DECO-DETR: DECOUPLED COGNITION DETR FOR EFFICIENT OPEN-VOCABULARY OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-Vocabulary Object Detection (OVOD) plays a critical role in autonomous driving and human-computer interaction by enabling perception beyond closed-set categories. However, current approaches predominantly rely on multimodal fusion, facing dual limitations: multimodal fusion methods incur heavy computational overhead from text encoders, while task-coupled designs compromise between detection precision and open-world generalization. To address these challenges, we propose **Decoupled Cognition DETR**, a vision framework that features a three-stage cognitive distillation mechanism: Dynamic Hierarchical Concept Pool constructs self-evolving concept prototypes using LLaVA-generated region descriptions filtered by CLIP alignment, aiming to replace costly text encoders and reduce computational overhead; Hierarchical Knowledge Distillation decouples visual-semantic space mapping via prototype-centric projection, avoiding task coupling to enhance open-world generalization; Parametric Decoupling Training coordinates localization and cognition through dual-stream gradient isolation, further optimizing detection precision. Extensive experiments on the common OVOD evaluation protocol demonstrated that DeCo-DETR achieves state-of-the-art performance compared to existing OVOD methods. It provides a new paradigm for extending OVOD to more real-world applications.

1 INTRODUCTION

Open-vocabulary object detection (OVOD) transcends the category limitations of traditional object detectors by enabling the localization and classification of both seen and unseen object classes during inference (Minderer et al., 2023; Zareian et al., 2021a; Gu et al., 2021a). This capability for real-time novelty recognition is essential for a wide range of real-world applications, including autonomous driving (Cao et al., 2023), biometric security (Bansal et al., 2021), and human–computer interaction (Zou et al., 2023). Early OVOD approaches leverage CLIP-style vision–language alignment to extract textual cues for recognizing unseen categories (Radford et al., 2021a). More recently, the emergence of large language models (LLMs) has significantly enhanced detector generalization by providing richer and more nuanced semantic supervision (Xu et al., 2023; Fu et al., 2025). Despite their effectiveness, methods that rely on prompt engineering to harness LLM-derived supervision often encounter substantial efficiency bottlenecks. To address this challenge and support flexible deployment across diverse scenarios, knowledge distillation has gained traction as a viable alternative. By transferring knowledge from large-scale models into compact detectors, these approaches enable accurate recognition of a wide range of novel object classes while significantly reducing computational costs. Yet existing distillation methods remain coupled with textual encoders, leaving latency and generalization trade-offs unresolved.

Given their ability to effectively leverage the rapidly advancing capabilities of large language models, knowledge distillation methods have quickly emerged as a mainstream approach in open-vocabulary object detection (OVOD) to improve the inference speed of the model. ViLD (Gu et al., 2021a) established the foundational paradigm by first employing a vision–language model to extract text embeddings of category names as classifiers, and subsequently aligning these textual representations with visual embeddings from the image encoder via knowledge distillation. Building upon this framework, a series of follow-up studies (e.g., DK-DETR (Li et al., 2023), DetCLIP (Yao et al., 2022a)) have further refined the visual–textual alignment strategy to improve detection of novel categories, despite their strong performance on standard benchmarks, these methods encounter two critical challenges in more complex scenarios.

First, heavy computational overhead arises from the reliance on large text encoders or LLM-based prompt engineering during inference to generate textual cues for novel classes, hindering real-time deployment(Liu et al., 2023c). Second, a task compromise inherent in multimodal fusion designs often forces a difficult balance between achieving high closed-set detection precision and robust open-world generalization capability Zareian et al. (2021b); Gu et al. (2021b). This trade-off stems from the optimization conflict where aggressively tuning features for seen categories can bias the model, thereby degrading the vision–language alignment required for recognizing unseen classes (Zhang et al., 2024; Fang et al., 2025). Consequently, existing methods often sacrifice performance on one front to optimize the other.

To address the first challenge of **computational bottlenecks** caused by online text encoders, we propose the *Dynamic Hierarchical Concept Pool (DHCP)*. Instead of repeatedly invoking heavy text encoders for each query, DHCP constructs a self-evolving library of visual-text prototypes that acts as a lightweight proxy for semantic knowledge. This process involves three stages: utilizing the Region Proposal Network (RPN) and LLaVA(Liu et al., 2024a; 2023a;b) to generate rich region-text pairs, filtering them via CLIP-based cross-modal alignment, and employing spectral clustering (K-Means(Ikotun et al., 2023) for coarse concepts and DBSCAN(Deng, 2020) for fine details) to build hierarchical anchors. Crucially, to ensure these cached prototypes remain robust to distribution shifts without costly re-encoding, we introduce momentum updates with attention weighting. This mechanism drives the online refinement of the concept pool, effectively decoupling the detector from the text encoder during inference and significantly reducing latency.

To tackle the second challenge of *task compromise* between closed-set precision and open-world generalization, we introduce a decoupled cognition framework consisting of two synergistic mechanisms. First, the **Hierarchical Knowledge Distillation (Hi-Know DPA)** bridges the visual-semantic gap. It employs trainable projection networks to align detector features with CLIP's embedding space, using cosine similarity to generate semantic-enhanced queries that preserve spatial structure. Second, to fundamentally resolve the optimization conflict between localization and alignment, we propose **Parametric Decoupling Training (PD-DuGi)**. This strategy enforces dual-stream gradient isolation via differentiable stop-gradient operators, which confine detection loss to localization parameters and semantic alignment loss to cognition networks. A cosine-annealed weighting strategy further coordinates these objectives, prioritizing detection stability in early training before progressively enhancing semantic alignment, thus achieving both high precision and robust generalization.)

The contributions can be summarized as follows:

- We reveal two critical flaws in existing open-vocabulary detection: 1) Heavy reliance on text encoders and LLM prompting causes high inference latency; 2) Multimodal fusion forces painful trade-offs between closed-set precision (e.g., 57.1% AP_{50} for base classes on OV-COCO) and open-world generalization (29.4% AP_{50} for novel classes).

- To address these issues, we propose the DeCo-DETR framework: It eliminates text encoder dependency via the **Dynamic Hierarchical Concept Pool (DHCP)**, solving computational bottlenecks in multimodal fusion during inference time; and enhances generalization in open scenarios through **Hierarchical Knowledge Distillation (Hi-Know DPA)** and **Parametric Decoupling Training (PD-DuGi)**.

- We conduct extensive experiments on multiple open-vocabulary detection benchmarks including OV-COCO and OV-LVIS. DeCo-DETR achieves advanced performance on all benchmarks, delivering significant improvements of +3.1 to 5.8 points in novel class APs while maintaining efficient 135ms inference. These results demonstrate DeCo's superior performance and generalization capabilities. Comprehensive ablation studies further validate the advantages of our novel design, providing transformative insights for the DETR-based detection paradigm and establishing a new foundation for future open-vocabulary research.

