ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447940

SWEEPMM: A HIGH-QUALITY MULTIMODAL DATASET FOR SWEEPING ROBOTS IN HOME SCENARIOS FOR VISION-LANGUAGE MODEL

Weichen Xu[1,†], *Xinxin Xu*[1,†], *Tianhao Fu*[1], *Jian Cao*[1,*], *Xiaoyang Xu*[1], *Yuetian Huang*[1], *Xixin Cao*[1], *Xing Zhang*[1,2,*]

[1]School of Software and Microelectronics, Peking University, Beijing, China
[2]Shenzhen Graduate School, Peking University, China

ABSTRACT

Embodied intelligence based on vision-language models aims to learn from interactions and derive general intelligence. However, existing generalized vision-language models cannot understand domain knowledge in home scenarios due to the lack of sweeping robot multimodal datasets. In this paper, we propose the first multimodal dataset for sweeping robots, called SweepMM. We create textual data such as room type, scene descriptions, and moving recommendations using various approaches including rule-based, manual-based, and off-the-shelf model-based methods. Based on this dataset, we fine-tune the first generative pretrained model for sweeping robots, called SweepGPM. This model enables human-robot dialogue and surpasses previous state-of-the-art methods by 0.8% in room type recognition, 0.4% in obstacle detection, and 8.0% in lost item search, demonstrating the potential of embodied intelligence in sweeping robots.

Index Terms— Sweeping robot, Benchmark dataset, Vision-language model, Embodied intelligence

1. INTRODUCTION

Embodied Intelligence [1] aims to study intelligent agents that interact and learn from the real world and is considered a promising path toward artificial general intelligence. Recently, many researchers [2, 3] have attempted to combine multimodal models with robots to assist them in handling embodied reasoning tasks. However, general-purpose embodied multimodal models perform poorly in typical service robots, such as sweeping robots, due to a lack of domain knowledge. Although relevant research [4, 5] has focused on detection datasets and perception tasks for sweeping robots, these studies have been limited to the visual domain, severely hindering the research on multimodal models.

To promote the development of embodied intelligence in sweeping robots for multimodal scenarios, we have constructed the first multimodal dataset for sweeping robots in home scenarios, called SweepMM, based on the Object Detection for Sweeping Robots in Home Scenes (ODSR-IHS) dataset [4]. Considering the needs of downstream tasks, the textual descriptions consist of three parts: room type, scene descriptions, and moving and cleaning recommendations. We have employed various approaches, including rule-based, manual-based, off-the-shelf model-based, and prompt engineering, to expedite the text generation process. Fig. 1 illustrates the research framework for fine-tuning the multimodal model. Our SweepMM dataset contains specialized domain knowledge, and we have developed a specialized multimodal model for sweeping robots through techniques

