Appendix for DetCLIPv2: Scalable Open-Vocabulary Object Detection Pre-training via Word-Region Alignment

A. Limitations

Our method provides a possible way to achieve open-world detection by scaling up image-text pairs. However, its localization capability still strongly relies on bounding box annotations provided in detection data. To improve the generalization of localization, designing architectures like [8] for robust open-world region proposals is a promising direction for future work. Furthermore, image-text pairs crawled from the Internet are noisy and suffer from severe incomplete descriptions, which undermines the learning efficiency of word-region alignment and requires further designs like [9] for ameliorating data quality. When further scale up image-text pairs to overwhelm detection data, imbalanced training can potentially hurt the performance, which also calls for a future exploration.

B. More Implementation Details

In this section, we provide more implementation details for both pre-training and fine-tuning experiments.

Pre-training details. We pre-train DetCLIPv2 with AdamW [11] optimizer. The leaning rate first warms up linearly to a peak value (2.8e-4/4e-4 for Swin-B/-L based models, respectively) and then decays following a cosine annealing schedule, where the peak lr values are obtained using a square root scaling rule: $lr = base\_lr \times \sqrt{\frac{\text{batchsize}}{16}}$, where $base\_lr$ = 1e-4. We initialize the text encoder with a pre-trained FILIP [14] model and reduce the learning rate of text encoder by a factor of 0.1 to preserve the language knowledge obtained in FILIP's pre-training. To save the GPU memory cost and allow a large batch size for contrastive learning, we adopt automatic mixed-precision [12] and gradient checkpointing [2] for training. Mmdetection [1] repository is used for implementation. Table 1 summarizes the detailed training settings.

Fine-tuning details. We fine-tune DetCLIPv2 on 2 datasets, i.e., LVIS [6] and ODinW13 [10]. For LVIS, we follow most settings of pre-training except that we use a smaller learning rate and the total epochs are set to 24 (i.e., 2x schedule). Table 3 summarizes the detailed setting of fine-tuning LVIS. For ODinW13, since the number of training samples of different datasets varies a lot, we cannot set the same training epoch for all datasets. To avoid tedious hyper-parameter turning and ensure a sufficient training for all datasets, we adopt a long training schedule with early stop mechanism. Specifically, we assign a maximum training epoch with an auto-step learning rate schedule. We monitor the performance and decay the learning rate by 0.1 when the performance reaches a plateau for a tolerance of t_1 epochs. If the learning rate reaches a given minimum value and there is no performance improvement for t_2 epochs, the training exits. We use the same learning rate configuration for all datasets and *do not* search optimal hyper-parameters for each dataset separately.

Config	Value
GPUs (V100)	32(T)/64(L)
training epochs	12
loss weight	$\alpha = 1, \beta = 2, \lambda = 0.1$
optimizer	AdamW [11]
optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.999$
lr for image encoder	2.8e-4(T)/4e-4(L)
lr for text encoder	2.8e-5(T)/4e-5(L)
weight decay	0.05
warmup iters	1000
learning rate schedule	cosine decay
batch size (det/grounding)	128(T)/256(L)
batch size (image text pairs)	6144
input resolution (det/grounding)	1333×800
input resolution (image-text pairs)	320×320
drop path of visual backbone	0.2
max text token length	16
# of concepts M (det)	150
# of concepts M (grounding)	100
label smooth for contrastive loss	0.1
augmentation	multi-scale training, random flip

Table 1. **Detailed pre-training settings** of DetCLIPv2. T/L in parentheses denote Swin-T/L models, respectively. Det/grounding mean detection and grounding data, respectively.

C. More Experimental Results

Effect of input resolution of image-text pairs. Reducing the input resolution of massive image-text pairs can significantly boost the training efficiency while may lead to performance degradation. Table 4 studies the effect of input resolution change of image-text pairs, where we conduct experiments with the Swin-T-based model on O365+CC3M

Config	Value
GPUs (V100)	16
training epochs	24
optimizer	AdamW [11]
optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.999$
lr for image encoder	4e-5
lr for text encoder	4e-6
weight decay	0.05
warmup iters	1000
learning rate schedule	cosine decay
batch size	64
input resolution	1333×800
drop path of visual backbone	0.2
# of concepts M	150
augmentation	multi-scale training, random flip

Table 2. **Detailed fine-tuning settings** for LVIS [6].

