Supplementary Material — Benefit From Seen: Enhancing Open-Vocabulary Object Detection by Bridging Visual and Textual Co-Occurrence Knowledge

In this supplementary material to the main paper, we provide experiment details, supplementary experiments (Appendix 1), and LLM prompt details (Appendix 2).

1. Experiment

1.1. Implementation Details

OV-COCO (COCO Benchmark) [3]:

- Architecture: Faster R-CNN with ResNet50-C4 backbone, following OVR-CNN [4].
- Training: No data augmentation.
- Baseline: "Base" method refers to Faster R-CNN trained on COCO's 48 known categories, using CLIP embeddings as the classifier head (aligned with VLDet [2]).

OV-LVIS (LVIS Benchmark) [1]:

- Architecture: CenterNet2 with ResNet50 backbone, following Detic [5].
- Training: Includes large-scale jittering and repeat factor sampling for data augmentation.
- Baseline: "Base" method is fully supervised on LVIS's 866 known categories (following VLDet [2]).

Textual Co-Occurrence Category Generation:

- Iteration: Textual co-occurrence categories are regenerated every 100 training iterations to refine contextual knowledge.
- Parameters: $N = 5$ Generates 5 co-occurring candidates per known object using each strategy (**Q1**-**Q3**). $\tau = 0.6$ Confidence threshold for pseudo-label selection, balancing precision and recall (Section 4.3.2).

1.2. Textual Co-Occurrence Description

Figures 1-4 exemplify CODet's textual co-occurrence generation strategies (Section 4.3.1): spatial proximity (**Q1**), functional correlation (**Q2**), and hierarchical relationship (**Q3**). For each anchored known object (e.g., *motorcycle*), LLMs generate distinct candidates: **Q1** prioritizes spatially adjacent items (*helmet*), **Q2** identifies functional analogs (*bicycle*), and **Q3** groups taxonomically related categories (*airplane*). While humans may conflate spatial and functional relationships, LLMs discern nuanced distinctions, e.g., *bench* → *trash* can (**Q1**) vs. *bench* → *couch* (**Q2**), validating their role in contextual reasoning. Similarly, hierarchical queries resolve ambiguous cases (*mouse* → *smartphone* under "interactive devices"), demonstrating CODet's ability to leverage LLMs for diverse, complementary co-occurrence cues critical for novel category detection.

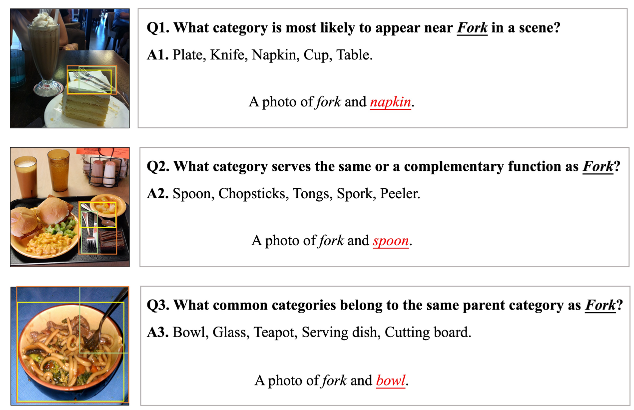

Figure 1. LLM-generated co-occurrence candidates for *Fork* (known, `green`) via spatial proximity (**Q1**: *napkin*), functional correlation (**Q2**: *spoon*), and hierarchical relationships (**Q3**: *bowl*). `Red` text denotes co-occurring categories validated by LLMs, aligning with real-world visual arrangements.

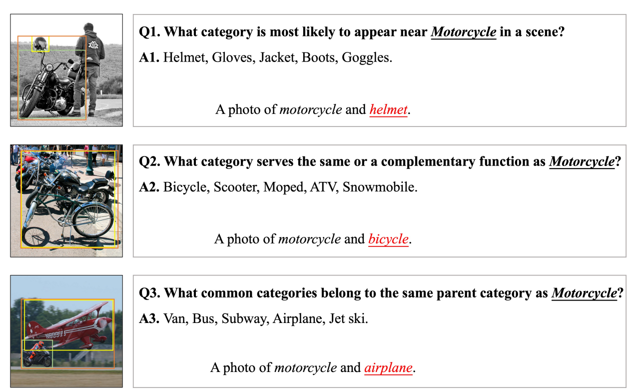

Figure 2. Co-occurrence validation for *Motorcycle* (known, `green`). **Q1** (*helmet*) captures spatial context, **Q2** (*bicycle*) reflects functional similarity, and **Q3** (*airplane*) leverages vehicular taxonomy. `Red` categories highlight LLM-guided semantic alignment.

