Seeing Clearly, Reasoning Confidently: Plug-and-Play Remedies for Vision Language Model Blindness

Supplementary Material

1. Related Work

Training-free Adaptation of VLMs: Training-free methods aim to adapt pretrained vision language models (VLMs) without finetuning, instead modifying inference while keeping all backbone weights frozen. Common approaches include score reweighting for zero-shot classification [7], prototype-based similarity retrieval for task-specific predictions [6], and compositional pipelines that leverage VLMs for complex reasoning and retrieval [2]. Recent studies [8, 9] extend this idea by performing "prompting in feature space"—injecting optimized latent prompts at test time to steer attention and decision boundaries without updating VLM parameters. Although yielding notable improvements under strict no-training constraints, these approaches struggle when VLMs are weak for rare objects. Our method follows this paradigm but differs by enhancing rare-object perception and reasoning by jointly refining visual token representations and textual prompts, keeping VLM frozen.

2. Additional baseline comparisons on the CODA-LM and GeoBench-VLM datasets

We report the performance of additional baseline VLMs, LLaVANext-7B and LLaVAOneVision-7B, on the CODA-LM and GeoBench-VLM datasets in Table 1 and Table 2. The results demonstrate that these two baseline models behave differently across the two datasets, while our refinement consistently improves both models on rare object recognition and reasoning. These results further indicate the strong generalization capability of our proposed method across different VLMs.

3. Ablation Study

3.1. Number k of Injected Object Hints.

Figure 1 further analyzes how many detected classes should be injected as hints. Here, we use accuracy rather than GPT score to more clearly demonstrate the detection accuracy of our multi-modal class embeddings (described in Sec 3.5) and the VLM–LLaVA prediction accuracy on referred objects. Here we don't refine visual tokens in LLaVA but with text hints. As k increases from 1 to 9, the detection accuracy and the VLM's trust rate (the ratio of VLM predictions aligned with the injected hints) steadily improve, indicating that our class embeddings yield increasingly reliable candidates and that VLM prefers our hints. However, VLM's prediction accuracy peaks around 1–3 and then gradually decreases when more hints are added. This indicates that, beyond a certain point, additional candidates start to confuse the model rather than help it. We therefore choose $k = 3$ as a good trade-off: it retains the peak VLM accuracy while supplying richer object-level information than a single object.

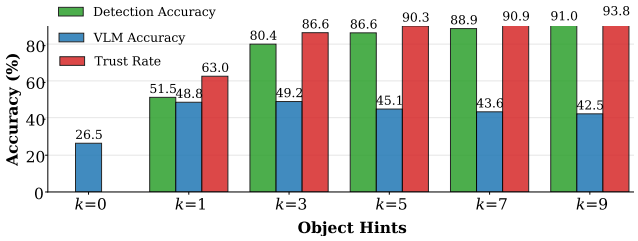

Figure 1. Comparison of different k for LLaVA-7B on CODA-LM. "Detection Accuracy" is the top-k detection accuracy of multi-modal class embeddings for objects. "VLM Accuracy" measures how VLMs recognize objects with/without our hints. "Trust Rate" is the ratio of VLMs' output that aligns with our hints.

3.2. Effectiveness of different components

To investigate which components contribute to the learned class embeddings, we experiment with different training strategies for the class embeddings, and the results are shown in Figure 2. We use detection accuracy as the metric to evaluate the fine-grained representations of the learned class embeddings.

First, we observe that DINOv3 performs best among the evaluated VFMs, mainly due to its strong spatial coherence. A detailed feature comparison is provided in Figure 3. The visual features from DINOv3 are more concentrated and coherent within object regions, indicating that DINOv3 is more sensitive to similar object concepts than CLIP.

We also study the effect of augmented text. In the "DINOv3 + Class" setting, we use only the class label as textual supervision and do not use any additional synonym variations described in Sec. 3.3.1 of the main paper. The results show that sufficient text augmentation is important for training class embeddings. Interestingly, "DINOv3 (w/o Text)" performs better than "DINOv3 + Class" but still lags behind our full setting with augmented texts. This suggests that using only VFM supervision can already produce reasonable representations for rare object classes, but incorporating augmented text makes the resulting class embeddings more discriminative than those trained with VFM supervision alone.

