6 Supplementary Material

6.1 Visualization

We visualize detection results of LaMI-DETR on LVIS novel categories (Figure 4).

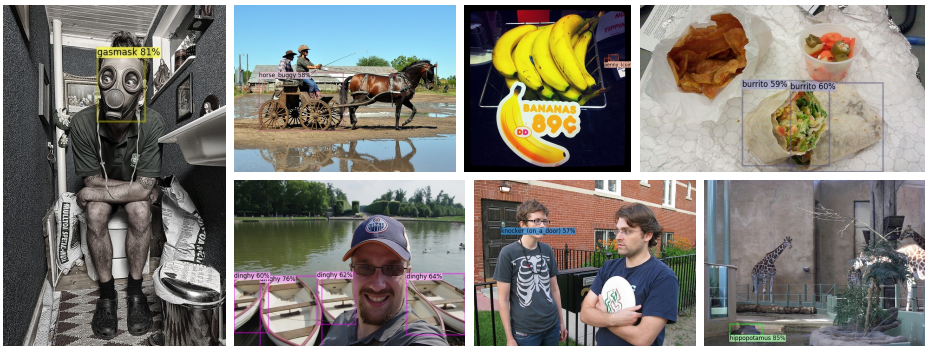

Fig. 4: Visualization of results by LaMI-DETR on OV-LVIS. For better clarity, we only display the prediction results for novel categories.

6.2 Ablation

In the OVD setting, there exist both base and novel categories during inference. The logits for novel classes are usually lower than those for base categories. This issue is commonly alleviated by rescoring novel categories [35]. We multiply the logit of novel classes by a factor of 5.0 during inference. We include results related to the factor in Table 7.

Table 7: Novel classes factor. †: results with factor.

Model	Cluster Encoder	Cluster Text	AP_r	AP
baseline	-	-	33.0	40.6
baseline+VCS	Instructor Embedding	name+visual desc.	34.2	41.7
baseline+VCS†	Instructor Embedding	name+visual desc.	40.1	40.5
baseline+LaMI	Instructor Embedding	name+visual desc.	41.7	41.1
baseline+LaMI†	Instructor Embedding	name+visual desc.	43.4	41.3

6.3 Further Analysis on generalization of LaMI

Figure 5 illustrates the base-to-novel generalization capability of LaMI. Specifically, it employs models trained on the OV-LVIS benchmark to generate proposals. We visualize proposals having an IoU > 0.5 with the nearest ground-truth box for novel categories in the LVIS validation set.

Fig. 5: Visualization of proposals generated by the model with and without LaMI. Sequentially from top to bottom, each row displays the results for the ground-truth, LaMI-DETR, and the baseline, respectively. For detailed examination, please zoom in.

6.4 Confusing Category Details

We provide a detailed description of the Confusing Category module pipeline in LaMI. Based on text embeddings from the CLIP text encoder, we identify visually similar categories for each inference category. Our method then constructs tailored prompts for GPT by incorporating disambiguating context about the confusable categories.

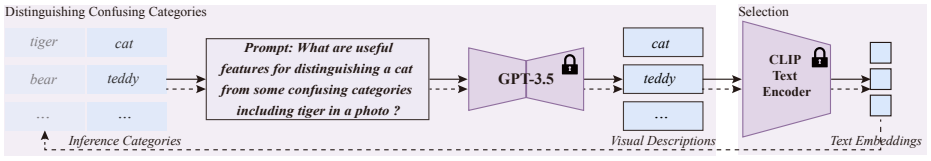

Fig. 6: Illustration of Confusing Category module.

6.5 Inference Time

Table 8: Zero-shot Evaluation on LVIS-minival. The FPS is evaluated on NVIDIA V100 GPU. To highlight our model's efficiency, we compare with methods using lighter backbones like Swin-T.

Method	Backbone	FPS↑
GLIP-T	Swin-T	0.12
GLIPv2-T	Swin-T	0.12
Grounding DINO-T	Swin-T	1.5
DetCLIP-T	Swin-T	2.3
LaMI-DETR	ConvNext-L	**4.5**

During inference, confusing categories are first selected using cosine similarity with sklearn. Next, API calls regenerate descriptions, followed by updating classifier weights. Finally, the model runs at 4.5 FPS. We report FPS reflecting wall-clock time in tab 8.