OV-VOD: Open-Vocabulary Video Object Detection

Zhihong Zheng
Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University
Xiamen, China
Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University
Xiamen, China
zhengzhihong@stu.xmu.edu.cn

Yang Cao
Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University
Xiamen, China
Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University
Xiamen, China
yangcao@stu.xmu.edu.cn

Junlong Gao
Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University
Xiamen, China
Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University
Xiamen, China
jlgao@xmu.edu.cn

Hanzi Wang*
Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University
Xiamen, China
Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University
Xiamen, China
hanzi.wang@xmu.edu.cn

Abstract

Traditional Video Object Detection (VOD) is limited by pre-defined closed-set categories, restricting its ability to detect novel objects in real-world scenarios. To address this limitation, we make three key contributions. First, we formally define *Open-Vocabulary Video Object Detection* (Open-Vocabulary VOD) as the task of detecting objects in video streams from open-set categories, including novel categories unseen during training. Second, we establish an evaluation benchmark by utilizing existing datasets (LV-VIS, BURST, and TAO) to bridge the data gap for this new task. Third, we propose OV-VOD, an Open-Vocabulary VOD method that detects objects in videos beyond pre-defined training categories and addresses the shortcomings of image-level open-vocabulary detectors, which generally neglect the essential temporal and spatial information. Specifically, we design a Semantic-Presence Memory Tracking (SPMT) module that propagates object features across frames through a memory bank to leverage temporal consistency. Moreover, we propose a Spatial Object Relationship Distillation loss ($\mathcal{L}_{SR}$) that captures inter-object spatial dependencies and enhances knowledge transfer during feature distillation. Experiments on multiple video datasets demonstrate that our OV-VOD exhibits superior zero-shot generalization capability compared to existing image-level open-vocabulary object detection methods.

CCS Concepts

• **Computing methodologies** → **Object detection**.

Keywords

Open-Vocabulary; Video Object Detection; Memory Tracking; Relationship Distillation

ACM Reference Format:
Zhihong Zheng, Yang Cao, Junlong Gao, and Hanzi Wang. 2025. OV-VOD: Open-Vocabulary Video Object Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755402

1 Introduction

Although traditional video object detection (VOD) methods [5, 22, 25, 31, 33, 37, 38, 51, 61, 64, 65] have made significant progress over the years, they remain fundamentally constrained by their reliance on a fixed set of training categories. This closed-set paradigm limits their ability to generalize to novel concepts and recognize previously unseen categories, thereby hindering their applicability in real-world scenarios where new object categories frequently appear [52]. This limitation is a key factor contributing to the difficulty of deploying traditional VOD methods in practical settings. To address this issue, recent research has focused on open-vocabulary object detection [6, 11, 13, 15, 35, 44, 49, 50, 53, 58, 62], which aims to detect and classify all objects in an image without being restricted

*Corresponding author: Hanzi Wang (hanzi.wang@xmu.edu.cn)

MM '25, October 27–31, 2025, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755402

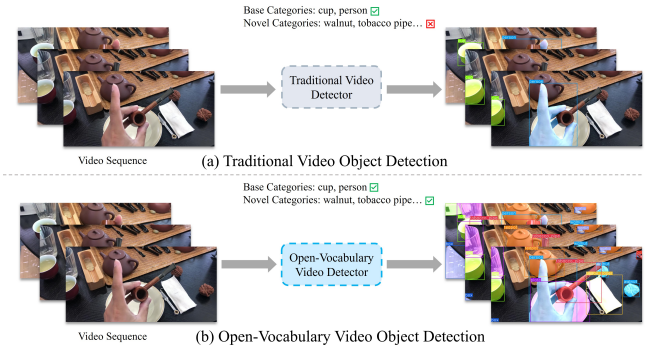


Figure 1: (a) Traditional VOD can only detect objects from training categories (e.g. cup and person); (b) Open-Vocabulary VOD aims to detect and classify both training categories and novel categories (unseen during training, e.g., walnut and tobacco pipe). Different colors in the figure represent different object instances.

to a pre-defined label set. However, most existing open-vocabulary object detection methods are designed for static images and fail to exploit the rich temporal information available in videos. Furthermore, they do not adequately address video-specific challenges such as object occlusion, atypical poses, and motion blur, which are either absent or significantly less severe in image-based datasets. To bridge this gap, we propose the task of *Open-Vocabulary Video Object Detection*, which seeks to detect and classify objects from an open set of categories across video frames. An illustration of this task is provided in Figure 1.

Accurate benchmarking of Open-Vocabulary VOD methods necessitates a video dataset that includes a large and diverse set of object categories. However, existing datasets commonly used in traditional VOD, such as ImageNet VID [41] and EPIC-KITCHENS-55 [8], lack sufficient category diversity, as illustrated in Table 1. This deficiency poses a significant barrier to the advancement of open-vocabulary techniques in the video domain, due to the absence of datasets designed for open-vocabulary video tasks.

To address this limitation, we draw inspiration from the fields of open-vocabulary tracking [32] and open-vocabulary video instance segmentation [7, 12, 48, 57, 63], adapting open-vocabulary video datasets from these domains to the detection setting. This cross-domain transfer is highly feasible, as the annotations used in tracking and segmentation tasks can be readily converted into the bounding boxes required for object detection. Specifically, we utilize one open-vocabulary video instance segmentation dataset, LV-VIS [48], and two open-vocabulary tracking datasets, BURST [1] and TAO [9], to benchmark Open-Vocabulary VOD methods. Among these, the LV-VIS dataset is particularly well-suited for evaluating the generalization ability of Open-Vocabulary VOD models to novel categories, as it includes a large number of object classes, most of which differ from those in widely used datasets such as MS-COCO [34] and LVIS [17]. Therefore, we adopt LV-VIS as the primary benchmark for assessing detection performance.

Table 1: Comparison of key characteristics between datasets in this paper and the traditional video object detection datasets. VID refers to the ImageNet VID dataset, while EPIC-55 refers to the EPIC-KITCHENS-55 dataset.

Dataset	VID [41]	EPIC-55 [8]	TAO [9]	BURST [1]	LV-VIS [48]
Videos	4417	272	1488	2914	4828
Frames	1298k	174k	51378	16089	25588
Instances	2005k	326k	168k	600k	544k
Categories	30	295	363	482	1196

A straightforward strategy for Open-Vocabulary VOD is to treat each video frame as an independent image and apply existing open-vocabulary object detection methods on a frame-by-frame basis. However, such image-level methods ignore the temporal information inherent in videos and fail to exploit inter-frame correlations. Moreover, video-specific challenges such as variations in object appearance and quality degradation caused by motion blur or occlusion further compromise the effectiveness of this naive method and lead to suboptimal detection performance.

