Appendix of Grounded Language-Image Pre-training

This appendix is organized as follows.

- In Section A, we provide more visualizations of our model's grounding predictions on the Conceptual Caption 12M dataset [1].
- In Section B (referred by Section 3.1), we discuss the equivalence between detection and grounding.
- In Section C.1 (referred by Section 4), we introduce the pre-training details of the models we use in Section 4.
- In Section C.2 (referred by Section 4), we introduce the evaluation details of experiments on COCO, LVIS, and Flickr30K.
- In Section C.3 (referred by Section 4), we discuss the difference between the public image-text data (Google Conceptual Captions,SBU) and the image-text data we collected.
- In Section D, we provide a detailed analysis on the computational cost and performance effect of the language-aware deep fusion.
- In Section E.1 (referred by Section 5), we introduce the 13 datasets in Object Detection in the Wild (ODinW).
- In Section E.2 (referred by Section 5), we detail the manual prompt design.
- In Section E.3 (referred by Section 5.1), we give the details for the data efficiency experiments.
- In Section E.4 (referred by Section 5.3), we give the details for the linear probing and prompt tuning experiments.
- In Section E.5, we present per-dataset results for all experiments in Section 5.

A. Visualization

We provide more visualizations of the predictions from our teacher model. Even given noise image-text pairs, our model is still capable of grounding semantic-rich phrases accurately.

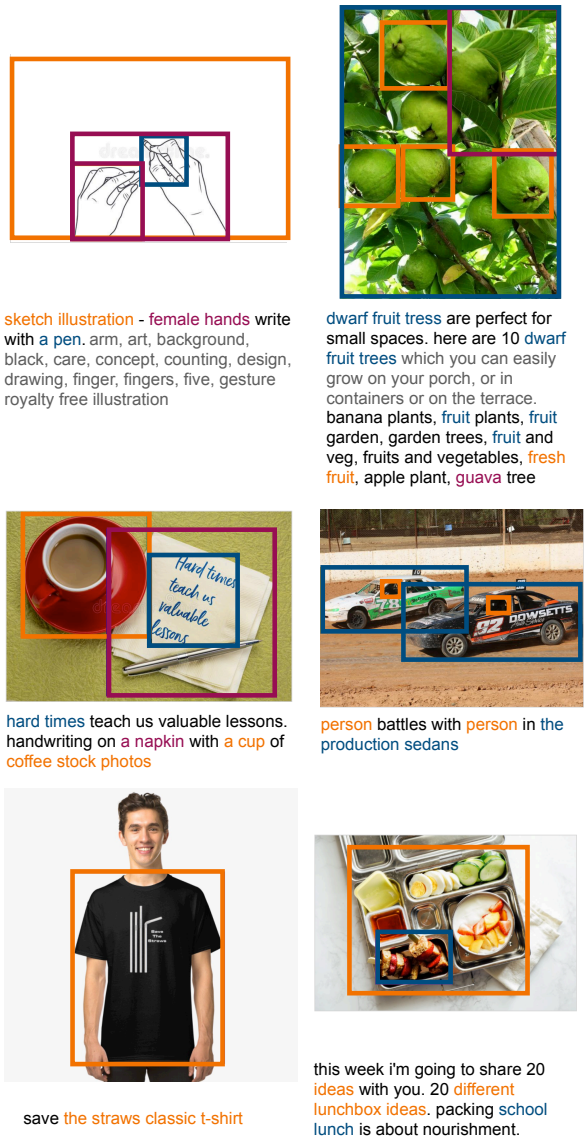

Figure 1. Predictions from the teacher model on 6 examples from Conceptual Captions 12M. Phrases and corresponding boxes are matched with the same colors.

B. Equivalence Discussion between Detection and Grounding

In Section 3.1 of the main paper, we discussed the equivalence between detection and grounding. We corroborate the discussion with empirical experiments.

When all object categories fit into a single prompt. We first confirm that when all categories fit into one prompt, our grounding formulation is equivalent to classical object detection. We conduct the experiments on COCO [6]. We first choose the SoTA detection model Dynamic Head (DyHead) [2] based on the Swin-Tiny Transformer backbone [7] as the base object detection mode. We then transform this model into a grounding model as described in Section 3.1: we concatenate the 80 class names with ". " into one prompt and replace DyHead's classification loss with our grounding loss. We use BERT (base-uncased) [3] to encode the text prompt. When concatenating the class names, we follow a fixed order.

We train the two models with the exact same hyperparameters as in [2]: we train with the standard 2x training configurations [4]. We train with batch size 32 and learning rate 1×10^{-4} (for the model with grounding reformulation, we use 1×10^{-5} for the BERT text encoder). We decay the learning rate at 67% and 89% of the total training steps.

The two models achieve the same performance on COCO 2017val: 49.4 AP. Their results are close to the 49.7 reported in the last row of Table 6 of Dai et al. [2] (the small difference is presumably due to the implementation difference). Thus, we conclude that when all categories can fit into a single prompt, grounding and detection tasks are equivalent.

When not all object categories can fit into a single prompt. The text encoder for the prompt has a limit on the input sentence length. For example, BERT can only encode sentences containing at most 512 tokens. In our implementation, to reduce computational costs, we limit the input length to 256. Thus, for certain datasets with a large vocabulary (e.g., Objects365 [8] has 365 object categories), we cannot fit all category names into one prompt. As a practical solution, we can split the category names into multiple prompts, during both training time and inference time. We find that this incurs minor performance drop. For example, in Table 2 in the main paper, DyHead-T pre-trained on Objects365 achieves 43.6 on COCO zero-shot, while GLIP-T (A) (the grounding reformulated model of DyHead) achieves 42.9 on COCO.

Pre-Train Data	COCO		LVIS minival				Flickr30K val		
	Zero-Shot	Fine-Tune	APr	APc	APf	AP	R@1	R@5	R@10
O365,GoldG,Cap4M	46.3	54.9	**20.8**	**21.4**	**31.0**	**26.0**	**85.7**	95.4	96.9
O365,GoldG,CC3M,SBU	**46.6**	**55.2**	20.1	21.3	31.1	25.9	85.3	**95.7**	**97.2**

Table 1. Comparison between public data and data crawled by us.

C. Transfer to Established Benchmarks

We introduce the implementation details of the models used in Section 4 and discuss the difference between public image-text data and the data crawled by us.

C.1. Pre-training Details

In Section 4, we introduced GLIP-T (A), GLIP-T (B), GLIP-T (C), GLIP-T, and GLIP-L. We introduce the implementation details in the following. We pre-train models based on Swin-Tiny models with 32 GPUs and a batch size of 64, and models based on Swin-Large with 64 GPUs and a batch size of 64. We use a base learning rate of 1×10^{-5} for the language backbone and 1×10^{-4} for all other parameters. The learning rate is stepped down by a factor of 0.1 at the 67% and 89% of the total training steps. We decay the learning rate when the zero-shot performance on COCO saturates. The max input length is 256 tokens for all models.

Prompt design for detection data. As noted in Section B, when we pre-train on datasets such as Objects365, we cannot fit all categories into one prompt. During pre-training, we randomly down-sample the categories and keep only the down-sampled categories in the prompt. We randomly shuffle the categories' order in the prompt. If a positive category is discarded and not kept in the prompt after down-sampling, we will also drop its corresponding boxes from the box labels.

The down-sampling is done randomly on the fly for each training example and serves as data augmentation. Specifically, for an example, we denote the positive classes that appear in the image as C_{pos} and the rest negative classes as C_{neg}. With a probability of 0.05, we sample one positive category from C_{pos}; with a probability of 0.05, we sample one negative category from C_{neg}; with a probability of 0.5, we keep all of C_{pos} and sample from C_{neg} till we have 85 categories in the prompt. For the rest of the time, we uniformly choose a number N from 1-85 and put N categories in the prompt; we always prioritize positive categories; but with a probability of 0.2, we might drop some positive categories from the prompt.

Augmentation for image-text data with generated boxes. When we pre-train the model on image-text data with generated boxes, we find it beneficial to increase the difficulty. We mix a few negative captions (that are from other examples and do not match with the image) with the positive caption (that is matched to the image) to form a longer text input. The model is trained to predict boxes and align them to the correct phrases in the positive caption. The model would need to first identify the positive caption among a few potential captions and then align the box to the correct phrases in the positive caption. This makes the grounding task more challenging and help the model learn a semantic-rich representation during pre-training. This augmentation is also done randomly on the fly. For each training example, with a probability of 0.3, we conduct such augmentation and mix in 19 negative captions; with a probability of 0.3, we mix in a random number (uniformly drawn between 1-19) of negative captions; for the rest of the time, we do not conduct such augmentation.

C.2. Evaluation Details

For fine-tuning on COCO, we use a base learning rate of 1×10^{-5} for pre-trained models.

For zero-shot evaluation on LVIS, since LVIS has over 1,000 categories and they cannot be fit into one text prompt, we segment them into multiple chunks, fitting 40 categories into one prompt and query the model multiple times with the different prompts. We find that models tend to overfit on LVIS during the course of pre-training so we monitor the performance on minival for all models and report the results with the best checkpoints.

For zero-shot evaluation on Flickr30K, models may also overfit during the course of pre-training so we monitor the performance on the validation set for all models and report the results with the best checkpoints.

C.3. Difference Between Public Data and Web-Crawled Data

For GLIP-T pre-trained with image-text data, as mentioned in Section 4, we train two versions, one with public data (CC3M,SBU) and another with data we crawled (Cap4M). Here we provide a comparison between the two models in Table 1.

