More Pictures Say More: Visual Intersection Network for Open Set Object Detection

Anonymous ACL submission

Abstract

Current approaches to open-set object detection heavily rely on vision-language fusion paradigms, yet this methodology faces an inherent challenge: many objects are difficult to describe accurately through language alone. While recent research has attempted to incorporate visual information to address this limitation, existing models still struggle with fine-grained object discrimination. In response, we introduce VINO (Visual Intersection Network for OSOD), a novel DETR-based pure vision model that constructs a multi-image visual bank to preserve semantic intersections across categories and facilitates the fusion of category and region semantics through a multi-stage mechanism. Furthermore, we implement a simple replacement strategy to ensure the model learns alignment capabilities rather than semantic approximation. With an image consumption of only 0.84M, VINO achieves competitive performance on par with vision-language models on benchmarks such as LVIS and ODinW35. Additionally, the successful integration of a segmentation head demonstrates the broad applicability of visual intersection across various visual tasks.

1 Introduction

Open-set object detection (OSOD) fundamentally aims to align region semantics with target object semantics. Current mainstream approaches (Li et al., 2022a; Zhang et al., 2022) leverage frozen large language models (LLMs) for their semantic generalization capabilities, encouraging visual extractors to align with LLMs' semantic space. However, this paradigm inherently constrains the model's object discrimination ability to the semantic resolution of LLMs, particularly struggling with objects that defy precise linguistic description. Moreover, bridging the modality gap between vision models and LLMs demands extensive pretraining, requiring substantial image consumption ranging from 11.52M (APE-A) to 200M (X-Decoder (Zou et al., 2022)).

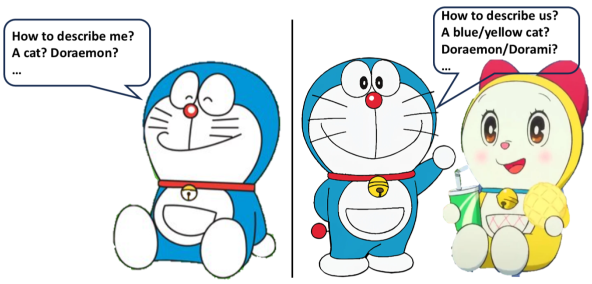

Figure 1: Illustration of the linguistic description challenge in fine-grained object detection, where similar visual characteristics make it difficult to distinguish between closely related objects using language or visual instructions in single image.

To address these semantic description limitations, several studies have explored the use of visual prompts. Some works (Xu et al., 2023; Kang et al., 2019) employ visual prompts as auxiliary information to enrich textual representations. However, these approaches rely heavily on visual-language fusion to perform cross-modal multi-head attention between high-dimensional words and regions, resulting in increased memory consumption and computational complexity. Other researchers (Jiang et al., 2024; Li et al., 2023; Ren et al., 2024) have investigated interactive visual instructions (e.g., points or boxes) to enhance detection performance. While these interactive approaches enable semantic learning through position-aware cross-attention, they are constrained to single-image scenarios, failing to capture semantic generalization across multiple images. Additionally, some methods (Li et al., 2022b; Zang et al., 2022) utilizing image-level prompts with siamese network architectures are primarily limited to few-shot learning scenarios.

We are motivated to work as shown in Figure 1. Object semantics can be effectively cap-