In this paper, we propose the first *Open-Vocabulary Video Object Detection* method, termed OV-VOD. To fully leverage the temporal information inherent in videos, we introduce a Semantic-Presence Memory Tracking (SPMT) module. By storing features in a memory bank and employing the Hungarian algorithm with an update factor, the module tracks objects across frames, effectively mitigating performance degradation caused by object disappearance, motion blur, or occlusion. Furthermore, to harness spatial contextual information and enhance knowledge transfer from pre-trained Vision-Language Models (VLMs) during distillation, we propose a novel Spatial Object Relationship Distillation loss ($\mathcal{L}_{SR}$). Inspired by similarity-based distillation techniques [14, 47], our method incorporates the spatial relationships among proposal-level features within the same frame as an additional constraint, thereby improving the effectiveness of the distillation process.

Our OV-VOD model is trained on the LVIS dataset and evaluated on three challenging video benchmarks: LV-VIS, BURST, and TAO. Without any dataset-specific fine-tuning, extensive experiments show that OV-VOD consistently outperforms state-of-the-art image-level open-vocabulary object detection methods in zero-shot generalization to novel object categories unseen during training.

In summary, the contributions of this paper are as follows:

- We define the task of Open-Vocabulary VOD, establish an evaluation benchmark, and introduce the first Open-Vocabulary VOD method, OV-VOD, which extends the traditional closed-set VOD framework to an open-set paradigm.
- We develop a Semantic-Presence Memory Tracking (SPMT) module that effectively leverages temporal information in videos, mitigating performance degradation caused by object disappearance, blur, or occlusion.
- We propose a Spatial Object Relationship Distillation loss ($\mathcal{L}_{SR}$), which captures spatial contextual relationships among proposal-level features within the same frame, facilitating more effective knowledge transfer during distillation.
- Experimental results demonstrate that OV-VOD respectively achieves AP_n of 12.8% on LV-VIS, 4.9% on BURST, and 5.0% on TAO,

surpassing existing image-level open-vocabulary object detection methods.

2 Related Work

2.1 Open-Vocabulary Object Detection

Open-Vocabulary Object Detection has progressed rapidly with the emergence of various large-scale models. ViLD [15] is the first approach to transfer knowledge from visual-language models (VLMs) [27, 39, 46, 56] into closed-set detectors [4, 23, 40] using a knowledge distillation framework. It introduces separate image and text branches to align visual features with the extensive textual information learned during the VLMs' pre-training stage, thereby pioneering the field of open-vocabulary object detection. DetPro [11] extends ViLD by integrating contextual learning [59, 60] to adapt static prompts to task-specific contexts, thereby enhancing detection performance. RegionCLIP [58] aligns image regions with textual descriptions, leveraging CLIP [39] to generate pseudo-labels and fine-tuning on manual detection datasets. Detic [62] tackles data imbalance in long-tail object detection by leveraging image-level supervision and a joint training strategy to enhance performance on novel categories. Furthermore, OV-DETR [54] introduces region-text alignment and conditional matching to enable end-to-end open-vocabulary object detection using language-based supervision in place of traditional annotations. Recently, BARON [53] enhances region-level alignment by encoding a bag of contextually related regions as textual representations, which are then aligned with visual embeddings from vision-language models.

Although these open-vocabulary object detection methods perform well on images, their direct application to videos is limited by an inability to fully exploit inherent temporal information, potentially yielding suboptimal results.

2.2 Video Object Detection

Video Object Detection typically makes use of abundant temporal information to enhance detection performance. Based on the manner in which temporal information is leveraged, current mainstream video object detection methods can generally be categorized into two groups: pos-processing and feature aggregation methods.

Post-processing methods first use a detector to extract bounding boxes from multiple frames and then employ linking or tracking techniques to connect these boxes into tubelets. For instance, Seq-NMS [22] forms high-confidence bounding box sequences across consecutive frames and utilizes them to enhance weaker detections. T-CNN [31] propagates bounding boxes using optical flow and integrates tracking algorithms to construct extended tubelet sequences. Feature aggregation methods [5, 16, 21, 28, 45, 55, 64] typically enhance the feature representation of the target frame by aggregating useful temporal information from multiple support frames. These methods can be classified into frame-level and proposal-level aggregation, depending on the stage at which features are aggregated. Frame-level aggregation methods, such as FGFA [64], use optical flow networks to guide the aggregation of features across frames. DFF [65] accelerates video inference by applying a large network to sparse key frames and propagating deep features to adjacent frames via optical flow. Early-stage feature aggregation enables end-to-end training but yields limited performance gains, whereas proposal-level methods aggregate features at the proposal stage. For example, SELSA [51] aggregates semantic features from the entire sequence rather than just adjacent frames, while MEGA [5] takes into account both global and local temporal information, utilizing a memory mechanism to aggregate features at the proposal level.

Although these methods achieve remarkable performance in traditional video object detection, their closed-set limitation hampers effective real-world application.

3 Setting of Open-Vocabulary VOD

Task Setting. Given a training dataset $\mathcal{D}_{train}$ consisting of instance-level candidate bounding box annotations for a set of training categories C_{train}, traditional VOD aims to train a model $f_\theta(\cdot)$. This model is designed to be evaluated on a test dataset $\mathcal{D}_{test} = \{\boldsymbol{V}_i\}_{i=1}^{L}$, where L denotes the number of videos, $\boldsymbol{V}_i \in \mathbb{R}^{T_i \times H_i \times W_i \times 3}$ represents a video clip of T_i frames with a spatial resolution of (H_i, W_i). The goal of $f_\theta(\cdot)$ is to predict the bounding boxes $\{\boldsymbol{b}_t\}_{t=1}^{T_i} \in \mathbb{R}^{T_i \times K_t \times 4}$, where K_t represents the total number of objects in the t-th frame, and corresponding class labels $c \in C_{train}(C_{base})$ for all objects in the video that belong to the base categories. Objects belonging to novel categories C_{novel} are ignored.

In contrast, Open-Vocabulary VOD aims to train a model on $\mathcal{D}_{train}$ and test it on $\mathcal{D}_{test}$ for both C_{train} and C_{novel}. Specifically, during inference, given a test video sequence $\boldsymbol{V}_i \in \mathbb{R}^{T_i \times H_i \times W_i \times 3}$, the trained model is expected to predict all object bounding boxes $\{\boldsymbol{b}_t\}_{t=1}^{T_i} \in \mathbb{R}^{T_i \times K_t \times 4}$ and the category label $c \in (C_{train} \cup C_{novel})$ for each bounding box in $\boldsymbol{V}_i$:

$$f_\theta(\boldsymbol{V}_i) = \{\{\hat{\boldsymbol{b}}_k, \hat{c}_k\}_{k=1}^{K_t}\}_{t=1}^{T_i}, \tag{1}$$

where the category $\hat{c}_k$ belongs to the union of the training categories and novel categories. Additionally, $\hat{\boldsymbol{b}}_k = \{x, y, w, h\}$ denotes the bounding box of the k-th object in the t-th frame of the i-th video. In the experimental section, the training categories are referred to as base classes, while the categories that do not overlap with the base classes are referred to as novel classes.