Table 1. Comparison on the CODA-LM dataset. "+ Ours" denotes our parameter-efficient refinement applied to frozen baseline VLMs.

Model / Metrics	Barrier↑	Other↑	Cone↑	Light↑	Sign↑	Vehicle↑	VRU↑	All↑
LLaVANext-7B [4]	46.7	56.2	65.8	75.5	65.1	59.9	50.5	57.6
LLaVANext-7B + Ours	67.6	69.1	83.5	78.1	69.7	67.2	58.8	69.6
LLaVAOneVision-7B [3]	64.1	68.2	71.6	67.7	75.1	66.1	56.5	66.2
LLaVAOneVision-7B + Ours	69.8	73.7	89.5	74.2	81.4	75.1	62.9	75.4

Table 2. Comparison on the GeoBench-VLM dataset. "+ Ours" denotes our parameter-efficient refinement applied to frozen baseline VLMs.

Model / Metrics	Aerial↑	Maritime↑	Vehicle↑	Sports↑	Construction↑	All↑
LLaVANext-7B [4]	10.5	23.1	16.5	24.8	12.2	19.8
LLaVANext-7B + Ours	22.8	50.7	18.3	35.3	13.1	35.3
LLaVAOneVision-7B [3]	16.0	43.1	23.5	25.9	11.1	29.8
LLaVAOneVision-7B + Ours	25.3	57.6	27.6	36.1	11.7	39.8

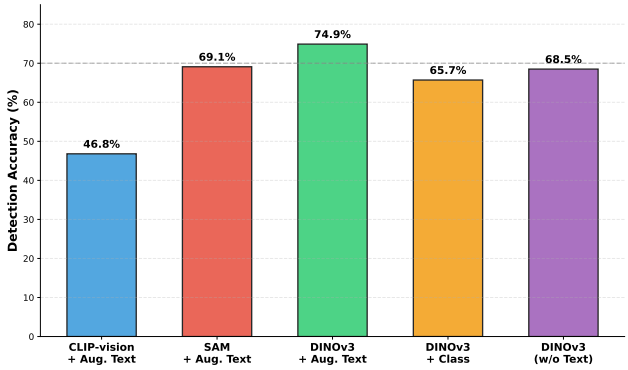

Figure 2. Ablation study of different VFMs and augmented text supervision to train the class embeddings. Among the three compared VFMs, DINOv3 achieves the best performance for constructing class embeddings. Augmented text supervision further improves the class embeddings compared with using only VFM supervision.

Table 3. Ablation study of the reconstruction loss for visual token enhancement for LLaVA-1.5-7B on the CODA-LM dataset.

$\mathcal{L}_{\text{rec}}$	Metrics							
	Barrier	Other	Cone	Light	Sign	Vehicle	VRU	All
✗	54.6	45.7	83.8	21.1	40.8	69.2	40.0	61.8
✓	58.1	59.0	84.7	62.2	50.5	77.7	58.9	70.2

3.3. Effectiveness of reconstruction loss

We introduce the reconstruction loss in Eq. 5 of the main paper to enforce consistency between the refined visual tokens $\hat{\mathbf{V}}$ and the original visual tokens $\mathbf{V}$. To assess its effectiveness, we conduct an ablation study in Table 3. The results show that LLaVA-1.5-7B performs markedly worse without the reconstruction loss, underscoring the importance of maintaining consistency between refined and original visual tokens.

Table 4. Ablation study of applying visual refinement at different decoder layers of LLaVA-1.5-7B on the CODA-LM dataset.

Metrics	0	8	16	24	32
Barrier	58.1	39.2	40.9	40.6	40.3
Other	59.0	43.4	42.1	40.8	42.5
Cone	84.7	54.1	53.5	53.9	54.0
Light	62.2	53.3	60.0	54.4	51.1
Sign	50.5	46.6	47.4	48.5	48.1
Vehicle	77.7	49.8	48.8	48.4	43.2
VRU	58.9	42.6	40.5	40.6	39.7
All	70.2	47.1	46.6	46.3	44.9

3.4. Effectiveness of visual refinements across different layers

As described in the main paper, our default setting applies refinement only at the first decoder layer of LLaVA-1.5-7B. We further apply refinement at different decoder layers, and the detailed results are shown in Table 4. Note that the results in this table are only with the visual refinement modules for our method. We observe that the overall performance drops steadily as the refined layer becomes deeper. We attribute this to the fact that visual tokens in deeper decoder layers no longer follow distributions similar to those of the visual features produced by the vision encoder. Since our class embeddings are primarily derived from VFM visual representations, the domain gap between the class embeddings and the visual tokens grows with depth, making our refinement less effective at deeper layers.

