Small Language Model Meets with Reinforced Vision Vocabulary

Haoran Wei[1,*] **Lingyu Kong**[2,*] **Jinyue Chen**[2] **Liang Zhao**[1]
Zheng Ge[1†] **En Yu**[3] **Jianjian Sun**[1] **Chunrui Han**[1] **Xiangyu Zhang**[1]
[1]MEGVII Technology [2]University of Chinese Academy of Sciences
[3]Huazhong University of Science and Technology
https://varytoy.github.io/

Abstract

Playing Large Vision Language Models (LVLMs) in 2023 is trendy among the AI community. However, the relatively large number of parameters (more than 7B) of popular LVLMs makes it difficult to train and deploy on consumer GPUs, discouraging many researchers with limited resources. Imagine how cool it would be to experience all the features of current LVLMs on an old GTX1080ti (our only game card). Accordingly, we present Vary-toy in this report, a small-size Vary along with Qwen-1.8B as the base "large" language model. In Vary-toy, we introduce an improved vision vocabulary, allowing the model to not only possess all features of Vary but also gather more generality. Specifically, we replace negative samples of natural images with positive sample data driven by object detection in the procedure of generating vision vocabulary, more sufficiently utilizing the capacity of the vocabulary network and enabling it to efficiently encode visual information corresponding to natural objects. For experiments, Vary-toy can achieve 65.6% ANLS on DocVQA, 59.1% accuracy on ChartQA, 88.1% accuracy on RefCOCO, and 29% on MMVet. The code will be publicly available on the homepage.

1 Introduction

Large Vision Language Model (LVLM) is one of the hottest research topics [1, 22, 26, 34, 48, 60] in the field of artificial intelligence among the last year. The exciting part is that one LVLM can achieve satisfactory performance in many downstream tasks [4, 24, 30, 32, 41, 45] guided by different prompts. However, there is still significant room for improvement in LVLM's overall image perception capacity. Intuitively, an advanced perceptual ability for visual concepts is essential to enhance the further development and implementation of a model. We deem that there are two main challenges to achieve that: 1) the shortcomings of the current vision vocabulary network [35, 48] in extracting rich visual information; 2) the huge model iteration cost in the optimization of a large number of parameters.

As aforementioned, current LVLMs demonstrate amazing ability in many tasks, especially the Computer Vision (CV) and Natural Language Processing (NLP) intersected ones (*e.g.,* image capion [24], VQA [41], memes understanding, scene OCR [32], *etc*), based on the almost perfect vision vocabulary network — CLIP [35]. The structures of popular LVLMs can be divided into two main streams: 1) image tokens as prefixes like MetaLM [14]; 2) cross-attention for feature fusion like Flamingo [1]. Regardless of which structure is used, the upper limit of the model may be hindered by the visual signals encoding efficiency of its vision vocabulary network. To break through the potential bottleneck, Vary [48] introduces a simple and effective manner to scale up the vision

