Supplementary Materials for

LLMDet: Learning Strong Open-Vocabulary Object Detectors under the Supervision of Large Language Models

Shenghao Fu[1,3,4], Qize Yang[3], Qijie Mo[1,4], Junkai Yan[1,4], Xihan Wei[3],
Jingke Meng[1,4], Xiaohua Xie[1,4,5]*, Wei-Shi Zheng[1,2,4,6]*
[1]School of Computer Science and Engineering, Sun Yat-sen University, China;
[2]Peng Cheng Laboratory, China; [3]Tongyi Lab, Alibaba Group;
[4]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China;
[5]Guangdong Province Key Laboratory of Information Security Technology, China;
[6]Pazhou Laboratory (Huangpu), China
fushh7@mail2.sysu.edu.cn, xiexiaoh6@mail.sysu.edu.cn, wszheng@ieee.org

LLM	AP	AP_r	AP_c	AP_f
Qwen2-0.5b-instruct [12]	44.4	36.4	39.2	50.5
LLaVA-OneVision-0.5b-ov [4]	44.5	38.6	39.3	50.3
Qwen2-1.5b-instruct [12]	44.6	35.3	39.5	50.8

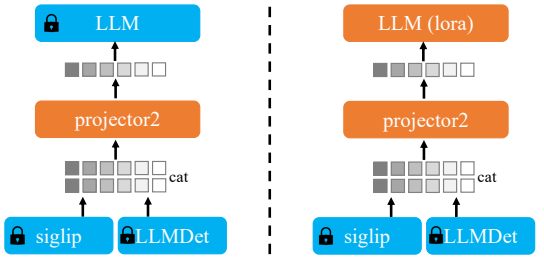
Table 1-1. Ablations on large language models.

1. More Ablation Studies

Effect of different large language models. By default, we use the LLM in LLaVA-OneVision-0.5b-ov [4], which is finetuned from Qwen2-0.5b-instruct [12]. Since the LLM in LLaVA-OneVision-0.5b-ov is pretrained with abundant multi-modal data but with a different vision encoder, the pretraining can still improve the performance, especially for rare classes (+2.2% AP_r), as shown in Table 1-1. But we find that increasing the size of the LLM only slightly improves the performance, perhaps larger language models mainly improve in reasoning ability which does not benefit the detector's visual representations.

2. LLMDet Builds a Stronger Large Vision-Language Model

In this subsection, we show that LLMDet can serve as a general vision foundation model and in turn gets a strong large multi-modal model. Recent large multi-modal models (LMM) are based on pretrained large language models and pretrained vision foundation models. Different vision foundation models will significantly affect the performance of LMMs [16]. Since LLMDet is enhanced under the supervision of long detailed image-level captions and pre-aligned with LLM, LLMDet inherits great potential to build a stronger LMM. Following recent advances [10, 11, 16], we build the LMM using a mixture of vision experts, *i.e.* a SigLIP [14] vision encoder and our LLMDet. As shown in Figure 2-1, the visual features from two vision encoders are concatenated along the channel dimension, and then a projector is utilized to map the features to the LLM's input space. We start from LLaVA-OneVision-0.5b-ov [4] and



	Step 3	Step 4
goal	Alignment	Training LMM
task	Image Captioning	Instruction Following
loss	llm loss	llm loss
dataset	LSC 558k recap	LLaVA 1.5 instruction tuning dataset (665K)

Figure 2-1. The multi-step training pipeline of using LLMDet to build a strong large multi-modal model. The large multi-modal model uses a mixture of vision encoders, including LLMDet and SigLIP. In each step, modules in orange color are tunable while modules in blue color are frozen. We first pretrain a new projector and then finetune the large multi-modal model with visual instruct tuning.

*: Corresponding authors are Xiaohua Xie and Wei-Shi Zheng.
Part of the work was done when Shenghao Fu was an intern at Alibaba.