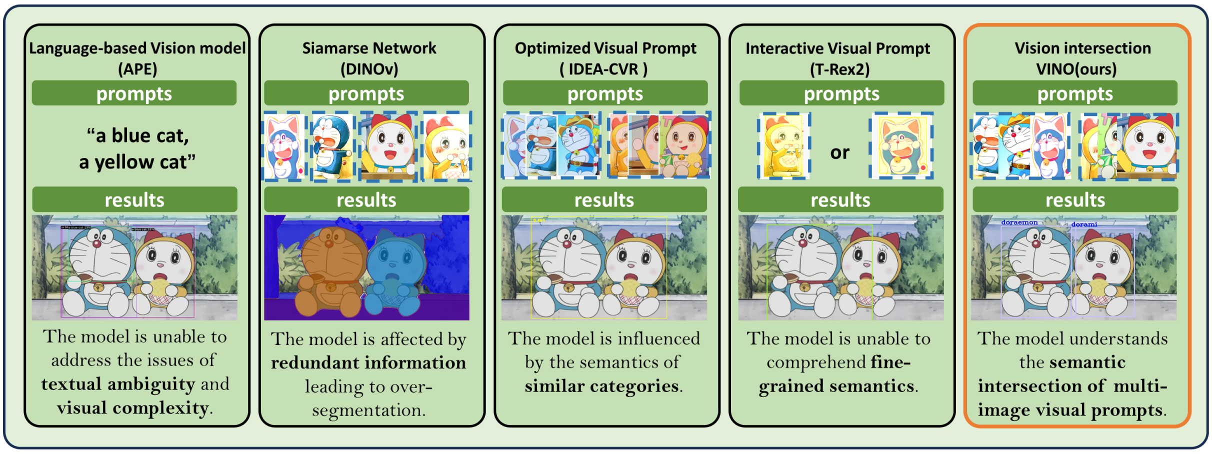

Figure 2: Comparison of various object detection models under visual and textual prompts. The figure highlights the challenges faced by existing models. In contrast, Vision Intersection Network (VINO) effectively addresses these challenges by leveraging the semantic intersection of multi-image visual prompts, enhancing detection accuracy and generalization in open-set environments.

tured through visual representations, spanning from coarse-grained categories (e.g., dog) to fine-grained distinctions (e.g., Corgi). By leveraging the semantic intersection of corresponding categories, we can circumvent the limitations of linguistic descriptions, cross-modal fusion, and single-image interaction, while naturally accommodating multi-granular object discrimination. An image is worth a thousand words. More images say more. Visual representations inherently contain richer semantic information than textual descriptions, particularly for fine-grained object recognition. As illustrated in Figure 2, our approach achieves fine-grained detection through detailed visual prompts, distinguishing it from previous methods.

To realize this vision, we propose VINO (Visual Intersection Network for Open Set Object Detection), a novel region classifier architecture that preserves visual information. At its core, we design a multi-image visual bank to maintain category semantic information across multiple time steps. However, limited images pose challenges in comprehensively describing target objects, and static object semantics during training can lead to overfitting. To address this, we introduce a novel mechanism for updating multi-image prompts, ensuring semantic quality and discriminability through careful image selection.

To enhance semantic matching capabilities and balance the disparity between inference ($<$10) and training ($>$ 1k) visual prompt numbers, we implement a simple yet effective replacement strategy. Our experiments demonstrate that this approach significantly improves semantic matching capability, achieving a 5.5-point improvement on Objects365v1. Furthermore, to minimize the feature discrepancy between CLIP-extracted small image features and EVA-CLIP-extracted large image features, we design a multi-stage fusion mechanism that facilitates effective integration of visual prompts and target image features.

By pre-training on the Objects365v1, ODinW-35 and LVIS datasets, VINO has achieved performance comparable to existing vision-language and vision-vision methods. To verify the general applicability of semantic intersections in enhancing label semantics, we added a segmentation head to the model. By pre-training VINO on the COCO dataset, the segmentation results are comparable to current methods, demonstrating the broad applicability of semantic intersections in visual tasks. In summary, our contributions are as follows:

- We pioneer the learning of semantic intersections from multiple images for OSOD, moving beyond traditional single-image or language-based representations. Our approach demonstrates its broad applicability across various visual tasks, as validated through extensive experiments including object detection and segmentation. This represents a fundamental shift in how semantic information is captured and utilized in open-set scenarios.
- We propose VINO (Visual Intersection Network for Open Set Detection). On the visual

prompt side, we construct a multi-image visual bank with a novel update mechanism to maintain and refine semantic information across time steps, on the target image side, we design a multi-stage fusion mechanism to effectively bridge the feature gap between visual prompts and target objects, facilitating robust semantic matching.

- We conduct extensive experiments and visualization analyses, demonstrating our model's ability to handle open-set object detection tasks. Specifically, VINO achieved an AP of 38.1 on Obj365 v1 , 29.2 on the LVIS v1 validation set, and 24.5 on the ODinW-35 validation set, comparable to current vision-language and vision-vision methods.