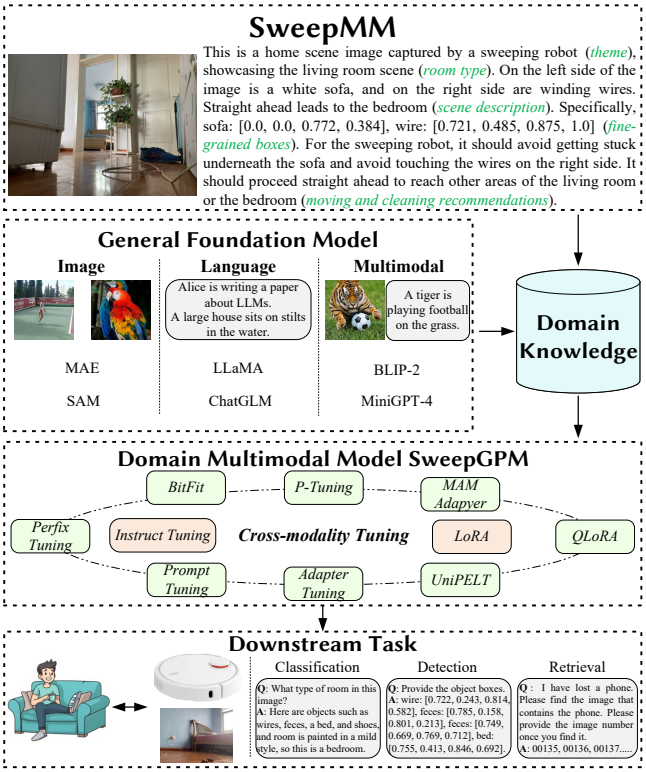


Fig. 1. The above SweepMM dataset contains various information such as room type, scene descriptions, and moving recommendations. Below is the research framework of the vision-language multimodal model SweepGPM.

* Corresponding author. † Equal contribution. This work was supported by Peking University-Delta Electronics Joint Industrial IoT and Intelligent Systems Laboratory Innovation Research Program.

such as Instruct Tuning [6] and LoRA [7]. Our generative pretrained model for sweeping robots, SweepGPM, has the ability to understand home scenarios. Based on prompts, it can be used to develop various valuable and fancy applications. We have already thought of and explored several applications, such as room type recognition based on objects and interior design style in images, replacing existing vision perception algorithms with the detection results output by the multimodal model, and using text to image retrieval tasks to find lost items in an open-vocabulary manner.

The main contributions of this paper are as follows:

- We have constructed the first multimodal dataset for sweeping robots, called SweepMM. It includes various texts such as room types, scene descriptions, and moving recommendations, paving the way for specialized multimodal artificial intelligence for sweeping robots.
- We fine-tuned Large Language Model (LLM) to obtain the multimodal model SweepGPM for sweeping robots, which can effectively understand domain knowledge and interact with humans.
- The superior performance of complex experiments in downstream demonstrates the potential of embodied intelligence for sweeping robots based on SweepGPM.

2. METHODOLOGY

2.1. Overall Pipeline

An overview of our proposed method is shown in Fig. 2. Fig. 2(a) illustrates the construction process of the SweepMM dataset. Descriptions specifically designed for sweeping robots enhance the understanding of upward-view images in home scenes. Fig. 2(b) depicts the network structure of SweepGPM. The Q-Former, fully connected layer, and LoRA parameters are used to align image and language features and adapt to the domain knowledge of the sweeping robot.

2.2. SweepMM Dataset Construction

The existing datasets for sweeping robots in home scenarios only consist of images, lacking textual descriptions of the image content. This severely hinders the research of multimodal models. Therefore, we set out to construct the first multimodal dataset for sweeping robots SweepMM, which paves the way for the development of specialized advanced artificial intelligence in home scenarios, as shown in Fig. 2(a). Unlike common images in ImageNet [8] and COCO [9], the images captured by sweeping robots exhibit significant differences: upward-view, unique categories, and focal perspective. These differences make it nearly impossible to retrieve suitable images from existing large-scale image-text datasets like LAION-400M [10]. Therefore, we produce specific descriptions based on ODSR-IHS [4]. Considering downstream tasks for sweeping robots, such as room type recognition, obstacle detection, lost item search, and obstacle avoidance, we constructed the dataset from the following three aspects:

Room Type Recognition: Considering that room types are not static attributes but are determined by the objects placed in the room, we utilized boxes to annotate room categories. Specifically, we initially defined six room categories: toilet, bedroom, living room, dining room, kitchen, and stock room. We then created a room type judgement table $\mathcal{T}_r$, where each item, t_i, provides the room category and objects that may be present in it, for example, {'toilet': 'closestool', 'trashcan', 'slippers', 'socks', 'carpet', 'feces'}. Each object in the image corresponds to multiple room types, resulting in the set $\mathcal{R}_j (j = 1, ..., m)$, where m is the number of objects. Finally, we took the intersection of the $\mathcal{R}_j$ to determine the room type. Manual judgement was used to resolve ambiguities and obtain the final room type annotations when the intersection is empty or not unique.