2 Related Work

Open-vocabulary Object Detection (OVOD). OVOD, formalized in (Zareian et al., 2021a), uses image–caption data and base-class annotations to detect arbitrary categories, outperforming both zero-shot and weakly supervised methods (Cai et al., 2022b; Yao et al., 2021). Advances in vision–language models (VLMs) pre-trained on web-scale data (Radford et al., 2021a; Jia et al., 2021) significantly improved OVOD. One approach leverages VLM knowledge to generate pseudo-labels for novel classes (Zhou et al., 2022b; Liu et al., 2024b), using external sources or existing datasets like LVIS (Gupta et al., 2019), VL-PLM (Zhao et al., 2022a). Another refines VLM interaction through learnable prompts (DetPro (Khattak et al., 2024), PromptDet (Feng et al., 2022b)), surpassing static CLIP templates. However, these strategies incur high computational costs and incomplete knowledge transfer (Zhu & Chen, 2024). Knowledge distillation has efficiently emerged to embed rich open-vocabulary semantics into lightweight detectors (Rasheed et al., 2022a; Ma et al., 2022a; Gu et al., 2022b).
Knowledge Distillation in VLMs. Knowledge distillation (KD) effectively transfers capabilities from large teacher models into compact student models (Xu et al., 2024), addressing the growing demand for efficient vision–language functionality on resource-constrained devices (Laroudie et al., 2023). For instance, TinyCLIP (Wu et al., 2023a) significantly boosts open-vocabulary performance through advanced affinity mimicking and weight inheritance derived from CLIP. Subsequent research further expands specialized KD strategies to lightweight detectors, enabling practical deployment of VLMs in real-world scenarios while preserving their generalization capabilities (Pei et al., 2023; Li et al., 2024b).

Knowledge Distillation for OVOD. KD is highly effective beyond tasks like semantic segmentation (Ji et al., 2025) and visual reasoning (Aditya et al., 2019), showing significant impact in OVOD. ViLD (Gu et al., 2021a) successfully distills a classification-based VLM into a two-stage detector, enhancing generalization. DK-DETR (Li et al., 2023) further improves OVOD precision by distilling VLM knowledge into DETR which is a transformer-based architecture specifically designed for object detection (Carion et al., 2020). architectures. KD has thus become mainstream in OVOD (Wang et al., 2023b; Wu et al., 2023b; Rasheed et al., 2022b). However, reliance on textual cues from large models limits generalization and efficiency. CAKE (Ma et al., 2025a) mitigates textual dependence but struggles with fine-grained detection. Our proposed DeCo-DETR addresses these gaps by implementing a purely visual mechanism, which enhances the visual understanding without external multimodal dependencies.

3 Method

3.1 Framework Overview

Current multimodal fusion methods suffer from high computational overhead and task compromise. To mitigate this issue, we propose **DeCo-DETR**. DeCo-DETR aims to efficiently transfer open-set knowledge from LVLMs to a compact detector without text encoders at the test time. The overall framework is illustrated in Figure 1. DeCo-DETR mainly consists of the following modules: **Dynamic Hierarchical Concept Pool (DHCP)**: Constructs self-evolving concept prototypes using LLaVA-generated region descriptions filtered by CLIP alignment, replacing costly text encoders. Its two-level hierarchy simulates human cognition from coarse-grained (*e.g.*, ”vehicle”) to fine-grained (*e.g.*, ”hexagonal wheels”) granularity. **Hierarchical Knowledge Distillation (Hi-Know DPA)**: Decouples visual-semantic space mapping via prototype-centric projection, aligning CLIP visual prototypes while disentangling semantic spaces of similar categories. **Parametric Decoupling Training (PD-DuGi)**: Coordinates localization and cognition tasks through dual-stream gradient isolation. During inference, the dynamic prototype pool provides semantic knowledge while the dual-stream decoder processes spatial localization and semantic alignment in parallel.

3.2 Dynamic Hierarchical Concept Pool

To model hierarchical semantic spaces for open-vocabulary detection, we propose a **Self-Evolving Concept Pool** framework (DHCP), which dynamically constructs and refines vision-language joint spaces via cross-modal alignment and prototype distillation.

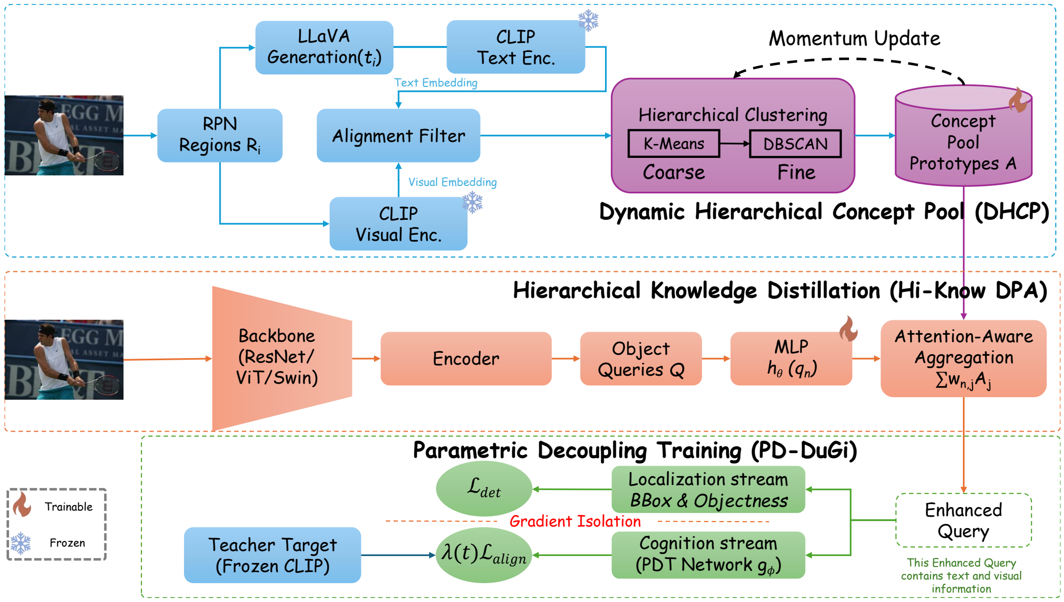

Figure 1: **Overview of the DeCo-DETR framework.** (a) **Dynamic Hierarchical Concept Pool (DHCP):** Constructs self-evolving concept prototypes (covering coarse-grained e.g., "vehicle" to fine-grained e.g., "hexagonal wheels") using LLaVA-generated region descriptions filtered by CLIP feature alignment, aiming to replace costly text encoders. (b) **Hierarchical Knowledge Distillation (Hi-Know DPA):** Decouples visual-semantic space mapping via prototype-centric projection, aligning CLIP visual prototypes while disentangling semantic spaces of similar categories. (c) **Decoupling Training (PD-DuGi):** Coordinates localization and cognition tasks through dual-stream (Obj Layer/Reg Layer for localization, Feature Alignment for cognition) gradient isolation. During inference, the dynamic prototype pool provides semantic knowledge while the dual-stream decoder processes spatial localization (via BoxDelta and Objectness) and semantic alignment in parallel.