3.5. Comparison of different text hints

To investigate the impact of text hints, we adopt different templates for injecting object hints, as summarized in Table 5. Our original method generates object hints according

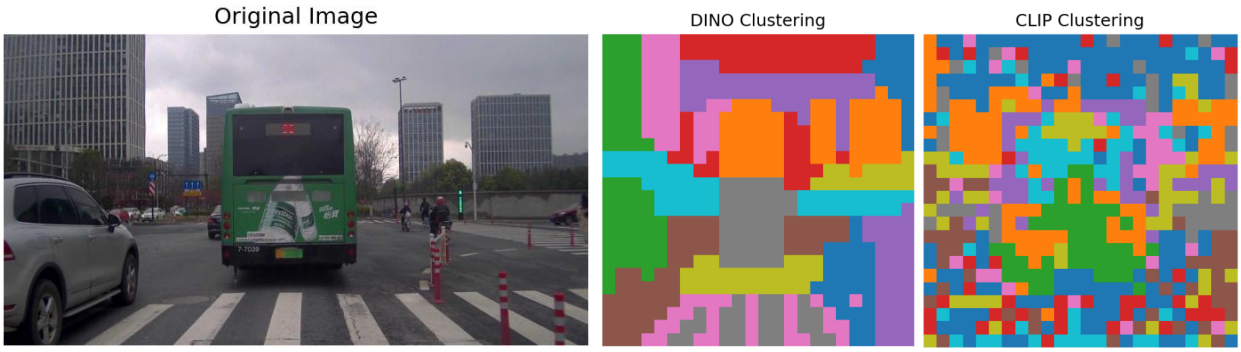

Figure 3. Feature distribution of CLIP and DINOv3. Unlike CLIP, whose feature representations can be more globally biased, DINOv3 demonstrates pronounced spatial coherence across different object regions, effectively capturing consistent local semantics and preserving structural relationships within objects. Note that the same color indicates the same object cluster.

Table 5. Ablation study of different formats for text hints with LLaVA-1.5-7B on the CODA-LM dataset.

Metrics	Original	Reverse	Random	With confidence
Barrier	47.9	50.5	49.5	45.7
Other	49.5	52.3	53.1	45.4
Cone	75.2	77.3	76.1	72.1
Light	61.9	61.1	55.5	52.2
Sign	42.1	36.5	40.2	40.5
Vehicle	58.7	57.5	56.1	60.5
VRU	42.7	40.8	41.1	43.4
All	55.8	56.6	55.9	55.5

to the confidence scores of detected objects and lists them in descending order. The "reverse" variant instead lists them in ascending order. For example, while our original augmented text hint is "It might be one in 'bollard, cyclist, bus"', the "reverse" version becomes "It might be one in 'bus, cyclist, bollard"'. We also evaluate hints presented in a "random" order and hints "with confidence" scores attached. Note that in this study, the test model does not include our visual refinement module. The results show that different templates can significantly change per-category scores but have little impact on overall performance.

3.6. Comparison with detection model

Figure 4 compares Grounding DINO [5] with our class embeddings. Grounding DINO tends to misclassify the "bus" region as "bollard", whereas our method more accurately localizes the position of the "bollard".

4. Qualitative results

We follow the interpretability analysis in the main paper and provide additional examples on both the CODA-LM and GeoBench-VLM datasets. Figures 5 and 6 compare attention weights between the original LLaVA-1.5-7B and LLaVA-1.5-7B with our refinement. The results clearly show that our method encourages the VLM to focus on the relevant object regions. Figures 7 and 8 further illustrate the semantic meanings of visual object tokens on CODA-LM and GeoBench-VLM. In both datasets, our refinement leads to visual tokens with more reasonable and consistent semantics.