Scene Descriptions: On the one hand, we use the coarse-grained initial boxes from the dataset as prompts and input them into the off-the-shelf SAM [11]. We segment the objects and obtain fine-grained boxes $\mathcal{B}_f$. We normalize the bounding boxes and directly use them as precise descriptions of the scene. On the other hand, since the boxes and their categories cannot describe the attributes of the objects, we utilize the off-the-shelf general multimodal model Flamingo [12] to generate descriptions for the image. Subsequently, we use CLIP [13] to retrieve the most similar text, obtaining fuzzy descriptions about the positions and attributes of the objects. The model-based descriptions include additional information about the object's color, shape, size, and even subclass descriptions. This enhances the model's comprehension of scenarios in an open-vocabulary manner, which is beneficial for text to image retrieval tasks such as lost item search.

Moving and Cleaning Recommendations: The moving and cleaning recommendations are unique descriptions for sweeping robots. Due to the lack of such images and descriptions during training, general multimodal models [12] cannot provide ideal results. Therefore, we exploited ChatGPT [14] through prompt engineering to mimic human annotations. Specifically, we manually annotated moving and cleaning recommendations $\mathcal{R}_{mc}$ for 100 images. Then, we input the following prompt to ChatGPT: *"The inputs are the objects and locations, and the outputs are the moving and cleaning recommendations. Please refer to the provided examples. Given the input, provide the output. Input:* $\mathcal{B}_f^1$, *Output:* $\mathcal{R}_{mc}^1$; *Input:* $\mathcal{B}_f^2$, *Output:* $\mathcal{R}_{mc}^2$, ..., *Input:* $\mathcal{B}_f^n$, *Output:"*. Through this process, we generated recommendations for a total of 6000 images. This enhances the network's understanding and reasoning capabilities for sweeping robots in home scenarios. Additionally, some potential descriptions about room type are also beneficial for room type recognition.

Fig. 3 presents some attribute analysis of the SweepMM dataset. In the descriptions, the majority of samples consist