Method	GQA [3]	POPE [6]			MME [2]	
		rand	pop	adv	perception	cognition
OneVision-0.5b	56.9	87.5	86.3	85.0	1238	240
OneVision-0.5b +MM-GDINO	**61.2**	**88.9**	**88.1**	**86.6**	1207	256
OneVision-0.5b +LLMDet	**61.2**	88.8	88.0	86.0	**1297**	**264**

Table 2-2. Multi-modal performance using different vision encoders. OneVision-0.5b is short for LLaVA-OneVision-0.5b-ov [4].

insert our LLMDet to it as shown in Figure 2-1. We first pretrain a new projector and then finetune the LLM with the LLaVA 1.5 [7] instruction tuning dataset which is only a small part of the dataset used in LLaVA-Onevision.

We select three representative benchmarks to evaluate the multi-modal performance of the LMM: the comprehensive understanding benchmark MME [2], the hallucination benchmark POPE [6] and the academic VQA benchmark GQA [3]. As shown in Table 2-2, combining the MM-GDINO to LLaVA-OneVision-0.5b-ov can improve the performance on GQA and POPE. As detectors excel at localizing objects in the image, the precise localization makes the LLM aware of the objects existed in images, which helps the LLM overcome hallucination and perform simple QA about objects in the image. The multi-modal perception and understanding ability can be further enhanced with a stronger LLMDet which is also pre-aligned [10] with the LLM in LLaVA-OneVision-0.5b-ov. The resulting LMM achieves the highest performance on the MME benchmark, validating the mutual benefits between the detector and the LMM.

3. Limitations

Although we provide detailed captions to train LLMs, we find that the LLM co-trained with detectors tends to output relatively short descriptions for the whole image, even given the prompts to describe the image in detail. We suppose the reason is that our region-level data is far more than the image-level data (one image has multiple regions).

Further, our region-level descriptions are too simple as they are just the grounding phrases of the regions. We believe collecting some high-informative data for regions like DetCLIPv3 can further improve the performance.

4. Implement Details of Zero-Shot Test on Referring Expression Comprehension Datasets

In this work, LLMDet is trained with phrase grounding loss and caption generation loss. In the phrase grounding task, the model is asked to detect each phrase in the given grounding text. For example, the model is expected to detect "the man" and "umbrella" in the text "the man with an umbrella".

To demonstrate the great open-vocabulary ability of LLMDet, we directly transfer LLMDet to the referring expression comprehension (REC) task, which is a task slightly different from the phrase grounding task. In REC, the model should only detect the single object referred by the given sentence. For example, the model should only detect "the man" in the text "the man with an umbrella", which means discrepancies exist between the pretraining task and the target task. Thus, we find that the model tends to predict the "umbrella" with the highest confidence. To minimize the discrepancies, we first use NLTK [1] tools to find the subject in the text and then select the box with the highest confidence corresponding to the subject as the answer.

5. Prompts for Calculating Detailedness and Hallucination Scores

In **??**, we utilize GPT-4o as a judge to give a comprehensive score for each caption-image pair. We referred to Halluci-Doctor [13] and adopted similar prompts as follows.

The prompt for calculating hallucination scores

Suppose you are a hallucination annotator who judges the degree of hallucination based on the number of errors in the description of objects, relations, and attributes. You should check each sentence in the description one by one.

{image}
Please carefully compare the image and the given caption below and provide the hallucination score (an integer value between 0 and 5) based on overall hallucinations in each sub-sentence, where the fewer descriptive errors in the caption, the lower the hallucination score given. Only output the score without any explanation.
Description: {caption}
Output:

The prompt for calculating detailedness scores

Suppose you are an image detail annotator who judges the degree of sentence detailedness based on the object types, textures and colors, parts of the objects, object actions, precise object locations, and texts.

{image}
Please carefully compare the image and the given caption below and provide the detailedness score (an integer value between 0 and 5) without any explanation, where caption with more factual content give a higher detailedness score. Only output the score without any explanation.
Description: {caption}
Output:

6. Detailed Zero-Shot Results

Detailed zero-shot results on ODinW35. Table 6-3 lists the detailed performance of Grounding-DINO-T [8], MM-GDINO-T [15], and our LLMDet on each dataset in ODinW35 [5]. The selected datasets in ODinW13 are also marked out.