5. Failure Cases

Figure 9 shows several failure cases on the CODA-LM and GeoBench-VLM datasets. In these examples, our method fails to guide the model to focus on relevant object regions and does not provide correct object hints because the referenced object is too small to be recognized by our class embeddings, and such small regions can not be well refined from the visual level, and the text hints are also wrong, ultimately leading to incorrect predictions.

6. Class details in CODA-LM and GeoBench-VLM

We report the class distribution for each category in the CODA-LM and GeoBench-VLM datasets in Tables 6 and 7. Among the CODA-LM categories, "VRU" and "Other" have the fewest cases per subclass, and "Barrier" also has fewer cases than "Traffic Light" in public datasets. Therefore, we treat "VRU", "Barrier", and "Other" as rare categories in CODA-LM. For the GeoBench-VLM dataset, "Sports" and "Construction" have the fewest cases per subclass, and we regard these two categories as rarer than the others.

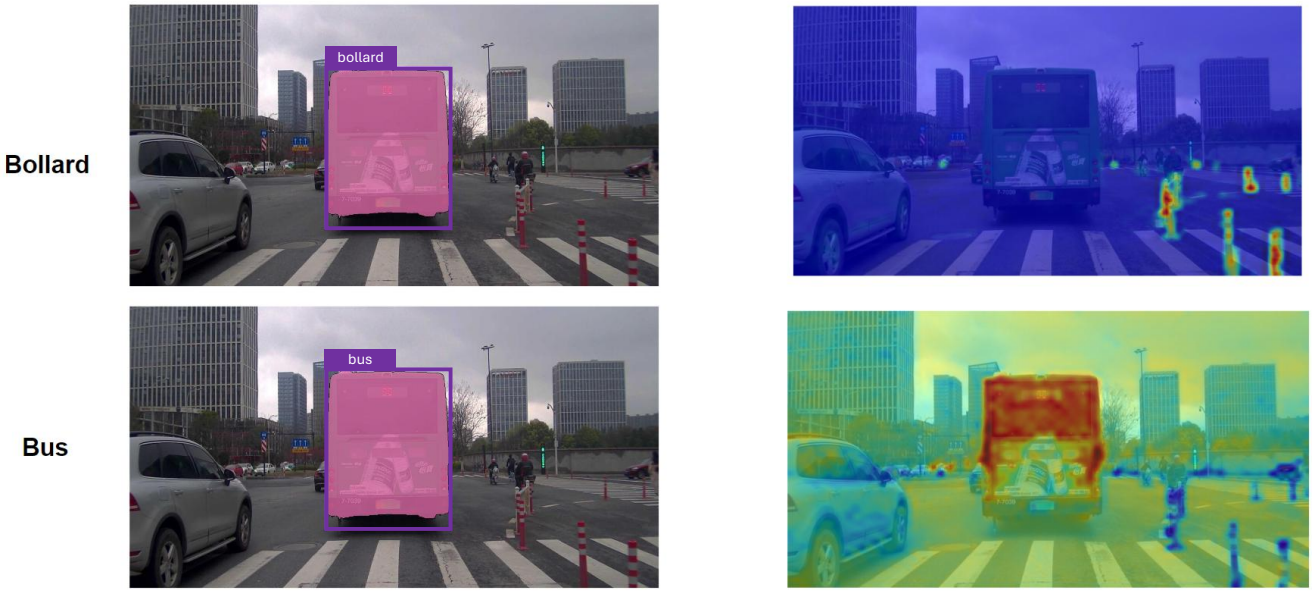

Figure 4. Comparison with detection models. The first column shows detection results from Grounding DINO, and the second column shows the saliency maps from our class embeddings.

Table 6. Subclasses for each category in the CODA-LM dataset.

Category	# Cases	Subclasses
Vehicle	4242	car, truck, tricycle, bus, construction vehicle
VRU	1126	pedestrian, cyclist, bicycle, moped, motorcycle, stroller, cart
Traffic Sign	388	traffic sign
Traffic Light	169	traffic light
Traffic Cone	1976	traffic cone
Barrier	1721	barrier, bollard
Other	1205	dog, sentry box, traffic island, debris, dustbin, concrete block, machinery, garbage, plastic bag, stone, suitcase, misc

VRU: Vulnerable Road Users

Table 7. Subclasses for each category in the GeoBench-VLM dataset.