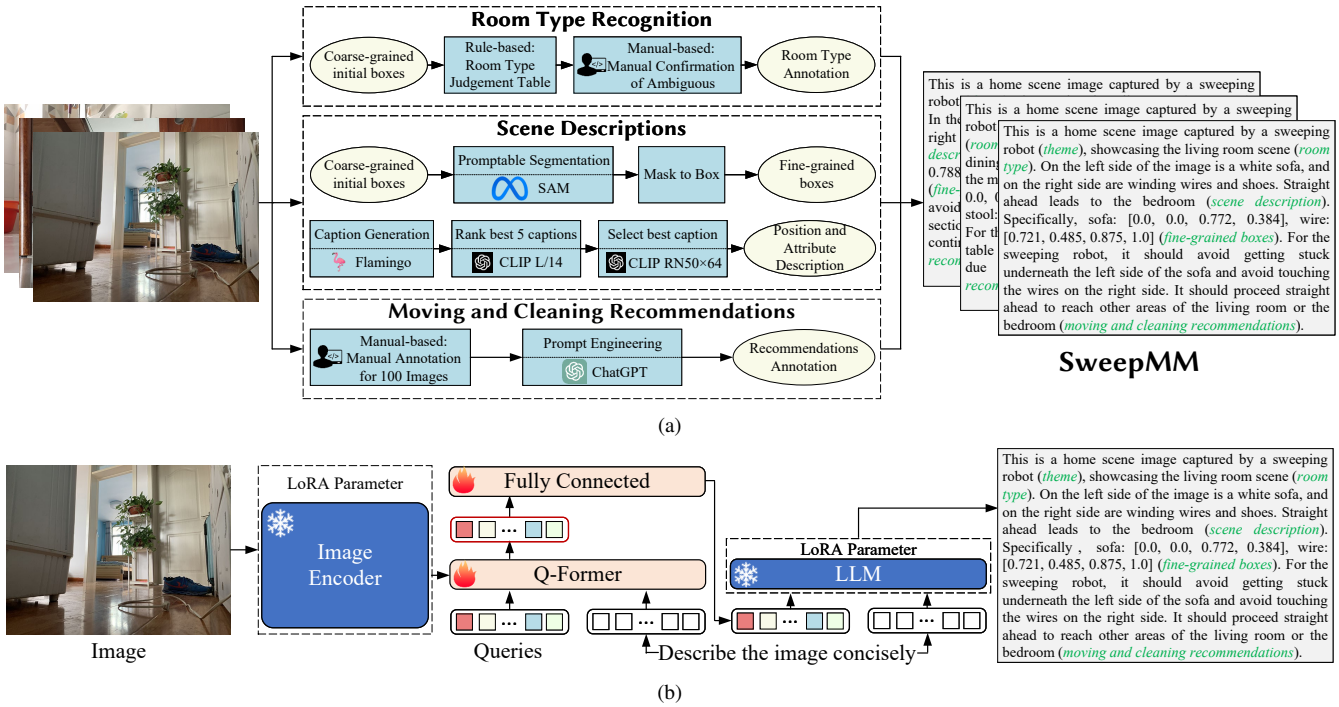


Fig. 2. (a) **The construction process of the SweepMM dataset** consists of three parts: room type recognition, scene descriptions, and moving and cleaning recommendations. The modules involving manual work and off-the-shelf models are marked with symbols. (b) **The fine-tuned network structure of SweepGPM**. The image encoder and LLM parameters are frozen, and the Q-Former, fully connected layer, and LoRA parameters are optimized to adapt the domain knowledge of sweeping robots.

of 8 sentences and nearly 90 words. Moreover, content words account for more than half of the descriptions, which is beneficial for networks to fine-tune the sweeping robot domain.

2.3. SweepGPM Network Structure

The proposed SweepGPM network is fine-tuned based on the structure of BLIP-2 [15], as shown in Fig. 2(b). Specifically, image tokens are extracted from the image using the image encoder CLIP ViT-L/14 [13]. These tokens serve as keys and are inputted into Q-Former [15] along with pre-defined queries and prompt tokens, mapping the image features into the text space. Subsequently, a fully connected layer is used to transform the dimensions of the queries. Unlike BLIP-2, we employ the high-performing ChatGLM-6B [16] as the LLM for text generation. In addition to freezing the image encoder and LLM, we add the LoRA [7] parameter to learn the domain knowledge of the sweeping robot. We train the model to generate text using autoregressive loss and enhance the ability of text-image matching using contrastive loss.

3. EXPERIMENTS AND DISCUSSIONS

3.1. Experimental Setup

Following [4], we divided the SweepMM dataset into a training set of 4800 images and a val set of 1200 images. We performed fine-tuning on the training set and evaluated the classification and detection tasks on the validation set. The retrieval performance was evaluated on the entire dataset.

We keep the image encoder and LLM frozen and optimize the Q-Former, fully connected layers, and LoRa parameters. To avoid catastrophic forgetting, we only train two layers of LoRA parameters, and the LoRA rank is set to 4.

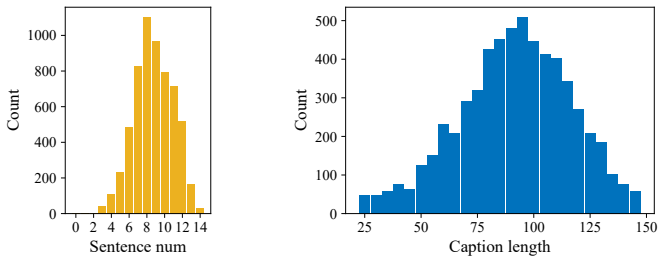
The network is initialized using VisualGLM-6B [16]. Then it is fine-tuned for 3000 iterations with a learning rate of 0.0001. We use the AdamW optimizer and cosine learning rate decay. We use 8 GeForce RTX 4090 GPUs.