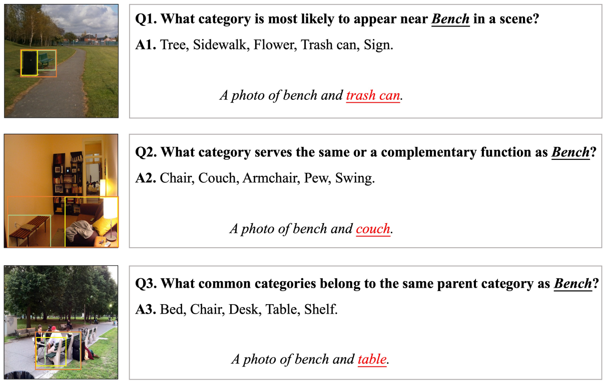

Figure 3. Cross-modal co-occurrence for *Bench* (known, `green`). **Q1** (*trash can*) reflects spatial adjacency, **Q2** (*couch*) emphasizes functional equivalence, and **Q3** (*table*) derives from furniture taxonomy. `Red` terms denote LLM-validated semantic matches.

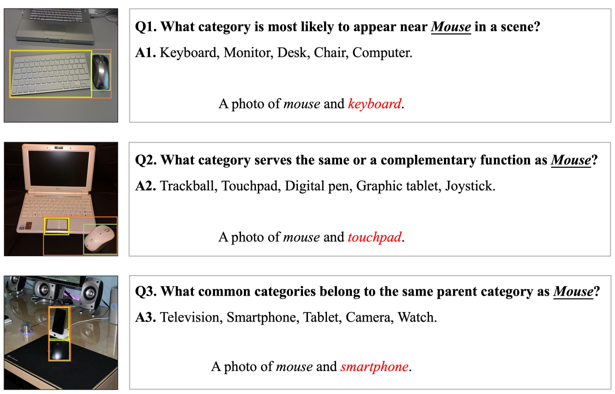

Figure 4. Co-occurrence patterns for *Mouse* (known, `green`). **Q1** (*keyboard*) captures spatial proximity, **Q2** (*touchpad*) identifies functional analogs, and **Q3** (*smartphone*) groups interactive devices hierarchically. `Red` labels signify LLM-aligned contextual relationships.

1.3. Signle vs Iterative Textual Description

Our iterative co-occurrence description strategy (§4.3.1) outperforms single-pass generation by +1.6 $\text{mAP}^{\text{Novel}}$ (VLDet: 32.3 $\rightarrow$ 33.9) and +1.4 $\text{mAP}^{\text{Novel}}$ (Detic: 28.4 $\rightarrow$ 29.8) (Table 1), highlighting three key benefits: (1) Single-pass methods yield unstable, contextually shallow relationships (e.g., *motorcycle* $\rightarrow$ *vehicle* vs. precise *helmet*); (2) Iteration dynamically refines candidates (e.g., *fork* $\rightarrow$ *knife*, *napkin* in Figure 1), expanding valid pairs; (3) Multi-cycle alignment tightens visual-textual correspondence, reducing spurious matches (e.g., *bench* $\rightarrow$ *tree* vs. *trash can*). By progressively aligning LLM-generated semantics with visual evidence, CODet bridges textual knowledge and scene-specific object distributions, ensuring robust generalization.

Table 1. Ablation study of single vs. iterative LLM interaction strategies on OV-COCO with VLDet and Detic baselines. We compare single-pass (static) and iterative (dynamic) textual co-occurrence generation, reporting novel ($\text{mAP}^{\text{Novel}}$) and overall ($\text{mAP}^{\text{All}}$) performance at IoU=0.5. Iterative refinement improves VLDet by +1.6 $\text{mAP}^{\text{Novel}}$ and Detic by +1.4 $\text{mAP}^{\text{Novel}}$, validating its role in enhancing contextual alignment.

Stractegy	VLDet [2]		Detic [5]	
	$\text{mAP}^{\text{Novel}}$	mAP^{All}	$\text{mAP}^{\text{Novel}}$	mAP^{All}
Signle	32.3	46.5	28.4	45.7
Iterative	33.9	47.6	29.8	46.8

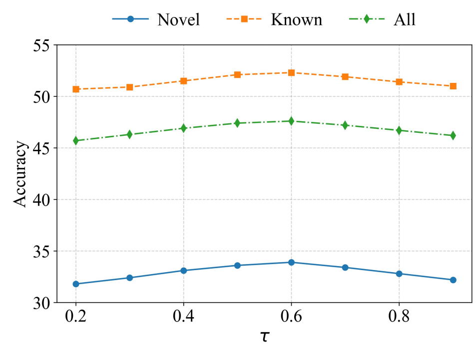

Figure 5. Impact of confidence threshold τ on OV-COCO performance using VLDet. Novel ($\text{mAP}^{\text{Novel}}$), known ($\text{mAP}^{\text{Known}}$), and overall ($\text{mAP}^{\text{All}}$) metrics (IoU=0.5) are evaluated across τ.