2 Related Work

2.1 Open-Vocabulary Object Detection

With the emergence of large pre-trained vision-language models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), methods based on vision and language (Kamath et al., 2021; Zhang et al., 2023) have gained significant popularity in the field of open-vocabulary object detection (OVOD). These methods locate objects using language queries while effectively handling open-set problems. OV-DETR is the first end-to-end Transformer-based open-vocabulary detector, combining CLIP embeddings from both images and text as object queries for the DETR decoder. GLIP treats object detection as a grounding problem and achieves significant success by semantically aligning phrases with regions. To address the limitations of single-stage fusion in GLIP, Grounding DINO (Liu et al., 2024) enhances feature fusion at three stages: neck, query initialization, and head phases, thus tackling the issue of incomplete multimodal information fusion. Furthermore, APE (Shen et al., 2023) scales the model prompts to thousands of category vocabularies and region descriptions, significantly improving the model's query efficiency for large-scale textual prompts. The language-based models aim to enhance the semantic description of language queries to adapt to various visual environments, achieving remarkable progress in zero-shot and few-shot settings. However, relying solely on text poses limitations due to language ambiguity and potential mismatches between textual descriptions and complex visual scenes. This underscores the ongoing need for improved integration of visual inputs to achieve more accurate and comprehensive results. Recent advancements suggest that incorporating richer visual prompts and enhancing multimodal fusion techniques are crucial for overcoming these challenges and pushing the boundaries of OVOD further.

2.2 Object Detection by Visual Queries

Building on language-based object detectors, some methods (Zhou et al., 2022a,b) have introduced visual elements to enhance detection accuracy and semantic richness. MQ-Det utilizes image examples as visual prompts to enhance textual semantics, thereby enabling more effective open-vocabulary object detection (OVOD). However, it remains constrained by textual semantics. Additionally, some methods explore the possibility of object detection using only visual prompts. This approach primarily addresses few-shot object detection and typically employs a two-branch Siamese network. For example, FCT (Han et al., 2022) uses a two-branch Siamese network to process target images and visual queries in parallel, computing the similarity between image regions and a few examples for few-shot object detection. OWL-ViT (Minderer et al., 2022) leverages CLIP's parallel paradigm and uses detection datasets for fine-tuning to adopt image examples for one-shot image-conditioned object detection. Similarly, DINOv expands on this concept by employing visual instructions (such as boxes, points, masks, doodles, and specified regions referencing another image) to handle open-set segmentation. These visual methods often adopt a Siamese network architecture, which has limitations in zero-shot learning capability. To address these limitations and improve semantic understanding, our goal is to learn the semantic intersection of multiple images. VINO enriches visual semantics by retaining semantic information in all time steps using a multi-image visual bank. This approach not only improves the model's ability to understand complex visual scenes but also enhances its robustness and generalization in open-set scenarios.

3 Method

This section presents VINO, our proposed DETR-based detection framework that preserves semantic intersections of visual prompts across temporal steps. By learning to match region features with semantic intersections derived from multiple

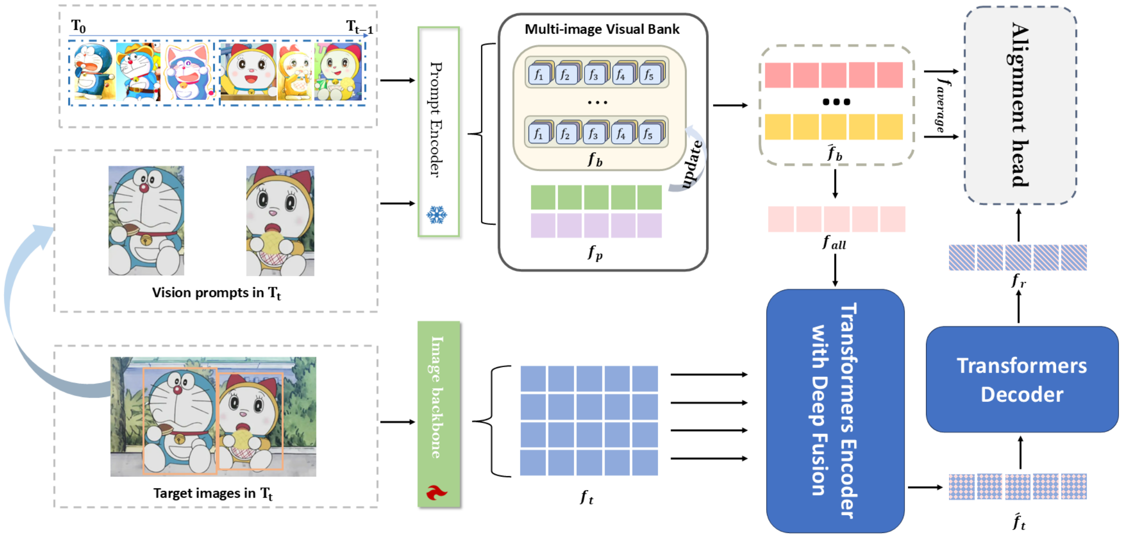

Figure 3: The model architecture of VINO with multi-image visual bank.

images, our approach enhances detection performance through improved category discrimination. We begin by introducing the cornerstone of our architecture, the multi-image visual bank, which serves as the fundamental building block for semantic intersection learning This is followed by a detailed overview of the overall architecture of VINO as Fig 3.