[*]Equal contribution
[†]Project leader

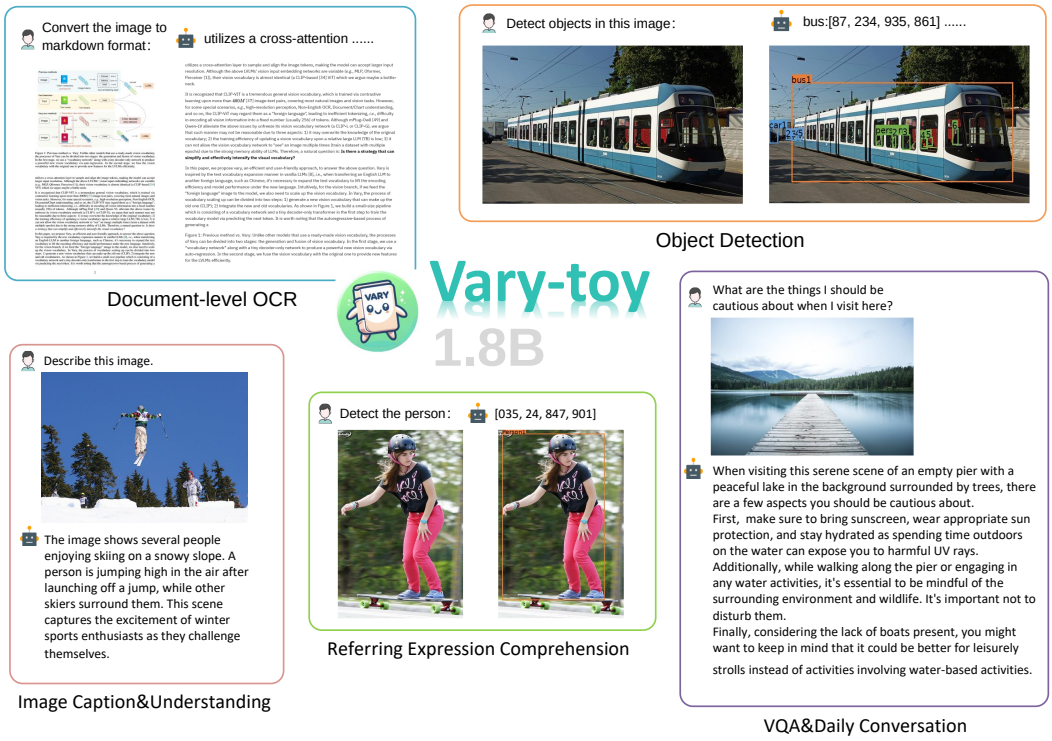

Figure 1: Features of Vary-toy. Based on a 1.8B language model, Vary-toy can achieve all features of vanilla Vary-base, including document OCR, image caption, VQA, general conversation, and so on. Besides, we introduce the natural object perception (location) ability for Vary-toy. Most importantly, with just only a single GTX1080ti GPU, you can experience all of the above.

vocabulary for an LVLM. The scaling law is to first train a new visual vocabulary network using a small auto-regressive model (OPT-125M [57]), and then merge the old and new vocabularies to form the final LVLM (Vary-base [48]). However, Vary suffers two drawbacks to being a user-friendly baseline: 1) The waste of network capacity in the new vision vocabulary (which in vanilla Vary is only used to compress text information in PDF images). 2) The Vary-base with 7B LLM takes high iteration costs (requiring multiple A100 machines to train).

In this report, we present a small-size Vary, *i.e.,* Vary-toy, to alleviate the aforementioned issues. Overall, Vary-toy enjoys the same pipeline as vanilla Vary, including a vision vocabulary generating and scaling up processes. Considering the original Vary masks natural images as negative samples during the creation of a new visual vocabulary. We believe this procedure, to some extent, wastes network capacity, leaving room for optimization. Instead, we regard the natural image as the object detection task [6, 19, 23, 37, 38, 49, 59]. Thus in processing the vision vocabulary, we incorporate both dense textual data (PDF) and natural object location data into the vocabulary network of Vary-toy, making it more universal. After completing the new and reinforced vocabulary, we merge it with the genuine (224×224) CLIP and then integrate them into a 1.8B language model [2].

In experiments, we report metrics on several challenging benchmarks, *i.e.,* DocVQA [30], ChartQA [29], MMvet [54], and RefCOCO [15]. Specifically, Vary-toy can achieve 65.6% ANLS on DocVQA, 59.1% accuracy on ChartQA, 29% accuracy on MMvet, and 88.1% accuracy on RefCOCO val. More specifically, it can gather on par performance compared to Qwen-VL-7B [3] on DocVQA and RefCOCO as well as a better accuracy than LLaVA-7B [26] on the general benchmark MMVet.

In conclusion, Vary-toy is a toy because it is at least three times smaller compared to popular LVLMs (>7B). Vary-toy is not a toy due to it demonstrates excellent potential in challenging tasks. We believe that Vary-toy still enjoys many improvement rooms and we hope that our small-size LVLM can encourage more attention in corresponding research and become a practical baseline, especially for those researchers with limited resources.