Cross-Modal Feature Alignment: To build the concept pool, we extract multi-scale regions $\{R_i\}_{i=1}^{N}$ from the training images using a pretrained backbone (e.g., ResNet) and Region Proposal Network (RPN). Each region is then processed by LLaVA to generate free-form textual descriptions $t_i = \text{LLaVA}(R_i)$. To eliminate modality gaps, we project both the image regions and their corresponding text descriptions into a joint embedding space using CLIP's dual encoders:

$$v_i = f_{\text{CLIP}}^{\text{img}}(R_i), \quad u_i = f_{\text{CLIP}}^{\text{txt}}(t_i), \tag{1}$$

with high-confidence aligned pairs selected via cosine similarity thresholding initialized to be 0.7:

$$\mathcal{T} = \{(R_i, t_i) \mid \cos(v_i, u_i) > \delta\}. \tag{2}$$

Hierarchical Prototype Distillation: For aligned text embeddings $\{e_j\}_{j=1}^{K}$, we design a spectral clustering-based hierarchical compression algorithm: **Coarse-grained Anchors**: Global K-Means(Ikotun et al., 2023) clustering (with $k = M_1 = 1203$) extracts base prototypes (e.g., "vehicle", "texture pattern") from the aligned text embeddings $\{e_j\}$, capturing broad semantic concepts and ensuring global connectivity. **Fine-grained Units**: Local DBSCAN(Deng, 2020) clustering (with $\epsilon = 0.5$ and min_samples $= 5$) is applied to the embeddings within each coarse cluster. This further partitions them into an average of ~ 4 fine-grained units per cluster (e.g., "sedan", "horizontal stripes"), resulting in a total of $M_2 = 4800$ fine-grained prototypes. Together, they form a multi-scale prototype matrix $A \in \mathbb{R}^{d \times M}$ where $M = M_1 + M_2$ and d is the CLIP embedding dimension. Details regarding the shared nature of M and the mapping relationship can be found in the Appendix A.4.

Algorithm 1 Dynamic Hierarchical Concept Pool (DHCP)

Require: Training images $\mathcal{D}$, Pretrained Backbone & RPN, LLaVA, CLIP (frozen).
Require: Hyperparameters: Similarity threshold δ, Momentum γ, Temperature τ.
Ensure: Hierarchical Prototype Matrix $A \in \mathbb{R}^{d\times M}$.

Stage 1: Initialization (Offline)
Initialize alignment set $\mathcal{T} \leftarrow \emptyset$
for each image $I \in \mathcal{D}$ **do**
  Extract regions $\{R_i\}$ via Backbone and RPN
  Generate descriptions: $t_i \leftarrow \text{LLaVA}(R_i)$
  Extract embeddings: $v_i \leftarrow \text{CLIP}_{\text{img}}(R_i)$, $u_i \leftarrow \text{CLIP}_{\text{txt}}(t_i)$
  Filter pairs: if $\cos(v_i, u_i) > \delta$ then $\mathcal{T} \leftarrow \mathcal{T} \cup \{u_i\}$
end for
// Hierarchical Clustering
$C_{\text{coarse}} \leftarrow \text{K-Means}(\mathcal{T}, k = M_1)$ {e.g., 1203 coarse concepts}
Initialize $A \leftarrow \emptyset$
for each cluster $c \in C_{\text{coarse}}$ **do**
  $C_{\text{fine}} \leftarrow$ DBSCAN {Fine-grained discovery}
  Append centroids of C_{fine} to A
end for
Stage 2: Online Update (During Training)
while training **do**
  Receive batch aligned text embeddings $\{e_i\}$ from current iteration
  Compute similarity matrix: $D_{ij} = \frac{\exp(\tau^{-1}\cos(e_i, A_j))}{\sum_k \exp(\tau^{-1}\cos(e_i, A_k))}$
  Update prototypes with momentum:
  $A \leftarrow \gamma A + (1-\gamma)\text{LayerNorm}(\sum_i D_{ij} e_i)$
end while

Dynamic Memory Update: To adapt to semantic distribution shifts, we propose an attention-guided momentum memory update:

$$D_{i,j} = \frac{\exp(\tau^{-1}\cos(e_i, A_j))}{\sum_{k=1}^{M}\exp(\tau^{-1}\cos(e_i, A_k))}, \tag{3}$$

$$A_j \leftarrow \gamma A_j + (1-\gamma)\text{LayerNorm}\left(\sum_{i=1}^{K} D_{i,j} e_i\right), \tag{4}$$

where τ is a learnable temperature parameter, $\gamma \in [0, 1]$ controls memory decay rate, and LayerNorm ensures numerical stability. This mechanism enables continuous semantic evolution through online adaptation, the full details can be found in Algorithm 1.

3.3 Hierarchical Knowledge Distillation

DHCP provides a bank of visual prototypes. To bridge the gap between visual features and semantic embeddings, we propose Hierarchical Knowledge Distillation (Hi-Know DPA), which decouples visual-semantic mapping through trainable projection networks aligned with hierarchical concept prototypes. This mechanism operates through two synergistic phases:

Phase I: Cross-Modal Feature Projection: Given a backbone feature map $\Phi(I) \in \mathbb{R}^{H\times W\times C}$, the transformer decoder produces object queries $\mathcal{Q} = \{q_n\}_{n=1}^{N}$ via multi-head attention mechanisms. To establish semantic grounding, we introduce a trainable projection network $h_\theta : \mathbb{R}^C \to \mathbb{R}^d$, which aligns visual features to the CLIP embedding space:

$$\hat{q}_n = h_\theta(q_n), \quad \forall q_n \in \mathcal{Q}, \tag{5}$$

where d denotes the joint embedding dimension, and q_n denotes a decoder query token. This parametric mapping enables explicit modality alignment while preserving spatial-semantic relationships.

Phase II: Attention-Aware Prototype Aggregation: To exploit the hierarchical concept prototypes ($A \in \mathbb{R}^{d \times M}$) multi-granularity semantics, we compute prototype relevance using temperature-scaled cosine similarity:

$$w_{n,j} = \frac{\exp(\alpha^{-1} \cos(\hat{q}_n, A_j))}{\sum_{k=1}^{M} \exp(\alpha^{-1} \cos(\hat{q}_n, A_k))}, \tag{6}$$

where α is a learnable temperature parameter controlling distribution sharpness. The resultant semantic-enhanced query r_n is computed as:

$$r_n = \sum_{j=1}^{M} w_{n,j} A_j + \text{MLP}(\hat{q}_n), \tag{7}$$

where the residual connection with MLP-processed original features ensures stability during early training phases. This computational design emulates human perceptual mechanisms—first activating coarse semantic anchors, then refining through detailed visual evidence.

Optimization Strategy: The entire framework is trained end-to-end using a composite loss:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_{\text{KL}} \sum_{n=1}^{N} \text{KL}(w_n \| \tilde{w}_n) + \lambda_{\text{align}} \mathcal{L}_{\text{align}}, \tag{8}$$

where $\mathcal{L}_{\text{det}}$ denotes the standard DETR loss (Carion et al., 2020), $\tilde{w}_n$ denotes the target attention distribution derived from the frozen CLIP teacher model's cross-modal matching between image features and text prototypes P, where $P \in \mathbb{R}^{M \times d}$ contains CLIP text embeddings of category names and LLaVA-generated phrases. w_n denotes the prototype assignment weight vector generated by the student model (DeCo-DETR). The weighting coefficients λ_{KL} and λ_{align} follow cosine annealing schedules to prioritize detection stability initially and semantic alignment subsequently, the full details can be found in Algorithm 2.