3.1 Multi-image Visual Bank

Rethinking Features in the Multi-image Visual Bank: Our approach addresses the limitations of single-timestep visual instructions in capturing comprehensive category semantics. To aggregate features across multiple timesteps, we construct a feature bank that preserves temporal semantic information. However, as instances of the same category accumulate, maintaining semantic representations for all categories becomes impractical due to memory constraints. A straightforward FIFO (first-in, first-out) approach would result in the loss of valuable semantic information from previous timesteps, compromising the integrity of category descriptions over time.

To overcome this challenge, we introduce a multi-image update mechanism that efficiently compresses and preserves critical semantic information across temporal steps while optimizing memory utilization. Leveraging the categorical distinctions within our Multi-image Visual Bank, our approach naturally facilitates multi-granular category discrimination through semantic intersection learning. While our visual prompts utilize ROI features, the framework remains compatible with investigated interactive visual instructions. Indeed, current interactive approaches can be viewed as special cases of our framework, equivalent to FIFO updates with a prompt number of one. Ablation studies demonstrate that our Multi-image Visual Bank effectively addresses the limitations inherent in single-timestep approaches.

Initialization and Updating of the Multi-image Visual Bank: During initialization, all entries in the multi-image visual bank are set to zero. Formally, the multi-image visual bank is represented as $f_b = (f_{I_1}, f_{I_2}, \ldots, f_{I_N})$, where $f_{I_i} \in R^{n \times d}$, $|I_N|$ represents the number of categories, n is the number of visual prompts, and d is the dimension of the visual features. This initial state ensures a clean slate, ready to incorporate meaningful features as they are processed. When new features f_p are received, they are integrated into the corresponding f_{I_i} based on their category I_i. The integration process(as Algorithm **??**) is carefully designed to ensure efficient and effective updating of the visual bank while maintaining the semantic intersections of each category.

Direct Replacement of Zero Entries: If any sub-feature in f_{I_i} is zero, it indicates that this slot is currently unused. The new feature f_p is directly placed into this slot, ensuring all slots are utilized as new data arrives.

Similarity-Based Updating: If all sub-features in f_{I_i} are non-zero, a more efficient approach is required to integrate the new feature without losing valuable information from previous time steps. To achieve this, we calculate the cosine similarity between f_p and each sub-feature in f_{I_i}. The cosine similarity s_m for the m-th sub-feature is computed as:

$$s_m = \cos(f_p, f_{[I_i,m]}) \quad m \in [1, n]. \tag{1}$$

This step identifies the sub-feature that is most similar to the new feature, indicating redundancy or relevance in the semantic space.

Averaging and Updating: Once the sub-feature with the highest cosine similarity is identified (denoted as $k = \arg\max(s_m)$), we update this sub-feature by averaging it with the new feature f_p.

$$\hat{f}_{[I_i,k]} = \text{average}(f_p, f_{[I_i,k]}). \tag{2}$$

This averaging process helps in retaining both the new and existing semantic information, thereby preserving temporal context and reducing noise.

To address the significant disparity between training and inference scenarios where category labels can number in the thousands during training but are limited to dozens or even single digits during inference, we implement an adaptive replacement strategy. Specifically, when the number of elements received by f_{I_i} exceeds a predetermined threshold, we directly substitute $f_{[I_i,k]}$ with f_p. This dynamic replacement mechanism ensures continuous evolution of category features during training, encouraging the model to learn semantic alignment capabilities between visual prompts and target images, rather than merely developing fixed closed-set classification abilities against static visual prompts.

3.2 The framework of VINO

Our model architecture comprises several key components designed to facilitate effective open-set object detection. Given a target image I_t, the framework incorporates: (1) the Image Backbone, a visual encoder that extracts rich feature representations from the target image; (2) the Prompt Encoder, which processes and encodes visual prompts; and (3) the Multi-image Visual Bank, a sophisticated memory mechanism that maintains visual prompt information for each category and synthesizes their semantic intersections. The architecture is further enhanced by (4) the DETR Encoder, which facilitates feature fusion between visual prompts and target images, and (5) the DETR Decoder, which identifies and localizes proposed regions while extracting their semantic information. Through aligning the semantic content of proposed regions with the synthesized semantic intersections from visual prompts, our model effectively assigns categorical labels to each detected region.

