V, our model demonstrates promising performance in image captioning by paying attention to textual information and capturing fine details within the images; its improved input resolution also enables remarkable performance in document images with dense text.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35:23716–23736, 2002.

[2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo:

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390, 2003.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966, 2003.

[4] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Tas¸ırlar. Introducing our multimodal models, 2003.

[5] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In Proceedings of the IEEE/CVF international conference on computer vision, pages 4291–4301, 2019.

[6] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. 2002.

[7] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. arXiv preprint arXiv:2305.18565, 2003c.

[8] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. Pali-3 vision language models: Smaller, faster, stronger. arXiv preprint arXiv:2310.09199, 2003d.

[9] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2003.

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2003.

[11] 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.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 8, Issue 3, 2019

[12] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. 2001.

[13] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3128–3137, 2015.

[14] 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.

[15] 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. arXiv preprint arXiv:2304.02643, 2003.