Evaluation Metrics. We follow the standard evaluation setup in MS-COCO [34] and LVIS [17], using Average Precision (AP) to assess the performance of both base and novel categories. Specifically, the average precision for i-th category across all video frames, denoted as AP_i, is defined as the area under the precision-recall curve plotted based on the category confidence scores. The value of AP_i is measured at 10 Intersection-over-Union (IoU) thresholds ranging from 0.5 to 0.95, with a step size of 0.05. Finally, the mean average precision is calculated separately for the base category set and the novel category set, denoted as AP_b and AP_n, respectively.

4 Structure of OV-VOD

After defining the Open-Vocabulary VOD task, we present our proposed method, OV-VOD, as illustrated in Figure 2. Overall, our OV-VOD are based on the existing open-vocabulary object detectors [11, 15] and contains two key improvements: a) To address the limitations of image-level object detection methods, which fail to leverage the rich temporal information inherent in videos, we propose a Semantic-Presence Memory Tracking (SPMT) module. This module enhances detection performance during inference by

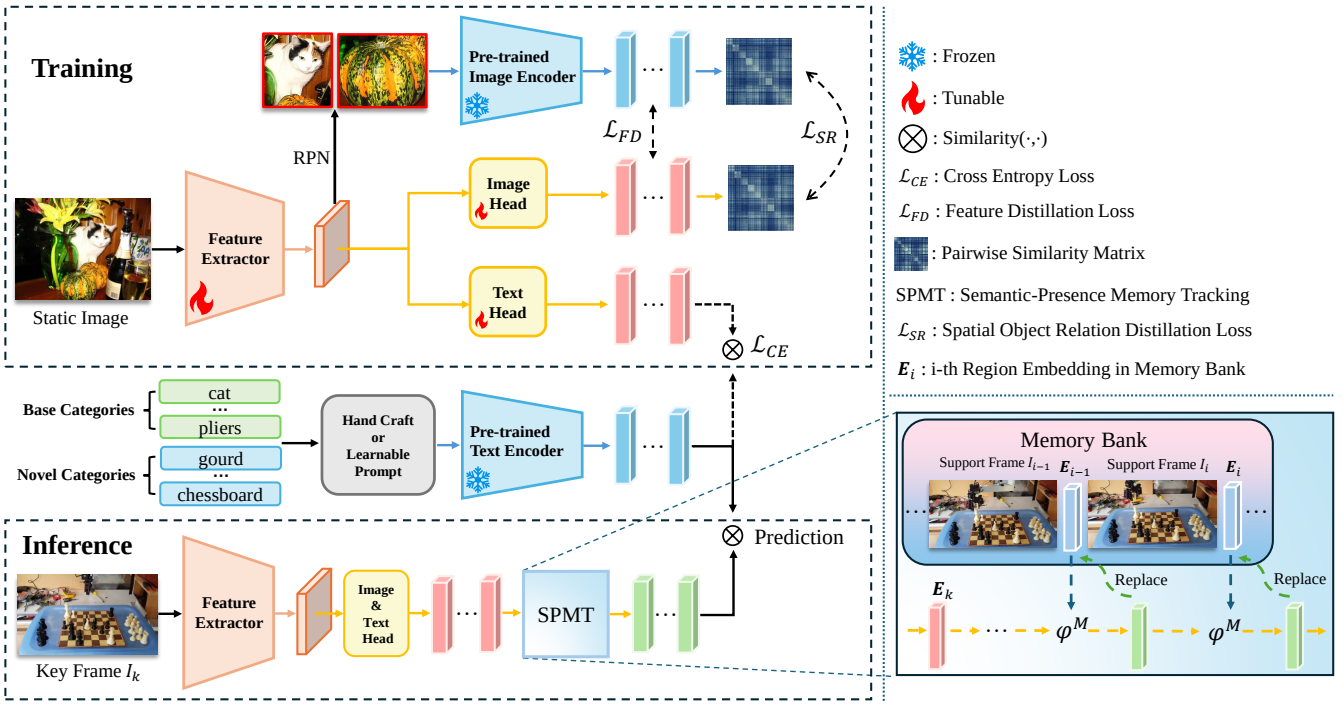


Figure 2: The overall architecture of the proposed OV-VOD. It first performs spatial relationship distillation through our proposed $\mathcal{L}_{SR}$ during training on the image-level dataset, and subsequently conducts inference on video datasets through our introduced Semantic-Presence Memory Tracking (SPMT) module.

tracking objects from support frames stored in a memory bank, utilizing the Hungarian algorithm and an adaptive update factor; b) In addition to conventional feature distillation, we introduce a Spatial Object Relationship Distillation loss ($\mathcal{L}_{SR}$), which facilitates more effective knowledge transfer from VLMs during training, thereby improving the model's generalization to unseen categories. Further details of these contributions are provided in the following sections.

4.1 Semantic-Presence Memory Tracking

Inspired by MinVIS [26], we adopt region embeddings stored in a memory bank to establish inter-frame object associations. While the Hungarian algorithm is employed to match object regions across frames, naive feature updates are inadequate for modeling long-term object dependencies in complex scenarios. OV2Seg [48] introduces an update frequency control factor S^{obj} to mitigate tracking failures caused by severe occlusions or temporary object disappearances. However, this approach remains suboptimal for open-vocabulary VOD, where both semantic confidence and bounding box quality play critical roles in overall detection performance.

As demonstrated in [15], the semantic quality of objects cannot be reliably assessed using semantic confidence scores alone, as accurate localization through bounding boxes is equally essential. This limitation is particularly pronounced in cross-frame object tracking, where the strong generalization ability of VLMs may result in high confidence scores for proposals that include only partial object regions. Such cases introduce false positives and propagate detection errors across successive frames (e.g., the baseline detection result shown in the second frame of Figure 5). To mitigate this issue, we propose a Semantic-Presence Memory Tracking (SPMT) module, including a novel update factor called Semantic-Presence that synergistically integrates semantic confidence evaluation with bounding box quality assessment.