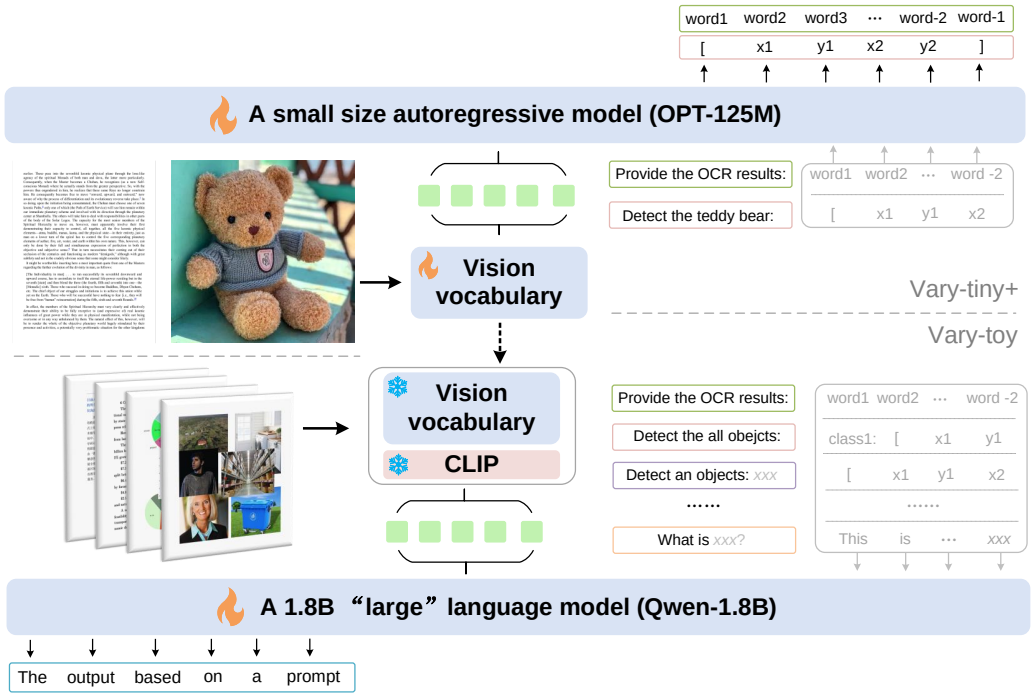

Figure 2: Architecture of the Vary-toy. We utilize the Vary-tiny+ pipeline to generate the new vision vocabulary of Vary-toy. Such vision vocabulary can efficiently encode dense text and natural object location information into tokens. Based on the improved vocabulary, Vary-toy not only possesses all the previous features (document OCR) but also handles object detection tasks well.

2 Related Works

Over the past years, Large Language Models (LLMs), such as the GPT family [5, 34, 36], LLaMA family [8, 42, 44], OPT [57], and the GLM family [55] gain significantly advanced performance in NLP tasks. With the help of LLMs' language reasoning abilities, Vision Language Models (VLMs) like Flamingo [1], BLIP2 [22], LLaVA [25, 26], Vary [48], *etc* [3, 12, 53, 58, 60] have achieved impressive results in various computer vision tasks such as image caption [24], VQA [4, 30, 32], image generation [12], visual grounding [3, 53, 60], document OCR [48] and so on. These models not only can follow human instructions but also possess remarkable few-shot and even zero-shot learning abilities, thereby driving the AI community toward the development of artificial general intelligence (AGI).

However, most popular open-source VLMs are parameter-heavy, with sizes like 7B (*e.g.,* Qwen-VL [3] and mPlUG-Owl [52]) or 13B [26], which to some extend hinder the participation of researchers with limited resources and poses challenges for the implementation of VLMs in resource-constrained environments like home computer. Recently, there has been a growing interest in and development of smaller language models, such as Phi-2 (2.7B) [31] and Qwen-1.8B [2] for NLP tasks, and Gemini-nano (1.8B/3.25B) [43], MobileVLM (1.4B/2.7B) [9] for vision-language tasks.

In this report, Vary-toy will be an open-source small model that possesses features of the most popular LVLMs and demonstrates exceptional potential in fine-grained perception tasks.

3 Method

In this section, we will delve into the details of how to devise Vary-toy. As shown in Figure 2, there are two main parts in implementing the model: 1) how to generate a more practical vision vocabulary based on the Vary-tiny+ pipeline. 2) how to utilize the new vision vocabulary to make the 1.8B Vary-toy gather new features on the premise of not harming the original model features.

3.1 Generating A Reinforced Vision Vocabulary Upon Vary-tiny+

Vary-tiny [48] is a tiny vision language model to generate a specific PDF-parsing vision vocabulary for Vary. The vision vocabulary network comprises a SAM-base [17] main body and paired convolutions to reshape the output, enjoying about 80M parameters. Experiments in Vary prove that using the SAM initializing to gain intensive text perception is effective. However, the vocabulary-generating procedure in vanilla Vary suffers the risk of forgetting SAM's original natural object perception ability. What's more, we also think that writing only the visual knowledge of dense text into an 80M network is wasteful. Thus we generate a new and more reasonable vision vocabulary upon the Vary-tiny+ pipeline.