Specifically, the model takes the target image $I_t \in R^{3\times h\times w}$ and the set of labels $R = \{r_1, r_2, \ldots, r_{|R|}\}$ as input. Here, $r_i = (x_1, y_1, x_2, y_2, I_i) \in R^5$ represents the coordinates of the top-left and bottom-right corners, along with the corresponding category label.

Feature Extraction and Region Proposal: For the target image I_t, the initial step involves feature extraction using the Image Backbone to produce the feature representation f_t: $f_t = \text{Image Backbone}(I_t)$, where $f_t \in R^{bs\times D}$, with bs representing the batch size and D denoting the dimensionality of the feature vectors.

To facilitate effective semantic fusion between target images and visual prompts, we introduce a multi-stage fusion mechanism. The process begins by computing a consolidated visual prompt representation f_{all} through averaging f_b across both quantity and category dimensions. We then implement a cross-attention mechanism where this aggregated representation f_{all} serves as the query, while the target image features f_t act as both key and value matrices. This cross-modal interaction is followed by a self-attention operation on f_t, yielding refined feature representations $\hat{f}_t$. Finally, we select the top-k elements from $\hat{f}_t$ based on feature magnitude, which serve as learnable tokens for the subsequent DETR Decoder stage in object detection.

The DETR-like decoder operates by decoding the features $\hat{f}_t$ into two outputs: the coordinates of the proposed regions $bbox \in R^{bs\times k\times 4}$ and the corresponding feature representations of these proposed regions $f_r \in R^{bs\times k\times D}$.To further validate the broad applicability of semantic intersections in visual tasks, we extend the model by incorporating a segmentation head. This addition allows the model to also output predicted masks $M \in R^{bs\times k\times h\times w}$.

Feature Fusion: For the set of labels $R = \{r_1, r_2, \ldots, r_{|R|}\}$, we first use the Prompt Encoder to extract the features from each region: $f_p = \text{Prompt Encoder}(R, I_t)$.

Next, we perform feature fusion by updating the multi-image visual bank $\hat{f}_{[I_i,k]}$ with the features extracted from the regions, aligning them with the same category in the visual bank, as described in the previous section. This fusion process integrates the new region features into the existing visual bank, ensuring that the updated bank retains and reflects the latest semantic information.

After the fusion, we average and align the dimensions to obtain the final average feature representation $f_{average}$:

$$f_{average} = \text{MLP}(\text{Average}(\hat{f}_{[I_i,k]})) \quad (3)$$

Label Assignment:Finally, we use the Alignment Head to match the features of the proposed regions f_r with the averaged features $f_{average}$ to determine the semantic labels:

$$I_{results} = \text{Softmax}(f_r @ f_{average}^T) \quad (4)$$

This step outputs $I_{results} \in R^{bs \times k \times |I_N|}$, assigning the most probable semantic labels to each proposed region.

Training Objective: Our model employs a unified loss function that accommodates both detection and segmentation tasks, with segmentation loss defaulting to zero when no segmentation task is present. The total loss function comprises classification, localization, and segmentation components, formulated as:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{class}} + \mathcal{L}_{\text{bbox}} + \mathcal{L}_{\text{giou}}}_{\text{encoder and decoder}} + \underbrace{\mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{dice}}}_{\text{last layer of decoder}} \quad (5)$$

where $\mathcal{L}_{\text{class}}$ employs Focal Loss to align the fused features of visual prompts with target image encodings. The localization component consists of $\mathcal{L}_{\text{bbox}}$ and $\mathcal{L}_{\text{giou}}$, utilizing L1 loss and Generalized IoU loss respectively for bounding box regression. For mask segmentation, $\mathcal{L}_{\text{mask}}$ and $\mathcal{L}_{\text{dice}}$ implement cross-entropy loss and dice loss respectively.