Detailed zero-shot results on COCO-O. COCO-O [9] is a dataset sharing the same 80 classes as COCO but in different domains including cartoon, handmake, painting, sketch, tattoo, and weather. Detailed performance on each domain is listed in Table 6-4.

7. Visualization

7.1. Visualizations of the Image-Level Captions in GroundingCap-1M

In this work, we collect a new GroundingCap-1M dataset which equips a standard grounding dataset with detailed image-level captions. The captions should contain as many details as possible, including object types, textures, colors, parts of the objects, object actions, precise object locations, and texts. And the captions should not contain imaginary contents. Figure 7-2 visualizes some examples in GroundingCap-1M. The captions shown depict the main entities in the pictures with great detail (demonstrated in green color) but also with some imaginary contents inevitably (also highlighted by underlines). The imaginary contents always start with speculative words, like "seemingly", "indicating", and "suggesting". We just find these pre-defined speculative words and delete the sub-sentences including them in an online manner.

7.2. Visualizations of the Captions Generated by LLMDet

In Figure 7-3, we visualize some examples of the generated image-level and region-level captions from the LLM co-trained with LLMDet. Images are selected from the COCO validation set. The LLM can generate precise class names for the objects in COCO (as we use the class names in COCO as the grounding text for deep fusion, only objects in COCO are detected out for caption generation). But we find that the image-level captions are relatively coarse-grained compared with the ones in GroundingCap-1M. We suppose the reason is that our region-level data is far more than the image-level data (one image has multiple regions) and the region-level data is overly simplistic.

References

[1] Steven Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, 2006. 2

[2] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 2

[3] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 2

[4] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2

[5] Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. In *NeurIPS*, 2022. 3, 4

[6] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023. 2

[7] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 2

[8] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 3, 4

[9] Xiaofeng Mao, Yuefeng Chen, Yao Zhu, Da Chen, Hang Su, Rong Zhang, and Hui Xue. Coco-o: A benchmark for object detectors under natural distribution shifts. In *ICCV*, 2023. 3, 4

[10] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024. 1, 2

Dataset	ODinW13	ODinW35	G-DINO-T	MM-GDINO-T	LLMDet
AerialMaritimeDrone large	✓	✓	0.173	0.155	0.153
AerialMaritimeDrone tiled		✓	0.206	0.201	0.174
AmericanSignLanguageLetters		✓	0.002	0.007	0.016
Aquarium	✓	✓	0.195	0.281	0.268
BCCD		✓	0.161	0.078	0.149
boggleBoards		✓	0.000	0.002	0.001
brackishUnderwater		✓	0.021	0.024	0.026
ChessPieces		✓	0.000	0.000	0.000
CottontailRabbits	✓	✓	0.806	0.788	0.797
dice		✓	0.004	0.001	0.004
DroneControl		✓	0.042	0.073	0.070
EgoHands generic	✓	✓	0.608	0.518	0.518
EgoHands specific		✓	0.002	0.003	0.010
HardHatWorkers		✓	0.046	0.109	0.178
MaskWearing		✓	0.004	0.009	0.004
MountainDewCommercial		✓	0.430	0.433	0.518
NorthAmericaMushrooms	✓	✓	0.471	0.747	0.749
openPoetryVision		✓	0.000	0.000	0.003
OxfordPets by breed		✓	0.003	0.004	0.006
OxfordPets by species		✓	0.011	0.016	0.024
PKLot		✓	0.001	0.007	0.034
Packages	✓	✓	0.695	0.706	0.717
PascalVOC	✓	✓	0.563	0.566	0.584
pistols	✓	✓	0.726	0.726	0.720
plantdoc		✓	0.005	0.011	0.005
pothole	✓	✓	0.215	0.164	0.175
Raccoons	✓	✓	0.549	0.533	0.519
selfdrivingCar		✓	0.089	0.082	0.083
ShellfishOpenImages	✓	✓	0.393	0.489	0.429
ThermalCheetah		✓	0.087	0.045	0.132
thermalDogsAndPeople	✓	✓	0.657	0.548	0.546
UnoCards		✓	0.006	0.005	0.010
VehiclesOpenImages	✓	✓	0.613	0.610	0.597
WildfireSmoke		✓	0.134	0.129	0.093
websiteScreenshots		✓	0.012	0.016	0.013
ODinW13 Average			0.514	0.525	0.521
ODinW35 Average			0.227	0.231	0.238