Category	# Cases	Subclasses
Aerial	67	airplane, plane, airport, helicopter
Maritime	140	ship, harbor
Vehicle	60	vehicle, small-vehicle, large-vehicle
Sports	64	soccer-ball-field, baseball-diamond, basketball-court, tennis-court, ground-track-field
Construction	30	bridge, roundabout, swimming-pool, storage-tank

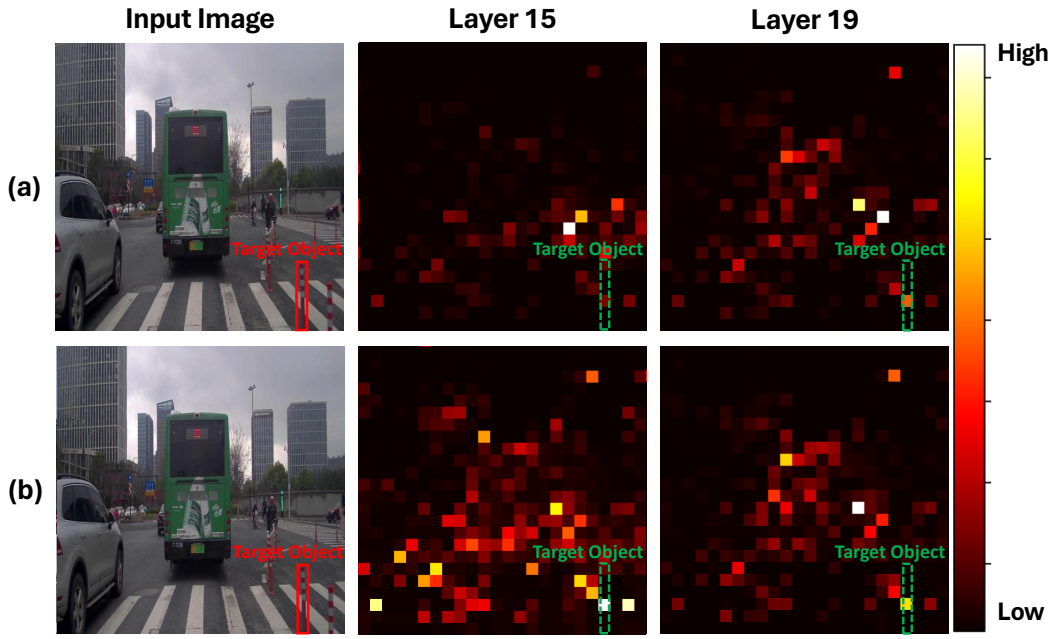

Figure 5. Attention-weight comparison between (a) LLaVA-1.5-7B and (b) LLaVA-1.5-7B + Ours on the CODA-LM dataset.