4 Experiments

4.1 Setup

Dataset and Settings. To evaluate our model's performance in open-set detection, we develop VINO-D, which is pre-trained on three large-scale datasets: COCO (Lin et al., 2015) (80 categories, 110K images), LVIS (Gupta et al., 2019) (1,203 categories, sharing images with COCO), and Objects365v1 (Dong et al., 2024) (365 categories, 600K images). The model is evaluated on ODinW35 (Li et al., 2022a), a collection of 35 diverse datasets specifically designed to assess model performance in real-world scenarios. To investigate the broader applicability of semantic intersections, we extend our framework to segmentation tasks by developing VINO-S with an additional segmentation head. VINO-S is pre-trained for both open-set detection and segmentation on the COCO dataset (110K images with object detection and panoramic segmentation annotations) and evaluated on the LVIS v1 validation set for both detection and segmentation tasks.

Training Details. Both VINO-D and VINO-S architectures incorporate APE-D weights for target image processing, with ViT-L as the backbone architecture. We employ a frozen CLIP-L model as the prompt encoder. The frameworks are configured with 5 prompts and 900 object queries. Model training is conducted on 2 $\times$ A100 GPUs with a batch size of 12, utilizing the AdamW optimizer with a learning rate of 5e-5. Both variants complete one epoch of training on their respective datasets. To mitigate the significant domain shift introduced by the prompt encoder processing cropped images (Li et al., 2023), we implement strict resolution controls for visual prompts: maintaining a minimum resolution of 2000 pixels for the initial prompt image and 1600 pixels for subsequent visual prompts.

4.2 Results on detection and segmentation

4.2.1 Object Detection

In **Table 1**, we present the detection results of our VINO-D model, which achieves comparable performance across the evaluated datasets(Du et al., 2024). Specifically, VINO-D attains an AP^b of 43.6 on the Objects365 dataset, 47.8 on LVIS v1 validation, and 24.5 on ODinW35.

When compared with current vision-language models such as GLIP and UNINEXT(Yan et al., 2023), VINO-D demonstrates highly competitive results. For instance, while GLIP achieves strong AP^b on Objects365 by leveraging language queries, VINO-D performs exceptionally well using vision-based queries, highlighting its capacity to learn robust semantic intersections from multiple images. This ability to model semantic intersections allows VINO-D to maintain high detection accuracy without relying on textual input, further showcasing its robustness in vision-dominated tasks.

In comparison with other vision-only methods, VINO-D significantly surpasses DINOv(L) by 8.8 points and MQ-GLIP-L by 0.6 points in terms of AP^b on the ODinW35 dataset. DINOv(L) emphasizes the challenges posed by domain shifts, partic-

Method	Backbone	Semantic Data	Type	objects365 AP^b	LVIS v1 val AP^b	Odinw35 val AP^b_{average}
GLIP	Swin-L	FourODs+...	Text Open-set	36.2	26.9	23.4
UNINEXT	ViT-H	O365v2+COCO+...	Text Open-set	23	14	-
OpenSeeD-L	Swin-L	O365v2+COCO+...	Text Open-set	-	23	15.2
MQ-GLIP-L	Swin-L	O365	Text and visual	-	34.7	23.9 (3-shot)
LaMI-DETR	ConVNext-L	object365+VG	Text and visual	21.9	41.3	-
DINOv (L)	Swin-L	SAM+COCO+...	Visual Prompt	-	-	15.7
VINO-D(ours)	ViT-L	COCO+O365+LVIS	Visual Prompt	43.6	47.8	24.5

Table 1: Open-set segmentation results for different methods."–" indicates that the work does not have a reported number.

Method	Backbone	Semantic Data	Type	COCO		LVIS v1 val
				AP^b	AP^m	AP^m
GLIPv2	Swin-H	O365+COCO+...	Text Open-set	64.1	47.4	-
UNINEXT	ViT-H	O365v2+COCO	Text Open-set	60.6	51.8	12.2
APE (D)	ViT-L	O365v2+COCO+...	Text Open-set	58.3	49.3	53
DINOv (L)	Swin-L	COCO+SA-1B	Visual Prompt	54.2	50.4	-
DINO-X Pro	ViT-L	Grounding-100M	visual Prompt	56.0	37.9	38.5
VIOSD-S(ours)	ViT-L	COCO+LVIS	Visual Prompt	62.9	53.7	41.4

Table 2: Open-set segmentation results for different methods."–" indicates that the work does not have a reported number.

ularly those arising from differences in resolution between cropped and target images. However, our approach addresses this issue by effectively learning multi-step semantic intersections across multiple images. On the other hand, while MQ-GLIP-L employs visual prompts to enhance text-based representations, its reliance on textual semantics introduces constraints, as evidenced by its lower performance in the 3-shot setting when compared to our zero-shot results.