Table 6-3. Detailed zero-shot results on ODinW35 [5].

Model	Cartoon	Handmake	Painting	Sketch	Tattoo	Weather	Average
Grounding-DINO [8]	40.2	30.2	43.1	37.6	29.8	44.8	37.6
MM-GDINO [15]	35.0	26.6	41.7	32.2	23.9	44.8	34.0
LLMDet	37.7	30.7	42.8	32.6	27.5	45.3	36.1

Table 6-4. Detailed zero-shot results on COCO-O [9].

[11] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 1

[12] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 1

[13] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wen-

tao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *CVPR*, 2024. 2

[14] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 1

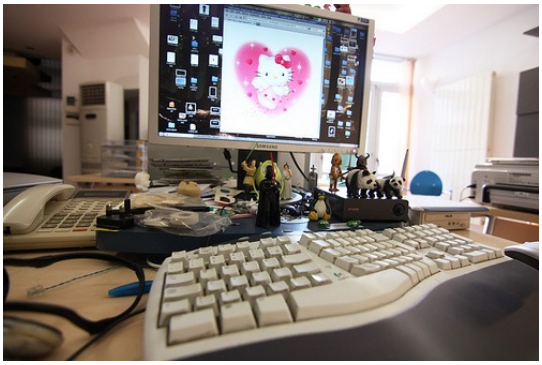
[15] Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*, 2024. 3, 4

[16] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. In *ECCV*, 2024. 1

The image depicts a sushi conveyor belt restaurant. A long, metallic conveyor belt runs horizontally across the image, carrying various dishes on red plates. The dishes include different types of sushi, sashimi, and other Japanese cuisine items. The conveyor belt is situated in the middle of the image, with diners seated on either side. On the left side of the image, a person is standing, wearing a white shirt. This person appears to be a server or a customer, and they are holding a plate with a piece of sushi on it. On the right side of the image, another person is seated at the conveyor belt. This individual is wearing a light-colored shirt and is holding a newspaper, seemingly reading it while waiting for their food. The person is seated on a red stool, which is typical in such restaurants. The overall setting suggests a casual dining environment with a focus on convenience and variety.

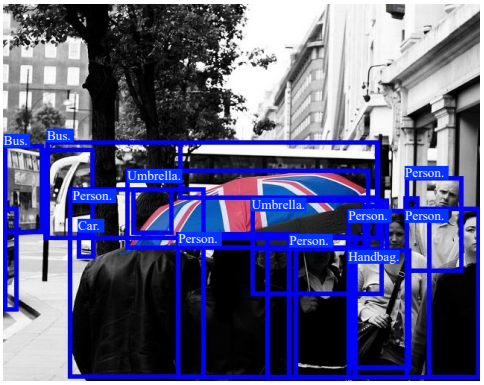
In the image, a young girl wearing a pink bathing suit is standing on a sandy beach. She has her right arm extended outward as if she's either throwing something or gesturing towards something in front of her. The girl is barefoot and looking towards the ocean. A seagull is captured mid-flight above the water, with its wings spread wide. The bird is positioned slightly to the left of the girl and appears to be flying parallel to the shoreline. The ocean itself is visible in the background, with gentle waves lapping at the shore. The overall scene suggests a sunny day at the beach, with the girl enjoying her time near the water.