Config	Value
GPUs (V100)	8
maximum training epochs	250
optimizer	AdamW [11]
optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.999$
lr for image encoder	4e-5
lr for text encoder	4e-7
weight decay	0.05
warmup iters	500
learning rate schedule	auto-step decay
lr decay tolerance t_1 (epochs)	5
training exit tolerance t_2 (epochs)	8
minimum lr to stop decay	1e-8
batch size	32
input resolution	1333×800
drop path of visual backbone	0.2
augmentation	multi-scale training, random flip

Table 3. **Detailed fine-tuning settings** for ODinW13 [10].

and vary the resolution of CC3M data from 224×224 to 384×384. Increasing the resolution from 256 to 320 leads to an obvious performance improvement (from 30.5 AP to 31.5 AP). However, further increasing it to 384 only brings limited performance gains and introduces considerable memory and training time overhead. Therefore, we choose 320×320 as our final setting.

Incorporating classification dataset. Our framework can be viewed as a more general design for weakly-supervised (WSOD) approaches, which eliminates the limit of predefined categories in traditional WSOD methods. By formulating classification data as a special type of image-text pair data, our method is capable of incorporating it into

Input res	GPU Memory	Training time (GPU hours)	AP
224×224	14.1 GB	697.6	30.5
256×256	16.0 GB	729.6	30.5
320×320	20.7 GB	793.6	31.3
384×384	24.2 GB	876.8	31.5

Table 4. **Input resolution change of image-text pairs.** We use Swin-T-based model trained with O365+CC3M. Zero-shot AP on LVIS minival5k is reported. Using resolution of 320×320 (marked in gray) achieves the best trade-off between computational cost and model performance.

#	Backbone	Pretrain-data	AP (r/c/f)
1	Swin-T	O365	28.6 (24.2/27.1/30.6)
2	Swin-T	O365+IN1k	30.4 (**32.2**/29.4/30.9)
3	Swin-T	O365+CC3M	**31.3** (29.4/**31.7**/**31.3**)
4	Swin-T	O365+GoldG+CC15M	40.4 (36.0/41.7/**40.0**)
5	Swin-T	O365+GoldG+CC15M +IN1k	**40.6** (**38.2**/**42.0**/39.9)
6	Swin-L	O365+GoldG+CC15M	**44.7** (43.1/**46.3**/43.7)
7	Swin-L	O365+GoldG+CC15M +IN1k	**44.7** (**43.8**/45.4/**44.3**)

Table 5. **Effect of incorporating IN21k data for training**. r/c/f indicate rare/common/frequent categories, respectively. Zero-shot AP on LVIS minival5k is reported.

training. Specifically, we use the category name as the caption for each image. To select region proposals, similar to image-text pair, we collect category names in a batch to calculate the similarity with a region and select the maximum value as the objectness score. Considering classification image typically contains only 1 main object, we select top $k = 1$ proposal. Finally, the contrastive loss for image-text pair is replaced with the cross entropy loss for classification, and we also set loss weight $\lambda = 0.1$.

We perform experiments on ImageNet1k [5] (denoted as IN1k) and show the results in Table 5. 2 settings are considered: 1. we train IN1k with only the detection data, i.e., O365; and 2. we incorporate IN1k into the final version of DetCLIPv2, i.e., all data including O365, CC15M, GoldG and IN1k are used. During training, we use the same image-text pair setting for IN1k and replicate IN1k by 3 times to make it have a similar size to CC3M.

First, incorporating classification data when using only detection data can significantly improve the performance from 28.6 to 30.4 (rows 1 and 2), especially for rare categories (from 24.4 to 32.2 AP, +8 AP), yet it is slightly worse than using CC3M in terms of overall AP (rows 2 and 3). However, when all data are used, the advantage of IN1k diminishes, i.e., it brings only 0.2 overall AP for Swin-T-based model (rows 4 and 5) and no performance gain is ob-

Method	Detector (Backbone)	LVIS (zero-shot)		LVIS (fine-tune)	
		AP	AP_r / AP_c / AP_f	AP	AP_r / AP_c / AP_f
Supervised	ATSS [16] (Swin-T)	-	- / - / -	28.4	18.9 / 27.3 / 33.6
GLIP-T [10]	DyHead [3] (Swin-T)	17.2	10.1 / 12.5 / 25.2	-	- / - / -
DetCLIP-T [13]	ATSS [16] (Swin-T)	28.4	25.0 / 27.0 / 31.6	-	- / - / -
DetCLIPv2-T (ours)	ATSS [16] (Swin-T)	**32.8**	**31.0 / 31.7 / 34.8**	**43.7**	**40.2 / 42.7 / 46.3**
Supervised	ATSS [16] (Swin-L)	-	- / - / -	38.3	28.5 / 38.1 / 42.9
GLIP-L [10]	DyHead [3] (Swin-L)	26.9	17.1 / 23.3 / 36.4	-	- / - / -
DetCLIP-L [13]	ATSS [16] (Swin-L)	31.2	27.6 / 29.6 / 34.5	-	- / - / -
DetCLIPv2-L (ours)	ATSS [16] (Swin-L)	**36.6**	**33.3 / 36.2 / 38.5**	**53.1**	**49.0 / 53.2 / 54.9**

Table 6. **Performance on LVIS [7] val split**. *Fixed* AP [4] is reported. DetCLIPv2 achieves SoTA performance.