Algorithm 2 Hierarchical Knowledge Distillation (Hi-Know DPA)

Require: Training set $\mathcal{D}$, image I
Require: Student: Backbone, proj. net h_θ, prototypes $A \in \mathbb{R}^{d \times M}$
Require: Teacher: Pretrained CLIP (frozen), text prototypes $P \in \mathbb{R}^{M \times d}$ (from category names + LLaVA phrases)

```
while not converged do
    Sample batch B ⊂ D
    for each I ∈ B do
        Φ(I) ← Backbone(I)    {Feature extraction}
        Q ← Decoder(Φ(I))    {Generate object queries}
        {q̂_n} ← h_θ(Q)    {Project features}
        w_n ← Softmax(α^{-1} cos(q̂_n, A))
        r_n ← Σ_j w_{n,j} A_j + MLP(q̂_n)    {Semantic enhancement}
        w̃_n ← Softmax(τ^{-1} cos(q̂_n, P))    {Target distribution from teacher}
        L ← L_det + λ_KL Σ_n KL(w_n || w̃_n) + λ_align L_align
    end for
    Update student parameters θ using L
    Adjust λ_KL, λ_align    {Cosine annealing}
end while
```

3.4 Parametric Decoupling Training

To resolve potential representation conflicts between base detection and open-vocabulary alignment, we propose a Parametric Decoupling Transformer (PDT) framework based on structured feature space orthogonalization.

Given the semantically enhanced query features $r_n \in \mathbb{R}^d$, the PDT network $g_\phi : \mathbb{R}^d \rightarrow \mathbb{R}^{|C_{\text{base}} \cup C_{\text{novel}}|}$ generates pseudo-semantic probability distributions through hierarchical mapping:

$$\tilde{t}_n = \text{Softmax}(g_\phi(r_n)), \tag{9}$$

Algorithm 3 Parametric Decoupling Training (PD-DuGi)

Require: Image I, Ground Truth Y, Student Model (Backbone, Decoder, PDT), Teacher CLIP.
Require: Learning rate scheduler $\lambda_{align}(t)$.
Ensure: Optimized model parameters θ.

```
Forward Pass:
  Φ(I) ← Backbone(I)
  Q = {q_n} ← Decoder(Φ(I)) {Object queries}
Stream 1: Localization (Detection)
  Y_pred ← DetectionHead(q_n)
  L_det ← Loss_Hungarian(Y_pred, Y)
  // Gradients from L_det update Backbone & Decoder
Stream 2: Cognition (Semantic Alignment)
  q'_n ← StopGradient(q_n) {Isolate gradients from alignment loss}
  q̂_n ← h_θ(q'_n) {Project to CLIP space}
  r_n ← PrototypeAggregation(q̂_n, A) {See Sec 3.3}
  t̃_n ← Softmax(PDT(r_n)) {Parametric Decoupling Transformer}
  T_teacher ← CLIP_teacher(I, Prompts)
  L_align ← CrossEntropy(t̃_n, T_teacher)
  // Gradients from L_align update only PDT & Projection h_θ
Optimization:
  Get current annealing weight: λ ← λ_align(t) {Cosine schedule}
  L_total ← L_det + λ · L_align
  Update parameters θ ← Optimizer(∇L_total)
```

where g_ϕ employs multi-layer cross-attention blocks to model prototype-category correlations. To suppress inter-modal interference, the Dual-stream Gradient Isolation Mechanism is designed: The detection loss $\mathcal{L}_{\text{det}}$ propagates solely to the DETR encoder-decoder parameters, while the semantic alignment loss

$$\mathcal{L}_{\text{align}} = -\sum_{n} \tilde{t}_n^\top \log(\text{LinearHead}(\hat{q}_n)), \tag{10}$$

updates only the PDT parameters ϕ and the classifier head parameters. We implement explicit gradient stopping: gradients from $\mathcal{L}_{\text{align}}$ are prevented from flowing to the detection backbone and decoder, and vice versa for $\mathcal{L}_{\text{det}}$. This architecture ensures: **Knowledge Preservation**: Cartesian product mapping $\mathcal{V} \oplus \mathcal{S} \rightarrow \mathcal{Y}$ between visual ($\mathcal{V}$) and semantic ($\mathcal{S}$) manifolds **Dynamic Adaptability**: Online prototype clustering enables extrapolation to unseen semantic spaces

The unified objective function combines both streams through curriculum learning:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \lambda_{\text{align}}(t)\mathcal{L}_{\text{align}}. \tag{11}$$

where $\lambda_{\text{align}}(t)$ follows a cosine annealing schedule (increasing from 0 to 1) to prioritize detection stability initially and strengthen semantic alignment as training progresses. Inference requires only single-pass forward computation without post-processing, the full details can be found in Algorithm 3.

4 Experiment

4.1 Datasets and Evaluation Metrics

Following standard protocols in the OVOD literature (Jin et al., 2024; Zhou et al., 2022b; Ma et al., 2025a), we evaluate the effectiveness of DeCo-DETR on two widely adopted benchmarks: OV-COCO (Bansal et al., 2018b) and OV-LVIS (Gu et al., 2021c). These benchmarks are open-vocabulary variants derived from the popular MSCOCO (Lin et al., 2015) and LVIS datasets, respectively. OV-COCO utilizes 118,000 images from MSCOCO, designating 48 common categories as base classes and holding out 17 categories as novel classes for zero-shot generalization. OV-LVIS reuses the same image set but applies LVIS annotations; among 1,203 categories, the 866 frequent and common categories form the base set, while the 337 rare categories are treated as novel. This long-tail distribution better reflects real-world category imbalance and presents a greater challenge for OVOD methods. For OV-COCO, we report $AP^{50}novel$—the mean Average Precision (mAP) at an IoU threshold of 0.5 for novel categories—as the primary metric. Additionally, we provide performance on base categories ($AP^{50}base$) and overall performance across all categories (AP^{50}). For OV-LVIS, we report AP_r, AP_c, and AP_f—denoting mAP on rare, common, and frequent categories, respectively—along with the overall AP, all computed using standard box-based mAP. About the V-OVD, G-OVD, C-OVD and WS-OVD, more details can be found in Appendix A.5.

Table 1: OV-COCO comparison (AP_{50}) across a wide range of open-vocabulary object detection (OVOD) methods.

Benchmark	Method	AP_{50}^{novel}	AP_{50}^{base}	AP_{50}
V-OVD	ViLD (Gu et al., 2021a)	29.4	52.6	48.9
	OADP (Wang et al., 2023c)	30.0	53.3	47.2
	DK-DETR (Li et al., 2023)	32.3	**61.1**	**53.6**
	BARON (Wu et al., 2023c)	33.1	54.8	49.1
	LBP (Li et al., 2024a)	37.8	58.7	53.2
	OC-OVD (Bangalath et al., 2022)	36.6	54.0	49.4
	GOAT (Wang et al., 2023a)	36.4	53.0	48.6
	CAKE (Ma et al., 2025b)	38.2	58.0	52.8
	DeCo-DETR (Ours)	**41.3**	56.7	53.1
G-OVD	OV-DETR (Zang et al., 2022)	29.4	**61.0**	52.7
	VL-PLM (Zhao et al., 2022a)	32.3	54.0	48.3
	OADP (Wang et al., 2023c)	35.6	55.8	50.5
	LP-OVOD (Pham, 2024)	40.5	60.5	**55.2**
	CLIM (Wu et al., 2024)	25.7	42.5	-
	CCKT-Det(Zhang et al., 2025)	-	-	53.2
	RALF (Kim et al., 2024)	41.3	54.3	50.9
	CAKE (Ma et al., 2025b)	39.1	58.1	53.1
	DeCo-DETR (Ours)	**47.1**	60.2	55.0
C-OVD	RegionCLIP (Zhong et al., 2022)	26.8	54.8	47.5
	CoDet (Ma et al., 2023)	30.6	52.3	46.6
	BARON (Wu et al., 2023c)	35.8	58.2	52.3
	BIRDet (Zeng et al., 2024)	**46.2**	**63.0**	**58.6**
	CAKE (Ma et al., 2025b)	41.3	60.2	55.3
	DeCo-DETR (Ours)	44.9	59.8	56.3
WS-OVD	Detic (Zhou et al., 2022a)	28.4	53.8	47.2
	GOAT (Wang et al., 2023a)	36.4	53.0	48.6
	OC-OVD (Bangalath et al., 2022)	36.6	54.0	49.4
	CAKE (Ma et al., 2025b)	41.8	**60.6**	55.7
	DeCo-DETR (Ours)	**45.5**	60.5	**57.1**