Specifically, we select K support frames, each of which contains N region embeddings, and maintain a set of updated region embeddings $\boldsymbol{E}^{\boldsymbol{M}} \in \mathbb{R}^{K\times N\times d}$ in the memory bank to model dependencies between video objects, where d denotes the dimensionality of the region embeddings, as shown in Figure 2. First of all, we calculate the inner product similarity between region embeddings $\boldsymbol{E}_i^M$ of each frame i and the previous region embeddings $\boldsymbol{E}_{i-1}^M$ of frame $i-1$ in memory. Each region embedding is associated with one region embedding from the previous frame through the Hungarian algorithm results on the similarity matrix [26]. The first frame $\boldsymbol{E}_0^M$ is initialized by the key frame. Subsequently, the update function $\varphi^M(\cdot,\cdot)$ is used to perform updates in the memory bank, gradually tracking video region embeddings to mitigate issues of object disappearance or occlusion. The function $\varphi^M(\cdot,\cdot)$ is defined as:

$$\begin{aligned}\boldsymbol{E}_i^M &= \varphi^M(\boldsymbol{E}_{i-1}^M, \boldsymbol{E}_i^*) \\ &= \alpha\cdot\beta^{\text{obj}}\cdot\boldsymbol{E}_i^* + (1-\alpha\cdot\beta^{\text{obj}})\cdot\boldsymbol{E}_{i-1}^M,\end{aligned} \tag{2}$$

where α is a hyperparameter controlling the frequency of embeddings updates, $\boldsymbol{E}_i^*$ represents the associated region embeddings after applying the Hungarian algorithm. β^{obj} is our introduced

Semantic-Presence factor, which is used to measure the semantic and bounding box quality of targets. β^{obj} is defined as:

$$\beta^{\text{obj}} = w_\beta \times S^{\text{obj}} + (1 - w_\beta) \times O^{\text{obj}}, \tag{3}$$

where S^{obj} represents the confidence score for each region embedding, O^{obj} represents the objectness score for each region, and w_β is the weight controlling the balance between semantic and bounding box quality. By introducing the SPMT module, tracking of regions with low bounding box quality but high semantic scores can be prevented. If a target disappears or becomes occluded during tracking, the introduced Semantic-Presence factor score tends toward a lower value, restricting the corresponding region embedding update to maintain the original high-quality semantic features. Through maintaining memory regions embeddings of length K, OV-VOD can efficiently track the same object across different frames over extended periods.

4.2 Spatial Object Relationship Distillation

Rethinking the ViLD [15] approach to open-vocabulary object detection reveals that its core strength lies in distilling knowledge from a pre-trained vision-language model into a student network, thereby enabling the student to acquire robust open-vocabulary classification capabilities. However, during the distillation process, ViLD adopts a simplistic region embedding strategy as the feature distillation loss to guide network training. This method treats objects within an image as independent entities, neglecting their spatial relationships. As a result, the effectiveness of knowledge transfer is limited, leading to suboptimal performance.

To address this issue, we are inspired by similarity-preserving knowledge distillation [2, 47] and propose a Spatial Object Relationship Distillation loss, denoted as $\mathcal{L}_{SR}$, to be integrated into the distillation process. While traditional similarity-preserving distillation is designed to capture inter-sample similarity relationships across different images within a mini-batch to guide the learning process, we extend this idea to the video object detection domain. In this context, different objects within a single image naturally exhibit stronger spatial and contextual relationships, which can be leveraged as additional supervisory signals to enhance the effectiveness of knowledge transfer.

$\mathcal{L}_{SR}$ is defined based on two paired similarity matrices extracted from both the student network (detector) and the teacher network (VLM). Specifically, let I denote the image, $\tilde{r}$ denote the pre-extracted proposals, $\mathcal{V}$ represent the pre-trained image encoder, and $\mathcal{R}$ represent the student network. We first obtain the region embeddings from $\mathcal{V}(\text{crop}(I, \tilde{r}))$ and $\mathcal{R}(I, \tilde{r})$, and then compute their corresponding similarity matrices of shape $N \times N$, denoted as $\mathcal{S}_V$ and $\mathcal{S}_R$, respectively. These matrices are normalized using the L_2 norm applied row-wise. Then, $\mathcal{L}_{SR}$ can be defined as:

$$\mathcal{S}_V = \frac{\mathcal{V}(\text{crop}(I, \tilde{r})) \cdot \mathcal{V}(\text{crop}(I, \tilde{r}))^T}{\| \mathcal{V}(\text{crop}(I, \tilde{r})) \cdot \mathcal{V}(\text{crop}(I, \tilde{r}))^T \|_2}, \tag{4}$$

$$\mathcal{S}_R = \frac{\mathcal{R}(I, \tilde{r}) \cdot \mathcal{R}(I, \tilde{r})^T}{\| \mathcal{R}(I, \tilde{r}) \cdot \mathcal{R}(I, \tilde{r})^T \|_2}, \tag{5}$$

$$\mathcal{L}_{SR} = \frac{1}{N^2} \| \mathcal{S}_V - \mathcal{S}_R \|_F^2, \tag{6}$$

where $\| \cdot \|_F$ denotes the Frobenius norm. The introduced loss enforces the student network to retain the spatial contextual relationships between different objects within the same image, as captured by the teacher network. This enhances the efficiency of knowledge transfer, ensuring that the student network learns more structured and meaningful object representations.

4.3 Training and Loss

Since the SPMT is a training-free post-processing module, we can train on the image-level dataset LVIS while still leveraging video-specific temporal information during inference on video data. This approach significantly reduces the training resource requirements for large-scale video datasets, thus enhancing training efficiency.

Regarding the training loss, following the foundation established by ViLD [15], we further optimize the objective function to enhance detection accuracy and efficiency. First, we replace the original classification loss with a text alignment loss $\mathcal{L}_{text}$ based on the CLIP [39] approach. Here, $\boldsymbol{E}_k^*$ is the feature used for classification in the key frame, $\boldsymbol{E}_{bg}$ is the text embedding of the background category, and $\boldsymbol{t}_i$ represents the text embedding of the i-th category obtained through a pre-trained text encoder. Note that $\boldsymbol{t}_i$ can be obtained through hand-crafted prompts as implemented in [15], or it can be obtained as learnable text embeddings as implemented in [11]. Then, $\mathcal{L}_{text}$ can be expressed as:

$$\mathbf{z}(t) = \left[sim(\boldsymbol{E}_k^*, \boldsymbol{E}_{bg}), sim(\boldsymbol{E}_k^*, \boldsymbol{t}_1), \cdots, sim(\boldsymbol{E}_k^*, \boldsymbol{t}_{|C_{train}|}) \right] \tag{7}$$

$$\mathcal{L}_{text} = \frac{1}{N} \sum_{r \in P} \mathcal{L}_{CE}(softmax(\mathbf{z}(r)/\tau), y_r), \tag{8}$$

where $sim(\mathbf{a},\mathbf{b}) = \mathbf{a}^T\mathbf{b}/(\|\mathbf{a}\| \cdot \|\mathbf{b}\|)$, y_r denotes the class label of region r, τ is the temperature, N is the number of proposals per image($|P|$), and $\mathcal{L}_{CE}$ is the cross entropy loss.

