CLIFF: Continual Latent Diffusion for Open-Vocabulary Object Detection
Supplementary Material

Wuyang Li[1], Xinyu Liu[1], Jiayi Ma[2], and Yixuan Yuan[1](✉)

[1] The Chinese University of Hong Kong
[2] Wuhan University
wymanbest@outlook.com, yxyuan@ee.cuhk.edu.hk

This appendix is organized as follows.

- Appendix A justifies the combinatorial challenge and gives proof of the implemented KL divergence.
- Appendix B presents the architectural design of the proposed diffusion module.
- Appendix C reports more quantitative results.
- Appendix D discusses technical highlights and limitations.
- Appendix E illustrates more qualitative comparisons.

A Experimental and Theoretical Justification

A.1 Justifying the Combinatorial Disagreement

During the training stage, existing OVD works [5,23] conduct the object-image and object-text alignment in a combinatorial manner, which ignores the disagreement between the image and text objectives. As illustrated in Fig. 1 Left, they reduce the L_1 distance between object embedding $\mathbf{x}_{\text{obj}}$ and CLIP image embedding $\hat{\mathbf{x}}_{\text{img}}$, and maximize the cosine similarity between object embedding $\mathbf{x}_{\text{obj}}$ and associated CLIP text embedding $\hat{\mathbf{x}}_{\text{txt}}$, *a.k.a.*, minimizing the angle between $\mathbf{x}_{\text{obj}}$ and $\hat{\mathbf{x}}_{\text{txt}}$. We can observe that there will be a significant disagreement between the object-text and object-image objective if the angle θ between $\hat{\mathbf{x}}_{\text{img}}$ and $\hat{\mathbf{x}}_{\text{txt}}$ is not equal to zero, indicating the non-ignorable inconsistency between $\hat{\mathbf{x}}_{\text{txt}}$ and $\hat{\mathbf{x}}_{\text{img}}$. Further, to study this phenomenon thoroughly, an experiment is conducted to justify the inconsistency.

We first randomly sample $N = 10,000$ region proposals, and then extract corresponding CLIP image embedding $\hat{\mathbf{x}}_{\text{img}} \in \mathbb{R}^{N\times D}$ ($D = 512$ [21]) by sending them to the CLIP image encoder, and generate the CLIP text embedding ($\hat{\mathbf{x}}_{\text{txt}} \in \mathbb{R}^{N\times D}$) by sending the associated class labels into text encoder. After that, we conduct statistic analysis in terms of the cosine similarity ($\in$ [-1,1]) between the image and the associated text embedding $cos(\hat{\mathbf{x}}_{\text{img}}, \hat{\mathbf{x}}_{\text{txt}})$, and plot the distribution in Fig. 1 Right. It can be observed that the average cosine similarity between $\hat{\mathbf{x}}_{\text{img}}$ and $\hat{\mathbf{x}}_{\text{txt}}$ is only 0.26 ($\theta \approx 75°$), indicating a significant inconsistency between the image and text embedding. Hence, combinatorically optimizing the object-image and object-text alignment will conflict with sub-optimal results.