Model	PascalVOC	AerialDrone	Aquarium	Rabbits	EgoHands	Mushrooms	Packages
GLIP-T	62.3	31.2	52.5	70.8	78.7	88.1	75.6
GLIPv2-T	66.4	30.2	52.5	74.8	80.0	88.1	74.3
DetCLIPv2-T (ours)	67.5	41.8	50.8	80.4	79.8	90.1	73.7
GLIP-L	69.6	32.6	56.6	76.4	79.4	88.1	67.1
GLIPv2-B	71.1	32.6	57.5	73.6	80.0	88.1	74.9
GLIPv2-H	74.4	36.3	58.7	77.1	79.3	88.1	74.3
DetCLIPv2-L (ours)	74.4	44.1	54.7	80.9	79.9	90	74.1

Model	Raccoon	Shellfish	Vehicles	Pistols	Pothole	Thermal	Avg
GLIP-T	61.4	51.4	65.3	71.2	58.7	76.7	64.9
GLIPv2-T	63.7	54.4	63.0	73.0	60.1	83.5	66.5
DetCLIPv2-T (ours)	70.8	54.8	66.5	77.7	54.8	82.2	**68.5**
GLIP-L	69.4	65.8	71.6	75.7	60.3	83.1	68.9
GLIPv2-B	68.2	70.6	71.2	76.5	58.7	79.6	69.4
GLIPv2-H	73.1	70.0	72.2	72.5	58.3	81.4	**70.4**
DetCLIPv2-L (ours)	69.4	61.2	68.1	80.3	57.1	81.1	**70.4**

Table 7. **Detailed fine-tuning AP (%) performance on ODinW13.**

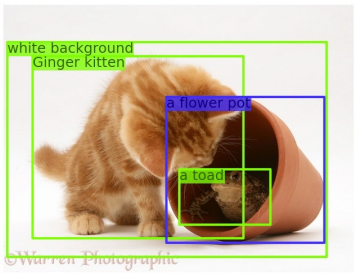
served for Swin-L-based model (rows 6 and 7). Therefore, we exclude the classification data from our method to keep it as neat as possible.

More results on LVIS. To make a comprehensive comparison with the existing methods, we also evaluate DetCLIPv2 with the complete validation set of LVIS [6] (including 20k images with 1203 categories), on both zero-shot and fine-tuning settings. Table 6 exhibits the results. DetCLIPv2 outperforms GLIP and DetCLIP by a large margin for both T/L models, e.g., DetCLIP-T surpasses GLIP-T/DetCLIP-T by 15.6/4.4 AP, respectively. Besides, by pre-training on large-scale hybrid data and fine-tuning on LVIS, DetCLIPv2 achieves significant improvements over the fully-supervised method, i.e., about 15 overall AP improvement can be observed for both T/L models.

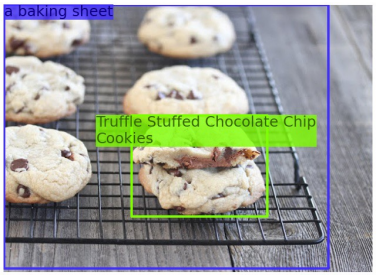
Detailed fine-tuning results for ODinW13. Table 7 reports the detailed fine-tuning performance for 13 datasets contained in ODinW13, and we make a comparison between DetCLIPv2 and GLIP [10]/GLIPv2 [15]. DetCLIPv2-L/T surpass their GLIP [10]/GLIPv2 [15] counterparts on average AP over 13 datasets.

D. More Visualization Results

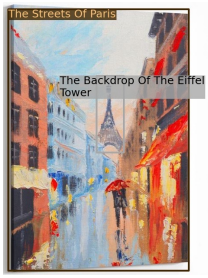
Figure 1-a and 1-b provide more visualization examples of word-region alignment learnt by DetCLIPv2, using the images from CC12M. As mentioned in the main paper, we find the optimal-match region in the image for each textual concept in the caption. As can be seen, DetCLIPv2 learns to recognize and locate various concepts with broad domain coverage, including comic objects (i.e., Monkey King, Santa Claus, etc.), abstract concepts (i.e., 'man's best friend' means a dog) and many concepts that are not covered by the detection/grounding data (i.e., tensioner, tattoos, lifebuoy and etc.), which demonstrates the effectiveness of learning from massive image-text pairs.



