Supplementary Materials: Weak-to-Strong Compositional Learning from Generative Models for Language-based Object Detection

Kwanyong Park[1], Kuniaki Saito[2], and Donghyun Kim[3⋆]

[1] ETRI
[2] OMRON SINIC X Corporation
[3] Korea University

Appendices

This supplementary material contains more details including:

A. Additional ablation study and analysis,
B. Limitations of our work,
C. Qualitative comparisons.

A Additional ablation study and analysis

Pseudo box generation strategy. As shown in the main paper, the strategy for generating pseudo-bounding boxes significantly influences the overall performance, with our proposed weak-to-strong methods yielding remarkable enhancements. For a more detailed understanding, we provide more experimental comparisons and analyses in this section.

We first assess the quality of various pseudo bounding boxes. Evaluating the quality of pseudo bounding boxes on a large set of synthetic images is challenging due to the absence of ground truth detection labels. For this reason, we manually annotated 100 randomly selected synthetic images and conducted direct evaluations of various pseudo bounding boxes. Our weak-to-strong method significantly improves the quality of bbox upon the grounding-based baseline, from 53.8AP to 65.0AP, with absolutely high accuracy (*i.e.*, See Qual in Table Ia).

⋆ Corresponding author

strategy	AP	AP-c	AP-d	AP-dS	AP-dL	Qual
Grounding-based	29.3	31.3	27.5	43.4	16.2	53.8
Weak-to-Strong	30.5	31.6	29.5	43.7	21.3	65.0

(a) pseudo box generation strategy

confidence threshold	AP	AP-c	AP-d	AP-dS	AP-dL	Recall
0.3	29.7	31.5	28.2	41.9	18.9	0.99
0.5	30.5	31.6	29.5	43.7	21.3	0.90
0.7	29.6	30.9	28.4	42.8	19.2	0.53

(b) confidence threshold

Table I: Additional ablation on pseudo label generation strategies.

diff. model	AP	AP-c	AP-d	AP-dp	AP-dS	AP-dM	AP-dL
Pixart	30.5	31.6	29.5	40.3	43.7	26.3	21.3
SDXL	30.3	31.2	29.4	39.8	44.7	26.3	20.0
SDXL-Turbo	29.9	31.0	28.9	39.5	43.5	25.7	19.9

(a) Diffusion Model

lang. model	AP	AP-c	AP-d	AP-dp	AP-dS	AP-dM	AP-dL
llama-70b	30.2	31.0	29.3	40.3	44.0	26.2	19.8
GPT3.5-turbo	30.5	31.6	29.5	40.3	43.7	26.3	21.3
GPT4	30.6	31.6	29.7	40.7	44.2	26.5	20.8

(b) Language Model

Table II: Additional analysis on choice of (a) the diffusion model and (b) the language model.

We further examine the impact of the thresholding hyperparameter p, which is used to filter out predictions with low confidence, as described in the main paper. We adjust p within the range of 0.3 to 0.7. As shown in Table Ib, optimal performance is observed at a threshold of 0.5, achieving a high recall rate for visual entities. Here, we treat noun phrases in descriptions as distinct visual entities and quantify their recall rate in the pseudo boxes. A higher parameter results in the exclusion of most predictions, leading to a significantly reduced recall rate. Conversely, setting the lower threshold increases the recall rate but also introduces noisy predictions into the pseudo labels, hindering the effectiveness of the learning process.

Choice of the diffusion model. We explore how the choice of text-to-image model influences the final performance of object detection. In this evaluation, we explore three state-of-the-art text-to-image models: Pixart [3], SDXL [12], and SDXL-Turbo [13]. Using these models, we generate varied sets of images for identical object descriptions, resulting in different collections of densely paired synthetic triplets. These triplets are then utilized to train the FIBER-B model and the experimental results are summarized in Table IIa. Our learning framework reliably enhances performance across the model, though the diffusion models exhibit variable results in terms of the visual quality of generated images and the accuracy of image-text correspondence. This highlights the robustness of our approach regardless of the diffusion model chosen. Pixart is selected as our default setting due to its marginally superior performance and fast inference speed.