As for the feature distilling loss $\mathcal{L}_{FD}$, following [15], we align region embeddings $\mathcal{R}(\phi(I), \tilde{r})$ to image embeddings $\mathcal{V}(crop(I, \tilde{r}))$. Note that the proposals $\tilde{r}$ here are extracted offline and contain objects in both C_{base} and C_{novel} . Thus, $\mathcal{L}_{FD}$ can be expressed as:

$$\mathcal{L}_{FD} = \frac{1}{M} \sum_{\tilde{r} \in \widetilde{P}} \|\mathcal{V}(\text{crop}(I, \tilde{r})) - \mathcal{R}(\phi(I), \tilde{r})\|_1, \tag{9}$$

where M denotes the number of the proposals per image($|\widetilde{P}|$). Therefore, the total training loss can be viewed as a weighted sum of multiple objective functions:

$$\mathcal{L} = \mathcal{L}_{text} + w_1 \cdot \mathcal{L}_{FD} + w_2 \cdot \mathcal{L}_{SR} + \mathcal{L}_{box}, \tag{10}$$

where w_1 and w_2 are hyperparameters that control the weights of feature distillation $\mathcal{L}_{FD}$ and spatial object relationship distillation $\mathcal{L}_{SR}$, respectively.

5 Experiments

5.1 Datasets

To ensure a fair comparison with existing open-vocabulary object detection methods [6, 11, 15, 53, 58, 62], we follow the training protocols used in most related works [11, 15, 53]. In particular, we train our model on the union of common and frequent categories from LVIS [17] and subsequently evaluate its zero-shot generalization performance on three video datasets: LV-VIS [48], BURST [1], and TAO [9]. It is important to note that neither our method nor the

Table 2: Zero-shot performance comparison on the LV-VIS validation set. AP_n, AP_b, and AP mean the average precision of novel categories, base categories, and overall categories. ViLD* denotes the reproduced ViLD version from DetPro. Note that for a fair comparison, all models use ResNet-50 [24] as the backbone, while YOLO-World-L employs YOLOv8-L [29] as its backbone, which has a parameter count comparable to ResNet-50.

Method	Pretraining Data	Vision Training Annotations	Detection			Instance Segmentation		
			AP_n	AP_b	AP	AP_n	AP_b	AP
ViLD* (ICLR'22) [15]	CLIP400M [39]	$LVIS_{Base}$	8.7	11.7	10.0	8.3	11.4	9.6
Detpro (CVPR'22) [11]	CLIP400M [39]	$LVIS_{Base}$	9.7	12.0	10.6	9.2	11.5	10.2
Baron (CVPR'23) [53]	CLIP400M [39]	$LVIS_{Base}$	9.1	8.4	8.7	9.2	8.0	8.5
Detic (ECCV'22) [62]	CC3M [43]	$LVIS_{Base}$+$Pseudo_{Novel}$	9.7	3.9	6.3	9.5	3.6	6.1
RegionCLIP (CVPR'22) [58]	CC3M [43]	$LVIS_{Base}$+$Pseudo_{Novel}$	12.3	6.6	9.0	10.4	5.9	7.8
YOLO-World-L (CVPR'24) [6]	O365 [42]+GoldG [30]+CC3M [43]	$LVIS_{Base}$+$Pseudo_{Novel}$	11.8	**14.7**	13.0	-	-	-
OV-VOD (Ours)	CLIP400M [39]	$LVIS_{Base}$	**12.8**	13.7	**13.2**	**12.3**	**13.1**	**12.6**

comparative methods have been fine-tuned on any video datasets, ensuring a fair comparison.

LVIS is a widely used image open-vocabulary object detection dataset that contains 1203 categories. Following the setting in ViLD [15], we treat the frequent and common categories as base categories while designating rare categories as novel categories.

LV-VIS is a recently introduced large-scale dataset for evaluating open-vocabulary video instance segmentation. It contains 1,196 categories, of which 641 are base categories following the LVIS split, and 555 are novel categories. Among the novel categories, there are not only rare categories from LVIS but also entirely new classes not present in LVIS. Therefore, LV-VIS is highly suitable for assessing the performance of Open-Vocabulary VOD methods.

TAO is a dataset designed for evaluating open-vocabulary tracking methods. Although the annotations for multi-object tracking are similar to those for detection, and may be somewhat incomplete, the diverse category set renders TAO suitable for evaluating Open-Vocabulary VOD methods. Following the LVIS setting, TAO comprises 363 categories, with 290 categorized as base and 73 as novel.

BURST is a recently released video dataset that extends TAO. It consists of 425 base categories and 57 novel categories as defined by the LVIS partitions.

5.2 Implementation Details

Baseline Model. We select ViLD [15], an open-vocabulary object detection method built on Mask-RCNN [23], as our baseline model. The pretrained image and text encoders are based on CLIP (ViT-B/32). Additionally, our baseline incorporates learnable text embeddings, as demonstrated in [11, 59, 60], to further boost detection performance.

OV-VOD. To ensure a fair comparison with the baseline, we employ the same backbone (ResNet-50 [24]) in all experiments. Consistent with the ViLD setting, the temperature coefficient τ is set to 0.01 during training and 0.007 during inference. The weighting coefficients for the losses $\mathcal{L}_{FD}$ and $\mathcal{L}_{SR}$ are assigned as $w_1 = 0.5$ and $w_2 = 1$, respectively. In our SPMT module, the update factor weight w_β is empirically determined as 0.5 to balance semantic information and bounding box quality. The size of the memory bank K is examined in detail in the ablation study section.

Training Details. For a fair comparison, both the baseline and OV-VOD models are trained on the LVIS dataset for 20 epochs using a batch size of 16 with the SGD optimizer [3]. The initial learning rate is set to 0.2 and decayed by a factor of ten at the 16th epoch. Momentum is set to 0.9 and weight decay to 0.000025. A warmup strategy is applied during the first 500 iterations. To maintain consistency, the same data augmentation strategy from [11] is adopted. Training is conducted on 4 RTX4090 GPUs over approximately 32 hours, while all inference is performed on a single RTX4090 GPU.

5.3 Results on the LV-VIS dataset

We compare the performance of our proposed OV-VOD with existing mainstream open-vocabulary object detection methods on the LV-VIS dataset, as shown in Table 2. Compared to existing methods, OV-VOD achieves optimal performance on the LV-VIS validation set with 12.8% AP_n^{bbox}, 13.2% AP^{bbox}, 12.3% AP_n^{mask}, and 12.6% AP^{mask}. This is primarily because existing image-level open-vocabulary object detection methods perform frame-by-frame detection on video data, ignoring the rich temporal information between frames, which limits their performance on video datasets.

Notably, methods such as Detic and RegionCLIP, despite achieving strong performance on the LVIS dataset following large-scale pretraining, exhibit limited zero-shot generalization capabilities on the LV-VIS video dataset. Although these two methods achieve relatively high AP_n scores, which is a key evaluation metric for open-vocabulary object detection, their performance on base classes is significantly lower. This discrepancy may stem from the introduction of pseudo-labels during their training phase, potentially causing the models to prioritize fitting features specific to novel categories. However, a central goal of open-vocabulary object detection is to successfully identify a wide range of categories in real-world scenarios. Therefore, an effective method should enhance novel category detection without compromising performance on base classes. While YOLO-World performs well on base classes, its detection accuracy for novel categories is lower than that achieved by our proposed method.