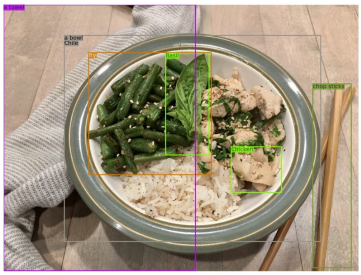
Ginger kitten inspecting a toad in a flower pot, white background



Truffle Stuffed Chocolate Chip Cookies on a baking sheet



Couple Walking On The Streets Of Paris Against The Backdrop Of The Eiffel Tower



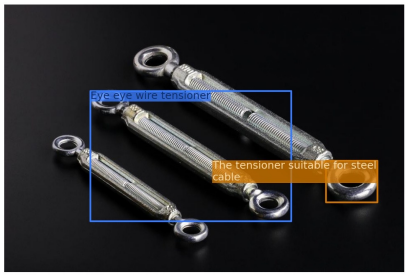
Chile, Basil & Chicken Stir fry served in a bowl with a towel and chop sticks

Playa Esmeralda in Holguin, Cuba. The view from the top of the beach. Beautiful Caribbean sea turquoise.

Little baby get an injection

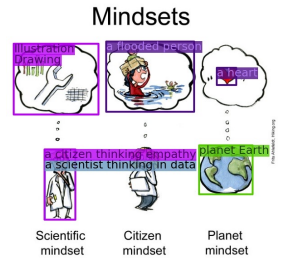
The Vibrant Tattoos

Gangway of the ship Sea. Lifebuoy hanging on a raised gangway of the ship going in the sea stock photography



Eye eye wire tensioner The tensioner suitable for steel cable

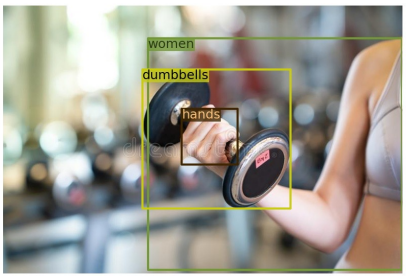
Featuring A Bold And Refined New Look The Completely Redesigned 2018 Chevrolet Traverse Offers Best



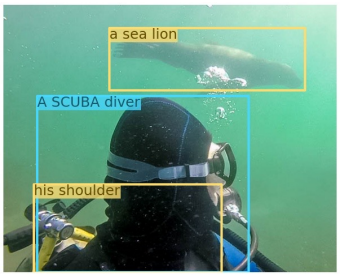
Drawing of a scientist thinking in data, a citizen thinking empathy to a flooded person and planet Earth thinking a heart.

Flat Vector image of the Coati on the Jungle Background Illustration

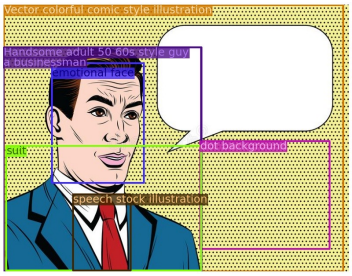
Diamonds Are a Girl's Best Friend: Herkimer Diamond Necklace



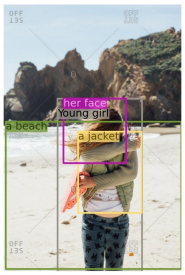
Close up women with dumbbells in hands the gym stock images



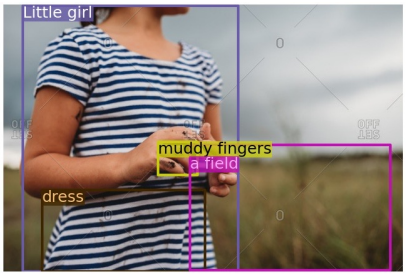
A SCUBA diver looks over his shoulder towards a sea lion



Vector colorful comic style illustration of a businessman with emotional face over dot background. Handsome adult 50 60s style guy in suit speaking with speech stock illustration