Choice of language model. We investigate the impact of selecting different large language models (LLMs) on object detection performance. In this study, we evaluate three LLMs: LLaMA2-70B [16], ChatGPT-3.5 Turbo [2], and ChatGPT-4 [1]. Similar to the above experiments, we generate varied collections of densely paired synthetic triplets and use them to train the detectors. The results are summarized in the table. Although superior language models slightly show improvements, the performance differences among them are marginal. Taking into account both performance and inference efficiency, we choose ChatGPT-3.5 Turbo as a default setting.

Freezing network components. In our main paper, we propose that freezing the visual backbone helps to prevent the model from overfitting to the synthetic distribution during training. To substantiate this claim more convincingly, we conduct a thorough exploration into the effects of freezing different components of the detector. Common language-based object detectors are comprised of three

key components: 1) a visual backbone for understanding the input image, 2) a language backbone for extracting linguistic features, and 3) fusion layers that fuse information from both modalities to detect objects according to the text query. We experiment with freezing each component individually and assess the impact on performance compared to a baseline model that is naively trained on generated triplets and the Objects365 [15] detection dataset.

The results, as presented in Table III, indicate that freezing the visual backbone yields better performance than freezing the other components or not applying any freezing technique at all (*i.e.*, w/o freeze). Moreover, freezing the language backbone shows degraded performance, particularly in description-based object detection. This reveals that the pre-trained image representations may generalize well, whereas the bottlenecks lie in the language component. Furthermore, compositional learning with synthetic triplets may degrade the generality of visual representation. Therefore, the optimal strategy is to teach the model to understand complex language queries while reading out high-quality pre-trained visual representations (*i.e.*, freezing visual backbone) for better compositional understanding.

learning method	AP	AP-c	AP-d	AP-dp	AP-dS	AP-dM	AP-dL
w/o freeze	26.3	30.2	23.3	34.2	41.0	19.7	11.5
Freeze Vis.	26.8	31.3	23.4	34.4	40.8	19.5	11.8
Freeze Lang.	26.1	30.6	22.8	35.9	38.5	19.6	12.1
Freeze Fuse.	26.4	30.1	23.5	34.5	41.2	19.9	11.7

Table III: Additional ablation on freezing network components.

Efficiency of the framework. Our framework brings minimal training costs. The generation of descriptions, images, and bounding boxes takes a total of 7.5 hours (0.5 hr, 6 hr, 1 hr for each) for 58K triplets, and the additional training requires only 3 hours. These costs are efficient, especially compared to the significant data curation cost of 1.3M data and the 72 hours of training time required for GLIP [11]. Our efficient framework supports the extension of data generation processes for novel classes. Most importantly, our framework significantly enhances detector performance for both novel classes (not covered in the data generation) and complex object descriptions, even with a relatively small number of generated triplets.

B Limitations of our work

While our framework significantly enhances the compositional understanding of language-based object detectors, there are several limitations within our proposal that could be interesting points for future research.

Firstly, despite our efforts to mitigate the effects of artifacts in generated triplets—such as freezing the visual backbone, employing real detection data as a regularizer, and training exclusively with detectable objects—implementing more sophisticated filtering techniques to exclude low-quality samples could be beneficial. The criteria for "low-quality" can vary, encompassing aspects like visual quality [9,17] and the accuracy of image-text correspondence [6,8]. Exploring the potential synergy between various filtering methods and our framework could yield insights, similar to previous studies [5].

Fig. I: Qualitative comparisons on OmniLabel [14] benchmark. We can observe clear improvements in compositional understanding against GLIP [11] and Desco-GLIP [10].

Moreover, while we instruct Large Language Models (LLMs) [1,2] to generate plausible descriptions of visual entities under a broad range of conditions, these prompts may not encompass all types of textual expressions. For example, LLMs typically describe objects based on their features but might not employ negations [7] (e.g., "A dog *without* dots"). Although our model demonstrates strong generalization capabilities regarding the concept of negation (See improved Abs scores in Table 1 of the main paper), curating synthetic triplets aimed at such specific cases could further enhance performance.