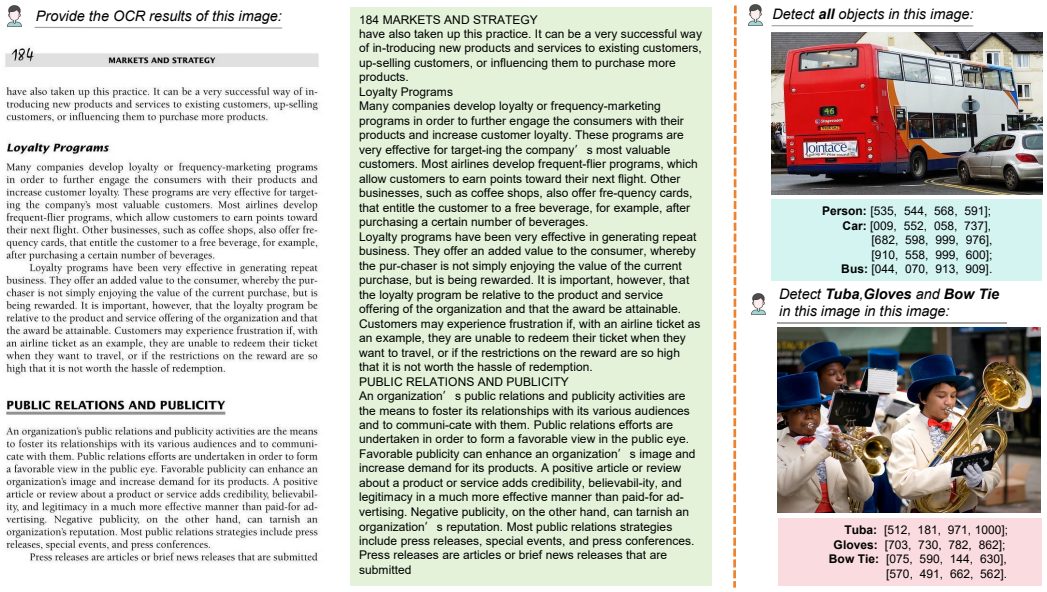

Figure 3: Visualization of image-text pairs used by Vary-tiny+. For PDF image-text pair, there is only one prompt, while for the object detection task, we utilize two types of prompts as shown in the right half of the figure because some images may have too many objects that exceed the maximum token length (4096) of the OPT125M after interpolation.

3.1.1 Data Engine

PDF data. We prepare about 4M PDF image-text pairs in this stage. Following Vary, we use the PDF processing packages to extract the texts of each PDF page, which we find many Python packages can realize (*e.g.,* pdfminer, pdfplumber, and fitz). Each page will be saved as a JPEG image and form an image-text pair with the corresponding text. In this way, we get 2M samples for English and 2M for Chinese. We use the sentence: "Provide the OCR results of this image." as the prompt for both English and Chinese tasks. The PDFs are mainly from arXiv, CC-MAIN-2021-31-PDF-UNTRUNCATED, and e-books. Figure 3 shows a sample of the PDF image-pair.

Object detection data. To fully utilize the capacity of the visual vocabulary network and obtain the natural image perception ability from SAM initialization, we introduce object detection data in the vision vocabulary generating process. We gather the samples from two large open-source datasets, *i.e.,* Object365 [40] and OpenImage [18]. Due to the low efficiency of coordinate (number texts) encoding in OPT's [57] text tokenizer, for images with too many objects, the number of tokens in the ground truth may exceed the maximum token length supported by OPT-125M (although we interpolate it to 4096). Therefore, we re-organize the annotations into two tasks: 1) **Object Detection**: If there are no more than 30 object-boxes in the image, we will allow the Vary-tiny+ detect all objects with the prompt: "*Detect all objects in this image*". 2) **REC**: If the object-box number is over 30, we will regard this image as a REC task using a prompt template: "*Detect class1, class2, ..., in this image*". The selected classes are random so one image can be used multiple times. Through the above manner, we obtain approximately 3M of detection data. Some samples can be seen in Figure 3.

3.1.2 Input Format

Different from the single input/output form of Vary-tiny, Vary-tiny+ needs various input formats to adapt to corresponding tasks due to it requires different prompts to guide the model output correct results. For simplicity, we use the template of Vicuna v1 [8] to construct all ground truth in a conversation format as USER: <img>"<image>"</img> "*texts input*" ASSITANT: "*texts output*" </s>. We add the "<img>" and "</img>" as special tokens of the text tokenizer of OPT-125M and we find that it can adapt very well to the Vicuna template. For the vision input branch, we don't utilize any augmentations and only resize the image to a fixed resolution, *i.e.,* 1024×1024.

3.2 Forge the Cost-Effective Vary-Toy

In this section, we depict the design details of Vary-toy, mainly including the structure of the network and the data construction utilized in the pre-training and SFT stages.

3.2.1 Architecture

As shown in Figure 2, we follow the Vary pipeline to devise the main body of Vary-toy but there are some minor differences. When fed an input image with a shape of H×W, the new vision vocabulary branch will directly resize the image to 1024×1024, while the CLIP [35] branch gains a 224×224 image by the center crop. Both the two branches output 256 tokens with channels of 1024. The dimension of the Qwen-1.8B's input channel is also 2048, so the simplest manner is to concatenate the image tokens in two branches directly as the input image tokens of the language model. In terms of code implementation, to maintain consistency with the Vary structure, we still add input embedding layers behind the vision vocabulary networks.

Task	Dataset	Sample	A prompt example
Cap.	Laion-COCO [39]	4M	Describe the content of this image in a sentence.
	BLIP558k [26]	558K	Describe the image with one saying.
PDF	Pure OCR	1M	Provide the OCR results of this image.
	Markdown	500K	Convert the image to markdown format.
Det.	COCO [24]	50K	Detect all objects in this image.
	RefCOCO	train set	Detect an object: the left woman.
NLP	ShareGPT	125K	*Original conversation*
	Baize [50]	112K	*Original conversation*
	Alpaca [42]	52K	*Original conversation*
VQA	DocVQA [30]	train set	*Qestion.*Answer using a single word or phrase.
	ChartVQA [29]	train set	*Qestion.*Answer using a single-word or phrase.