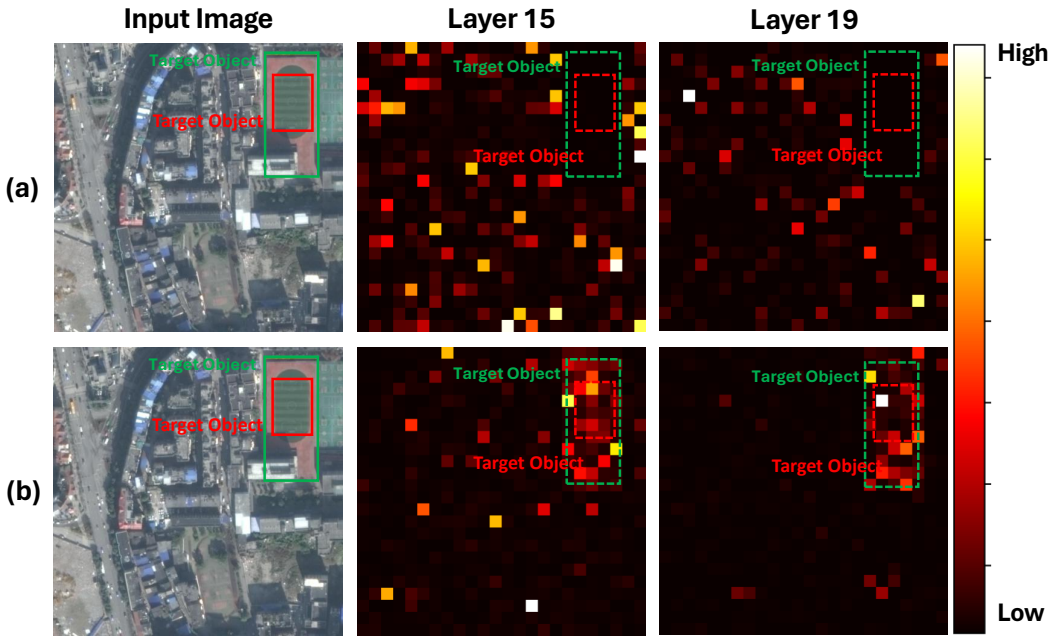

Figure 6. Attention-weight comparison between (a) LLaVA-1.5-7B and (b) LLaVA-1.5-7B + Ours on the GeoBench-VLM dataset.

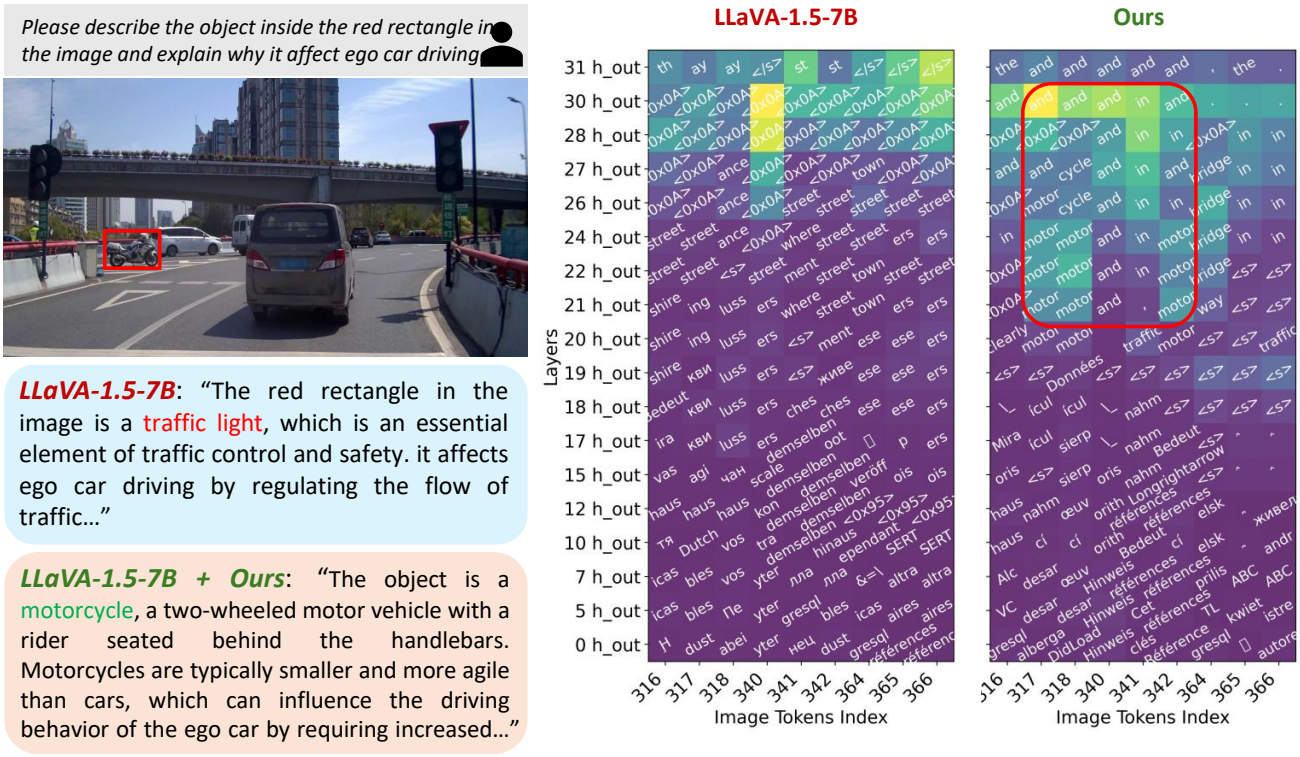

Figure 7. Interpretation of image hidden states for the object class "motorcycle" in LLaVA-1.5-7B via logit lens [1] on the CODA-LM dataset.

