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

Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, Wei-Shi Zheng

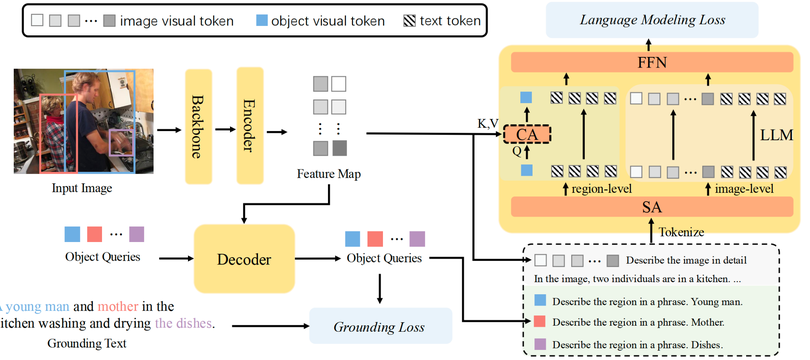
ArXiv: *https://arxiv.org/abs/2501.18954* ***Code***: *https://github.com/iSEE-Laboratory/LLMDet* ***Email***: *fushh7@mail2.sysu.edu.cn*

CVPR Nashville JUNE 11-15, 2025

Motivation & Contribution

- In this work, we show that an open-vocabulary detector **co-training with a large language model** by generating **image-level detailed captions** for each image can further improve performance.
- Compared with region-level data, image-level detailed captions:
  - ✓ **Contain more details of the image,** e.g., object types, textures, colors, parts of the objects, object actions, precise object locations, and texts.
  - ✓ **Provide a more comprehensive understanding of the image**. It aligns all elements in the image as a whole.
  - ✓ **Are more scalable**, which are easy to collect. (GroundingCap-1M)
- The fully-pretrained large language model is naturally open-vocabulary. Using an LLM to generate captions with great details makes the detector align with it, thus inheriting a strong generalization ability.
- We also show that the improved LLMDet can serve as a strong vision foundation model and in turn build a better LMM, achieving **mutual benefits**.

Design & Methodology



Training Data: (Img, $T_{Grounding}$, T_{Image}, B)

Two training objectives:

- Grounding Loss: the main task (B, $T_{Grounding}$)
- Language Modeling Loss: the auxiliary task
  - Image-level caption generation (T_{Image})
  - Region-level caption generation ($T_{Grounding}$)

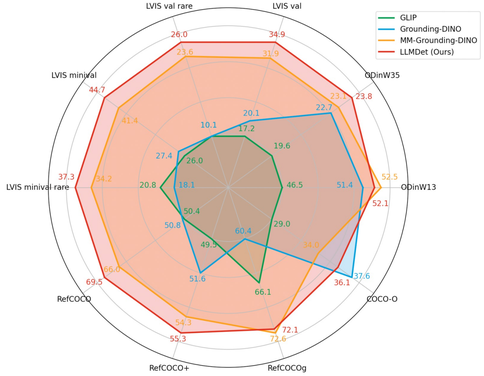
Why it works?

- Since the output answers include various details and the comprehensive understanding of the image, these visual cues should be modeled in the visual features so that the LLM can minimize the training loss and generate the captions correctly.

Experimental Result

Method	Backbone	Pre-training data	LVIS^{minival} AP	LVIS^{minival} AP_r	LVIS^{minival} AP_c	LVIS^{minival} AP_f	LVIS AP	LVIS AP_r	LVIS AP_c	LVIS AP_f
GLIP [27]	Swin-T	O365,GoldG,Cap4M	26.0	20.8	21.4	31.0	17.2	10.1	12.5	25.2
GLIPv2 [62]	Swin-T	O365,GoldG,Cap4M	29.0	–	–	–	–	–	–	–
CapDet [38]	Swin-T	O365,VG	33.8	29.6	32.8	35.5	–	–	–	–
Grounding-DINO [36]	Swin-T	O365,GoldG,Cap4M	27.4	18.1	23.3	32.7	20.1	10.1	15.3	29.9
OWL-ST [41]	CLIP B/16	WebLI2B	34.4	38.3	–	–	28.6	30.3	–	–
Desco-GLIP [25]	Swin-T	O365,GoldG,CC3M	34.6	30.8	30.5	39.0	26.2	19.6	22.0	33.6
DetCLIP [56]	Swin-T	O365,GoldG,YFCC1M	35.9	33.2	35.7	36.4	28.4	25.0	27.0	28.4
DetCLIPv2 [57]	Swin-T	O365,GoldG,CC15M	40.4	36.0	41.7	40.4	32.8	31.0	31.7	34.8
DetCLIPv3 [58]	Swin-T	O365,V3Det,GoldG,GranuCap50M	47.0	45.1	47.7	46.7	38.9	37.2	37.5	41.2
YOLO-World-L [6]	YOLOv8-L	O365,GoldG,CC3M	35.4	27.6	34.1	38.0	–	–	–	–
T-Rex2 [19]	Swin-T	10M data from various resources	42.8	37.4	39.7	46.5	34.8	29.0	31.5	41.2
OV-DINO [49]	Swin-T	O365,GoldG,CC1M	40.1	34.5	39.5	41.5	32.9	29.1	30.4	37.4
MM-GDINO [65]	Swin-T	O365,GoldG,GRIT,V3Det	41.4	34.2	37.4	46.2	31.9	23.6	27.6	40.5
LLMDet	Swin-T	GroundingCap-1M	44.7	37.3	39.5	50.7	34.9	26.0	30.1	44.3
GLIP [27]	Swin-L	FourODs,GoldG,Cap24M	37.3	28.2	34.3	41.5	26.9	17.1	23.3	36.4
GLIPv2 [62]	Swin-H	FiveODs,GoldG,CC15M,SBU	50.1	–	–	–	–	–	–	–
Grounding-DINO [36]	Swin-L	O365,OI,GoldG,Cap4M,COCO,RefC	33.9	22.2	30.7	38.8	–	–	–	–
OWL-ST [41]	CLIP L/14	WebLI2B	40.9	41.5	–	–	35.2	36.2	–	–
DetCLIP [56]	Swin-L	O365,GoldG,YFCC1M	38.6	36.0	38.3	39.3	28.4	25.0	27.0	31.6
DetCLIPv2 [57]	Swin-L	O365,GoldG,CC15M	44.7	43.1	46.3	43.7	36.6	33.3	36.2	38.5
DetCLIPv3 [58]	Swin-L	O365,V3Det,GoldG,GranuCap50M	48.8	**49.9**	**49.7**	47.8	41.4	**41.4**	**40.5**	42.3
MM-GDINO [65]	Swin-B	O365,GoldG,V3Det	44.5	37.5	39.9	49.9	34.9	26.7	30.4	43.5
MM-GDINO [65]	Swin-L	O365V2,OpenImageV6,GoldG	36.8	28.1	31.8	42.8	29.1	19.7	25.6	37.2
LLMDet	Swin-B	GroundingCap-1M	48.3	40.8	43.1	54.3	38.5	28.2	34.3	47.8
LLMDet	Swin-L	GroundingCap-1M	**51.1**	45.1	46.1	**56.6**	**42.0**	31.6	38.8	**50.2**

Table 2. Zero-shot fixed AP [8] on LVIS val [15] and minival [20]. LLMDet achieves state-of-the-art performance with much less data.



Any question? Feel free to open an issue on our GitHub or email me.

Paper **Code**