Table 1: Multi-task training data. We introduce 5 types of data in the pretrain stage, including weakly supervised pair data, PDF image-text pair data, detection data, pure text auto-regressive data, and VQA data. All data annotations are reorganized to a conversation format.

3.2.2 Data Details

Intuitively, the sensitivity of the 1.8B model to data quantity and ratio is higher than that of the 7B or above models, so we put more effort into the data processing aspect for Vary-toy.

Pre-training & SFT data. For Vary-toy, the pretrain stage is actually a multi-task training stage, wherein we prepare a large amount of image-text pairs in various formats. As summarized in Table 1, we mainly focus on a total of 5 types of data in such stage, containing weakly annotated image caption, PDF dense OCR, object detection, pure text conversation, and VQA. Specifically, for natural images, we sample 4M image-text pair in the Laion-COCO [39] dataset, and we also use the BLIP-558K data proposed in LLaVA [26]. For PDF image-text pair, we prepare two types of data following Vary. One is pure dense text OCR, and the other is a task that converts the PDF image to a markdown format. The previous type of data is randomly sampled from the PDF data used in Vary-tiny+ and the last

one is obtained via *LaTeX* rendering. Compared to vanilla Vary, we reduce the proportion of PDF data to maintain universal capability. For the detection data, we gather images from the COCO [24] dataset. We sample 50K images with fewer objects included for the pure object detection task and use all train data of RefCOCO for the REC task. We normalize the coordinates of each box and then magnify them to 1000 times. To prevent the language ability of the LLM from deteriorating, we also introduce pure NLP conversation data, including ShareGPT, Baize [50], and Alpaca [42]. For the last downstream VQA tasks, we choose two challenge datasets (DocVQA and ChartQA [29]) to monitor the text perception and reasoning performance of Vary-toy for artificial data. There are at least 10 prompts made through GPT3.5 [5] for each task, and Table 1 shows one example of them.

In the SFT stage, we only use the LLaVA-80K [26] to instruction tuning the model. LLaVA-80K is a dataset with detailed descriptions and prompts of various types of images, produced by GPT4 [26, 33].

3.2.3 Data Format

In Vary-toy, we are pleased to keep the Chinese PDF-parsing feature to some extent because there is very little exploration in this area, which is also one of the reasons that we select Qwen-1.8B [2] as our base language model (due to the relatively comprehensive text vocabulary). The data input to Qwen-1.8B follows the vanilla Vary [48] format. That is: <|im_start|>user: <img>"<image>"</img> "*human prompts*"<|im_end|> <|im_start|>assistant: "*model outputs*" <|im_end|>.

4 Experiments

4.1 Evaluation Metrics

We report the accuracy of Vary-toy on four popular and challenging benchmarks: DocVQA [30], ChartQA [29], RefCOCO [15], and MMVet [54]. Wherein, the DocVQA and ChartQA can measure the text perception and reasoning ability of the model in manual images, RefCOCO can be used to test the model's ability to locate natural objects, while MMVet, including 6 measurement areas, can be utilized to monitor the general ability of Vary-toy. We use the evaluation metrics introduced in their original paper for fair comparison. Specifically, we utilize ANLS, relaxed accuracy, accuracy under 0.5 IoU, and GPT4 scoring as the metrics for the above four datasets.

4.2 Implementation Details

For Vary-tiny+, we unfreeze all the parameters and train the whole model with a batch size of 512 for 2 epochs. We select the AdamW [28] optimizer with a cosine annealing scheduler [27]. The initial learning rate is set to 5e-5 and the end is 0. It is worth noting that the Vary-tiny is initialized by the weights of Vary-tiny for faster convergence.

For Vary-toy, following vanilla Vary, we freeze all weights of two vision vocabulary networks and only optimize the parameters of the input embedding layers and language model (Qwen-1.8B). In the multi-task training (pre-training) stage, we set the start learning rate to be 5e-5 while it is set to 2e-5 in SFT. We train the model with a batch size of 512 for only 1 epoch in both two stages.

Method	**Size**	DocVQA		ChartQA		
		val	**test**	**human**	**augmented**	**Average**
Dessurt [10]	-	46.5	63.2	-	-	-
Donut [16]	-	-	67.5	-	-	41.8
Pix2Sturct [20]	-	-	72.1	30.5	81.6	56.0
mPLUG-DocOwl [52]	7B	62.2	-	-	-	57.4
Qwen-VL-chat [2]	7B	65.1	-	-	-	65.7
Vary-toy	1.8B	65.6	65.0	33.4	84.8	59.1

Table 2: Performance comparison to popular methods on DocVQA and ChartQA. Vary-toy can achieve 65.6% ANLS on DocVQA which is on par with the 7B Qwen-VL-chat and 59.1% accuracy on ChartQA which is higher than 7B-size mPLUG-DocOwl.