The two models differ only slightly, with the Cap4M version better on LVIS while the CC3M+SBU version better on COCO. We conjecture that this is potentially because the public data is more extensively screened and contains more common categories and less rare concepts. Thus it performs slightly better on COCO while lags slightly on LVIS.

Model	Fusion	Inference (P100)		Train (V100)	
		Speed	Memory	Speed	Memory
GLIP-T	✗	4.84 FPS	1.0 GB	2.79 FPS	11.5 GB
	✓	2.52 FPS	2.4 GB	1.62 FPS	16.0 GB
GLIP-L	✗	0.54 FPS	4.8 GB	1.27 FPS	19.7 GB
	✓	0.32 FPS	7.7 GB	0.88 FPS	23.4 GB

Table 2. Computational cost of language-aware deep fusion. For speed, we report FPS, which is the number of images processed per second per GPU (higher is better). For memory consumption, we report the GPU memory used in GB (lower is better). Deep fusion brings less than 1x additional computational cost.

D. Computation Cost and Performance Analysis of Deep Fusion

In this section, we provide a more detailed ablation on the computational cost and performance effect of the language-aware deep fusion proposed in Section 3.

D.1. Computational Cost

We test the additional computational cost of the language-aware deep fusion for both GLIP-T and GLIP-L. For inference, we test on a P100 GPU with batch size 1. Note that for inference with GLIP without deep fusion, we could cache the language embeddings of the prompts; thus the inference time of GLIP without deep fusion is equivalent to that of DyHead [2].

For training, we test on a standard DGX-2 machine with 16 V100 GPUs (we test under the multi-GPU setting as it mimics the actual training environment): for GLIP-T models, we use 2 images per batch and for GLIP-L models, we use 1 images per batch. As the fusion module invovles multi-head attention over a large number of input elements, we turn on gradient checkpointing[1] for the deep fusion module, which increases training time but reduces GPU memory consumption.

Table 2 shows that the language-aware deep fusion brings less than 1x additional computational cost overall.

D.2. Performance

We provide an analysis on the effect of language-aware deep fusion when different kinds of pre-training data are used. We pre-train four variants of GLIP-T and show the results In Table 3. Deep fusion is beneficial for testing on 1) common categories (i.e., COCO); 2) grounding tasks (i.e., Flickr30K), and 3) low-resource transfer to real-world downstream tasks (i.e., ODinW).

However, on LVIS, the effect of deep fusion seems unclear: when only detection data are used, deep fusion seems

[1] https://pytorch.org/docs/stable/checkpoint.html

Deep Fusion	Data	COCO		LVIS minival				Flickr30K val			ODinW					
		Zero-Shot	Fine-Tune	APr	APc	APf	AP	R@1	R@5	R@10	0-Shot	1-Shot	3-Shot	5-Shot	10-Shot	Full-Shot
✗	O365	42.9	52.9	**14.2**	**13.9**	**23.4**	**18.5**	46.4	63.2	66.9	28.7	43.5	48.8	50.4	54.1	**63.6**
✓	O365	**44.9**	**53.8**	13.5	12.8	22.2	17.8	41.4	57.7	61.0	**33.2**	**48.0**	**52.0**	**53.2**	**54.9**	62.7
✗	O365,GoldG	41.6	52.9	15.8	**23.0**	30.8	**26.1**	82.4	94.7	**96.6**	35.5	47.2	51.9	53.8	54.3	**65.1**
✓	O365,GoldG	**46.7**	**55.1**	**17.7**	19.5	**31.0**	24.9	**84.8**	**94.9**	96.3	**44.4**	**49.6**	**53.8**	**54.8**	**57.2**	63.9

Table 3. Language-aware fusion benefits most tasks. We reported the full-model tuning performance for ODinW few-shot results. For models trained with only O365, performance on Flickr30K (grey numbers) is significantly worse because the models are not trained to ground natural language captions.

to degrades performance (row 1 v.s. row 2); when grounding data are present, deep fusion degrades common category performance but improves rare category performance. Our assumption is that when GLIP is only trained with detection data (e.g., O365), the language model could "overfit" to the categories in O365 and does not generalize to novel categories well (i.e., outputs out-of-distribution text representation). The deep fusion could "amplify" such overfit as the visual representation is conditioned on the language model. Thus, when tested on prompts containing novel categories (e.g., LVIS), deep fusion could degrade performance. When grounding data are used, such overfit could be mitigated.

E. Object Detection in the Wild

In this section, we provide the details and additional results for the experiments in Section 5.

E.1. Dataset Details

We use 13 datasets from Roboflow[2]. Roboflow hosts over 30 datasets and we exclude datasets that are too challenging (e.g., detecting different kinds of chess pieces) or impossible to solve without specific domain knowledge (e.g., understanding sign language).

We provide the details of the 13 datasets we use in Table 4. We include the PASCAL V0C 2012 dataset as a reference dataset, as public baselines have been established on this dataset. For PascalVOC, we follow the convention and report on validation set. For Pistols, there are no official validation or test sets so we split the dataset ourselves.

E.2. Manual Prompt Tuning

As discussed in Section 5, we find it beneficial to manually design some prompts to provide language guidance. We provide the prompts we use in Table 5. We design the prompts for 6 datasets. Since some prompts are sentences, we only apply these prompts for models trained with grounding data (GLIP-T (C), GLIP-T, and GLIP-L). For GLIP-T (A) and GLIP-T (B), we find it beneficial to use prompts for the Rabbits and Mushrooms datasets, as the prompts there are just single word or short phrases. Overall, using prompts improves AP without any model re-training (e.g., the AP improves from 22.1 to 50.0 for EgoHands).

[2] https://public.roboflow.com/object-detection

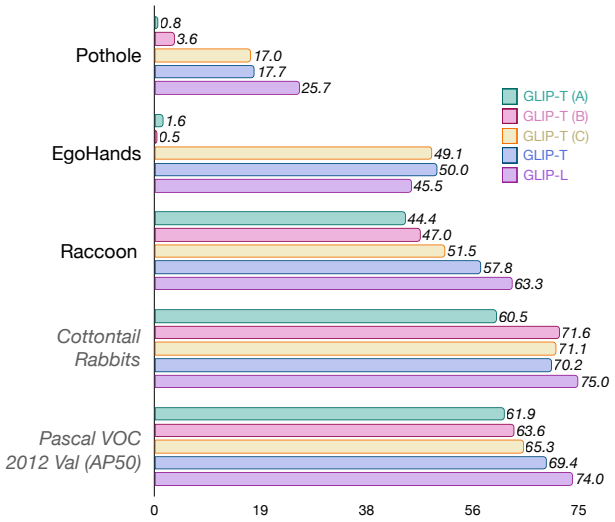

Figure 2. Per dataset zero-shot performance. The first 3 datasets contain novel categories not present in the Objects365 vocabulary while the last 2 datasets' categories are covered by Obj365 data. Grounding data bring significant benefit to novel categories.

E.3. Data Efficiency

We provide details for the experiments in Section 5.1. We train with batch size 4, learning rate 1×10^{-4} (for the model with grounding reformulation, we use 1×10^{-5} for the BERT text encoder), and weight decay of 0.05. We do not find that increasing batch size improves performance significantly. For computational reasons, we use a batch size of 4. Following convention, we freeze the bottom 2 layers of the backbone during fine-tuning. We monitor the performance on validation and decay the learning rate by 0.1 when the validation performance plateaus. In X-shot settings, we randomly sample the dataset such that there are at least X examples per category [5]. We change the random seeds (and thus change the sampled data) and conduct 3 independent runs for each X-shot experiment. We pro-

Dataset	Objects of Interest	Train/Val/Test	URL
PascalVOC	Common objects (PascalVOC 2012)	13690/3422/-	https://public.roboflow.com/object-detection/pascal-voc-2012
AerialDrone	Boats, cars, etc. from drone images	52/15/7	https://public.roboflow.com/object-detection/aerial-maritime
Aquarium	Penguins, starfish, etc. in an aquarium	448/127/63	https://public.roboflow.com/object-detection/aquarium
Rabbits	Cottontail rabbits	1980/19/10	https://public.roboflow.com/object-detection/cottontail-rabbits-video-dataset
EgoHands	Hands in ego-centric images	3840/480/480	https://public.roboflow.com/object-detection/hands
Mushrooms	Two kinds of mushrooms	41/5/5	https://public.roboflow.com/object-detection/na-mushrooms
Packages	Delivery packages	19/4/3	https://public.roboflow.com/object-detection/packages-dataset
Raccoon	Raccoon	150/29/17	https://public.roboflow.com/object-detection/raccoon
Shellfish	Shrimp, lobster, and crab	406/116/58	https://public.roboflow.com/object-detection/shellfish-openimages
Vehicles	Car, bus, motorcycle, truck, and ambulance	878/250/126	https://public.roboflow.com/object-detection/vehicles-openimages
Pistols	Pistol	2377/297/297	https://public.roboflow.com/object-detection/pistols/1
Pothole	Potholes on the road	465/133/67	https://public.roboflow.com/object-detection/pothole
Thermal	Dogs and people in thermal images	142/41/20	https://public.roboflow.com/object-detection/thermal-dogs-and-people

Table 4. 13 ODinW dataset statistics. We summarize the objects of interest for each dataset and report the image number of each split.