Furthermore, the aforementioned methods are all pretrained on large-scale grounding, image-text, and detection datasets before being fine-tuned on LVIS, incurring significant data and training costs. In contrast, our method employs knowledge distillation solely

Table 3: Ablation Study of $\mathcal{L}_{SR}$ and SPMT. Note that AP_r has been regarded as the metric for novel categories in the LVIS dataset in previous open-vocabulary object detection works.

Modules		LVIS		LV-VIS		
$\mathcal{L}_{SR}$	SPMT	AP_r	AP	AP_n	AP_b	AP
		16.8	26.8	9.7	12.0	10.6
✓		19.9$_{\uparrow 3.1}$	28.5$_{\uparrow 1.7}$	10.3$_{\uparrow 0.6}$	12.7$_{\uparrow 0.7}$	11.3$_{\uparrow 0.7}$
	✓	-	-	12.1$_{\uparrow 2.4}$	13.4$_{\uparrow 1.4}$	12.7$_{\uparrow 2.1}$
✓	✓	-	-	12.8$_{\uparrow 3.1}$	13.7$_{\uparrow 1.7}$	13.2$_{\uparrow 2.6}$

on the LVIS dataset, thereby reducing data dependency and training overhead while still delivering superior performance.

5.4 Ablation Studies

We conduct comprehensive ablation studies to validate the effectiveness of the key components of OV-VOD.

Effectiveness of the $\mathcal{L}_{SR}$ and SPMT. We evaluate the effectiveness of the proposed Spatial Object Relationship Distillation loss ($\mathcal{L}_{SR}$) and Semantic-Presence Memory Tracking (SPMT) module on the LVIS and LV-VIS validation sets. As shown in Table 3, introducing $\mathcal{L}_{SR}$ yields improvements of 3.1% AP_r and 1.7% AP on the LVIS validation set over the baseline. On the video dataset LV-VIS, it further improves AP_n by 0.6% and AP by 0.7%, indicating that spatial relation distillation effectively captures contextual dependencies, enhances knowledge transfer, and boosts image-level open-vocabulary detection. Additionally, incorporating SPMT leads to substantial gains on LV-VIS, improving AP_n by 2.4%, AP_b by 1.4%, and AP by 2.1%. These improvements highlight SPMT's ability to leverage temporal cues and mitigate the limitations of image-based detectors in video scenarios. When combined, $\mathcal{L}_{SR}$ and SPMT are highly complementary: the final model achieves a 3.1% gain in APn and 2.6% in AP on LV-VIS. The stronger spatial-aware detector enabled by $\mathcal{L}_{SR}$ further enhances SPMT's robustness, particularly under occlusion, disappearance, or motion blur.

Impact of the Support Frame Selection Strategy. Consistent with prior findings [18, 20, 36], the support frame selection strategy and the memory bank size remain crucial factors influencing the performance of conventional video object detection. Accordingly, we analyze these factors in our proposed method. As in prior work [10, 36, 38], support frames are defined as frames sampled from the same video and stored in the SPMT memory bank, from which corresponding region embeddings are retrieved. We analyze variations in AP_{50} and AP_n on the LV-VIS validation set by varying memory bank size and employing different support frame selection strategies, as illustrated in Figure 3. Irrespective of selection strategy, both AP_{50} and APn demonstrate consistent improvement as memory bank size K increases. When $K = 0$ (i.e., SPMT is not utilized for video-level inference), the model achieves AP50 of 17.0% and AP_n of 17.3%. Performance steadily improves as K increases from 0 to 6, yet saturates at $K = 14$, indicating a performance bottleneck. This trend corroborates findings from prior video object detection research [18–20, 37, 38], demonstrating that increasing support frame count enables SPMT to track objects across extended temporal windows, thereby enhancing key frame

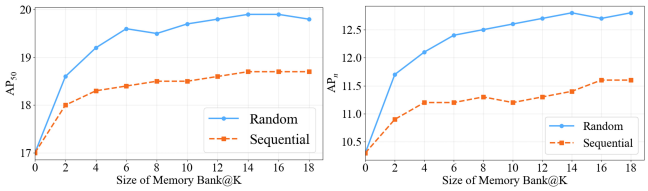


Figure 3: AP_{50} and AP_n under different size of memory bank K and the selection strategy of support frames on the LV-VIS validation set. *Random* denotes selecting support frames randomly from the entire video sequence, whereas *Sequential* refers to selecting the past K frame set $\{I_{t-K}, \cdots, I_{t-2}, I_{t-1}\}$ before the current time step t.

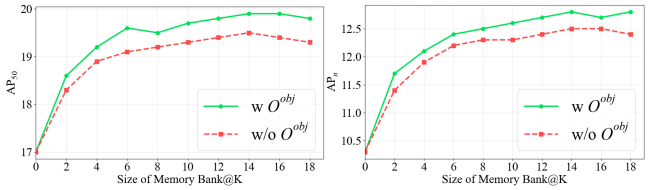


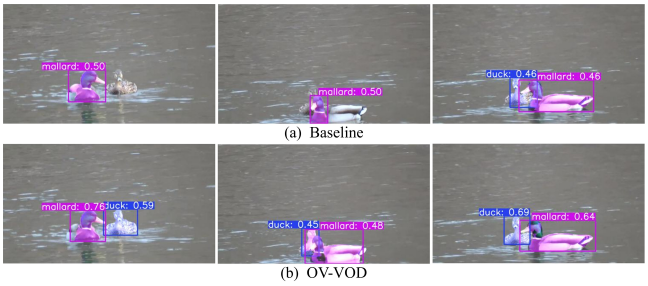
Figure 4: AP_{50} and AP_n under different size of memory bank K and the update factor of SPMT on the LV-VIS validation set. w O^{obj} and w/o O^{obj} denote the inclusion and removal of the objectness score in the update factor, respectively.

detection through leverage of support frames with superior visual features. However, once K exceeds a certain threshold, marginal or negligible additional benefits are observed, suggesting an upper bound on achievable performance improvement. Furthermore, models employing random selection strategy consistently outperform those using sequential strategy, consistent with observations in [5, 36]. The random strategy facilitates global temporal information aggregation, particularly advantageous for handling occluded or blurred objects by incorporating support frames with clearer visual appearances. Conversely, the sequential strategy constrains support frames to those preceding the key frame, which proves suboptimal under low video sampling rates and complicates object tracking in complex scenarios. Optimal performance is achieved using random selection strategy with $K = 14$, yielding AP_{50} of 19.9% and AP_n of 12.8% on the LV-VIS validation set.