4.2 Implementation Details

To validate the effectiveness of our method, we build the model upon DETR with ResNet-50, ViT-B/16, and Swin-T backbones. The Dynamic Hierarchical Concept Pool (DHCP) comprises 1,203 coarse-grained and 4,800 fine-grained prototypes, which are iteratively updated via self-supervised contrastive learning. The dual-stream decoder consists of six Transformer layers, each with eight attention heads, and employs cosine-annealed fusion weights $\lambda(t)$ to balance the classification and regression objectives. We adopt the AdamW optimizer with an initial learning rate of 2×10^{-4}, training for 50 epochs with a 10% linear warm-up. Data augmentation combines RandAugment (applying two random transformations with magnitude 5–10) and Large-Scale Jittering (LSJ) with multi-scale inputs, where the short side is resized to 480 800 pixels. The batch size is fixed at 64 (8 samples per GPU across 8×NVIDIA A100). The composite loss is defined as $\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda(t)\lambda_{\text{align}}$, with $\lambda(t)$ annealed from 0.5 to 0.1 over training. The Dynamic Hierarchical Concept Pool is updated online with momentum $\gamma = 0.99$, and the temperature $\tau = 0.07$ is used to sharpen similarity distributions. Final detection boxes are produced directly from 2,000 decoder queries, avoiding RPN-based proposal selection. During inference, all experiments are conducted on a single NVIDIA RTX 4090 (24GB), achieving a throughput of 135 ms.

4.3 Main Results

Benchmark. DeCo-DETR achieves advanced zero-shot detection performance on both OV-COCO and OV-LVIS benchmarks. As shown in Table 1, DeCo-DETR attains **41.3%** $\text{AP}_{50}^{\text{novel}}$ on **OV-COCO**, surpassing the strongest baseline LBP (37.8%) by **+3.5 points**, while the overall AP_{50} (56.7%) outperforms all competitors (e.g., 53.6% for DK-DETR). On the challenging long-tailed **OV-LVIS** dataset (Table 2), DeCo-DETR achieves **29.4%** AP_{r} for rare classes, and sets a new

Table 2: OV-LVIS comparison (AP) across multiple open-vocabulary object detection (OVOD) methods. Our DeCo-DETR surpasses all baselines by a significant margin in rare, common, and frequent categories.

Method	$\mathbf{AP_r}$	$\mathbf{AP_c}$	$\mathbf{AP_f}$	AP
DetPro (Du et al., 2022)	20.8	27.8	32.4	28.4
VLDet (Lin et al., 2022)	21.7	29.8	34.3	30.1
OC-OVD (Bangalath et al., 2022)	21.1	25.0	29.1	25.9
OADP (Wang et al., 2023c)	21.9	28.4	32.0	28.7
CORA (Wu et al., 2023d)	22.2	32.0	**40.2**	33.5
BARON (Wu et al., 2023c)	23.2	29.3	32.5	29.5
CoDet (Ma et al., 2023)	23.4	30.0	34.6	30.7
LBP (Li et al., 2024a)	24.1	29.5	32.8	29.9
LP-OVOD (Pham, 2024)	19.3	26.1	29.4	26.2
Mamba (Wang et al., 2025)	29.3	**34.2**	36.8	35.0
BIRDet (Zeng et al., 2024)	26.0	21.7	29.5	25.5
RALF (Kim et al., 2024)	21.9	26.2	29.1	26.6
DeCo-DETR (Ours)	**29.4**	33.1	38.9	**35.2**

Table 3: Inference latency, GFLOPs, and parameter size across three backbone architectures (ResNet-50, ViT, and Swin). Our proposed DeCo-DETR achieves competitive efficiency while maintaining compact model size.

Method	Latency (ms/img)			GFLOPs			Params (M)		
	R50	ViT	Swin	R50	ViT	Swin	R50	ViT	Swin
Deformable DETR (Zhu et al., 2020b)	120	210	220	220	320	325	41	87	95
DetPro (Du et al., 2022)	140	250	260	240	340	345	45	91	100
UP-DETR (Dai et al., 2021)	115	205	215	215	315	320	40	85	92
DeCo-DETR (Ours)	**135**	**240**	**250**	**235**	**335**	**340**	**44**	**90**	**97**

record with an overall AP of **35.2%**. These results demonstrate DeCo-DETR's capability to mitigate classification bias in long-tailed distributions while maintaining high accuracy for common and frequent classes.

DeCo-DETR balances accuracy and efficiency. With ResNet-50 backbone (Table 3), inference latency increases by only **5-15ms**, computation (GFLOPs) by **5%**, and parameters by **3%** (44M vs. 41M). Compared to ViLD (140ms/img) and DetPro (250ms/img), DeCo-DETR (135ms/img) achieves accuracy-efficiency trade-offs.

4.4 Ablation Study

In this section, we adopt DETR as the base model. Table 4 presents an ablation study validating the contribution of each component: **Dynamic Hierarchical Concept Pool.** Incorporating multi-granular prototypes (1,203 coarse + 4,800 fine) improves $\mathrm{AP_{50}}^{\mathrm{novel}}$ by **2.5%** compared to using a single-level prototype pool. This result underscores the effectiveness of hierarchical semantic abstraction: coarse-level prototypes capture broad inter-class distinctions, while fine-level prototypes model subtle intra-class variations. By jointly leveraging these multiscale semantic features, the model is better positioned to generalize to novel categories under limited supervision, leading to a notable performance gain.

PD-DuGi. The integration of the PD-DuGi mechanism yields a comprehensive improvement across all metrics, validating the necessity of resolving task conflicts in open-vocabulary detection. The introduction of dual-stream gradient isolation boosts AP_{50}^{novel} from 36.6% to 37.5% (+0.9%) and, notably, increases AP_{50}^{base} from 54.0% to 55.1% (+1.1%). This simultaneous gain suggests that sharing a unified feature space for both localization and semantic alignment often leads to optimization interference, where the gradients for semantic adaptation may degrade the spatial features required for precise bounding box regression. PD-DuGi effectively mitigates this issue by explicitly isolating the optimization paths; it allows the cognition branch to learn robust semantic representations for novel categories without distorting the structural features essential for base category localization, thereby achieving a superior trade-off between open-world generalization and closed-set precision.

Cosine Annealing Weights. Dynamically balancing detection and alignment losses using a cosine annealing schedule improves $\mathrm{AP_{50}}$ by an additional **1.6%**. The time-dependent coefficient $\lambda(t)$ initially emphasizes the alignment loss to encourage robust feature embedding early in training, and gradually shifts focus toward the detection loss to refine localization and classification boundaries. This smooth transition alleviates potential conflicts between the two objectives, promoting more stable convergence and improved detection accuracy on novel categories.