Dataset	Original Prompt	AP	Manually Designed Prompts	AP
Aquarium	*penguin* *puffin* *stingray*	17.7	*penguin*, which is black and white *puffin* with orange beaks *stingray* which is flat and round	18.4
Rabbits	*Cottontail-Rabbits*	68.0	*rabbit*	70.2
EgoHands	*hand*	22.1	*hand* of a person	50.0
Mushrooms	*Cow. Chanterelle*	13.6	*flat mushroom. yellow mushroom*	73.8
Packages	*package*	50.0	there is a *package* on the porch	72.3
Pothole	*pothole*	17.8	there are some *holes* on the road	17.7

Table 5. Manually designed prompts for 6 datasets. Words in *italic* are the objects of interest. The prompts either provide attributes, specify the category name in more common words, or provide language contexts. They can improve AP (CLIP-T) without any annotation or model re-training. Specifically for Pothole, although the changed prompt does not improve the AP of CLIP-T, we find it effective for CLIP-T (C) so we still apply the prompt.

Model	Zero Shot	Full Tuning				
		1	3	5	10	All
DyHead-T COCO	-	$31.9_{\pm 4.1}$	$44.2_{\pm 0.4}$	$44.7_{\pm 2.1}$	$50.1_{\pm 2.0}$	63.2
DyHead-T O365	-	$33.8_{\pm 4.3}$	$43.6_{\pm 1.2}$	$46.4_{\pm 1.4}$	$50.8_{\pm 1.6}$	60.8
GLIP-T (A)	28.7	$43.5_{\pm 1.5}$	$48.8_{\pm 0.4}$	$50.4_{\pm 0.7}$	$54.1_{\pm 0.5}$	63.6
GLIP-T (B)	33.2	$48.0_{\pm 0.8}$	$52.0_{\pm 0.4}$	$53.2_{\pm 0.9}$	$54.9_{\pm 0.7}$	62.7
GLIP-T (C)	44.4	$49.6_{\pm 0.3}$	$53.8_{\pm 0.2}$	$54.8_{\pm 1.0}$	$57.2_{\pm 1.1}$	63.9
GLIP-T	46.5	$51.1_{\pm 0.1}$	$54.9_{\pm 0.3}$	$56.4_{\pm 0.5}$	$58.4_{\pm 0.2}$	64.9
GLIP-L	52.1	$59.9_{\pm 1.7}$	$62.1_{\pm 0.8}$	$64.2_{\pm 0.4}$	$64.9_{\pm 0.9}$	68.9

Table 6. Zero-shot and full fine-tuning performance. GLIP models exhibit superior data efficiency.

vide two DyHead-T variants as baselines, one trained on COCO and one trained on Objects365. We report the full zero-shot results in Table 9 and few-shot results in Table 6.

We further plot the zero-shot performance of GLIP variants on 5 different datasets in Figure 2. We find that the introduction of grounding data brings significant improvement on certain tasks that test novel concepts, e.g., on Pothole and EgoHands, models without grounding data (A&B) performs terribly, while models with grounding data (C)

Model	Linear Probing				
	1	3	5	10	All
DyHead-T COCO	$22.7_{\pm 1.1}$	$32.7_{\pm 1.4}$	$30.5_{\pm 2.9}$	$34.1_{\pm 1.4}$	43.1
DyHead-T COCO-Cosine	$21.8_{\pm 4.4}$	$30.6_{\pm 2.2}$	$33.3_{\pm 1.2}$	$35.5_{\pm 1.2}$	43.5
DyHead-T O365	$30.7_{\pm 3.3}$	$36.2_{\pm 3.3}$	$39.6_{\pm 0.4}$	$40.0_{\pm 2.7}$	48.2
DyHead-T O365-Cosine	$25.2_{\pm 2.6}$	$37.6_{\pm 0.5}$	$38.9_{\pm 0.7}$	$41.5_{\pm 0.5}$	49.4
GLIP-T (A)	$34.6_{\pm 0.7}$	$35.9_{\pm 0.2}$	$37.6_{\pm 0.1}$	$37.9_{\pm 0.2}$	44.1
GLIP-T (B)	$40.9_{\pm 0.3}$	$42.8_{\pm 0.4}$	$44.0_{\pm 0.2}$	$44.4_{\pm 0.3}$	51.8
GLIP-T (C)	$43.9_{\pm 0.1}$	$45.4_{\pm 0.1}$	$45.9_{\pm 0.2}$	$46.7_{\pm 0.3}$	52.7
GLIP-T	$48.9_{\pm 0.2}$	$50.5_{\pm 0.1}$	$50.4_{\pm 0.3}$	$51.2_{\pm 0.2}$	55.1
GLIP-L	$54.1_{\pm 0.3}$	$54.7_{\pm 0.2}$	$55.0_{\pm 0.0}$	$55.9_{\pm 0.4}$	59.2

Table 7. Linear probing performance.

Model	Prompt Probing				
	1	3	5	10	All
GLIP-T (A)	$34.0_{\pm 0.1}$	$37.0_{\pm 0.6}$	$40.0_{\pm 0.4}$	$39.2_{\pm 1.0}$	43.3
GLIP-T (B)	$46.4_{\pm 0.5}$	$49.0_{\pm 0.9}$	$50.6_{\pm 0.5}$	$52.7_{\pm 0.1}$	58.5
GLIP-T (C)	$50.6_{\pm 0.5}$	$52.9_{\pm 0.5}$	$53.9_{\pm 0.7}$	$55.8_{\pm 1.1}$	62.8
GLIP-T	$49.9_{\pm 0.7}$	$53.7_{\pm 1.6}$	$55.5_{\pm 0.6}$	$56.6_{\pm 0.3}$	62.4
GLIP-L	$59.5_{\pm 0.4}$	$61.4_{\pm 0.4}$	$62.4_{\pm 0.6}$	$64.1_{\pm 0.6}$	67.9

Table 8. Prompt tuning performance.

outperform them with ease. Detailed results for all datasets are available in Table 9.

E.4. One Model for All Tasks

In Section 5.2, we conduct experiments with respect to deployment efficiency: tuning the least amount of parameters for the best performance. For all models, we experiment with the linear probing setting; for GLIP models, we also experiment with the prompt tuning setting. For linear probing, we try both the vanilla approach (simply tune the classification and localization head) and the cosine scale approach [9]. Below we provide the implementation details.

For the vanilla linear probing, we train with a learning

rate of 1×10^{-4}, batch size of 4, and weight decay of 0.05. For linear probing with the cosine scale, we use a scale of 20.0 per suggestions of Wang et al. [9], learning rate of 0.01, batch size of 4, and weight decay of 0.05. For prompt tuning, we train with a learning rate of 0.05, batch size of 4, and weight decay of 0.25. We have conducted preliminary searches for the hyper-parameters.

Results are present in Table 7 (linear probing) and Table 8 (prompt tuning). Comparing them with full-tuning results (Table 6), we see prompt tuning performance of GLIP is competitive, showing the deployment efficiency. Contrary to Wang *et al*. [9] who report that linear probing can deliver competitive performance for classical detection models, we find that linear probing does not work well compared to full tuning. We find that the reason could be the transfer datasets (ODinW) in our case contain a lot of novel tasks and domains, while experiments in Wang *et al*. focus on transferring to common domains (e.g., PascalVOC and COCO). In Table 10, we report the per-dataset performance. We find that for some common tasks or domains (e.g., PascalVOC and Vehicles), linear probing of DyHead COCO performs competitively with full fine-tuning but the gap is large for some other tasks of a novel domain (e.g., AerialDrone).

E.5. All Results

We report the per-dataset performance under 0,1,3,5,10-shot and full data as well as linear probing, prompt tuning, and full-model tuning in Table 9, Table 10, and Table 11 (on the next pages).

Model	PascalVOC	AerialDrone	Aquarium	Rabbits	EgoHands	Mushrooms	Packages	Raccoon	Shellfish	Vehicles	Pistols	Pothole	Thermal	Avg
GLIP-T (A)	47.7	9.8	16.8	60.5	1.6	13.7	48.5	44.4	20.4	52.4	25.3	0.8	32.3	28.8
GLIP-T (B)	50.6	4.9	19.4	71.6	0.5	21.8	29.7	47.0	21.4	56.0	47.4	3.6	57.1	33.2
GLIP-T (C)	51.6	8.1	22.6	71.1	49.1	69.4	65.6	51.5	29.3	49.9	42.7	17.0	49.2	44.4
GLIP-T	56.2	12.5	18.4	70.2	50.0	73.8	72.3	57.8	26.3	56.0	49.6	17.7	44.1	46.5
GLIP-L	61.7	7.1	26.9	75.0	45.5	49.0	62.8	63.3	68.9	57.3	68.6	25.7	66.0	52.1

Table 9. Zero-shot performance on 13 ODinW datasets.