3.2. Performance of Downstream Tasks in Classification

Table 1. Comparison to SoTA classification and multimodal models on SweepMM val set for room type recognition.

method	Toilet	Bedroom	Living room	dining room	Kitchen	Stock room	**mean Acc.**
ConvNeXt [17]	91.3	82.4	83.0	**77.4**	**93.8**	75.5	83.5
ViT [18]	84.4	75.8	74.9	69.2	82.9	66.3	76.3
Flamingo [12]	72.5	66.1	68.2	62.3	72.2	43.2	63.1
BLIP-2 [15]	76.6	69.8	72.5	66.6	74.9	58.6	70.2
LLaVA [19]	80.6	68.5	77.6	72.5	77.2	44.6	69.7
VisualGLM [16]	74.8	61.3	74.1	61.3	75.5	49.0	66.5
SweepGPM (Ours)	**94.6**	**88.9**	**85.5**	68.5	91.4	**77.1**	**84.3**
Improvements	+3.3	+6.5	+2.5	-8.9	-2.4	+1.6	+0.8

In Fig. 2(a), the SweepMM dataset explicitly indicates the room type of the image, and provides suggestions for adjacent room types in terms of moving and cleaning recommenda-



(a) (b)

attributes	count
Number of captions	6000
Number of words	503261
Average number of sentences	8
Average number of words	88
Average number of content words	47
Average number of words in Scene Descriptions	35

(c)

Fig. 3. **Attribute analysis of the SweepMM dataset**. (a) Frequency histogram indexed by the number of sentences. (b) Frequency histogram indexed by the caption length. (c) Other attributes, including average number of sentences, etc.

tions on the other hand. Using *"Give the room type in the image."* as a prompt, SweepGPM can recognize the category of the room. Table 1 presents the comparison results with other vision classification and multimodal networks. It can be observed that, due to fine-tuning with domain knowledge from the sweeping robot, SweepGPM outperforms other multimodal models and performs on par with the advanced vision classification model ConvNeXt [17].

3.3. Performance of Downstream Tasks in Detection

In SweepMM dataset, on one hand, fine-grained bounding boxes for each object in the image are explicitly indicated, and on the other hand, descriptions about object positions and attributes are generated by off-the-shelf multimodal models. This enables the fine-tuned multimodal model SweepGPM to detect obstacles. Using *"Provide fine-grained bounding boxes for all objects in the image."* as a prompt, SweepGPM can demonstrate perception capabilities similar to vision detection models. Table 2 presents the comparison results with other SoTA detection and multimodal models. It can be observed that, due to the lack of corresponding descriptions in training dataset, Flamingo [12] fails to complete the detection task, and multimodal models like BLIP-2 [15] perform poorly in detecting objects in upward-view images due to the lack of domain knowledge from the sweeping robot. Our SweepGPM model can effectively detect obstacles and has the potential to replace traditional vision detection models.

Table 2. Comparison to other SoTA vision detection methods and multimodal models on SweepMM val set. For presentation, abbreviations of category names e.g. ST are used.

method	ST	CA	TA	TR	CU	SO	BE	WI	**mAP**
Cascade R-CNN [20]	87.5	77.9	80.7	86.9	79.5	80.2	91.6	48.6	77.3
DETR [21]	83.7	71.8	72.5	80.5	75.1	78.6	87.5	34.8	72.8
YOLOX [22]	92.2	**89.5**	91.0	**95.2**	**85.2**	87.4	94.2	63.5	85.7
Flamingo [12]	2.2	1.3	1.6	0.8	0.2	2.8	2.4	0.3	0.9
BLIP-2 [15]	58.6	52.6	63.8	42.0	45.8	59.6	66.9	36.3	55.1
LLaVA [19]	62.9	53.0	59.4	58.6	47.6	65.7	70.2	37.2	54.9
VisualGLM [16]	71.4	66.1	66.2	54.7	55.1	63.0	68.4	40.8	58.6
SweepGPM (Ours)	**93.3**	84.2	**94.3**	88.6	81.9	**89.4**	**94.8**	**67.8**	**86.1**
Improvements	+1.1	-5.3	+3.3	-6.6	-3.3	+2.0	+0.6	+4.3	+0.4