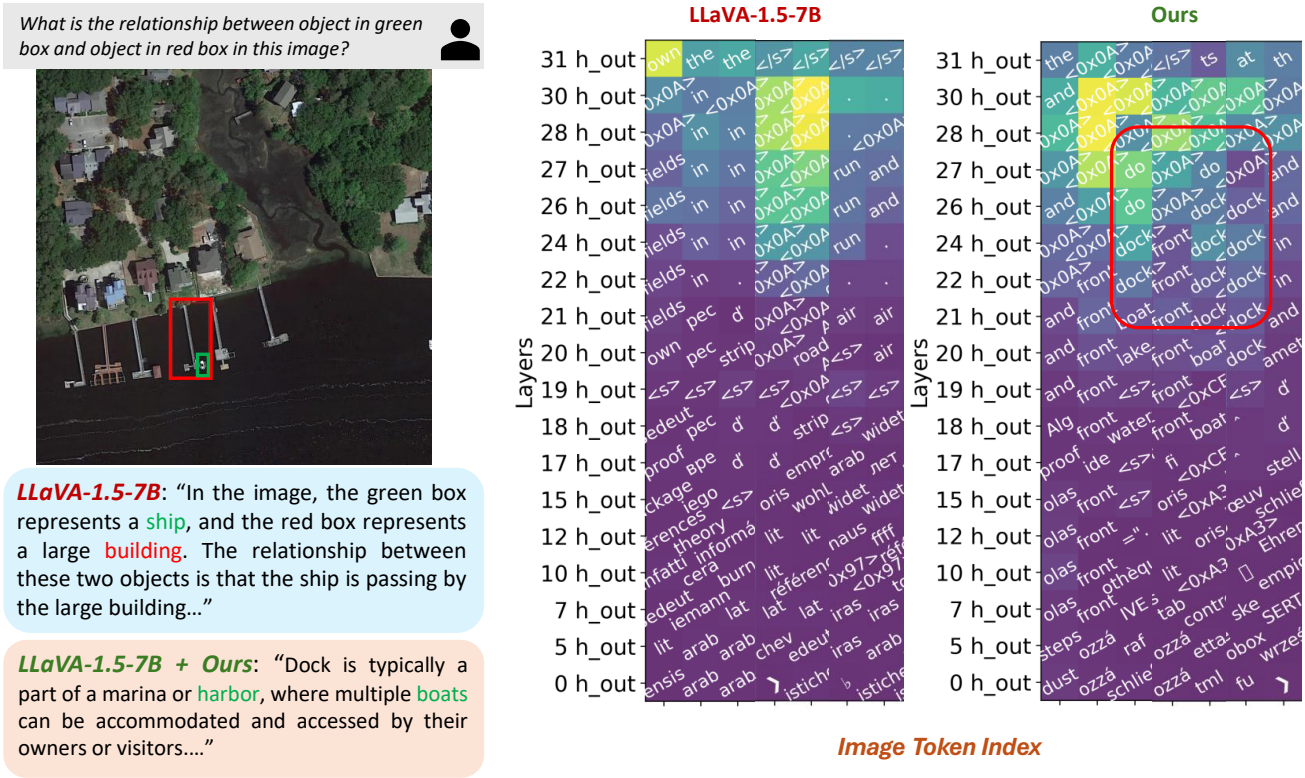

Figure 8. Interpretation of image hidden states for the object class "harbor" in LLaVA-1.5-7B via logit lens [1] on the GeoBench-VLM dataset. It is observed that the semantic meaning of visual tokens in our method demonstrates "dock", a synonym of "harbor", which doesn't affect the final prediction for "harbor". Both methods exhibit semantic meaning for "ship", thus we don't show it here.