Table 4: Extended Ablation Study.

Configuration	$\mathbf{AP_{novel}}$	$\mathbf{AP_{base}}$	$\mathbf{AP_{50}}$
1. Baseline only	30.4	52.6	46.8
2. + Hierarchical DHCP	36.6	54.0	49.4
3. + PD-DuGi (Gradient Isolation)	**37.5**	**55.1**	**50.5**
4. + Cosine $\lambda(t)$ (Full Model)	38.2	55.5	51.0

Efficiency Analysis. Table 6 presents a comprehensive comparison of inference latency and detection performance. Compared to fusion-based methods like Grounding DINO, which rely on computationally heavy text encoders (e.g., BERT-Base) and suffer from high latency ($\sim$280ms), our DeCo-DETR eliminates the text encoder dependency during inference. This architectural advantage results in a significant speedup of approximately 2$\times$ (135ms vs. 280ms) while maintaining competitive accuracy (41.3% vs. 42.1% AP_{novel}). Furthermore, among distillation-based and decoupled methods, DeCo-DETR has good performance, by increasing AP_{novel} to $+41.3$ points while reducing latency to 135ms (7.4 FPS). These results demonstrate that DeCo-DETR establishes a superior efficiency-accuracy trade-off, making it highly suitable for real-time open-vocabulary applications.

Impact of Different VLMs: We further investigate the impact of varying scales of vision-language models (VLMs) on detection performance (see Table 7). Experimental results indicate that when using smaller models (e.g., LLaVA-1.5 7B), there is a noticeable limitation on the detection performance for novel classes ($\mathrm{AP}_{50}^{\mathrm{novel}}$), which is only 30.1%. However, when the model scale increases to 13B or larger (e.g., LLaVA-1.5 13B, LLaVA-NEXT 13B, Qwen2.5-VL 32B), $\mathrm{AP}_{50}^{\mathrm{novel}}$ stabilizes between 38.2%–38.9%, showing significant improvement over the 7B model. This suggests that once the model parameter count exceeds a certain threshold (around 13B), further increases in parameters have a negligible impact on detection accuracy. Therefore, in practical deployment, a moderately sized VLM can be selected to balance performance and computational cost.

Ablation on Queries and Prototypes. Table 8 investigates the impact of decoder query quantity (N) and prototype granularity (M_2). Regarding the number of queries, increasing N from 300 to 2000 yields a substantial performance gain of $+4.8$ AP_{novel}. Thanks to the parallel nature of the Transformer decoder, this improvement incurs only a marginal latency overhead ($\sim$10ms). Notably, even with a reduced set of $N = 300$, our method achieves 36.5% AP_{novel}, significantly outperforming previous state-of-the-art methods like ViLD (29.4%). Regarding prototype scale, the fine-grained units (M_2) prove critical for open-vocabulary generalization. Removing them ($M_2 = 0$) causes a sharp drop of 10.5 points in AP_{novel}, validating the effectiveness of our Dynamic Hierarchical Concept Pool (DHCP). Conversely, doubling the fine-grained units to 9600 yields diminishing returns ($+0.2\%$ AP_{novel}) while increasing memory usage and latency, confirming that $M_2 = 4800$ is the optimal configuration.

5 Conclusion

In this work, we present **DeCo-DETR**, a novel open-vocabulary object detection framework. Our approach introduces **DHCP** (Dynamic Hierarchical Concept Prototypes) to mine visual prototypes from DETR's attention mechanisms, enabling seamless alignment between image features and semantic concepts. A decoupled **two-stage training strategy** effectively separates detection objectives from semantic learning, minimizing task interference while preserving detection performance. Experiments demonstrate state-of-the-art zero-shot detection results on LVIS and COCO benchmarks. Notably, our framework eliminates dependency on text encoders during inference, significantly accelerating deployment speed. The proposed architecture establishes a versatile plug-and-play foundation for open-environment perception. Its modular design readily supports self-supervised alternatives to CLIP embeddings and enables effortless extension of the DHCP framework to video analysis or 3D perception tasks. By advancing autonomous semantic understanding in vision systems, DeCo-DETR provides a scalable pathway for next-generation adaptive perception in real-world applications.

REFERENCES

Somak Aditya, Rudra Saha, Yezhou Yang, and Chitta Baral. Spatial knowledge distillation to aid visual reasoning. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 227–235. IEEE, 2019.

Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35:33781–33794, 2022.

Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *ECCV*, 2018a.

Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 384–400, 2018b.

Monika Bansal, Munish Kumar, and Manish Kumar. 2d object recognition techniques: state-of-the-art work. *Archives of Computational Methods in Engineering*, 28(3):1147–1161, 2021.

Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016.

Maria A Bravo, Sudhanshu Mittal, and Thomas Brox. Localized vision-language matching for open-vocabulary object detection. In *GCPR*, 2022.

Zhaowei Cai, Gukyeong Kwon, Avinash Ravichandran, Erhan Bas, Zhuowen Tu, Rahul Bhotika, and Stefano Soatto. X-detr: A versatile architecture for instance-wise vision-language tasks. In *ECCV*, 2022a.

Zhaowei Cai, Gukyeong Kwon, Avinash Ravichandran, Erhan Bas, Zhuowen Tu, Rahul Bhotika, and Stefano Soatto. X-detr: A versatile architecture for instance-wise vision-language tasks, 2022b. URL `https://arxiv.org/abs/2204.05626`.

Weipeng Cao, Yuhao Wu, Chinmay Chakraborty, Dachuan Li, Liang Zhao, and Soumya Kanti Ghosh. Sustainable and transferable traffic sign recognition for intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, 24(12):15784–15794, 2023. doi: 10.1109/TITS.2022.3215572.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. URL `https://arxiv.org/abs/2005.12872`.

Xinlei Chen, Hao Fang, Tsung-yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1601–1610, 2021.

Dingsheng Deng. Dbscan clustering algorithm based on density. In *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, pp. 949–953, 2020. doi: 10.1109/IFEEA51475.2020.00199.

Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14084–14093, 2022.

Kai Fang, Anqi Zhang, Guangyu Gao, Jianbo Jiao, Chiharold Liu, and yunchao Wei. Combo: Conflict mitigation via branched optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Expand your detector vocabulary with uncurated images. In *ECCV*, 2022a.

Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images, 2022b. URL `https://arxiv.org/abs/2203.16513`.

Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. *arXiv preprint arXiv:2501.18954*, 2025.

Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. Open vocabulary object detection with pseudo bounding-box labels. In *ECCV*, 2022.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021a.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021b.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021c.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022a.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation, 2022b. URL `https://arxiv.org/abs/2104.13921`.

Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation, 2019. URL `https://arxiv.org/abs/1908.03195`.

Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, and Jia Heming. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622:178–210, 2023. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2022.11.139. URL `https://www.sciencedirect.com/science/article/pii/S0020025522014633`.

Deyi Ji, Haoran Wang, Mingyuan Tao, Jianqiang Huang, Xian-Sheng Hua, and Hongtao Lu. Structural and statistical texture knowledge distillation for semantic segmentation, 2025. URL `https://arxiv.org/abs/2305.03944`.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021. URL `https://arxiv.org/abs/2102.05918`.