Model	Shot	Tune	PascalVOC	AerialDrone	Aquarium	Rabbits	EgoHands	Mushrooms	Packages	Raccoon	Shellfish	Vehicles	Pistols	Pothole	Thermal	Avg
DyHead COCO	1	Linear	48.2$_{\pm 2.4}$	2.7$_{\pm 2.0}$	8.5$_{\pm 1.5}$	57.8$_{\pm 3.2}$	9.7$_{\pm 3.4}$	30.2$_{\pm 18.3}$	13.2$_{\pm 9.4}$	30.2$_{\pm 4.0}$	9.9$_{\pm 4.0}$	42.5$_{\pm 4.1}$	5.7$_{\pm 7.1}$	2.6$_{\pm 2.0}$	34.2$_{\pm 19.7}$	22.7$_{\pm 0.9}$
DyHead COCO	3	Linear	55.6$_{\pm 0.6}$	2.7$_{\pm 3.0}$	12.3$_{\pm 0.5}$	57.4$_{\pm 3.1}$	15.4$_{\pm 2.1}$	57.1$_{\pm 1.6}$	30.6$_{\pm 16.9}$	55.4$_{\pm 1.6}$	14.8$_{\pm 1.4}$	51.0$_{\pm 3.9}$	22.8$_{\pm 3.1}$	8.7$_{\pm 1.0}$	41.5$_{\pm 11.1}$	32.7$_{\pm 1.1}$
DyHead COCO	5	Linear	56.4$_{\pm 0.2}$	2.7$_{\pm 2.4}$	14.1$_{\pm 0.9}$	54.7$_{\pm 4.9}$	8.8$_{\pm 6.6}$	47.1$_{\pm 12.6}$	24.6$_{\pm 22.9}$	51.6$_{\pm 2.9}$	17.0$_{\pm 0.6}$	46.6$_{\pm 3.0}$	20.3$_{\pm 13.9}$	7.8$_{\pm 2.1}$	44.3$_{\pm 4.2}$	30.5$_{\pm 2.4}$
DyHead COCO	10	Linear	57.4$_{\pm 0.3}$	7.4$_{\pm 0.7}$	16.0$_{\pm 2.2}$	59.8$_{\pm 0.8}$	18.6$_{\pm 0.3}$	55.0$_{\pm 0.8}$	30.8$_{\pm 17.1}$	53.0$_{\pm 4.0}$	16.7$_{\pm 0.7}$	50.7$_{\pm 0.9}$	27.8$_{\pm 1.9}$	3.1$_{\pm 4.3}$	47.5$_{\pm 3.1}$	34.1$_{\pm 1.2}$
DyHead COCO	All	Linear	61.3	10.3	21.6	61.4	39.0	55.4	54.4	57.3	23.1	60.7	47.9	14.9	53.5	43.1
DyHead COCO	1	Full	31.7$_{\pm 3.1}$	14.3$_{\pm 2.4}$	13.1$_{\pm 2.0}$	63.6$_{\pm 1.4}$	40.9$_{\pm 7.0}$	67.0$_{\pm 3.6}$	34.6$_{\pm 12.1}$	45.9$_{\pm 3.8}$	10.8$_{\pm 5.0}$	34.0$_{\pm 3.3}$	12.0$_{\pm 10.4}$	6.1$_{\pm 1.3}$	40.9$_{\pm 7.4}$	31.9$_{\pm 3.3}$
DyHead COCO	3	Full	44.1$_{\pm 0.7}$	19.2$_{\pm 3.0}$	22.6$_{\pm 1.3}$	64.8$_{\pm 1.7}$	54.4$_{\pm 2.5}$	78.9$_{\pm 1.3}$	61.6$_{\pm 10.3}$	50.0$_{\pm 2.1}$	20.8$_{\pm 3.5}$	44.9$_{\pm 1.9}$	34.4$_{\pm 11.1}$	20.6$_{\pm 2.4}$	57.9$_{\pm 2.3}$	44.2$_{\pm 0.3}$
DyHead COCO	5	Full	44.9$_{\pm 1.5}$	22.2$_{\pm 3.0}$	31.7$_{\pm 1.0}$	65.2$_{\pm 1.5}$	55.6$_{\pm 3.7}$	78.7$_{\pm 3.9}$	50.1$_{\pm 13.7}$	48.7$_{\pm 4.8}$	22.8$_{\pm 3.3}$	52.0$_{\pm 1.2}$	39.8$_{\pm 6.7}$	20.9$_{\pm 1.5}$	48.0$_{\pm 2.8}$	44.7$_{\pm 1.7}$
DyHead COCO	10	Full	48.4$_{\pm 1.2}$	27.5$_{\pm 1.4}$	39.3$_{\pm 2.7}$	62.1$_{\pm 5.9}$	61.6$_{\pm 1.4}$	81.7$_{\pm 3.4}$	58.8$_{\pm 9.0}$	52.9$_{\pm 3.2}$	30.1$_{\pm 3.2}$	54.1$_{\pm 3.3}$	44.8$_{\pm 4.9}$	26.7$_{\pm 2.4}$	63.4$_{\pm 2.8}$	50.1$_{\pm 1.6}$
DyHead COCO	All	Full	60.1	27.6	53.1	76.5	79.4	86.1	69.3	55.2	44.0	61.5	70.6	56.6	81.0	63.2
DyHead O365	1	Linear	45.2$_{\pm 3.0}$	10.8$_{\pm 3.6}$	13.8$_{\pm 0.7}$	61.4$_{\pm 0.7}$	8.9$_{\pm 6.3}$	52.6$_{\pm 8.7}$	58.7$_{\pm 3.7}$	44.0$_{\pm 10.4}$	14.9$_{\pm 2.9}$	40.0$_{\pm 0.4}$	6.9$_{\pm 5.0}$	1.7$_{\pm 1.2}$	39.8$_{\pm 7.2}$	30.7$_{\pm 2.7}$
DyHead O365	3	Linear	54.6$_{\pm 0.4}$	12.4$_{\pm 3.0}$	22.3$_{\pm 1.5}$	64.0$_{\pm 2.4}$	10.5$_{\pm 6.8}$	53.6$_{\pm 10.6}$	49.1$_{\pm 16.3}$	60.5$_{\pm 1.6}$	20.6$_{\pm 2.2}$	51.3$_{\pm 2.3}$	25.5$_{\pm 0.9}$	8.2$_{\pm 1.1}$	38.9$_{\pm 12.6}$	36.3$_{\pm 2.7}$
DyHead O365	5	Linear	56.1$_{\pm 0.4}$	13.6$_{\pm 1.8}$	24.8$_{\pm 1.1}$	63.1$_{\pm 5.5}$	15.3$_{\pm 1.6}$	55.2$_{\pm 10.3}$	70.2$_{\pm 2.8}$	60.1$_{\pm 2.4}$	23.0$_{\pm 1.4}$	53.5$_{\pm 0.9}$	26.1$_{\pm 2.1}$	6.8$_{\pm 2.3}$	46.9$_{\pm 3.5}$	39.6$_{\pm 0.4}$
DyHead O365	10	Linear	57.5$_{\pm 0.3}$	8.2$_{\pm 3.0}$	28.2$_{\pm 0.8}$	65.4$_{\pm 3.2}$	17.5$_{\pm 0.6}$	68.0$_{\pm 0.8}$	49.8$_{\pm 17.3}$	60.3$_{\pm 2.1}$	22.9$_{\pm 1.0}$	56.4$_{\pm 0.8}$	28.0$_{\pm 2.2}$	7.6$_{\pm 0.9}$	50.3$_{\pm 0.5}$	40.0$_{\pm 2.2}$
DyHead O365	All	Linear	63.0	18.9	33.7	69.2	36.3	70.9	52.4	66.7	26.6	60.6	48.2	16.1	64.6	48.2
DyHead O365	1	Full	25.8$_{\pm 3.0}$	16.5$_{\pm 1.8}$	15.9$_{\pm 2.7}$	55.7$_{\pm 6.0}$	44.0$_{\pm 3.6}$	66.9$_{\pm 3.9}$	54.2$_{\pm 5.7}$	50.7$_{\pm 7.7}$	14.1$_{\pm 3.6}$	33.0$_{\pm 11.0}$	11.0$_{\pm 6.5}$	8.2$_{\pm 4.1}$	43.2$_{\pm 10.0}$	33.8$_{\pm 3.5}$