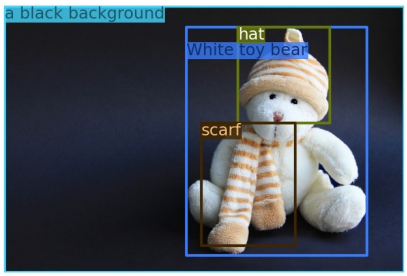


Young girl in a jacket standing on a beach shielding her face from the wind



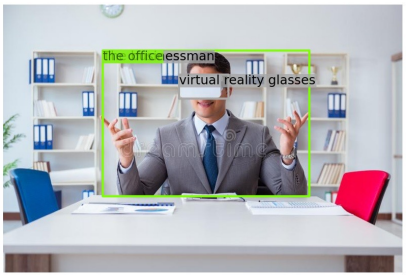
Little girl with muddy fingers and dress in a field

Figure 1-a. **More visualizations for word-region alignment**. DetCLIPv2 learns word-region alignment with broad domain coverage.



White toy bear in **hat** and **scarf** on **a black background**

Ice cream cookie sandwich in **a paper container** that reads ``Cheat Day''



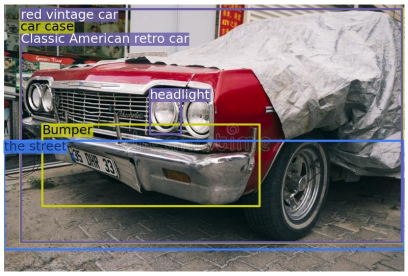
The businessman with virtual reality glasses in the office. Businessman with virtual reality glasses in the office stock photos

Brazilian Tan Stone Horse Sculpture on **an Clear Acrylic Base**, 1980s For Sale

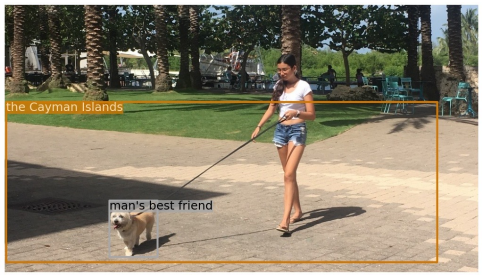
The Mask Loki Pendant Tippy Taste Jewelry

The Art Cards iPhone Case

``I'm ready to party like it's going out of style!'' **Monkey Business Fantasy Inspiration**, **Character Inspiration**, **Character Art**, **Character Design**, Writing **Inspiration**, **Monkey Art**, **Monkey King**, **Cyberpunk Character**, **Cyberpunk Art**



Classic American retro car under **car case** on **the street**. **Bumper** and **headlight** of **red vintage car**. Turkey, Cappadocia stock images



In the Cayman Islands, there are many opportunities to have fun with **man's best friend**.

Golden statue of the god from Hinduism seated stock images

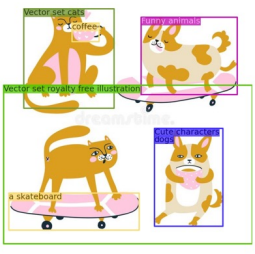
Silhouette of a deer head with **big horns** royalty free illustration

4 pc dream catcher hollow **silver tassel** indian feather all match women necklace **stud earring bracelet** & ring **bohemia style fashion fine**

Surrealistic illustration of a **hatching shaman**. Trying to please **a giant park bird**, detailed **intricate colorful drawing**, outlined royalty free illustration

A rainbow appears at Manning **Park** prior to **the event**.

Christmas illustration with **Santa Claus**. **New year vector illustration**. **Hand** drawn. **Funny Santa Claus** on **skis** and **a skateboard** royalty free illustration



Funny animals drink **coffee** and ride **a skateboard**. **Vector set cats** and **dogs**. **Cute characters**. **Vector set royalty free illustration**

Figure 1-b. **More visualizations for word-region alignment (cont.).** DetCLIPv2 learns word-region alignment with broad domain coverage.

References

[1] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[2] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. preprint arXiv:1604.06174, 2016.

[3] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *CVPR*, pages 7373–7382, 2021.

[4] Achal Dave, Piotr Dollár, Deva Ramanan, Alexander Kirillov, and Ross Girshick. Evaluating large-vocabulary object detectors: The devil is in the details. *arXiv preprint arXiv:2102.01066*, 2021.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[6] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364, 2019.

[7] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, pages 1780–1790, 2021.

[8] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *IEEE Robotics and Automation Letters*, 7(2):5453–5460, 2022.

[9] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

[10] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. *arXiv preprint arXiv:2112.03857*, 2021.

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[12] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. In *ICLR*, 2018.

[13] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. *arXiv preprint arXiv:2209.09407*, 2022.

[14] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2022.

[15] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022.

[16] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *CVPR*, pages 9759–9768, 2020.