4.3 Manual Image Understanding Ability

We evaluate the fine-grained text perception and reasoning ability via the DocVQA [30] and ChartQA [29]. As shown in Table 2, along with the only 1.8B language model, Vary-toy can achieve 65.6% ANLS on DocVQA and 59.1% accuracy on ChartQA. For DocVQA, the Vary-toy enjoys comparable performance to the 7B-size Qwen-VL-chat, proving the excellent document-level text perception ability of the model and also proving that the new vision vocabulary is available on tokenizing PDF images. For ChartQA, Vary-toy can achieve 59.1% average accuracy, which is better than the 7B size mPLUG-DocOwl, demonstrating the effectiveness of our model further.

Type	Method	Size	RefCOCO		
			val	**testA**	**testB**
Traditional	OFA-L [46]	-	80.0	83.7	76.4
	TransVG [11]	-	81.0	82.7	78.4
	VILLA [13]	-	82.4	87.5	74.8
	UniTAB [51]	-	86.3	88.8	80.6
LLM-based	VisionLLM-H [47]	-	-	86.7	-
	Shikra-7B [7]	7B	87.0	90.6	80.2
	Shikra-13B [7]	13B	87.8	91.1	81.7
	Qwen-VL-chat [2]	7B	88.6	92.3	84.5
	Next-chat [56]	7B	85.5	90.0	77.9
	Vary-toy	1.8B	88.1	90.6	85.7

Table 3: Comparison with popular methods on RefCOCO. Benefiting from the new vision vocabulary, Vary-toy can achieve 88.1% accuracy on RefCOCO val, which is on par with the 7B Qwen-VL-chat.

4.4 Natural Object Perception Ability

The vision vocabulary network generated by Vary-tiny+ should enjoy two main advanced perception abilities: one for dense text and the other for natural objects. In this part, We test the latter ability of Vary-toy after accessing the improved vision vocabulary. It is worth noting that a center crop operation processes the input image of the CLIP branch. Therefore, it can be ruled out that the model uses CLIP for object localization.

As shown in Table 3, Vary-toy can get 88.1% accuracy@0.5 on the RefCOCO validation set, which is also on par with Qwen-VL-chat (7B) and even better than the Shikra-13B. The results show that under the knowledgeable vision vocabulary, Vary-toy gathers great natural object perception ability, proving the effectiveness of using the Vary-tiny+ architecture to build a vision vocabulary, allowing us to further reflect on the necessity of CLIP if we add a large amount of weakly labeled image caption data, *e.g.,* Laion-400M [39], during the new vocabulary generating process.

Method	MM-Vet						
	Rec	**OCR**	**Know**	**Gen**	**Spat**	**Math**	**Total**
BLIP-2 [22]	27.5	11.1	11.8	7.0	16.2	5.8	22.4
LLaVA-7B [26]	28.0	17.1	16.3	18.9	21.2	11.5	23.8
MiniGPT-4 [60]	29.9	16.1	20.4	22.1	22.2	3.8	24.4
Otter [21]	27.3	17.8	14.2	13.8	24.4	3.8	24.7
OpenFlamingo [1]	28.7	16.7	16.4	13.1	21.0	7.7	24.8
LLaVA1.5-7B [25]	-	-	-	-	-	-	30.5
Vary-toy (1.8B)	33.4	20.3	19.9	17.5	24	10.8	29.0

Table 4: Comparison with popular LVLMs on MMVet. With only a 1.8B language model, Vary-toy can get a promising 29.0% accuracy. The abbreviations represent Rec: Recognition; Know: Knowledge; Gen: Language generation; Spat: Spatial awareness.

4.5 General Ability

We report the accuracy of Vary-toy on MMVet [54] to test its general performance. As summarized in table 4, Vary-toy gains 29.0% of the total accuracy, which is higher than the classical LLaVA-7B [26] (23.8%), BLIP-2 [22], and MiniGPT-4 [60], demonstrating Vary-toy enjoys excellent general knowledge. For a 1.8B model, the generation ability is relatively poor and needs to be strengthened.