DyHead O365	3	Full	40.4$_{\pm 1.0}$	20.5$_{\pm 4.0}$	26.5$_{\pm 1.3}$	57.9$_{\pm 2.0}$	53.9$_{\pm 2.5}$	76.5$_{\pm 2.3}$	62.6$_{\pm 13.3}$	52.5$_{\pm 5.0}$	22.4$_{\pm 1.7}$	47.4$_{\pm 2.0}$	30.1$_{\pm 6.9}$	19.7$_{\pm 1.5}$	57.0$_{\pm 2.3}$	43.6$_{\pm 1.0}$
DyHead O365	5	Full	43.5$_{\pm 1.0}$	25.3$_{\pm 1.8}$	35.8$_{\pm 0.5}$	63.0$_{\pm 1.0}$	56.2$_{\pm 3.9}$	76.8$_{\pm 5.9}$	62.5$_{\pm 8.7}$	46.6$_{\pm 3.1}$	28.8$_{\pm 2.2}$	51.2$_{\pm 2.2}$	38.7$_{\pm 4.1}$	21.0$_{\pm 1.4}$	53.4$_{\pm 5.2}$	46.4$_{\pm 1.1}$
DyHead O365	10	Full	46.6$_{\pm 0.3}$	29.0$_{\pm 2.8}$	41.7$_{\pm 1.0}$	65.2$_{\pm 2.5}$	62.5$_{\pm 0.8}$	85.4$_{\pm 2.2}$	67.9$_{\pm 4.5}$	47.9$_{\pm 2.2}$	28.6$_{\pm 5.0}$	53.8$_{\pm 1.0}$	39.2$_{\pm 4.9}$	27.9$_{\pm 2.3}$	64.1$_{\pm 2.6}$	50.8$_{\pm 1.3}$
DyHead O365	All	Full	53.3	28.4	49.5	73.5	77.9	84.0	69.2	56.2	43.6	59.2	68.9	53.7	73.7	60.8
GLIP-T	1	Linear	57.1$_{\pm 0.0}$	15.0$_{\pm 0.3}$	21.2$_{\pm 0.3}$	68.3$_{\pm 1.6}$	59.5$_{\pm 0.1}$	72.7$_{\pm 0.3}$	72.3$_{\pm 0.0}$	65.2$_{\pm 0.2}$	26.5$_{\pm 0.1}$	57.6$_{\pm 0.1}$	54.1$_{\pm 0.4}$	18.2$_{\pm 0.1}$	47.3$_{\pm 0.2}$	48.9$_{\pm 0.1}$
GLIP-T	3	Linear	58.9$_{\pm 0.1}$	15.3$_{\pm 0.1}$	26.0$_{\pm 0.3}$	70.1$_{\pm 0.5}$	61.6$_{\pm 0.4}$	74.7$_{\pm 0.1}$	72.3$_{\pm 0.0}$	64.6$_{\pm 0.2}$	25.9$_{\pm 0.0}$	60.1$_{\pm 0.1}$	51.0$_{\pm 0.2}$	20.9$_{\pm 0.1}$	55.5$_{\pm 0.2}$	50.5$_{\pm 0.1}$
GLIP-T	5	Linear	59.0$_{\pm 0.1}$	15.5$_{\pm 0.4}$	27.6$_{\pm 0.9}$	69.7$_{\pm 0.8}$	61.8$_{\pm 0.1}$	75.1$_{\pm 0.4}$	72.3$_{\pm 0.0}$	62.8$_{\pm 0.5}$	25.4$_{\pm 0.4}$	62.5$_{\pm 0.6}$	51.4$_{\pm 0.3}$	19.6$_{\pm 0.6}$	52.7$_{\pm 1.2}$	50.4$_{\pm 0.2}$
GLIP-T	10	Linear	60.1$_{\pm 0.1}$	14.1$_{\pm 0.1}$	29.6$_{\pm 0.8}$	69.5$_{\pm 0.3}$	62.4$_{\pm 0.2}$	76.8$_{\pm 0.1}$	72.3$_{\pm 0.0}$	61.1$_{\pm 0.3}$	25.8$_{\pm 0.2}$	63.4$_{\pm 0.6}$	51.0$_{\pm 0.1}$	23.3$_{\pm 0.3}$	55.8$_{\pm 1.3}$	51.2$_{\pm 0.1}$
GLIP-T	All	Linear	65.5	14.1	36.5	68.2	67.2	76.6	70.2	63.8	29.1	65.5	63.5	29.9	66.5	55.1
GLIP-T	1	Prompt	54.4$_{\pm 0.9}$	15.2$_{\pm 1.4}$	32.5$_{\pm 1.0}$	68.0$_{\pm 3.2}$	60.0$_{\pm 0.7}$	75.8$_{\pm 1.2}$	72.3$_{\pm 0.0}$	54.5$_{\pm 3.9}$	24.1$_{\pm 3.0}$	59.2$_{\pm 0.9}$	57.4$_{\pm 0.6}$	18.9$_{\pm 1.8}$	56.9$_{\pm 2.7}$	49.9$_{\pm 0.6}$
GLIP-T	3	Prompt	56.8$_{\pm 0.8}$	18.9$_{\pm 3.6}$	37.6$_{\pm 1.6}$	72.4$_{\pm 0.5}$	62.8$_{\pm 1.3}$	85.4$_{\pm 2.8}$	64.5$_{\pm 4.6}$	69.1$_{\pm 1.8}$	22.0$_{\pm 0.9}$	62.7$_{\pm 1.1}$	56.1$_{\pm 0.6}$	25.9$_{\pm 0.7}$	63.8$_{\pm 4.8}$	53.7$_{\pm 1.3}$
GLIP-T	5	Prompt	58.5$_{\pm 0.5}$	18.2$_{\pm 0.1}$	41.0$_{\pm 1.2}$	71.8$_{\pm 2.4}$	65.7$_{\pm 0.7}$	87.5$_{\pm 2.2}$	72.3$_{\pm 0.0}$	60.6$_{\pm 2.2}$	31.4$_{\pm 4.2}$	61.0$_{\pm 1.8}$	54.4$_{\pm 0.6}$	32.6$_{\pm 1.4}$	66.3$_{\pm 2.8}$	55.5$_{\pm 0.5}$
GLIP-T	10	Prompt	59.7$_{\pm 0.7}$	19.8$_{\pm 1.6}$	44.8$_{\pm 0.9}$	72.1$_{\pm 2.0}$	65.9$_{\pm 0.6}$	87.4$_{\pm 1.1}$	72.3$_{\pm 0.0}$	57.5$_{\pm 1.2}$	30.0$_{\pm 1.4}$	62.1$_{\pm 1.4}$	57.8$_{\pm 0.9}$	33.5$_{\pm 0.1}$	73.1$_{\pm 1.4}$	56.6$_{\pm 0.2}$
GLIP-T	All	Prompt	66.4	27.6	50.9	70.6	73.3	88.1	67.7	64.0	40.3	65.4	68.3	50.7	78.5	62.4
GLIP-T	1	Full	54.8$_{\pm 2.0}$	18.4$_{\pm 1.0}$	33.8$_{\pm 1.1}$	70.1$_{\pm 2.9}$	64.2$_{\pm 1.8}$	83.7$_{\pm 3.0}$	70.8$_{\pm 2.1}$	56.2$_{\pm 1.8}$	22.9$_{\pm 0.2}$	56.6$_{\pm 0.5}$	59.9$_{\pm 0.4}$	18.9$_{\pm 1.3}$	54.5$_{\pm 2.7}$	51.1$_{\pm 0.1}$
GLIP-T	3	Full	58.1$_{\pm 0.5}$	22.9$_{\pm 1.3}$	40.8$_{\pm 0.9}$	65.7$_{\pm 1.6}$	66.0$_{\pm 0.2}$	84.7$_{\pm 0.5}$	65.7$_{\pm 2.8}$	62.6$_{\pm 1.4}$	27.2$_{\pm 2.7}$	61.9$_{\pm 1.8}$	60.7$_{\pm 0.2}$	27.1$_{\pm 1.2}$	70.4$_{\pm 2.5}$	54.9$_{\pm 0.2}$
GLIP-T	5	Full	59.5$_{\pm 0.4}$	23.8$_{\pm 0.9}$	43.6$_{\pm 1.4}$	68.7$_{\pm 1.3}$	66.1$_{\pm 0.6}$	85.4$_{\pm 0.4}$	72.3$_{\pm 0.0}$	62.1$_{\pm 2.0}$	27.3$_{\pm 1.2}$	61.0$_{\pm 1.8}$	62.7$_{\pm 1.6}$	34.5$_{\pm 0.5}$	66.6$_{\pm 2.3}$	56.4$_{\pm 0.4}$