1.4. Parameter Analysis of Selection Threshold

Our confidence-based threshold τ (Section 4.3.2) critically balances pseudo-label quality and coverage during co-occurrence alignment. As shown in Figure 5, performance peaks at $\tau = 0.6$ (33.9 $\text{mAP}^{\text{Novel}}$ for VLDet on OV-COCO), degrading at lower/higher values. This reveals: (1) Low τ (< 0.6) introduces noise via under-filtered pairs (e.g., *bench* $\rightarrow$ *tree*); (2) $\tau = 0.6$ optimally balances diversity (capturing *fork* $\rightarrow$ *knife*) and precision (rejecting *motorcycle* $\rightarrow$ *cloud*); (3) High τ (> 0.6) over-filters valid relationships (*mouse* $\rightarrow$ *keyboard*), reducing recall. The threshold thus acts as a tunable gatekeeper, ensuring robust alignment between LLM-derived semantics and visual context while mitigating noise.

1.5. Textual Co-Occurring Category Candidates N

Our analysis of co-occurring candidate count N (Table 2) reveals that novel category detection (AP^{Novel}) peaks at $N = 5$ (33.9), while overall performance (AP^{All}) optimizes at $N = 10$ (47.8), reflecting a trade-off between noise reduction and contextual coverage. Lower N (e.g., $N = 2$) limits diversity, underutilizing co-occurrence relationships (32.6 AP^{Novel}), whereas higher N (e.g., $N = 20$) introduces noisy candidates, degrading novel detection (32.1 AP^{Novel}).

This suggests $N = 5$ balances precision for novel categories, while $N = 10$ accommodates broader context for known ones, emphasizing the need for task-specific tuning to harmonize diversity and accuracy.

Table 2. Impact of the number of textual co-occurring category candidates N on OV-COCO performance with VLDet. We report Novel (AP^{Novel}) and overall (AP^{Novel}) at IoU=0.5.

N	2	3	5	10	20
AP^{Novel}	32.6	33.4	33.9	33.6	32.1
AP^{All}	46.7	47.4	47.6	47.8	45.9

2. Prompt for Generating Textual Category Candidates

2.1. Prompt Design of LLMs

Objective: Generate valid co-occurring categories for known objects via GPT-4, adhering to dataset-specific constraints, including the number of outputs, output granularity, and repeated outputs, et al.

Input and Examples:

{examples} have the following content and in the following format: [train, handbag, bottle, boat, bed, toothbrush, skis, remote, ...]

Outputs Restrictions:

1. The answer should be a specific external object or type of external object, not an object modified by terms like 'some','other',or similar.
2. The output object should not include any objects from {examples}.
3. Output the [N] most relevant objects in examples format.
4. The output of Q2 cannot be identical to the output of Q1.
5. The output of Q3 cannot be identical to the output of Q2.

Examples (Category: *boat*, $N = 10$):

- **Q1.** *What category is most likely to appear near boat in a scene?*
  A1. *dock, life jacket, anchor, sail, paddle, compass, fishing rod, buoy, life ring, harbor.*
- **Q2.** *What category serves the same or a complementary function as boat?*
  A2. *ship, canoe, kayak, submarine, raft, yacht, sailboat, dinghy, catamaran, ferry.*
- **Q3.** *What common categories belong to the same parent category as boat?*
  A3. *train, car, motorcycle, bicycle, truck, skis, surfboard, airplane, helicopter, scooter.*

Examples (Category: *toothbrush*, $N = 5$):

- **Q1.** *What category is most likely to appear near toothbrush in a scene?*
  A1. *toothpaste, mirror, sink, towel, cup.*
- **Q2.** *What category serves the same or a complementary function as toothbrush?*
  A2. *floss, mouthwash, tongue scraper, electric toothbrush, dental pick.*
- **Q3.** *What common categories belong to the same parent category as toothbrush?*
  A3. *tray, ladle, napkin, plate, tea kettle.*

2.2. Ablation of LLMs

In Table 3, we conduct ablation study of LLMs for co-occurrence generation on OV-COCO (VLDet baseline). GPT-4 achieves the highest performance in novel (33.9 AP^{Novel}) and overall (47.6 AP^{All}) detection at IoU=0.5, outperforming Qianwen and Llama 2-13B. Results underscore the critical role of LLM capability in generating semantically meaningful co-occurrence category candidates, with GPT-4's superior contextual alignment driving significant gains.

Table 3. Ablation study of LLMs. We conduct experiments using different LLMs to generate textual co-occurrence category candidates, on the OV-COCO dataset with VLDet baseline, and report Novel (AP^{Novel}) and overall (AP^{Novel}) at IoU=0.5.

	GPT-4	Qianwen	Llama-2-13B
AP^{Novel}	33.9	33.7	32.0
AP^{All}	47.6	47.3	45.6

References

[1] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 1

[2] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. *arXiv preprint arXiv:2211.14843*, 2022. 1, 2

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 1

[4] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14393–14402, 2021. 1

[5] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European conference on computer vision*, pages 350–368. Springer, 2022. 1, 2