Impact of the Semantic-Presence factor. To further evaluate the impact of the Semantic-Presence factor introduced in Equation 2, we analyze performance variations on the LV-VIS validation set by altering the size of memory bank K and removing O^{obj} in the update factor, as illustrated in Figure 4. The impact of the size of memory bank K on performance is consistent with the analysis in the previous subsection. Regardless of the update factor used, increasing K from 0 to 6 results in significant performance improvement, after which the gains gradually plateau. The best performance is achieved when $K = 14$ and the update factor includes O^{obj}. Notably, incorporating O^{obj} into the update factor consistently leads to higher AP_{50} and AP_n across all values of K. In particular, when performance saturates with increasing K, the inclusion of

Table 4: Zero-shot generalization on the validation sets of BURST and TAO. We report the detection and segmentation metrics on the BURST validation set, denoted as AP_*^{bbox} and AP_*^{mask}, respectively. Additionally, due to the low quality of instance segmentation annotations in TAO, we only report the detection metrics for it, represented as AP_*^{bbox}.

Method	Backbone	BURST						TAO		
		AP_n^{bbox}	AP_b^{bbox}	AP^{bbox}	AP_n^{mask}	AP_b^{mask}	AP^{mask}	AP_n^{bbox}	AP_b^{bbox}	AP^{bbox}
ViLD* (ICLR'22) [15]	ResNet-50	3.8	7.7	7.0	3.4	7.0	6.4	3.6	7.7	6.9
Detpro (CVPR'22) [11]	ResNet-50	3.8	7.7	7.0	3.1	7.0	6.3	3.9	7.8	7.1
Baron (CVPR'23) [53]	ResNet-50	3.0	6.2	5.6	3.5	5.7	5.3	3.1	6.5	5.9
RegionCLIP (CVPR'22) [58]	ResNet-50	4.5	6.8	6.5	3.3	5.7	5.3	4.8	6.9	6.6
Detic (ECCV'22) [62]	ResNet-50	4.6	7.9	7.3	4.4	7.3	6.7	4.8	7.6	7.3
YOLO-World-L (CVPR'24) [6]	YOLOv8-L	4.3	**9.1**	**8.2**	-	-	-	4.6	**9.6**	**8.8**
OV-VOD (Ours)	ResNet-50	**4.9**	8.4	7.7	**4.5**	**7.9**	**7.2**	**5.0**	8.5	7.8



(a) Baseline

(b) OV-VOD

Figure 5: Qualitative comparison between baseline (a) and OV-VOD (b) on the LV-VIS val set. Blue represents base category objects, while pink denotes novel category objects.

O^{obj} results in a 0.5% and 0.4% improvement in AP_{50} and AP_n, respectively. This strongly validates the effectiveness of the proposed Semantic-Presence factor. Moreover, it demonstrates that combining the confidence score and the objectness score enables the model to account for both semantic information and bounding box quality when tracking objects across frames. By effectively suppressing the update of key frame features when encountering support frames with high semantic relevance but low quality, it imposes stricter control over the tracked object quality, thereby facilitating the key frame update with visually clearer support objects.

5.5 Zero-shot Generalization

To further assess the performance of existing object detection methods and our approach on a broader range of video datasets, we evaluate them on two open-vocabulary tracking datasets, BURST and TAO, as shown in Table 4. It is important to note that, due to the highly specialized nature of the tracking domain, annotations in these datasets are inherently incomplete and may not fully reflect the true performance in real-world scenarios. Our model achieves robust detection performances of 4.9% AP_n on BURST and 5.0% AP_n on TAO, surpassing the baseline method Detpro by 1.1% and decisively outperforming existing mainstream open-vocabulary object detection methods. However, our method does not achieve the highest AP across all categories. Based on our detailed statistical analysis, this is primarily attributed to the fact that novel class annotations constitute only a mere 3.6% of the total annotations in the validation sets of BURST and TAO. Therefore, it conclusively indicates that both datasets are heavily skewed toward base class annotations, making the overall AP an insufficient metric for accurately assessing the generalization capability of Open-Vocabulary VOD methods on novel categories.

5.6 Qualitative Analysis

The qualitative comparison results presented in Figure 5 illustrate distinct advantages of our OV-VOD over the baseline approach through three representative video frames. Although both methods show the capability to detect objects from base and novel categories, OV-VOD delivers significantly superior detection performance under challenging scenarios. For instance, in the first and third frames, OV-VOD achieves much higher confidence scores and exhibits more stable detection performance for both novel and base categories. Moreover, the incorporation of the SPMT module plays a critical role, as evidenced in the second frame, where the foreground occlusion poses a significant challenge. Thanks to the SPMT module, OV-VOD effectively leverages support frames stored in the memory bank to provide supplementary temporal and contextual cues, thereby enhancing the key frame detection. In contrast, the baseline method fails to detect these occluded objects entirely, illustrating the substantial advantage afforded by the additional temporal information. These qualitative results substantiate that the proposed OV-VOD effectively mitigates visual degradation in video sequences, particularly overcoming persistent issues such as object occlusion and transient object disappearance.

6 Conclusion

In this paper, we introduce a novel Open-Vocabulary VOD task aimed at detecting objects from open-set categories in videos. To facilitate the evaluation of open-vocabulary VOD methods, we repurpose annotations from existing video datasets to establish a comprehensive benchmark. Furthermore, we propose a new method, OV-VOD, which leverages video-specific temporal information via a training-free Semantic-Presence Memory Tracking (SPMT) module and enhances knowledge transfer by incorporating a Spatial Object Relationship Distillation loss ($\mathcal{L}_{SR}$) during the distillation process. Our proposed OV-VOD demonstrates significantly stronger zero-shot generalization capabilities compared to existing image-level open-vocabulary object detection methods across video datasets.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant U21A20514; and in part by the Major Science and Technology Plan Project on the Future Industry Fields of Xiamen City under Grant 3502Z20241027.

References

[1] Ali Athar, Jonathon Luiten, Paul Voigtlaender, Tarasha Khurana, Achal Dave, Bastian Leibe, and Deva Ramanan. 2023. Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1674–1683.

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

[3] Léon Bottou. 2010. Large-scale machine learning with stochastic gradient descent. In *International Conference on Computational Statistics*. 177–186.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*. 213–229.

[5] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. 2020. Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10337–10346.

[6] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. 2024. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16901–16911.

[7] Zesen Cheng, Kehan Li, Li Hao, Peng Jin, Xiawu Zheng, Chang Liu, and Jie Chen. 2025. Aligning Instance Brownian Bridge with Texts for Open-Vocabulary Video Instance Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 2482–2490.

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling egocentric vision: The dataset. In *Proceedings of the European Conference on Computer Vision*. 753–771.

[9] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. 2020. TAO: A large-scale benchmark for tracking any object. In *Proceedings of the European Conference on Computer Vision*. 436–454.

[10] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. 2019. Relation distillation networks for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7023–7032.

[11] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. 2022. 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*. 14084–14093.