GLIP-T	10	Full	59.1$_{\pm 1.3}$	26.3$_{\pm 1.1}$	46.3$_{\pm 1.6}$	67.3$_{\pm 1.5}$	67.1$_{\pm 0.7}$	87.8$_{\pm 0.5}$	72.3$_{\pm 0.0}$	57.7$_{\pm 1.7}$	34.6$_{\pm 1.7}$	65.4$_{\pm 1.4}$	61.6$_{\pm 1.0}$	39.3$_{\pm 1.0}$	74.7$_{\pm 2.3}$	58.4$_{\pm 0.2}$
GLIP-T	All	Full	62.3	31.2	52.5	70.8	78.7	88.1	75.6	61.4	51.4	65.3	71.2	58.7	76.7	64.9
GLIP-L	1	Linear	63.7$_{\pm 0.1}$	7.6$_{\pm 0.3}$	28.1$_{\pm 0.2}$	74.6$_{\pm 0.0}$	60.3$_{\pm 0.0}$	41.3$_{\pm 3.1}$	70.2$_{\pm 1.3}$	67.0$_{\pm 1.0}$	71.0$_{\pm 0.0}$	60.5$_{\pm 0.3}$	67.9$_{\pm 0.1}$	24.8$_{\pm 0.0}$	66.1$_{\pm 0.0}$	54.1$_{\pm 0.3}$
GLIP-L	3	Linear	64.8$_{\pm 0.1}$	8.5$_{\pm 0.1}$	33.7$_{\pm 0.2}$	74.3$_{\pm 0.2}$	64.1$_{\pm 0.2}$	37.0$_{\pm 0.2}$	69.3$_{\pm 0.0}$	66.6$_{\pm 1.9}$	71.2$_{\pm 0.3}$	63.2$_{\pm 0.3}$	68.0$_{\pm 0.1}$	24.8$_{\pm 0.0}$	65.9$_{\pm 0.4}$	54.7$_{\pm 0.2}$
GLIP-L	5	Linear	65.0$_{\pm 0.1}$	8.8$_{\pm 0.1}$	33.4$_{\pm 0.3}$	74.1$_{\pm 0.1}$	63.8$_{\pm 0.0}$	37.2$_{\pm 0.0}$	69.3$_{\pm 0.0}$	69.2$_{\pm 0.6}$	71.5$_{\pm 0.1}$	64.2$_{\pm 0.3}$	68.0$_{\pm 0.1}$	25.3$_{\pm 0.2}$	65.2$_{\pm 0.5}$	55.0$_{\pm 0.0}$
GLIP-L	10	Linear	65.2$_{\pm 0.3}$	11.5$_{\pm 2.3}$	35.1$_{\pm 0.4}$	74.0$_{\pm 0.0}$	64.7$_{\pm 0.0}$	38.0$_{\pm 1.0}$	71.7$_{\pm 1.7}$	66.7$_{\pm 0.3}$	72.5$_{\pm 0.3}$	65.6$_{\pm 1.1}$	67.9$_{\pm 0.0}$	25.8$_{\pm 0.2}$	67.2$_{\pm 0.3}$	55.8$_{\pm 0.4}$
GLIP-L	All	Linear	70.9	9.6	42.3	75.3	70.5	39.4	69.3	71.6	73.9	69.7	72.1	33.2	72.3	59.2
GLIP-L	1	Prompt	62.8$_{\pm 0.4}$	18.0$_{\pm 1.8}$	37.4$_{\pm 0.3}$	71.9$_{\pm 2.4}$	68.9$_{\pm 0.1}$	81.8$_{\pm 3.4}$	65.0$_{\pm 2.8}$	63.9$_{\pm 0.4}$	70.2$_{\pm 1.2}$	67.0$_{\pm 0.4}$	69.3$_{\pm 0.1}$	27.6$_{\pm 0.4}$	69.8$_{\pm 0.6}$	59.5$_{\pm 0.4}$
GLIP-L	3	Prompt	65.0$_{\pm 0.5}$	21.4$_{\pm 1.0}$	43.6$_{\pm 1.1}$	72.9$_{\pm 0.7}$	70.4$_{\pm 0.1}$	91.4$_{\pm 0.7}$	57.7$_{\pm 3.7}$	70.7$_{\pm 1.1}$	69.7$_{\pm 0.9}$	62.6$_{\pm 0.8}$	67.7$_{\pm 0.4}$	36.2$_{\pm 1.1}$	68.8$_{\pm 1.5}$	61.4$_{\pm 0.3}$
GLIP-L	5	Prompt	65.6$_{\pm 0.3}$	19.9$_{\pm 1.6}$	47.7$_{\pm 0.7}$	73.7$_{\pm 0.7}$	70.6$_{\pm 0.3}$	86.8$_{\pm 0.5}$	64.6$_{\pm 0.7}$	69.4$_{\pm 3.3}$	68.0$_{\pm 1.3}$	67.8$_{\pm 1.5}$	68.3$_{\pm 0.3}$	36.6$_{\pm 1.6}$	71.9$_{\pm 0.6}$	62.4$_{\pm 0.5}$
GLIP-L	10	Prompt	65.9$_{\pm 0.2}$	23.4$_{\pm 2.6}$	50.3$_{\pm 0.7}$	73.6$_{\pm 0.7}$	71.8$_{\pm 0.3}$	86.5$_{\pm 0.3}$	70.5$_{\pm 1.1}$	69.0$_{\pm 0.5}$	69.4$_{\pm 2.4}$	70.8$_{\pm 1.2}$	68.8$_{\pm 0.6}$	39.3$_{\pm 0.9}$	74.9$_{\pm 2.1}$	64.2$_{\pm 0.4}$
GLIP-L	All	Prompt	72.9	23.0	51.8	72.0	75.8	88.1	75.2	69.5	73.6	72.1	73.7	53.5	81.4	67.9$_{\pm 0.0}$
GLIP-L	1	Full	64.8$_{\pm 0.6}$	18.7$_{\pm 0.6}$	39.5$_{\pm 1.2}$	70.0$_{\pm 1.5}$	70.5$_{\pm 0.2}$	69.8$_{\pm 18.0}$	70.6$_{\pm 4.0}$	68.4$_{\pm 1.2}$	71.0$_{\pm 1.3}$	65.4$_{\pm 1.1}$	68.1$_{\pm 0.2}$	28.9$_{\pm 2.9}$	72.9$_{\pm 4.7}$	59.9$_{\pm 1.4}$
GLIP-L	3	Full	65.6$_{\pm 0.6}$	22.3$_{\pm 1.1}$	45.2$_{\pm 0.4}$	72.3$_{\pm 1.4}$	70.4$_{\pm 0.4}$	81.6$_{\pm 13.3}$	71.8$_{\pm 0.3}$	65.3$_{\pm 1.6}$	67.6$_{\pm 1.0}$	66.7$_{\pm 0.9}$	68.1$_{\pm 0.3}$	37.0$_{\pm 1.9}$	73.1$_{\pm 3.3}$	62.1$_{\pm 0.7}$
GLIP-L	5	Full	66.6$_{\pm 0.4}$	26.4$_{\pm 2.5}$	49.5$_{\pm 1.1}$	70.7$_{\pm 0.2}$	71.9$_{\pm 0.2}$	88.1$_{\pm 0.0}$	71.1$_{\pm 0.6}$	68.8$_{\pm 1.2}$	68.5$_{\pm 1.7}$	70.0$_{\pm 0.9}$	68.3$_{\pm 0.5}$	39.9$_{\pm 1.4}$	75.2$_{\pm 2.7}$	64.2$_{\pm 0.3}$
GLIP-L	10	Full	66.4$_{\pm 0.7}$	32.0$_{\pm 1.4}$	52.3$_{\pm 1.1}$	70.6$_{\pm 0.7}$	72.4$_{\pm 0.3}$	88.1$_{\pm 0.0}$	67.1$_{\pm 3.6}$	64.7$_{\pm 3.1}$	69.4$_{\pm 1.4}$	71.5$_{\pm 0.8}$	68.4$_{\pm 0.7}$	44.3$_{\pm 0.6}$	76.3$_{\pm 1.1}$	64.9$_{\pm 0.7}$
GLIP-L	All	Full	69.6	32.6	56.6	76.4	79.4	88.1	67.1	69.4	65.8	71.6	75.7	60.3	83.1	68.9