C Qualitative comparisons

In this section, we present qualitative comparisons against previous methods. The Fig. I compares our model with the pre-trained GLIP [11] and the language-augmentation-based method, Desco-GLIP [10]. Additionally, Fig. II provides qualitative comparisons between our model, FIBER [4], and Desco-FIBER [10]. As illustrated in both figures, our model successfully identifies and locates the target object only, disregarding any confusable objects in the image based on the given descriptions.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv

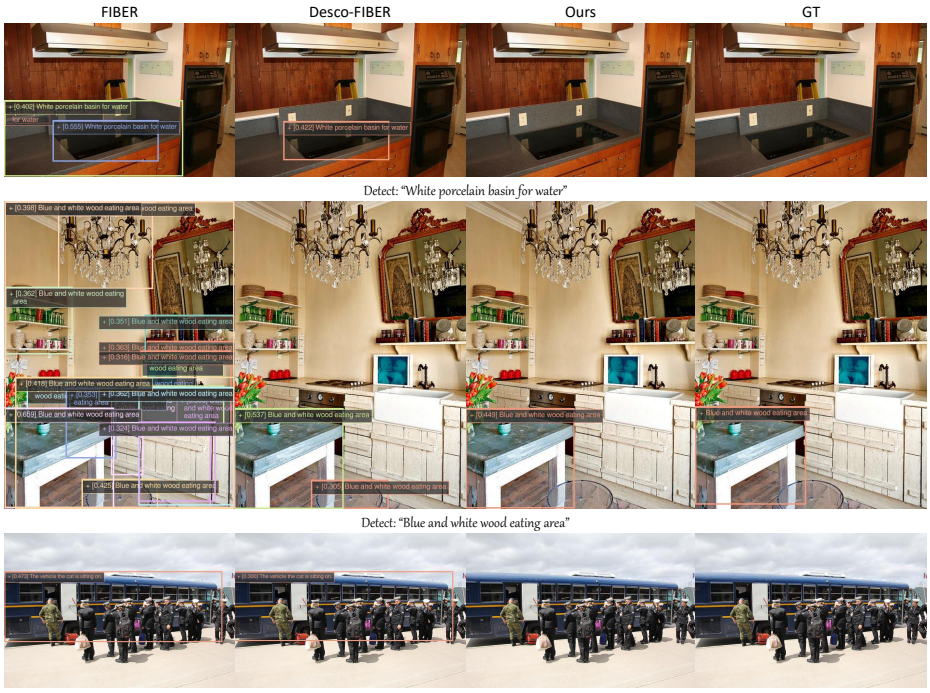

Fig. II: Qualitative comparisons on OmniLabel [14] benchmark. We can observe clear improvements in compositional understanding against FIBER [4] and Desco-FIBER [10].

preprint arXiv:2303.08774 (2023) 2, 4
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 2, 4
3. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al.: Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. arXiv preprint arXiv:2310.00426 (2023) 2
4. Dou, Z.Y., Kamath, A., Gan, Z., Zhang, P., Wang, J., Li, L., Liu, Z., Liu, C., LeCun, Y., Peng, N., et al.: Coarse-to-fine vision-language pre-training with fusion in the backbone. Advances in neural information processing systems **35**, 32942–32956 (2022) 4, 5
5. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al.: Datacomp: In search of the next generation of multimodal datasets. Advances in Neural Information Processing Systems **36** (2024) 3
6. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021) 3

7. Hosseini, A., Reddy, S., Bahdanau, D., Hjelm, R.D., Sordoni, A., Courville, A.: Understanding by understanding not: Modeling negation in language models. arXiv preprint arXiv:2105.03519 (2021) 4
8. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems **36** (2024) 3
9. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in Neural Information Processing Systems **36** (2024) 3
10. Li, L., Dou, Z.Y., Peng, N., Chang, K.W.: Desco: Learning object recognition with rich language descriptions. Advances in Neural Information Processing Systems **36** (2024) 4, 5
11. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022) 3, 4
12. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 2
13. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. arXiv preprint arXiv:2311.17042 (2023) 2
14. Schulter, S., Suh, Y., Dafnis, K.M., Zhang, Z., Zhao, S., Metaxas, D., et al.: Omnilabel: A challenging benchmark for language-based object detection (2023) 4, 5
15. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8430–8439 (2019) 3
16. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 2
17. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023) 3