A. Qualitative Examples

We present some of the qualitative examples in Figure 1. We observed that our method assigns higher attention weights to both relevant images and text in the k-NN list, indicating that RetFormer can capture effective relationships from two different modalities. We found that even in the absence of any relevant images, our method still benefits from related shared knowledge.

B. Results on CIFAR-100-LT

Table 1 shows the results for CIFAR-100-LT. The results clearly show that RetFormer outperforms other methods including PEL, LiVT, BALLAD, and various ab initio training methods. Our method is the best among all methods that use extra data, demonstrating the potential of retrieval-augmented in vision tasks. The advantage of our method is more obvious when the dataset is unbalanced, which is due to the fact that the retrieval module makes the model focus on the tail classes appropriately. In addition, we do not compare models using ViT pre-trained on the ImageNet-21K dataset due to data leakage that would introduce unfair comparisons.

References

[1] Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. Lpt: long-tailed prompt tuning for image classification. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[2] Fei Du, Peng Yang, Qi Jia, Fengtao Nan, Xiaoting Chen, and Yun Yang. Global and local mixture consistency cumulative learning for long-tailed visual recognitions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15814–15823, 2023. 2

[3] Yuting Li, Yingyi Chen, Xuanlong Yu, Dexiong Chen, and Xi Shen. Sure: Survey recipes for building reliable and robust deep networks. *arXiv preprint arXiv:2403.00543*, 2024. 2

[4] Teli Ma, Shijie Geng, Mengmeng Wang, Jing Shao, Jiasen Lu, Hongsheng Li, Peng Gao, and Yu Qiao. A simple long-tailed recognition baseline via vision-language model. *arXiv preprint arXiv:2111.14745*, 2021. 2

[5] Jiang-Xin Shi, Tong Wei, Zhi Zhou, Jie-Jing Shao, Xin-Yan Han, and Yu-Feng Li. Long-tail learning with foundation model: Heavy fine-tuning hurts. In *Forty-first International Conference on Machine Learning*, 2024. 2

[6] Zhengzhuo Xu, Ruikang Liu, Shuo Yang, Zenghao Chai, and Chun Yuan. Learning imbalanced data with vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15793–15803, 2023. 2

[7] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6908–6917, 2022. 2

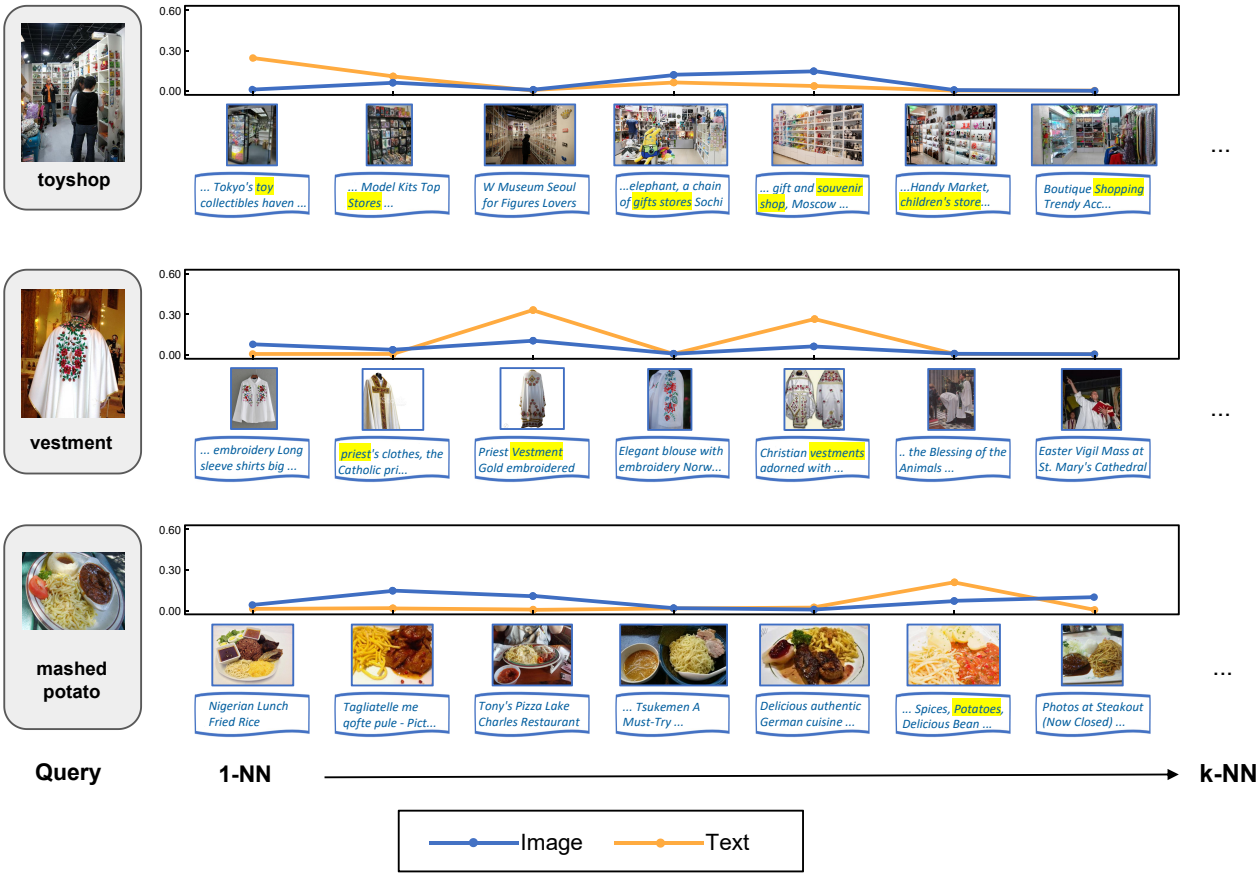

Figure 1. **Qualitative Examples.** We visually demonstrate how the retrieval cross-fusion module copes with tailed classes. Query images of tailed classes are shown on the left. The k-NN images from the knowledge base are shown on the right and sorted from left to right. We display the attention weight assigned to each k-NN above the corresponding image.

Methods	Extra Data	Backbone	Imbalance Ratio 100	50	10
Training from scratch					
BCL [7]	×	ResNeXt-50	51.93	56.59	64.87
GLMC [2]	×	ResNeXt-50	55.88	61.08	70.74
SURE [3]	×	ResNet-32	51.60	58.57	71.13
Fine-tuning pre-trained model					
LiVT [6]	×	ViT-B/16	58.2	82.0	69.2
BALLAD [4]	✓	ViT-B/16	77.8	-	-
LIFT [5]	✓	ViT-B/16	80.3	82.0	83.8
Ours	✓	ViT-B/16	**81.4**	**83.0**	**84.5**
Fine-tuning pre-trained model from ImageNet-21K					
LPT [1]	✓	ViT-B/16	89.1	90.0	91.0
LIFT [5]	✓	ViT-B/16	89.1	90.2	91.3

Table 1. Test top-1 accuracy (%) on CIFAR-100-LT with varying imbalance ratios. RetFormer outperforms prior arts when using a similar backbone network. Pre-trained model from ImageNet-21K has several classes related to CIFAR-100, which potentially leads to data leakage.