Table 10. Per-dataset performance of DyHead, GLIP-T, and GLIP-L. For PascalVOC, we report the mAP (IoU=0.50:0.95) using the COCO evaluation script, to be consistent with other 12 datasets. "Linear" denotes linear probing. "Prompt" denotes prompt tuning. "Full" denotes full-model tuning.

Model	Shot	Tune	PascalVOC	AerialDrone	Aquarium	Rabbits	EgoHands	Mushrooms	Packages	Raccoon	Shellfish	Vehicles	Pistols	Pothole	Thermal	Avg
GLIP-T (A)	1	Linear	52.9$_{\pm 0.1}$	13.2$_{\pm 0.3}$	21.3$_{\pm 3.2}$	65.0$_{\pm 2.0}$	23.1$_{\pm 0.3}$	11.4$_{\pm 0.1}$	57.3$_{\pm 4.6}$	53.5$_{\pm 0.7}$	16.8$_{\pm 0.0}$	54.1$_{\pm 0.1}$	34.5$_{\pm 0.2}$	5.8$_{\pm 0.1}$	40.8$_{\pm 0.4}$	34.6$_{\pm 0.6}$
GLIP-T (A)	3	Linear	54.6$_{\pm 0.2}$	13.4$_{\pm 0.1}$	28.3$_{\pm 0.1}$	65.4$_{\pm 1.0}$	26.0$_{\pm 0.3}$	11.4$_{\pm 0.0}$	50.8$_{\pm 0.7}$	58.8$_{\pm 0.3}$	15.8$_{\pm 0.7}$	56.1$_{\pm 1.0}$	34.4$_{\pm 0.9}$	6.5$_{\pm 0.0}$	45.8$_{\pm 0.3}$	35.9$_{\pm 0.2}$
GLIP-T (A)	5	Linear	55.3$_{\pm 0.1}$	14.0$_{\pm 0.3}$	28.5$_{\pm 0.1}$	65.2$_{\pm 1.3}$	28.4$_{\pm 0.2}$	11.7$_{\pm 0.0}$	63.9$_{\pm 0.0}$	59.2$_{\pm 0.8}$	16.9$_{\pm 0.2}$	56.6$_{\pm 0.2}$	36.9$_{\pm 0.5}$	9.3$_{\pm 0.0}$	43.2$_{\pm 0.3}$	37.6$_{\pm 0.1}$
GLIP-T (A)	10	Linear	56.8$_{\pm 0.2}$	14.3$_{\pm 0.2}$	29.0$_{\pm 0.1}$	67.0$_{\pm 0.1}$	29.2$_{\pm 0.1}$	11.6$_{\pm 0.1}$	64.5$_{\pm 0.3}$	59.7$_{\pm 0.7}$	16.6$_{\pm 0.7}$	56.9$_{\pm 0.0}$	33.2$_{\pm 1.5}$	7.4$_{\pm 0.1}$	46.2$_{\pm 0.8}$	37.9$_{\pm 0.2}$
GLIP-T (A)	All	Linear	62.0	15.1	32.2	66.1	40.9	12.1	66.9	60.5	22.5	62.4	49.8	17.1	65.7	44.1$_{\pm 0.0}$
GLIP-T (A)	1	Prompt	52.1$_{\pm 0.5}$	11.4$_{\pm 0.2}$	23.7$_{\pm 0.6}$	66.6$_{\pm 0.2}$	21.0$_{\pm 0.2}$	8.6$_{\pm 0.6}$	46.7$_{\pm 0.1}$	53.2$_{\pm 0.2}$	17.1$_{\pm 0.7}$	58.8$_{\pm 0.2}$	37.9$_{\pm 0.3}$	6.0$_{\pm 0.2}$	38.3$_{\pm 0.4}$	34.0$_{\pm 0.1}$
GLIP-T (A)	3	Prompt	54.9$_{\pm 0.1}$	13.4$_{\pm 2.5}$	25.9$_{\pm 0.2}$	65.9$_{\pm 0.5}$	22.7$_{\pm 0.1}$	33.6$_{\pm 1.4}$	46.6$_{\pm 0.0}$	53.7$_{\pm 0.4}$	18.5$_{\pm 0.8}$	58.2$_{\pm 0.6}$	38.1$_{\pm 0.5}$	6.2$_{\pm 0.1}$	42.4$_{\pm 0.2}$	36.9$_{\pm 0.5}$
GLIP-T (A)	5	Prompt	55.6$_{\pm 0.2}$	13.6$_{\pm 0.4}$	26.1$_{\pm 0.4}$	65.7$_{\pm 1.5}$	24.5$_{\pm 0.4}$	56.9$_{\pm 2.6}$	60.5$_{\pm 0.6}$	55.2$_{\pm 0.2}$	19.0$_{\pm 1.5}$	57.0$_{\pm 0.8}$	36.4$_{\pm 1.4}$	6.3$_{\pm 0.1}$	43.2$_{\pm 0.1}$	40.0$_{\pm 0.3}$
GLIP-T (A)	10	Prompt	56.6$_{\pm 0.1}$	15.8$_{\pm 0.8}$	26.2$_{\pm 0.1}$	68.0$_{\pm 0.6}$	24.4$_{\pm 0.1}$	41.2$_{\pm 12.5}$	60.3$_{\pm 0.9}$	55.9$_{\pm 0.4}$	19.6$_{\pm 1.6}$	57.5$_{\pm 1.0}$	36.1$_{\pm 0.3}$	6.0$_{\pm 0.1}$	42.4$_{\pm 1.2}$	39.2$_{\pm 0.9}$
GLIP-T (A)	All	Prompt	58.8	16.4	28.7	69.5	28.8	56.9	60.9	56.3	20.5	60.7	43.3	10.4	51.2	43.3
GLIP-T (A)	1	Full	44.8$_{\pm 0.7}$	16.9$_{\pm 1.2}$	28.0$_{\pm 1.0}$	64.6$_{\pm 1.6}$	54.1$_{\pm 1.5}$	64.1$_{\pm 12.0}$	55.8$_{\pm 0.6}$	55.6$_{\pm 1.8}$	21.6$_{\pm 0.9}$	53.4$_{\pm 1.3}$	43.8$_{\pm 0.9}$	10.9$_{\pm 1.2}$	52.3$_{\pm 4.7}$	43.5$_{\pm 1.2}$
GLIP-T (A)	3	Full	49.5$_{\pm 0.6}$	23.3$_{\pm 1.4}$	36.7$_{\pm 1.2}$	62.5$_{\pm 1.6}$	59.9$_{\pm 1.1}$	84.1$_{\pm 1.3}$	60.2$_{\pm 1.1}$	45.0$_{\pm 2.6}$	26.5$_{\pm 1.9}$	54.4$_{\pm 0.7}$	44.6$_{\pm 3.7}$	23.6$_{\pm 0.7}$	63.5$_{\pm 2.7}$	48.8$_{\pm 0.3}$
GLIP-T (A)	5	Full	50.8$_{\pm 0.5}$	25.3$_{\pm 0.7}$	41.2$_{\pm 0.8}$	62.4$_{\pm 0.9}$	60.4$_{\pm 0.9}$	86.4$_{\pm 2.3}$	59.2$_{\pm 8.5}$	44.7$_{\pm 2.5}$	28.2$_{\pm 0.7}$	55.6$_{\pm 2.0}$	51.7$_{\pm 0.8}$	27.0$_{\pm 0.8}$	62.1$_{\pm 6.0}$	50.4$_{\pm 0.6}$
GLIP-T (A)	10	Full	51.7$_{\pm 0.3}$	29.9$_{\pm 2.4}$	44.3$_{\pm 0.8}$	67.8$_{\pm 2.7}$	64.1$_{\pm 0.3}$	87.9$_{\pm 0.3}$	71.3$_{\pm 2.0}$	47.0$_{\pm 4.2}$	28.8$_{\pm 2.0}$	56.9$_{\pm 0.9}$	52.3$_{\pm 0.4}$	29.1$_{\pm 2.9}$	72.7$_{\pm 2.2}$	54.1$_{\pm 0.4}$
GLIP-T (A)	All	Full	55.1	35.3	50.9	78.0	78.0	86.3	75.2	54.8	44.1	61.4	69.3	57.3	80.6	63.6
GLIP-T (B)	1	Linear	54.0$_{\pm 0.1}$	6.6$_{\pm 0.0}$	17.2$_{\pm 0.0}$	73.3$_{\pm 0.7}$	23.7$_{\pm 0.7}$	63.6$_{\pm 0.2}$	51.5$_{\pm 0.0}$	51.8$_{\pm 0.2}$	25.5$_{\pm 0.1}$	56.4$_{\pm 0.1}$	45.2$_{\pm 1.0}$	6.7$_{\pm 0.1}$	56.5$_{\pm 0.4}$	40.9$_{\pm 0.2}$
GLIP-T (B)	3	Linear	54.9$_{\pm 0.0}$	6.6$_{\pm 0.0}$	25.2$_{\pm 0.2}$	73.1$_{\pm 0.3}$	29.3$_{\pm 0.4}$	63.3$_{\pm 0.1}$	55.3$_{\pm 3.6}$	56.1$_{\pm 0.4}$	24.8$_{\pm 0.4}$	57.5$_{\pm 0.6}$	44.8$_{\pm 0.1}$	6.9$_{\pm 0.2}$	58.5$_{\pm 0.3}$	42.8$_{\pm 0.3}$
GLIP-T (B)	5	Linear	56.0$_{\pm 0.5}$	6.6$_{\pm 0.0}$	25.7$_{\pm 0.3}$	72.9$_{\pm 0.8}$	28.4$_{\pm 0.1}$	62.7$_{\pm 0.2}$	70.5$_{\pm 1.2}$	56.1$_{\pm 0.3}$	25.4$_{\pm 0.5}$	58.6$_{\pm 0.2}$	46.8$_{\pm 0.5}$	9.4$_{\pm 0.9}$	52.8$_{\pm 0.4}$	44.0$_{\pm 0.2}$
GLIP-T (B)	10	Linear	57.3$_{\pm 0.2}$	6.6$_{\pm 0.0}$	27.8$_{\pm 0.9}$	75.8$_{\pm 0.5}$	30.1$_{\pm 0.2}$	62.8$_{\pm 0.4}$	67.8$_{\pm 1.3}$	53.2$_{\pm 0.2}$	24.0$_{\pm 0.1}$	61.5$_{\pm 1.4}$	43.9$_{\pm 0.3}$	7.6$_{\pm 0.1}$	58.4$_{\pm 0.5}$	44.4$_{\pm 0.3}$
GLIP-T (B)	All	Linear	64.3	6.6	35.6	73.9	44.9	62.8	73.6	63.9	34.2	65.0	61.8	20.5	66.6	51.8
GLIP-T (B)	1	Prompt	52.7$_{\pm 0.4}$	16.1$_{\pm 0.8}$	25.2$_{\pm 0.3}$	72.5$_{\pm 0.4}$	56.4$_{\pm 0.5}$	74.5$_{\pm 1.0}$	56.2$_{\pm 4.5}$	56.5$_{\pm 1.3}$	22.3$_{\pm 1.5}$	55.0$_{\pm 0.8}$	53.0$_{\pm 1.3}$	7.1$_{\pm 0.5}$	54.9$_{\pm 0.8}$	46.4$_{\pm 0.4}$
GLIP-T (B)	3	Prompt	54.7$_{\pm 0.9}$	16.6$_{\pm 0.6}$	33.8$_{\pm 0.3}$	76.7$_{\pm 1.0}$	55.9$_{\pm 0.6}$	77.2$_{\pm 4.2}$	59.5$_{\pm 5.6}$	55.7$_{\pm 2.7}$	24.2$_{\pm 1.2}$	56.9$_{\pm 0.7}$	51.3$_{\pm 1.4}$	18.4$_{\pm 0.6}$	56.6$_{\pm 1.7}$	49.0$_{\pm 0.7}$
GLIP-T (B)	5	Prompt	57.4$_{\pm 0.3}$	20.0$_{\pm 1.5}$	35.9$_{\pm 1.3}$	76.0$_{\pm 0.4}$	58.2$_{\pm 0.8}$	78.7$_{\pm 4.2}$	61.4$_{\pm 1.2}$	56.5$_{\pm 1.5}$	27.2$_{\pm 0.8}$	55.0$_{\pm 4.7}$	53.6$_{\pm 1.8}$	21.4$_{\pm 0.3}$	56.4$_{\pm 1.0}$	50.6$_{\pm 0.4}$