4.2.2 Object Segmentation

In **Table 2**, we present the segmentation results of our VINO-S model, designed with an integrated segmentation head. VINO-S achieves an AP^m of 53.7 on the COCO dataset, outperforming UNINEXT by 1.9 points and DINOv(L) by 3.3 points, thereby achieving comparable or superior performance relative to current leading vision-language and vision-only methods. On the LVIS v1 validation set, VINO-S achieves an AP^m of 41.4, significantly outperforming UNINEXT and delivering results comparable to advanced models such as DINO-X Pro and APE.

These results underscore the effectiveness of our model design, where the semantic intersections enabled by the multi-image visual bank yield substantial improvements for segmentation tasks. Overall, the VINO framework demonstrates its capability to advance both object detection and segmentation by leveraging robust multi-image visual representations without relying heavily on external text-based prompts, bridging the gap between vision-dominated tasks and real-world deployment scenarios.

4.3 Ablations Experiments

Type	
Prompt Num	**AP^b on COCO**
1(FIFN)	53.72
1	62.61
5	62.73
10	62.75
20	62.86
Update Mechanisms	**AP^b on COCO**
FIFN	55.64
Average(no update)	60.92
Average(with update)	62.73
Updata Threshold	**AP^b on COCO**
50	62.81
100	62.73
200	62.50
None	60.92
Reduce Visual Tokens	**AP^b on COCO lvis**
MLP	63.03 6.63
Sliding Convolution	62.81 8.44
Average	62.73 9.92

Table 3: Ablations Experiments

The ablation experimental results on the COCO dataset after one round of fine-tuning are shown in Table 3. Through our ablation studies, we investigated several crucial aspects:

Single-Image vs. Multi-Image Visual Interac-

tion: With vision prompts number set to 1 and FIFN update strategy, the model is limited to single-image visual interaction, resulting in the lowest AP. The introduction of averaging mechanism breaks through the single-image limitation, significantly enhancing detection performance. However, after adopting the averaging strategy, increasing the number of vision prompts (from 1 to 20) only yields a marginal improvement of 0.27 in AP^b.

Visual Semantic Enhancement: With vision prompts number fixed at 5, although FIFN strategy overcomes the single-image constraint, it underperforms in semantic fusion, showing a 5.28-point decrease compared to Average(no update) strategy. Without an update mechanism, the continuous accumulation of vision prompts leads to excessive semantic similarity in the multi-image vision bank, compromising the model's discriminative ability. The update mechanism effectively addresses this issue, transitioning the model from simple semantic approximation to more precise semantic alignment.

Impact of Visual Semantic Redundancy: The experiments demonstrate a consistent performance degradation as the update cycle decreases from 50 to no update mechanism. While moderately reducing multi-image visual semantic redundancy can enhance model performance, excessive reduction (as in FIFN strategy) proves detrimental. Our findings suggest that maintaining moderate semantic variation rates while keeping low semantic similarity is crucial for improving detection performance.

Visual Tokens Compression Mechanism: The compression of visual tokens has garnered significant attention across various domains. In our work, visual token compression is specifically implemented during the feature fusion process of multiple instances within the same category in the multi-image visual bank. We conducted experiments involving one epoch of training on the COCO dataset, followed by zero-shot evaluation on the LVIS v1 validation set. While sophisticated mechanisms such as MLP and Sliding Convolution enhance model alignment capabilities, they significantly compromise the model's zero-shot generalization ability. Notably, the simple yet efficient feature averaging strategy demonstrates superior performance in preserving semantic information, suggesting that architectural simplicity can often lead to more robust and generalizable solutions in visual feature fusion tasks.

Figure 4: The Visualization of VINO-D.

4.4 Visualization

The qualitative results presented in Figure 4 demonstrate our model's effectiveness across diverse scenarios. The examples showcase: (1) accurate single-prompt detection capabilities, (2) robust multi-instance detection across various categories, and (3) precise discrimination between semantically similar categories.

5 Conclusion

By dynamically integrating and updating multi-image visual prompts, VINO not only addresses the limitations associated with textual descriptions and single-image interaction but also effectively narrows the contextual gap between cropped and full images. This ongoing refinement of feature representations ensures that VINO adapts flexibly to new information, achieving robust generalization capabilities even with unseen objects. Experimental results show that VINO exhibits strong performance in open set object detection, achieving results comparable to current vision-language and vision-only methods. We hope that more studies will explore the application of semantic intersections in visual tasks, further expanding the capabilities and understanding of visual models in diverse environments.