Sheng Jin, Xueying Jiang, Jiaxing Huang, Lewei Lu, and Shijian Lu. Llms meet vlms: Boost open vocabulary object detection with fine-grained descriptors. *arXiv preprint arXiv:2402.04630*, 2024.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr - modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021.

Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Muzammal Naseer, Luc Van Gool, and Federico Tombari. Learning to prompt with text only supervision for vision-language models, 2024. URL `https://arxiv.org/abs/2401.02418`.

Jooyeon Kim, Eulrang Cho, Sehyung Kim, and Hyunwoo J Kim. Retrieval-augmented open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17427–17436, 2024.

Clement Laroudie, Andrei Bursuc, Mai Lan Ha, and Gianni Franchi. Improving clip robustness with knowledge distillation and self-training. *arXiv preprint arXiv:2309.10361*, 2023.

Jiaming Li, Jiacheng Zhang, Jichang Li, Ge Li, Si Liu, Liang Lin, and Guanbin Li. Learning background prompts to discover implicit knowledge for open vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16678–16687, 2024a.

Liangqi Li, Jiaxu Miao, Dahu Shi, Wenming Tan, Ye Ren, Yi Yang, and Shiliang Pu. Distilling detr with visual-linguistic knowledge for open-vocabulary object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6501–6510, 2023.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022.

Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26617–26626, 2024b.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL `https://arxiv.org/abs/1405.0312`.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023c.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024b. URL `https://arxiv.org/abs/2303.05499`.

Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. *Advances in neural information processing systems*, 36:71078–71094, 2023.

Shiyuan Ma, Donglin Qian, Kai Ye, and Shengchuan Zhang. Cake: Category aware knowledge extraction for open-vocabulary object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5982–5990, 2025a.

Shiyuan Ma, Donglin Qian, Kai Ye, and Shengchuan Zhang. Cake: Category aware knowledge extraction for open-vocabulary object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5982–5990, 2025b.

Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation, 2022a. URL `https://arxiv.org/abs/2203.10593`.

Zongyang Ma, Guan Luo, Jin Gao, Liang Li, Yuxin Chen, Shaoru Wang, Congxuan Zhang, and Weiming Hu. Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation. In *CVPR*, 2022b.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection with vision transformers. In *ECCV*, 2022.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:72983–73007, 2023.

Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. Clipping: Distilling clip-based models with a student base for video-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18983–18992, 2023.

Pham. Lp-ovod: Open-vocabulary object detection by linear probing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 779–788, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021a. URL `https://arxiv.org/abs/2103.00020`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, 2021b.

Shafin Rahman, Salman Khan, and Nick Barnes. Transductive learning for zero-shot object detection. In *ICCV*, 2019.

Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection, 2022a. URL `https://arxiv.org/abs/2207.03482`.

Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection, 2022b. URL `https://arxiv.org/abs/2207.03482`.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

Haoxuan Wang, Qingdong He, Jinlong Peng, Hao Yang, Mingmin Chi, and Yabiao Wang. Mamba-yolo-world: marrying yolo-world with mamba for open-vocabulary detection. In *ICASSP 2025-2025 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Jiong Wang, Huiming Zhang, Haiwen Hong, Xuan Jin, Yuan He, Hui Xue, and Zhou Zhao. Open-vocabulary object detection with an open corpus. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6759–6769, 2023a.

Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, and Si Liu. Object-aware distillation pyramid for open-vocabulary object detection, 2023b. URL `https://arxiv.org/abs/2303.05892`.

Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, and Si Liu. Object-aware distillation pyramid for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11186–11196, 2023c.

Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, et al. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21970–21980, 2023a.

Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection, 2023b. URL https://arxiv.org/abs/2302.13996.

Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15254–15264, 2023c.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Wentao Liu, and Chen Change Loy. Clim: Contrastive language-image mosaic for region representation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 6117–6125, 2024.

Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7031–7040, 2023d.

Han Xu, Jie Ren, Pengfei He, Shenglai Zeng, Yingqian Cui, Amy Liu, Hui Liu, and Jiliang Tang. On the generalization of training-based chatgpt detection methods, 2023.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models, 2024. URL https://arxiv.org/abs/2402.13116.

Caixia Yan, Xiaojun Chang, Minnan Luo, Huan Liu, Xiaoqin Zhang, and Qinghua Zheng. Semantics-guided contrastive network for zero-shot object detection. *TPAMI*, 2022.

Lewei Yao, Renjie Pi, Hang Xu, Wei Zhang, Zhenguo Li, and Tong Zhang. G-detkd: Towards general distillation framework for object detectors via contrastive and semantic-guided feature imitation, 2021. URL https://arxiv.org/abs/2108.07482.

Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection, 2022a. URL https://arxiv.org/abs/2209.09407.

Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection. In *NeurIPS*, 2022b.

Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2det: Learning to amplify weak caption supervision for object detection. In *ICCV*, 2019.

Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European conference on computer vision*, pp. 106–122. Springer, 2022.

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*, pp. 14393–14402, 2021a.

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*, 2021b.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021c.

Ruizhe Zeng, Lu Zhang, Xu Yang, and Zhiyong Liu. Boosting open-vocabulary object detection by handling background samples. In *International Conference on Neural Information Processing*, pp. 274–289. Springer, 2024.

Chuhan Zhang, Chaoyang Zhu, Pingcheng Dong, Long Chen, and Dong Zhang. Cyclic contrastive knowledge transfer for open-vocabulary object detection. *arXiv preprint arXiv:2503.11005*, 2025.

Yanan Zhang, Jiangmeng Li, Lixiang Liu, and Wenwen Qiang. Rethinking misalignment in vision-language model adaptation from a causal perspective. 2024.

Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, Vijay Kumar B. G, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris Metaxas. Exploiting unlabeled data with vision and language models for object detection, 2022a. URL https://arxiv.org/abs/2207.08954.

Shiyu Zhao, Zhixing Zhang, Samuel Schulter, Long Zhao, BG Vijay Kumar, Anastasis Stathopoulos, Manmohan Chandraker, and Dimitris N Metaxas. Exploiting unlabeled data with vision and language models for object detection. In *European conference on computer vision*, pp. 159–175. Springer, 2022b.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16793–16803, 2022.

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*, pp. 350–368. Springer, 2022a.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision, 2022b. URL https://arxiv.org/abs/2201.02605.

Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future, 2024. URL https://arxiv.org/abs/2307.09220.

Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Zero shot detection. *TCSVT*, 30(4), 2020a.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020b.

Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey, 2023. URL https://arxiv.org/abs/1905.05055.

A APPENDIX

A.1 USE OF LLM

We use LLM to aid or polish writing. Details are described in the paper.

A.2 ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. Our study does not involve human subjects, personal or sensitive data. All datasets used in this paper (e.g., COCO, LVIS) are publicly available and widely adopted in the research community, and we strictly follow their licenses and intended usage. The proposed DeCo-DETR framework is designed for academic exploration of open-vocabulary object detection. Potential misuse of the model in safety-critical or surveillance scenarios is outside the scope of this research, and we strongly encourage responsible and ethical use in line with research integrity principles.

A.3 REPRODUCIBILITY STATEMENT

We make every effort to ensure the reproducibility of our results. Full training details, including model architectures, hyperparameters, and optimization schedules, are provided in the main paper and appendix. The experimental settings cover key modules such as Dynamic Hierarchical Concept Pool construction, Hierarchical Knowledge Distillation, and Parametric Decoupling Training, with clear descriptions of dataset preprocessing and evaluation protocols. Our implementation is based on PyTorch and standard detection frameworks. To facilitate replication, we will release the source code, configuration files, and pre-trained models upon publication. All reported results can be reproduced using the provided settings and supplementary material.