GLIP-T (B)	10	Prompt	57.8$_{\pm 0.6}$	22.5$_{\pm 0.7}$	39.1$_{\pm 0.8}$	74.7$_{\pm 1.3}$	58.8$_{\pm 0.8}$	85.6$_{\pm 1.3}$	59.6$_{\pm 0.0}$	56.7$_{\pm 1.5}$	32.4$_{\pm 0.8}$	59.3$_{\pm 1.8}$	52.4$_{\pm 0.5}$	20.7$_{\pm 1.0}$	66.1$_{\pm 1.8}$	52.8$_{\pm 0.1}$
GLIP-T (B)	All	Prompt	64.6	18.2	47.3	71.3	70.1	85.6	59.6	65.0	37.9	61.3	64.6	39.0	76.4	58.5
GLIP-T (B)	1	Full	48.4$_{\pm 1.9}$	16.6$_{\pm 0.6}$	31.8$_{\pm 1.7}$	70.9$_{\pm 1.4}$	55.3$_{\pm 0.4}$	78.8$_{\pm 2.7}$	66.3$_{\pm 1.6}$	48.1$_{\pm 6.9}$	23.3$_{\pm 1.3}$	57.0$_{\pm 0.8}$	52.9$_{\pm 0.6}$	12.9$_{\pm 0.4}$	61.0$_{\pm 1.8}$	48.0$_{\pm 0.7}$
GLIP-T (B)	3	Full	51.7$_{\pm 0.8}$	23.4$_{\pm 2.6}$	37.2$_{\pm 1.0}$	69.5$_{\pm 1.0}$	59.6$_{\pm 0.7}$	85.4$_{\pm 0.4}$	62.4$_{\pm 1.1}$	56.5$_{\pm 1.5}$	30.0$_{\pm 1.0}$	57.6$_{\pm 1.2}$	54.7$_{\pm 1.6}$	24.5$_{\pm 1.3}$	64.3$_{\pm 1.8}$	52.1$_{\pm 0.4}$
GLIP-T (B)	5	Full	52.9$_{\pm 0.7}$	27.4$_{\pm 0.7}$	41.5$_{\pm 0.6}$	68.4$_{\pm 1.4}$	61.9$_{\pm 0.5}$	81.0$_{\pm 3.3}$	69.3$_{\pm 3.5}$	61.2$_{\pm 2.6}$	26.9$_{\pm 1.9}$	58.1$_{\pm 0.3}$	57.4$_{\pm 1.7}$	28.3$_{\pm 1.5}$	57.3$_{\pm 2.4}$	53.2$_{\pm 0.7}$
GLIP-T (B)	10	Full	53.9$_{\pm 1.1}$	28.2$_{\pm 1.3}$	43.1$_{\pm 0.8}$	69.0$_{\pm 2.1}$	65.4$_{\pm 1.4}$	87.3$_{\pm 0.6}$	65.1$_{\pm 2.1}$	52.3$_{\pm 3.2}$	30.6$_{\pm 0.7}$	60.2$_{\pm 2.2}$	53.0$_{\pm 2.5}$	34.2$_{\pm 1.9}$	71.8$_{\pm 2.3}$	54.9$_{\pm 0.6}$
GLIP-T (B)	All	Full	56.9	28.7	54.0	68.3	78.4	88.1	72.7	57.7	41.2	63.8	69.0	59.8	75.8	62.7
GLIP-T (C)	1	Linear	57.0$_{\pm 0.2}$	6.4$_{\pm 0.1}$	21.1$_{\pm 0.4}$	74.2$_{\pm 0.0}$	60.9$_{\pm 0.1}$	24.6$_{\pm 0.1}$	64.0$_{\pm 0.0}$	52.0$_{\pm 0.1}$	21.2$_{\pm 0.1}$	55.6$_{\pm 0.2}$	50.9$_{\pm 0.3}$	14.6$_{\pm 0.0}$	68.7$_{\pm 0.6}$	43.9$_{\pm 0.1}$
GLIP-T (C)	3	Linear	59.0$_{\pm 0.1}$	8.2$_{\pm 0.4}$	28.4$_{\pm 0.2}$	74.2$_{\pm 0.0}$	61.5$_{\pm 0.1}$	24.2$_{\pm 0.3}$	64.0$_{\pm 0.0}$	57.8$_{\pm 0.6}$	20.9$_{\pm 0.1}$	57.1$_{\pm 0.5}$	49.3$_{\pm 0.2}$	15.7$_{\pm 0.1}$	69.5$_{\pm 0.5}$	45.4$_{\pm 0.0}$
GLIP-T (C)	5	Linear	59.6$_{\pm 0.0}$	6.5$_{\pm 0.1}$	29.9$_{\pm 0.6}$	74.1$_{\pm 1.5}$	61.9$_{\pm 0.0}$	24.9$_{\pm 0.1}$	64.9$_{\pm 1.3}$	52.0$_{\pm 0.3}$	21.7$_{\pm 0.4}$	63.4$_{\pm 0.3}$	48.5$_{\pm 1.2}$	22.2$_{\pm 0.3}$	67.6$_{\pm 0.7}$	45.9$_{\pm 0.1}$
GLIP-T (C)	10	Linear	60.8$_{\pm 0.2}$	7.6$_{\pm 0.5}$	31.6$_{\pm 0.1}$	74.3$_{\pm 1.2}$	63.2$_{\pm 0.1}$	25.3$_{\pm 0.2}$	65.8$_{\pm 0.6}$	58.2$_{\pm 2.8}$	22.6$_{\pm 0.3}$	62.6$_{\pm 0.3}$	46.0$_{\pm 0.1}$	20.0$_{\pm 0.4}$	69.4$_{\pm 1.1}$	46.7$_{\pm 0.2}$
GLIP-T (C)	All	Linear	66.4	8.2	38.2	71.0	68.5	37.7	64.0	59.7	32.5	66.1	62.4	32.4	78.2	52.7
GLIP-T (C)	1	Prompt	52.6$_{\pm 1.0}$	13.3$_{\pm 0.8}$	30.8$_{\pm 1.5}$	70.4$_{\pm 0.9}$	60.3$_{\pm 0.4}$	74.5$_{\pm 3.1}$	71.1$_{\pm 1.4}$	58.8$_{\pm 0.2}$	24.8$_{\pm 1.4}$	58.4$_{\pm 1.1}$	51.8$_{\pm 1.5}$	22.8$_{\pm 1.1}$	68.2$_{\pm 0.1}$	50.6$_{\pm 0.4}$
GLIP-T (C)	3	Prompt	57.4$_{\pm 0.2}$	18.9$_{\pm 1.3}$	36.2$_{\pm 1.2}$	74.0$_{\pm 2.6}$	64.0$_{\pm 1.1}$	84.6$_{\pm 0.7}$	64.1$_{\pm 3.7}$	59.2$_{\pm 3.8}$	23.0$_{\pm 2.6}$	61.2$_{\pm 1.4}$	53.1$_{\pm 1.7}$	27.0$_{\pm 1.1}$	65.5$_{\pm 1.4}$	52.9$_{\pm 0.4}$
GLIP-T (C)	5	Prompt	58.8$_{\pm 1.0}$	20.2$_{\pm 0.7}$	41.3$_{\pm 1.3}$	73.2$_{\pm 1.4}$	64.6$_{\pm 1.8}$	82.3$_{\pm 2.7}$	69.1$_{\pm 5.1}$	58.0$_{\pm 2.7}$	27.2$_{\pm 4.0}$	59.2$_{\pm 1.9}$	53.7$_{\pm 0.5}$	26.2$_{\pm 2.3}$	66.5$_{\pm 2.5}$	53.9$_{\pm 0.5}$
GLIP-T (C)	10	Prompt	59.8$_{\pm 0.6}$	21.9$_{\pm 3.1}$	42.8$_{\pm 0.7}$	73.1$_{\pm 0.9}$	66.9$_{\pm 0.5}$	85.7$_{\pm 3.6}$	69.9$_{\pm 2.1}$	58.5$_{\pm 1.8}$	25.7$_{\pm 1.4}$	61.3$_{\pm 1.1}$	54.1$_{\pm 0.4}$	30.1$_{\pm 3.7}$	74.9$_{\pm 0.2}$	55.8$_{\pm 0.9}$
GLIP-T (C)	All	Prompt	67.3	24.8	49.0	72.2	73.2	82.5	72.2	61.1	42.6	64.5	68.8	51.8	80.7	62.4
GLIP-T (C)	1	Full	52.5$_{\pm 0.4}$	16.2$_{\pm 1.2}$	34.5$_{\pm 1.3}$	68.9$_{\pm 1.1}$	64.2$_{\pm 1.2}$	80.9$_{\pm 1.3}$	65.9$_{\pm 3.9}$	51.9$_{\pm 1.2}$	22.3$_{\pm 3.1}$	56.3$_{\pm 1.3}$	55.7$_{\pm 1.2}$	20.8$_{\pm 1.3}$	55.0$_{\pm 4.2}$	49.6$_{\pm 0.2}$
GLIP-T (C)	3	Full	57.1$_{\pm 0.4}$	23.9$_{\pm 0.2}$	39.2$_{\pm 0.1}$	68.2$_{\pm 0.7}$	65.9$_{\pm 0.6}$	85.4$_{\pm 0.3}$	68.3$_{\pm 0.2}$	52.0$_{\pm 2.9}$	30.8$_{\pm 1.8}$	59.0$_{\pm 1.3}$	54.9$_{\pm 1.1}$	29.5$_{\pm 3.3}$	64.8$_{\pm 3.0}$	53.8$_{\pm 0.1}$
GLIP-T (C)	5	Full	57.6$_{\pm 0.7}$	27.6$_{\pm 1.1}$	43.6$_{\pm 0.3}$	67.8$_{\pm 2.0}$	66.4$_{\pm 0.4}$	84.2$_{\pm 0.4}$	67.6$_{\pm 2.6}$	55.4$_{\pm 2.7}$	27.1$_{\pm 5.2}$	60.4$_{\pm 2.7}$	59.8$_{\pm 0.8}$	37.8$_{\pm 1.1}$	57.0$_{\pm 6.3}$	54.8$_{\pm 0.8}$
GLIP-T (C)	10	Full	57.1$_{\pm 0.4}$	31.9$_{\pm 1.3}$	47.9$_{\pm 1.0}$	66.7$_{\pm 4.1}$	67.7$_{\pm 0.4}$	86.1$_{\pm 2.8}$	63.2$_{\pm 3.4}$	52.2$_{\pm 4.3}$	35.5$_{\pm 1.1}$	61.2$_{\pm 0.7}$	58.6$_{\pm 0.9}$	38.9$_{\pm 1.6}$	75.8$_{\pm 3.6}$	57.1$_{\pm 0.9}$
GLIP-T (C)	All	Full	62.3	29.1	53.8	72.7	78.4	85.8	68.6	60.7	43.6	65.9	72.2	55.9	81.1	63.9

Table 11. Per-dataset performance of GLIP-T (A, B, and C). For PascalVOC, we report the mAP (IoU=0.50:0.95) using the COCO evaluation script, to be consistent with other 12 datasets. "Linear" denotes linear probing. "Prompt" denotes prompt tuning. "Full" denotes full-model tuning.

References

[1] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 1

[2] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7373–7382, 2021. 2, 3

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[5] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8420–8429, 2019. 4

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 2

[8] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 8430–8439, 2019. 2

[9] Xin Wang, Thomas E Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. *arXiv preprint arXiv:2003.06957*, 2020. 5, 6