References

Bingcheng Dong, Yuning Ding, Jinrong Zhang, Sifan Zhang, and Shenglan Liu. 2024. More pictures say more: Visual intersection network for open set object detection. *Preprint*, arXiv:2408.14032.

Penghui Du, Yu Wang, Yifan Sun, Luting Wang, Yue Liao, Gang Zhang, Errui Ding, Yan Wang, Jingdong Wang, and Si Liu. 2024. Lami-detr: Open-vocabulary detection with language model instruction. *Preprint*, arXiv:2407.11335.

Agrim Gupta, Piotr Dollár, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. *Preprint*, arXiv:1908.03195.

Guangxing Han, Jiawei Ma, Shiyuan Huang, Long Chen, and Shih-Fu Chang. 2022. Few-shot object detection with fully cross-transformer. *Preprint*, arXiv:2203.15021.

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

Qing Jiang, Feng Li, Zhaoyang Zeng, Tianhe Ren, Shilong Liu, and Lei Zhang. 2024. T-rex2: Towards generic object detection via text-visual prompt synergy. *Preprint*, arXiv:2403.14610.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790.

Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. 2019. Few-shot object detection via feature reweighting. *Preprint*, arXiv:1812.01866.

Feng Li, Qing Jiang, Hao Zhang, Tianhe Ren, Shilong Liu, Xueyan Zou, Huaizhe Xu, Hongyang Li, Chunyuan Li, Jianwei Yang, Lei Zhang, and Jianfeng Gao. 2023. Visual in-context prompting. *Preprint*, arXiv:2311.13601.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. 2022a. Grounded language-image pre-training. *Preprint*, arXiv:2112.03857.

Xiangyu Li, Xu Yang, Kun Wei, Cheng Deng, and Muli Yang. 2022b. Siamese contrastive embedding network for compositional zero-shot learning. *Preprint*, arXiv:2206.14475.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

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. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *Preprint*, arXiv:2303.05499.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. 2022. Simple open-vocabulary object detection with vision transformers. *Preprint*, arXiv:2205.06230.

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. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, Xingyu Chen, Zhuheng Song, Yuhong Zhang, Hongjie Huang, Han Gao, Shilong Liu, Hao Zhang, Feng Li, Kent Yu, and Lei Zhang. 2024. Dino-x: A unified vision model for open-world object detection and understanding. *Preprint*, arXiv:2411.14347.

Yunhang Shen, Chaoyou Fu, Peixian Chen, Mengdan Zhang, Ke Li, Xing Sun, Yunsheng Wu, Shaohui Lin, and Rongrong Ji. 2023. Aligning and prompting everything all at once for universal visual perception. *Preprint*, arXiv:2312.02153.

Yifan Xu, Mengdan Zhang, Chaoyou Fu, Peixian Chen, Xiaoshan Yang, Ke Li, and Changsheng Xu. 2023. Multi-modal queried object detection in the wild. *Preprint*, arXiv:2305.18980.

Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. 2023. Universal instance perception as object discovery and retrieval. *Preprint*, arXiv:2303.06674.

Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. 2022. *Open-Vocabulary DETR with Conditional Matching*, page 106–122. Springer Nature Switzerland.

Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. 2023. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1020–1031.

Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. 2022. Glipv2: Unifying localization and vision-language understanding. *Preprint*, arXiv:2206.05836.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. 2022. Generalized decoding for pixel, image, and language. *Preprint*, arXiv:2212.11270.

A Limitations

A1. Did you describe the limitations of your work?: Yes, I reported the gap between it and the Vision language paradigm in the experiment.

A2. Did you discuss any potential risks of your work?: No.

B1. Did you cite the creators of artifacts you used?: Yes.

B2. Did you discuss the license or terms for use and / or distribution of any artifacts?: No, We build our model based on open-source models.

B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?: Yes,in section 1.

B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it? : No, We use open dataset.

B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?: No, we don't need to use it.

B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created?: Yes, in section 4.

C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?: Yes, in section 4.

C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?: Yes, in section 4.

C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run? : Yes, in section 4.

C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, Spacy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?: Yes, in section 4.

D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?: No. We don't need to use it.

D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?: No. We don't need to use it.

D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating?: No. We don't need to use it.

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?: No. We don't need to use it.

D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?: No. We don't need to use it.

E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?: No.