Table 3. Comparison to multimodal models on SweepMM all set for text to image retrieval in the lost item search task.

method	Phone	Key	Wallet	Glasses	Watch	**mean Recall**
CLIP [13]	69.7	**76.6**	67.2	**87.5**	68.4	72.2
Flamingo [12]	48.5	29.8	58.6	68.8	42.1	49.3
BLIP-2 [15]	66.7	48.9	58.6	56.3	31.6	58.7
VisualGLM [16]	56.1	44.7	62.1	59.4	31.6	54.2
SweepGPM (Ours)	**87.1**	68.1	**75.9**	81.3	**73.7**	**80.2**
Improvements	+17.4	-8.5	+8.7	-6.2	+5.3	+8.0

3.4. Performance of Downstream Tasks in Retrieval

Text to image retrieval is one of the typical applications of multimodal models. For sweeping robots, a promising task is lost item search. Unlike detection, the text is open-vocabulary and not limited to fixed categories. We annotated images in SweepMM dataset with five classes: phone, key, wallet, glasses, and watch. However, these annotations were not used as input during training. Table 3 presents the recall of CLIP [13] and other multimodal models for retrieval. As shown, our method significantly outperforms general multimodal models in retrieval due to its domain knowledge of sweeping robots.

4. CONCLUSIONS

In this paper, we observed that existing research on sweeping robots, including datasets and perception approaches, has been predominantly focused on visual aspects, which severely hinders the development of multimodal models and embodied intelligence. To bridge this gap, we propose the first multimodal dataset for sweeping robots, called SweepMM. On one hand, this dataset is annotated considering potential downstream tasks. On the other hand, it contains high-quality descriptions with an average of 8 sentences and nearly 90 words, which are beneficial for real-world applications. Based on this dataset, we fine-tune the first multimodal model for sweeping robots, SweepGPM, which not only enables human-robot dialogue but also performs downstream tasks. This paves the way for multimodal interaction and embodied intelligence in sweeping robots. In future work, we plan to further expand the dataset and conduct research on model compression.

5. REFERENCES

[1] Jiafei Duan, Samson Yu, Hui Li Tan, et al., "A survey of embodied ai: From simulators to research tasks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 6, no. 2, pp. 230–244, 2022.

[2] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al., "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[3] Sai Vemprala, Rogerio Bonatti, et al., "Chatgpt for robotics: Design principles and model abilities," *Microsoft Auton. Syst. Robot. Res*, vol. 2, pp. 20, 2023.

[4] Yong Lv, Yuemei Fang, Wenzheng Chi, Guodong Chen, and Lining Sun, "Object detection for sweeping robots in home scenes (odsr-ihs): a novel benchmark dataset," *IEEE Access*, vol. 9, pp. 17820–17828, 2021.

[5] Richard Bormann, Xinjie Wang, Jiawen Xu, and Joel Schmidt, "Dirtnet: Visual dirt detection for autonomous cleaning robots," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2020.

[6] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le, "Finetuned language models are zero-shot learners," *arXiv preprint:2109.01652*, 2021.

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, et al., "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, et al., "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, 2014.

[10] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.

[11] 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*, 2023.

[12] 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*, 2022.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, 2021.

[14] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[16] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang, "GLM: general language model pretraining with autoregressive blank infilling," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2022.

[17] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A convnet for the 2020s," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[19] Haotian Liu, Chunyuan Li, et al., "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.

[20] Zhaowei Cai and Nuno Vasconcelos, "Cascade r-cnn: High quality object detection and instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1483–1498, 2019.

[21] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, 2020.

[22] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.