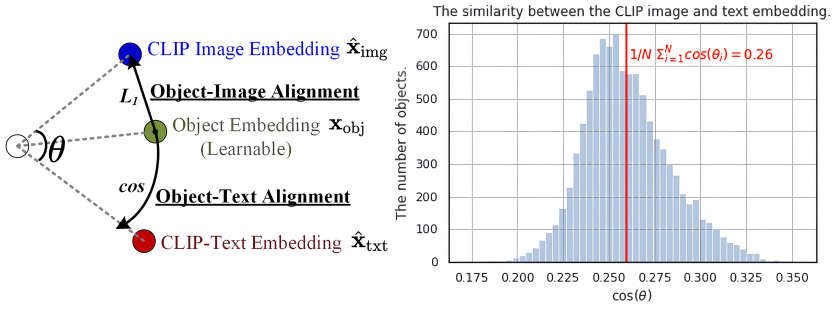

Fig. 1: Left: Illustration of the optimization disagreement between the object-image and object-text alignment in existing combinatorial OVD works [5, 23]. This is caused by the inconsistent embedding of $\hat{\mathbf{x}}_{\text{img}}$ and $\hat{\mathbf{x}}_{\text{txt}}$. Right: The distribution of the cosine similarity between the CLIP image embedding $\hat{\mathbf{x}}_{\text{img}}$ and the associated CLIP text embedding $\hat{\mathbf{x}}_{\text{txt}}$ of region proposals, justifying the inconsistency between $\hat{\mathbf{x}}_{\text{img}}$ and $\hat{\mathbf{x}}_{\text{txt}}$.

This critical observation can also explain why directly using CLIP in other fine-grained tasks [3, 4], *e.g.*, detection [4, 17] and segmentation [3], has unsatisfactory performance due to this inconsistency.

A.2 Proof of the Eq. 7 in the Main Paper

Based on the variational inference [9], we establish a probabilistic object space $p_\phi(\mathbf{x}_T|\mathbf{x}_{\text{obj}}) = \mathcal{N}(\mathbf{x}_T; \mu, \sigma^2 I)$ with multivariate Gaussian assumption and encourage it to satisfy the normally distributed constraint $\mathcal{N}(0, I)$ for the following diffusion process. The detailed proof of the KL divergence [9] is as follows,

$$
\begin{aligned}
& KL\left(p_\phi(\mathbf{x}_T|\mathbf{x}_{\text{obj}}) \| \mathcal{N}(0, I)\right) \\
=& \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \log \frac{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}} \, \mathrm{d}x \\
=& \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \log \frac{1}{\sqrt{\sigma^2}} \times e^{\frac{x^2}{2} - \frac{(x-\mu)^2}{2\sigma^2}} \, \mathrm{d}x \\
=& \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left[-\frac{1}{2}\log\sigma^2 + \frac{1}{2}x^2 - \frac{1}{2}\frac{(x-\mu)^2}{\sigma^2} \right] \mathrm{d}x \\
=& \frac{1}{2} \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left[-\log\sigma^2 + x^2 - \frac{(x-\mu)^2}{\sigma^2} \right] \mathrm{d}x \\
=& \frac{1}{2} \left(-\log\sigma^2 + \mathbb{E}\left[x^2\right] - \frac{1}{\sigma^2}\mathbb{E}\left[(x-u)^2\right] \right) \\
=& \frac{1}{2} \left(-\log\sigma^2 + \sigma^2 + \mu^2 - 1 \right),
\end{aligned}
$$

where μ and σ are the mean and s.d. of the object embedding space learned via Eq. 6 of the main paper.

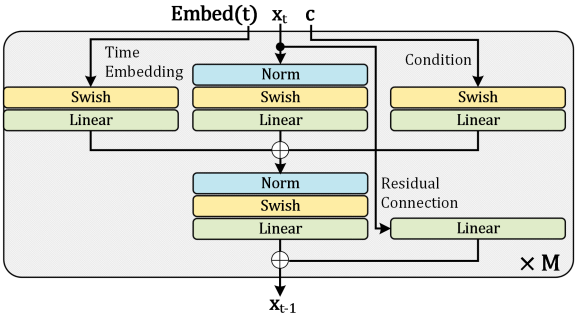

Fig. 2: Illustration of the proposed Continual Diffusion Module (CDM). CDM contains M = 2 basic blocks. Embed(t), $\mathbf{x}_t$, and $\mathbf{c}$ indicate the time embedding, latent embedding corrupted by t-step noise, and the guided condition, respectively.

B The Design of the Diffusion Module

Different from the data-level [8] and latent-space [25] diffusion model diffusing on the RGB image and image features, the proposed Continual Diffusion Module (CDM) conducts latent diffusion among N high-level embedding $\in \mathbb{R}^{N\times D}$ ($D = 512$ [21]) with multi-modalities.