In the image, there is a woman wearing a blue raincoat and hat, walking on a wet sidewalk. She is using a white cane for support as she navigates the slippery surface. The woman appears to be elderly and is dressed appropriately for the weather conditions. Nearby, there is a metal pole with a unicycle leaning against it. The unicycle has a black seat and large wheels, indicating that it is designed for stability and balance. The wet sidewalk suggests recent rainfall, which could make the surface slippery and challenging to walk on. The relative position of the objects shows the woman walking past the unicycle, which is positioned closer to the camera than the woman.

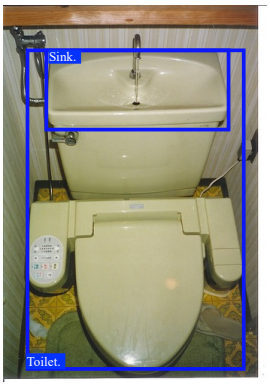


The image depicts a computer desk setup with various objects. The primary focus is on the keyboard, which is prominently placed in the foreground. It is a standard QWERTY keyboard with white keys and black lettering. Behind the keyboard, there is a computer monitor displaying an image of a pink heart with a Hello Kitty design inside it. The monitor is positioned slightly to the left of center. To the left of the keyboard, there is a telephone with a corded handset. The phone has a traditional design with a push-button keypad. On the right side of the keyboard, there are several small figurines or toys, including what appears to be a panda bear and other characters. These toys are arranged in a somewhat cluttered manner, suggesting that they might be personal items or collectibles. In the background, there is a printer situated on the right side of the desk. The printer is a standalone unit with a paper tray visible at the top. There is also a window with blinds partially drawn, allowing some natural light to enter the room. The overall setting suggests a home office or personal workspace, with a mix of work-related equipment and personal items.

Figure 7-2. Visualizations of the image-level captions in GroundingCap-1M, which are rich in detail. The great details are marked in green color. But the captions still contain some imaginary contents, which are also highlighted by underlines.



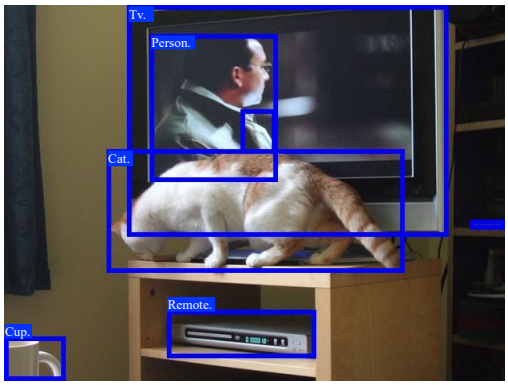
In the image, a group of people are standing on a street. The image captures a moment in time in an urban setting. It's a snapshot of life in the city. In the background, there are trees and buildings. There are also cars parked on the side of the road. Overall, this image gives you a glimpse into the daily life of a city dweller.



In the image, there is a toilet with its lid up. The toilet is white and has a handle on the side of the toilet itself. There is also a small amount of water coming out of it. The image also shows a pipe running along the wall. It is located in the bottom left corner. Overall, this image captures a moment in a bathroom setting.

In the image, there is a dining table with a bowl of fruit. The image also shows a kitchen area with cabinets and a refrigerator. There is also a chair visible in the background. Overall, this image gives the impression of a well-organized and inviting kitchen space.

In the image, there is a toilet that is beside a sink. The walls of the bathroom are white. There is also a towel hanging on the wall. The photo is taken from inside a bathroom.



In the image, a cat is on a table. There is also a tv in the background of the scene. The image also contains a cup. In the foreground, there is a coffee cup with a spoon in it. It has a handle and a spout.

In the image, there is a television that is on a table. There is also a bookshelf with books on it. The image shows a room with a bed that has a mattress and a pillow. The television is turned on.

Figure 7-3. Visualizations of the image-level and region-level generated captions from the LLM co-trained with LLMDet. Image-level captions are placed under the corresponding images and region-level captions are placed beside the bounding boxes. Only object queries with scores higher than 0.3 are visualized in the images.