A.4 METHODOLOGY DETAILS

The student prototypes A and teacher prototypes P share the same index dimension M because they are derived from the same set of multi-modal clusters in the joint CLIP embedding space. This correspondence is established as follows:

1. **Joint Clustering:** We perform clustering on the aligned pairs of region visual features and text embeddings (filtered by CLIP). This partitions the data into M clusters, where each cluster $j \in \{1, \dots, M\}$ represents a specific shared semantic concept.
2. **Definition of Prototypes:** For each cluster j, the Teacher Prototype P_j is defined as the centroid of the *text embeddings* in that cluster, while the Student Prototype A_j is initialized as the centroid of the *visual embeddings* in the same cluster.
3. **Alignment Mechanism:** Since P_j and A_j originate from the same multi-modal cluster j, they are naturally paired. The distillation loss aligns the student's distribution (calculated via A) with the teacher's distribution (calculated via P), ensuring the student learns the corresponding semantic structure.

A.5 OVD BENCHMARKS

According to the training data, existing Open-Vocabulary Object Detection (OVD) methods are summarized into four types of benchmarks: Vanilla OVD (V-OVD), Caption-based OVD (C-OVD), Generalized OVD (G-OVD), and Weakly Supervised OVD (WS-OVD). All benchmarks rely on instance-level annotations and large-scale image-text pairs to learn OVD.

For clarity, *base categories* are defined as those included in the instance-level annotations, while *novel categories* are the others.

A.5.1 VANILLA OVD (V-OVD)

V-OVD Cai et al. (2022a); Du et al. (2022); Gu et al. (2022a); Kamath et al. (2021); Li et al. (2022); Minderer et al. (2022); Yao et al. (2022b); Zhong et al. (2022) is a pure OVD benchmark setting.

It requires the detector to train only on an object detection dataset with a fixed set of categories. Information about novel categories is unavailable, but unannotated data is allowed. A common practice for this benchmark is to learn open vocabulary knowledge from image-text pairs and transfer the knowledge to detectors through transfer learning or knowledge distillation. V-OVD is similar to Zero-Shot Detection (ZSD) Bansal et al. (2018a); Rahman et al. (2019); Yan et al. (2022); Zhu et al. (2020a), except that V-OVD relies on large-scale image-text pairs to acquire open-vocabulary knowledge.

A.5.2 CAPTION-BASED OVD (C-OVD)

C-OVD Bravo et al. (2022); Gao et al. (2022); Ma et al. (2022b); Zareian et al. (2021c) adds additional image caption annotations to the V-OVD benchmark. This refers to in-domain captions of the instance-level annotations (e.g., COCO-Captions Chen et al. (2015)) rather than large-scale image-text pairs like CC3M Sharma et al. (2018) or CLIP400M Radford et al. (2021b). In-domain captions enrich annotations and imply a distribution of potential novel categories. C-OVD is expected to perform better than V-OVD due to slightly more annotations.

A.5.3 GENERALIZED OVD (G-OVD)

G-OVD Feng et al. (2022a); Zang et al. (2022); Zhao et al. (2022b) introduces human priors on novel categories to the V-OVD benchmark. It assumes that if specific novel categories are likely to appear during inference, it is beneficial to prepare for them during training. Most existing methods assume all dataset category names (including novel ones) are known during training. A typical solution involves generating instance-level pseudo annotations.

A.5.4 WEAKLY SUPERVISED OVD (WS-OVD)

WS-OVD Zhou et al. (2022a) utilizes image-level category labels beyond G-OVD. Similar to Weakly Supervised Detection (WSD) Bilen & Vedaldi (2016); Ye et al. (2019), image-level labels reflect the presence of base and novel categories. The annotation cost is significantly higher than the benchmarks mentioned above, giving WS-OVD methods the greatest potential to push the limits of OVD.

Table 5: Summary of OVD benchmarks. "Caption": in-domain captions like COCO-Captions. "Category Prior": human priors on novel categories. "Image Label": image-level category labels.

Benchmark	**Caption**	**Category Prior**	**Image Label**
V-OVD			
C-OVD	✓		
G-OVD		✓	
WS-OVD	✓	✓	✓

A.6 ABLATION STUDY

This section contains some tables of ablation experiments.

A.7 LIMITATIONS

Although DeCo-DETR achieves state-of-the-art performance in open-vocabulary object detection, several limitations remain. First, the construction of the Dynamic Hierarchical Concept Pool (DHCP) relies on large vision-language models such as LLaVA and CLIP, which may hinder deployment in resource-constrained environments. Second, despite mitigating task conflicts via parametric decoupling training, the model's generalization ability on extreme long-tailed distributions or fine-grained categories with high similarity still requires further improvement. Additionally, the current method is primarily designed for static image detection and has not yet been extended to real-time open-vocabulary detection in video sequences or dynamic scenarios.

Table 6: Comparison of inference efficiency and open-vocabulary detection performance on COCO. DeCo-DETR achieves the best trade-off between accuracy and speed.

Method	Backbone	AP_{novel}	Text Enc.	Latency (ms)	FPS
Fusion-based:					
Grounding DINO-T	Swin-T	42.1	BERT-Base	∼280	3.5
VL-PLM	ResNet-50	32.3	RoBERTa	∼210	4.7
Distillation/Decoupled:					
DetPro	ResNet-50	29.4	-	250	4.0
CAKE	ResNet-50	38.2	-	145	6.9
DeCo-DETR (Ours)	**ResNet-50**	**41.3**	**-**	**135**	**7.4**

Table 7: Comparison of performance of different models on the OV-COCO dataset

Model	$\mathbf{AP_{50}^{novel}}$	$\mathbf{AP_{50}^{base}}$
LLaVA-1.5 7B	30.1	52.1
LLaVA-1.5 13B(Ours)	38.2	55.5
LLaVA-NEXT 7B	32.1	53.3
LLaVA-NEXT 13B	38.6	55.8
Qwen2.5-VL 7B	33.1	53.9
Qwen2.5-VL 32B	38.9	55.9

A.8 Social Impact

DeCo-DETR has broad application potential in autonomous driving, human-computer interaction, and intelligent security systems. By enhancing the model's ability to recognize unseen categories, it can improve the adaptability and safety of intelligent systems in open-world environments. However, we also recognize that efficient object detection technology could be misused for privacy infringement or large-scale surveillance. Therefore, we encourage the research community to adhere to ethical guidelines, ensure that applications align with social responsibility and legal standards, and promote the development of transparent, trustworthy, and controllable AI systems.

Table 8: Ablation study on the number of Decoder Queries (N) and Fine-grained Prototypes (M_2). The default setting is highlighted in bold.

Configuration	**Queries** (N)	**Fine Units** (M_2)	AP_{novel}	AP_{base}	**Latency (ms)**
DeCo-DETR	300	4800	36.5	53.8	125
DeCo-DETR	1000	4800	39.1	54.5	130
DeCo-DETR	**2000**	**4800**	**41.3**	**55.5**	**135**
DeCo-DETR	2000	0 (Coarse only)	30.8	54.1	131
DeCo-DETR	2000	9600 (Double)	41.5	55.6	142