Hence, the continual diffusion module (CDM) is designed in an MLP-based architecture, which contains some essential layers: Layer Norm [1], Swish activation [22], and linear projection. Specifically, CDM contains M = 2 basic blocks, as shown in Fig. 2. Each block follows the philosophical designing idea of the block in UNet-based [10, 26] diffusion model [8] with residual connections and the fusion of the time embedding and condition knowledge. The optimal number of used blocks is justified in Table 3.

C Quantitative Analysis

All experiments about object detection are conducted on the COCO benchmark with a 2× schedule for a fair comparison.

C.1 Extension to Generic Classification

Considering that the proposed object-to-text diffusion process can be used as a generic classification head with a probabilistic nature, we further extend it on the ResNet-18 [7] backbone and compare the classification performance with Bayesian neural networks (BNNs). For the implementation, we only use the Generative objective (the first term) in Eq. 12 of the main paper since the NMS is not needed in classification. For the performance comparison, we borrow the benchmark from the state-of-the-art diffusion-based classification method, CARD [6] (NeurIPS-22), as shown in Table 1. Our CLIFF outperforms all existing BNN-based classifiers

Table 1: Comparison of classification accuracy (%) with other Bayesian Neural Networks (BNNs). The benchmark is from CARD [6]

Model	CMV-MF-VI	CV-MF-VI	MF-VI	MC Dropout	MAP	CARD	CLIFF (ours)
Accuracy	86.25 ± 0.06	79.78 ± 0.30	77.08 ± 1.14	83.64 ± 0.28	84.69 ± 0.35	90.93 ± 0.02	$\mathbf{95.02 \pm 0.02}$

by a large margin and surpasses the latest CARD with 4.09% accuracy, pushing the performance of probabilistic classifiers to a new level. Moreover, our method is trained end-to-end, different from [6] relying on multi-stage training. Hence, our method has great potential to be a new learning paradigm parallel to existing probabilistic calssifier [16].

Table 2: Analysis of the loss weight in the overall optimization objective, including α, β_1, and β_2 controlling the loss term of $\mathcal{L}_{KL}$, $\mathcal{L}_{o2i}$ and $\mathcal{L}_{o2t}$, respectively.

α	0.01	0.1	1.0	2.0	5.0
mAP_n	39.1	40.6	41.2	**41.3**	**41.3**
β_1	1.0	5.0	10.0	15.0	20.0
mAP_n	40.5	40.7	40.9	**41.3**	41.2
β_2	0.1	0.5	1.0	2.0	5.0
mAP_n	38.2	40.1	**41.3**	41.0	39.2

C.2 Analysis on the Loss Weight

As shown in Table 2, we analyze each loss weight in the overall loss function (Eq. 7 in the main paper): $\mathcal{L}_{CLIFF} = \mathcal{L}_{rpn} + \mathcal{L}_{reg} + \alpha\mathcal{L}_{KL} + \beta_1\mathcal{L}_{o2i} + \beta_2\mathcal{L}_{o2t}$. $\mathcal{L}_{rpn}$ and $\mathcal{L}_{reg}$ are consistent with the base detector without reweighing in CLIFF for a fair comparison. The observation and analysis are shown below in three aspects. First, for the α controlling the probabilistic space, we observe a significant performance drop when the intensity is insufficient, *e.g.*, 39.1% with $\alpha = 0.01$, and notice a relatively stable performance with a large enough value. Hence, the Gaussian constraint is critical in generating effective object-centric noise for the diffusion process, without which the module may suffer from failure with limited performance gains. Second, for the β_1 controlling the object-to-image diffusion, the performance is relatively stable, favoring the more significant intensity, revealing our effective design. Third, for the β_2 controlling the object-to-text diffusion, setting a too large ($\beta_1 = 5.0$ with 40.0%) and small ($\beta_1 = 0.1$ with 38.2%) value both give negative effect. The reason may lie in the disharmony relation with the other optimization objectives in the same role, *e.g.*, the regression loss.