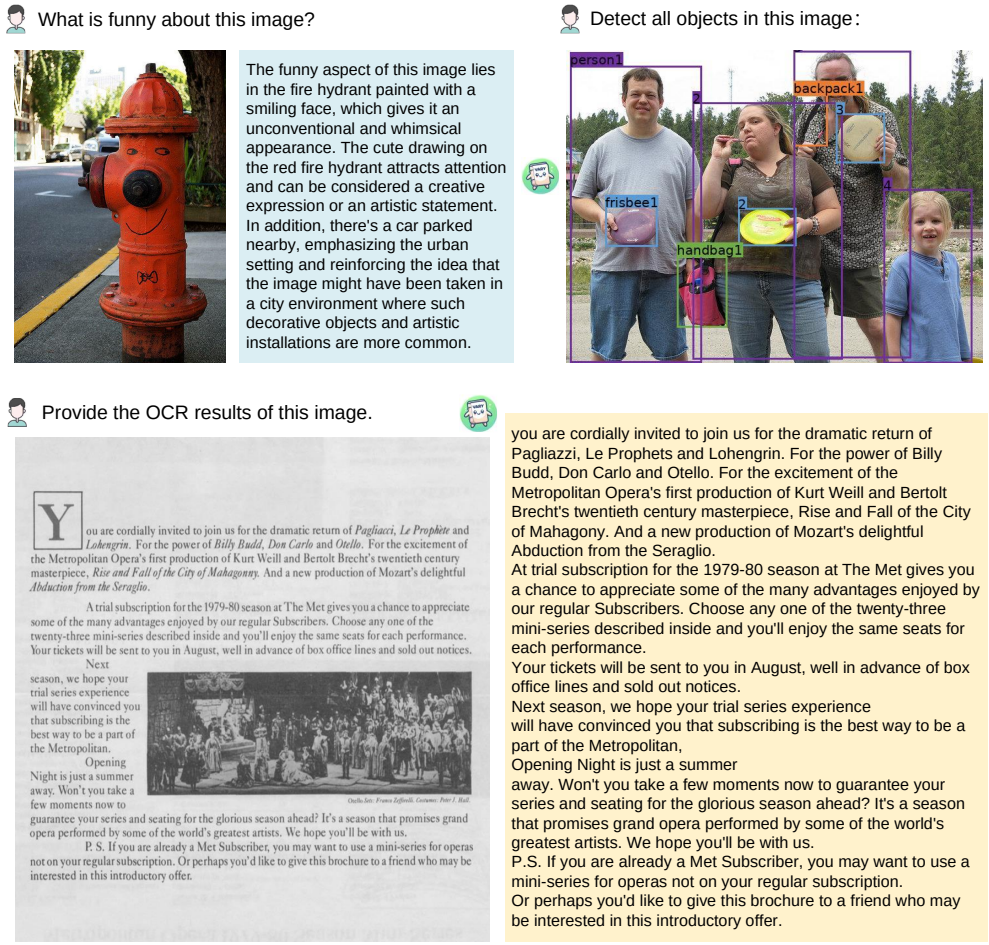

Figure 4: Visualization of high-quality results of our model in four common fields. We can see that Vary-toy has satisfactory general ability and enjoys strong text and object perception abilities.

4.6 Visualization

Figure 4 shows high-quality results of Vary-toy on four different downstream fields. We can see that the model enjoys good vision concept understanding and localization capacities, indicating that a reinforced vision vocabulary with a small language model can also perform well in multimodal tasks.

5 Conclusion

In this report, we propose a small LVLM — Vary-toy, which can be deployed on a GTX1080ti GPU and enjoys fine performance in many downstream tasks. What's more, we generate a new and more comprehensive vision vocabulary for the presented model, which is the key to the success of Vary-toy. We hope the promising and user-friendly Vary-toy can become a new baseline in such fields as well as draw more attention to LVLM, especially for researchers who previously lacked computing resources. We also encourage researchers to use our reinforced vision vocabulary for more downstream tasks. Finally, we firmly confirm that the Vary-toy will evolve beyond just a toy.

References

[1] Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J.L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: NeurIPS (2022) 1, 3, 7

[2] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., Zhu, T.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023) 2, 3, 6, 7

[3] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023) 2, 3

[4] Biten, A.F., Litman, R., Xie, Y., Appalaraju, S., Manmatha, R.: Latr: Layout-aware transformer for scene-text vqa. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16548–16558 (2022) 1, 3

[5] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 3, 6

[6] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 213–229. Springer (2020) 2

[7] Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023) 7

[8] Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna/ (2023) 3, 5

[9] Chu, X., Qiao, L., Lin, X., Xu, S., Yang, Y., Hu, Y., Wei, F., Zhang, X., Zhang, B., Wei, X., Shen, C.: Mobilevlm: A fast, strong and open vision language assistant for mobile devices (2023) 3

[10] Davis, B., Morse, B., Price, B., Tensmeyer, C., Wigington, C., Morariu, V.: End-to-end document recognition and understanding with dessurt. In: European Conference on Computer Vision. pp. 280–296. Springer (2022) 6

[11] Deng, J., Yang, Z., Chen, T., Zhou, W., Li, H.: Transvg: End-to-end visual grounding with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1769–1779 (2021) 7

[12] Dong, R., Han, C., Peng, Y., Qi, Z., Ge, Z., Yang, J., Zhao, L., Sun, J., Zhou, H., Wei, H., et al.: Dreamllm: Synergistic multimodal comprehension and creation. arXiv preprint arXiv:2309.11499 (2023) 3

[13] Gan, Z., Chen, Y.C., Li, L., Zhu, C., Cheng, Y., Liu, J.: Large-scale adversarial training for vision-and-language representation learning. Advances in Neural Information Processing Systems **33**, 6616–6628 (2020) 7