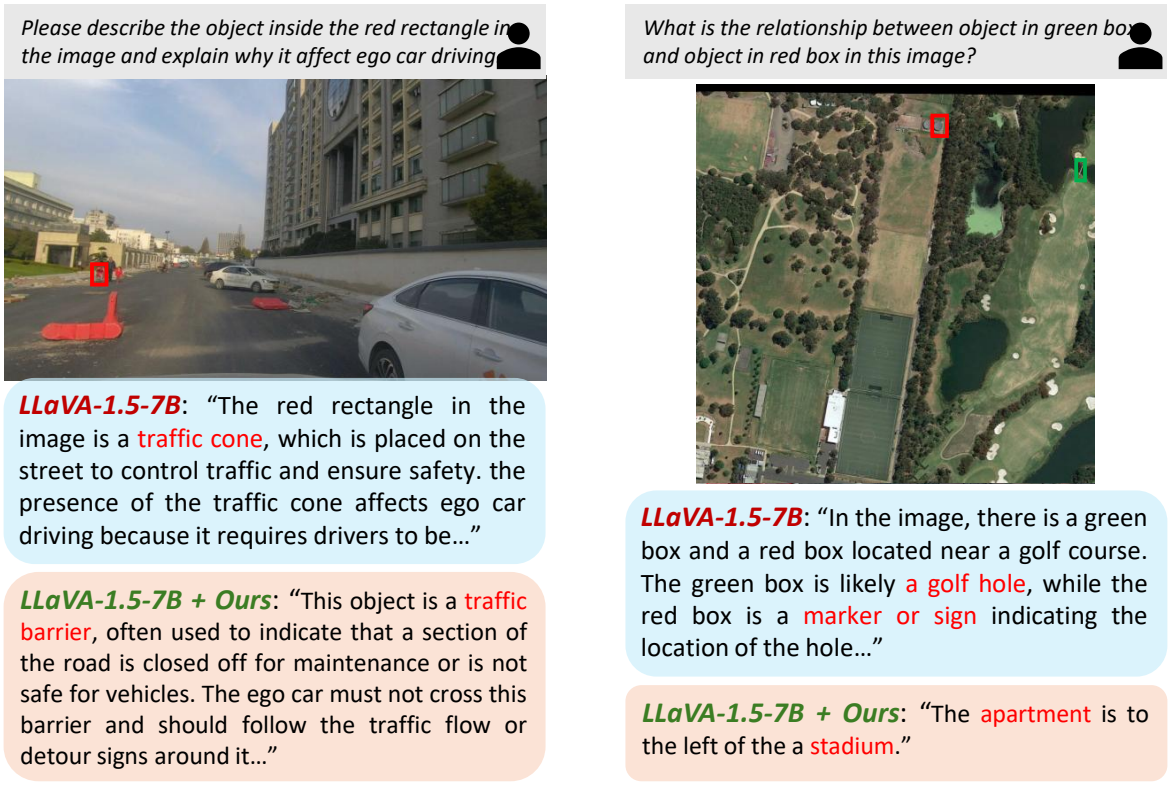

Figure 9. Failure cases on (a) the CODA-LM dataset and (b) the GeoBench-VLM dataset. The provided object hints are "debris, barrier, stone" for (a) and "ground-track-field, soccer-ball-field, harbor" for (b). The referred objects are very small, which brings challenges to both class embeddings and VLMs. Such small regions could be well refined by class embeddings and given wrong text hints, leading to final wrong predictions.

References

[1] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25004–25014, 2025. 6, 7

[2] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. Vision-by-language for training-free compositional image retrieval. *arXiv preprint arXiv:2310.09291*, 2023. 1

[3] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2

[4] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 2

[5] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 3

[6] Sifan Long, Linbin Wang, Zhen Zhao, Zichang Tan, Yiming Wu, Shengsheng Wang, and Jingdong Wang. Training-free unsupervised prompt for vision-language models. *arXiv preprint arXiv:2404.16339*, 2024. 1

[7] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2725–2736, 2023. 1

[8] Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. Controlmllm: Training-free visual prompt learning for multimodal large language models. *Advances in Neural Information Processing Systems*, 37:45206–45234, 2024. 1

[9] Jinjin Zhang, Guodong Wang, Yizhou Jin, and Di Huang. Towards training-free anomaly detection with vision and language foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15204–15213, 2025. 1