Table 3: Analysis of the number of basic blocks in CDM.

M	1	2	3	4	5
mAP_n	38.2	**41.3**	40.2	39.0	38.2

Table 4: Model efficiency comparison among Para: model parameter (M); Inf: inference time (s/img); TrTime: training time (h); FPS: frame per second; Acc: novel-class performance (mAP_n) with state-of-the-art methods.

Method	Para↓	Inf↓	TrTime↓	FPS↑	Acc↑
OCD [23]	42.9	0.1169	**26.2**	8.5	36.6
BARON [27]	104.8	0.1310	32.7	7.6	41.0
CLIFF (ours)	48.5	0.1289	29.6	7.8	43.2

C.3 Analysis on the Diffusion Module Design

Table 3 illustrates the analysis of the number of the basic blocks (refer to Fig. 2) used in the proposed CDM. We observe that using a single block gives significantly bad performance (38.2% mAP_n), caused by the limited module capacity in learning an optimal diffusion process. Moreover, we find that using too many blocks ($M > 3$) gives consistent performance decline, which is a similar phenomenon in extending the stage in Cascade RCNN [2]. The reason may be the overfitting of the diffusion feature, which memorizes the noise schedule instead of learning the denoising effect.

C.4 Model Efficiency

We further conduct experiments to compare with the two latest works, BARON [27] and OCD [23], on the model parameters (Para), inference speed (Inf), and frame per second (FPS), and show the novel-class performance (Acc), as illustrated in Table 4. All experiments are conducted with the COCO benchmark setting (Faster RCNN C4 backbone) for a fair comparison. Compared with the latest work BARON, CLIFF demonstrates significantly better performance in parameter efficiency (48.5 M over 104.8 M), inference efficiency (7.8 FPS over 7.6 FPS), and also gives better model performance (43.2 mAP_n over 41.0 mAP_n). While BARON relies on a heavy CLIP text encoder for inference, our CLIFF only requires an efficient diffusion head with a few diffusion steps (10+3 steps), resulting in better model efficiency. Compared to OCD, our MLP-based diffusion module achieves comparable inference speed while yielding significant accuracy gains, providing clear evidence of the effectiveness and efficiency of CLIFF.

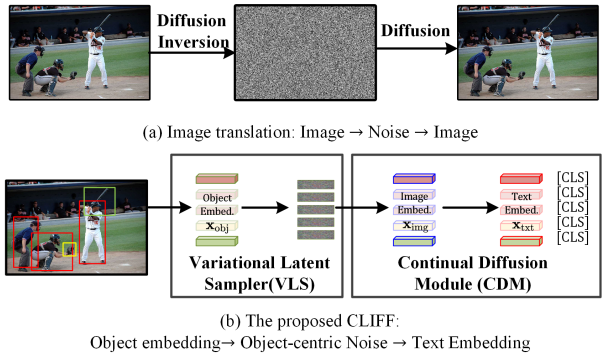

Fig. 3: Comparison of the design with diffusion-based image-to-image translation [19,20].

D Discussion

D.1 Comparison with Diffusion-based Image-to-image Translation

This work models a distribution transfer [11–15] among the object, CLIP image, and text embedding with diffusion. Hence, we compare with the existing diffusion-based image-to-image translation paradigm [19, 20], which also models a distribution transfer among different sub-spaces, as illustrated in Fig. 3. Compared with the image translation shown in Fig. 3 (a), the proposed CLIFF has three technical highlights. First, instead of conducting a diffusion inversion, we propose a simple but effective mechanism with reparameterization (VLS) to obtain object-centric noise, which does not need the heavy diffusion model with good efficiency. Second, unlike the single diffusion process, CLIFF formulates a continual diffusion among three multi-modal sub-spaces in CDM. Third, instead of diffusing in the data space [8] and feature space [25], CLIFF models a latent diffusion in the region-based embedding space, enabling the computation efficiency and effectiveness [25].