[14] Hao, Y., Song, H., Dong, L., Huang, S., Chi, Z., Wang, W., Ma, S., Wei, F.: Language models are general-purpose interfaces. arXiv preprint arXiv:2206.06336 (2022) 1

[15] Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 787–798 (2014) 2, 6

[16] Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: European Conference on Computer Vision. pp. 498–517. Springer (2022) 6

[17] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 4

[18] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision **128**(7), 1956–1981 (2020) 4

[19] Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 734–750 (2018) 2

[20] Lee, K., Joshi, M., Turc, I.R., Hu, H., Liu, F., Eisenschlos, J.M., Khandelwal, U., Shaw, P., Chang, M.W., Toutanova, K.: Pix2struct: Screenshot parsing as pretraining for visual language understanding. In: International Conference on Machine Learning. pp. 18893–18912. PMLR (2023) 6

[21] Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023) 7

[22] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023) 1, 3, 7, 8

[23] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017) 2

[24] Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: ECCV. pp. 740–755 (2014) 1, 3, 5, 6

[25] Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023) 3, 7

[26] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023) 1, 2, 3, 5, 6, 7, 8

[27] Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016) 6

[28] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019) 6

[29] Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv preprint arXiv:2203.10244 (2022) 2, 5, 6, 7

[30] Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021) 1, 2, 3, 5, 6, 7

[31] Microsoft: Phi-2: The surprising power of small language models. https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/ (2023) 3

[32] Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. In: 2019 international conference on document analysis and recognition (ICDAR). pp. 947–952. IEEE (2019) 1, 3

[33] OpenAI: Gpt-4 technical report (2023) 6

[34] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: NeurIPS (2022) 1, 3

[35] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 1, 5

[36] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019) 3

[37] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016) 2

[38] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015) 2

[39] Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021) 5, 7

[40] Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8430–8439 (2019) 4

[41] Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019) 1

[42] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca (2023) 3, 5, 6

[43] Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023) 3

[44] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 3

[45] Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S.: Coco-text: Dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv:1601.07140 (2016) 1

[46] Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: International Conference on Machine Learning. pp. 23318–23340. PMLR (2022) 7

[47] Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. arXiv preprint arXiv:2305.11175 (2023) 7

[48] Wei, H., Kong, L., Chen, J., Zhao, L., Ge, Z., Yang, J., Sun, J., Han, C., Zhang, X.: Vary: Scaling up the vision vocabulary for large vision-language models. arXiv preprint arXiv:2312.06109 (2023) 1, 2, 3, 4, 6

[49] Wei, H., Liu, C., Guo, P., Zhu, Y., Fu, J., Wang, B., Wang, P.: Corner affinity: A robust grouping algorithm to make corner-guided detector great again. In: Raedt, L.D. (ed.) Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. pp. 1458–1464. International Joint Conferences on Artificial Intelligence Organization (7 2022). https://doi.org/10.24963/ijcai.2022/203, https://doi.org/10.24963/ijcai.2022/203, main Track 2

[50] Xu, C., Guo, D., Duan, N., McAuley, J.: Baize: An open-source chat model with parameter-efficient tuning on self-chat data. arXiv preprint arXiv:2304.01196 (2023) 5, 6

[51] Yang, Z., Gan, Z., Wang, J., Hu, X., Ahmed, F., Liu, Z., Lu, Y., Wang, L.: Unitab: Unifying text and box outputs for grounded vision-language modeling. In: European Conference on Computer Vision. pp. 521–539. Springer (2022) 7

[52] Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Dan, Y., Zhao, C., Xu, G., Li, C., Tian, J., et al.: mplug-docowl: Modularized multimodal large language model for document understanding. arXiv preprint arXiv:2307.02499 (2023) 3, 6

[53] Yu, E., Zhao, L., Wei, Y., Yang, J., Wu, D., Kong, L., Wei, H., Wang, T., Ge, Z., Zhang, X., et al.: Merlin: Empowering multimodal llms with foresight minds. arXiv preprint arXiv:2312.00589 (2023) 3

[54] Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 (2023) 2, 6, 8

[55] Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., et al.: Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:2210.02414 (2022) 3

[56] Zhang, A., Zhao, L., Xie, C.W., Zheng, Y., Ji, W., Chua, T.S.: Next-chat: An lmm for chat, detection and segmentation. arXiv preprint arXiv:2311.04498 (2023) 7

[57] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022) 2, 3, 4

[58] Zhao, L., Yu, E., Ge, Z., Yang, J., Wei, H., Zhou, H., Sun, J., Peng, Y., Dong, R., Han, C., et al.: Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. arXiv preprint arXiv:2307.09474 (2023) 3

[59] Zhou, X., Zhuo, J., Krahenbuhl, P.: Bottom-up object detection by grouping extreme and center points. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 850–859 (2019) 2

[60] Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023) 1, 3, 7, 8