D.2 Potential Limitations

The proposed CLIFF framework formulates a novel continual distribution transfer among the object, CLIP image, and CLIP text embeddings via a latent diffusion model. While achieving satisfactory performance, it may be limited in two aspects. First, due to the involved frozen CLIP model, the proposed method is bounded by the quality of the latent embedding provided by the CLIP model while training the diffusion process. Second, we replace the discriminative classification head with a probabilistic diffusion head but preserve the conventional regression head. Due to the feature-sharing property between the classification and regression, we assume that the diffusion feature may not be optimal for the vanilla regression head [24]. Hence, our future work will explore regression-based diffusion to maintain consistent probabilistic feature usage with classification.

E Qualitative Comparison

As shown in Fig. 4, we present more qualitative comparisons among (a) the baseline model [23] trained via extra MAVL proposals [18], (b) the proposed CLIFF, and (c) the ground-truth labels. For a better and clearer view, only the novel-class prediction is shown to evaluate the open-vocabulary capacity. It can be observed that the proposed method gives more accurate novel-class predictions, especially for the occlusion situations, *e.g.,* the cup, snowboard, elephant, and the keyboard in the first row. This verifies the effectiveness and great potential of the proposed new probabilistic paradigm.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. ArXiv:1607.06450 (2016)
2. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: CVPR (2018)
3. Ding, J., Xue, N., Xia, G.S., Dai, D.: Decoupling zero-shot semantic segmentation. In: CVPR (2022)
4. Gao, M., Xing, C., Niebles, J.C., Li, J., Xu, R., Liu, W., Xiong, C.: Open vocabulary object detection with pseudo bounding-box labels. In: ECCV (2022)
5. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: ICLR (2022)
6. Han, X., Zheng, H., Zhou, M.: Card: Classification and regression diffusion models. In: NeurIPS (2022)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
9. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. ArXiv:1312.6114 (2013)
10. Li, C., Liu, X., Li, W., Wang, C., Liu, H., Yuan, Y.: U-kan makes strong backbone for medical image segmentation and generation. ArXiv:2406.02918 (2024)
11. Li, W., Guo, X., Yuan, Y.: Novel scenes & classes: Towards adaptive open-set object detection. In: ICCV. pp. 15780–15790 (2023)
12. Li, W., Liu, X., Yao, X., Yuan, Y.: Scan: Cross domain object detection with semantic conditioned adaptation. In: AAAI. pp. 1421–1428 (2022)
13. Li, W., Liu, X., Yuan, Y.: Scan++: Enhanced semantic conditioned adaptation for domain adaptive object detection. TMM (2022)
14. Li, W., Liu, X., Yuan, Y.: Sigma: Semantic-complete graph matching for domain adaptive object detection. In: CVPR (2022)
15. Li, W., Liu, X., Yuan, Y.: Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. TPAMI (2023)
16. Liang, C., Wang, W., Miao, J., Yang, Y.: Gmmseg: Gaussian mixture based generative semantic segmentation models. ArXiv:2210.02025 (2022)
17. Liu, X., Li, W., Yamaguchi, T., Geng, Z., Tanaka, T., Tsai, D.P., Chen, M.K.: Stereo vision meta-lens-assisted driving vision. ACS Photonics (2024)
18. Maaz, M., Rasheed, H., Khan, S., Khan, F.S., Anwer, R.M., Yang, M.H.: Multi-modal transformers excel at class-agnostic object detection. ArXiv:2111.11430 (2021)

19. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. ArXiv:2211.09794 (2022)
20. Parmar, G., Singh, K.K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. ArXiv:2302.03027 (2023)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
22. Ramachandran, P., Zoph, B., Le, Q.V.: Searching for activation functions. ArXiv:1710.05941 (2017)
23. Rasheed, H.A., Maaz, M., Khattak, M.U., Khan, S., Khan, F.: Bridging the gap between object and image-level representations for open-vocabulary detection. In: NeurIPS (2022)
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: NeurIPS (2015)
25. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
26. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
27. Wu, S., Zhang, W., Jin, S., Liu, W., Loy, C.C.: Aligning bag of regions for open-vocabulary object detection. ArXiv:2302.13996 (2023)

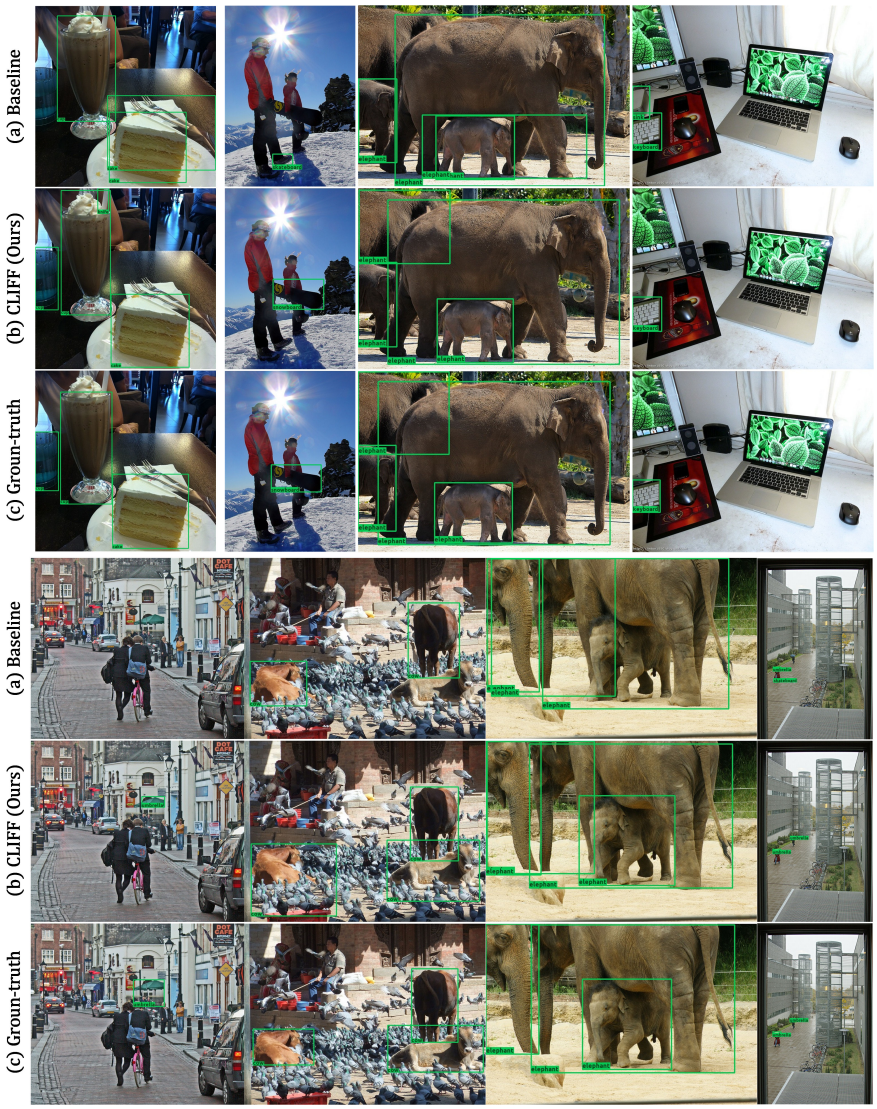

Fig. 4: Qualitative comparison among (a) the baseline [23], (b) our CLIFF, and (c) ground-truth. We only visualize the novel-class predictions for a clear evaluation of open-vocabulary generalization.