[12] Hao Fang, Peng Wu, Yawei Li, Xinxin Zhang, and Xiankai Lu. 2024. Unified embedding alignment for open-vocabulary video instance segmentation. In *European Conference on Computer Vision*. Springer, 225–241.

[13] Mingfei Gao, Chen Xing, Juan Carlos Niebles, Junnan Li, Ran Xu, Wenhao Liu, and Caiming Xiong. 2022. Open vocabulary object detection with pseudo bounding-box labels. In *Proceedings of the European Conference on Computer Vision*. 266–282.

[14] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* 129, 6 (2021), 1789–1819.

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

[16] Chaoxu Guo, Bin Fan, Jie Gu, Qian Zhang, Shiming Xiang, Veronique Prinet, and Chunhong Pan. 2019. Progressive sparse local attention for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3909–3918.

[17] Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5356–5364.

[18] Liang Han, Pichao Wang, Zhaozheng Yin, Fan Wang, and Hao Li. 2020. Exploiting better feature aggregation for video object detection. In *Proceedings of the ACM International Conference on Mmultimedia*. 1469–1477.

[19] Liang Han, Pichao Wang, Zhaozheng Yin, Fan Wang, and Hao Li. 2021. Class-aware feature aggregation network for video object detection. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 12 (2021), 8165–8178.

[20] Liang Han, Pichao Wang, Zhaozheng Yin, Fan Wang, and Hao Li. 2021. Context and structure mining network for video object detection. *International Journal of Computer Vision* 129, 10 (2021), 2927–2946.

[21] Liang Han and Zhaozheng Yin. 2022. Global memory and local continuity for video object detection. *IEEE Transactions on Multimedia* 25 (2022), 3681–3693.

[22] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang. 2016. Seq-nms for video object detection. *arXiv preprint arXiv:1602.08465* (2016).

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*. 2961–2969.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[25] Lu He, Qianyu Zhou, Xiangtai Li, Li Niu, Guangliang Cheng, Xiao Li, Wenxuan Liu, Yunhai Tong, Lizhuang Ma, and Liqing Zhang. 2021. End-to-end video object detection with spatial-temporal transformers. In *Proceedings of the ACM International Conference on Multimedia*. 1507–1516.

[26] De-An Huang, Zhiding Yu, and Anima Anandkumar. 2022. Minvis: A minimal video instance segmentation framework without video-based training. *Advances in Neural Information Processing Systems* 35 (2022), 31265–31277.

[27] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning*. 4904–4916.

[28] Zhengkai Jiang, Yu Liu, Ceyuan Yang, Jihao Liu, Peng Gao, Qian Zhang, Shiming Xiang, and Chunhong Pan. 2020. Learning where to focus for efficient video object detection. In *Proceedings of the European Conference on Computer Vision*. 18–34.

[29] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. 2023. *Ultralytics YOLO*. https://github.com/ultralytics/ultralytics

[30] 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*. 1780–1790.

[31] Kai Kang, Hongsheng Li, Junjie Yan, Xingyu Zeng, Bin Yang, Tong Xiao, Cong Zhang, Zhe Wang, Ruohui Wang, Xiaogang Wang, and Wanli Ouyang. 2018. T-CNN: Tubelets with convolutional neural networks for object detection from videos. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 10 (2018), 2896–2907.

[32] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. 2023. Ovtrack: Open-vocabulary multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5567–5577.

[33] Lijian Lin, Haosheng Chen, Honglun Zhang, Jun Liang, Yu Li, Ying Shan, and Hanzi Wang. 2020. Dual semantic fusion network for video object detection. In *Proceedings of the ACM International Conference on Multimedia*. 1855–1863.

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*. 740–755.

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

[36] Qiang Qi, Tianxiang Hou, Yang Lu, Yan Yan, and Hanzi Wang. 2023. DGRNet: A dual-level graph relation network for video object detection. *IEEE Transactions on Image Processing* 32 (2023), 4128–4141.

[37] Qiang Qi, Tianxiang Hou, Yan Yan, Yang Lu, and Hanzi Wang. 2023. TCNet: A novel triple-cooperative network for video object detection. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 8 (2023), 3649–3662.

[38] Qiang Qi, Zhenyu Qiu, Yan Yan, Yang Lu, and Hanzi Wang. 2024. IMC-Det: Intra–inter modality contrastive learning for video object detection. *International Journal of Computer Vision* 133, 2 (2024), 1–20.

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. 8748–8763.

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2017), 1137–1149.

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115 (2015), 211–252.

[42] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8430–8439.

[43] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*. 2556–2565.

[44] Hengcan Shi, Munawar Hayat, and Jianfei Cai. 2023. Open-vocabulary object detection via scene graph discovery. In *Proceedings of the ACM International*

Conference on Multimedia. 4012–4021.

[45] Guanxiong Sun, Yang Hua, Guosheng Hu, and Neil Robertson. 2021. Mamba: Multi-level aggregation via memory bank for video object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2620–2627.

[46] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. 2024. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13019–13029.

[47] Frederick Tung and Greg Mori. 2019. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1365–1374.

[48] Haochen Wang, Cilin Yan, Keyan Chen, Xiaolong Jiang, Xu Tang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. 2024. OV-VIS: Open-vocabulary video instance segmentation. *International Journal of Computer Vision* 132, 11 (2024), 1–18.

[49] Jiong Wang, Huiming Zhang, Haiwen Hong, Xuan Jin, Yuan He, Hui Xue, and Zhou Zhao. 2023. Open-vocabulary object detection with an open corpus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6759–6769.

[50] Tao Wang. 2023. Learning to detect and segment for open vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7051–7060.

[51] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. 2019. Sequence level semantics aggregation for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9217–9225.

[52] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, et al. 2024. Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 7 (2024), 5092–5113.

[53] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. 2023. Aligning bag of regions for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15254–15264.

[54] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. 2022. Open-vocabulary DETR with conditional matching. In *Proceedings of the European Conference on Computer Vision*. 106–122.

[55] Bingqing Zhang, Sen Wang, Yifan Liu, Brano Kusy, Xue Li, and Jiajun Liu. 2023. Object detection difficulty: Suppressing over-aggregation for faster and better video object detection. In *Proceedings of the ACM International Conference on Multimedia*. 1768–1778.

[56] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46 (2024), 5625,5644.

[57] Tao Zhang, Xingye Tian, Yikang Zhou, Shunping Ji, Xuebo Wang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, and Yu Wu. 2025. DVIS++: Improved Decoupled Framework for Universal Video Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47, 7 (2025), 5918–5929. doi:10.1109/TPAMI.2025.3552694

[58] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16793–16803.

[59] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16816–16825.

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

[61] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. 2022. TransVOD: End-to-end video object detection with spatial-temporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 6 (2022), 7853–7869.

[62] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. 2022. Detecting twenty-thousand classes using image-level supervision. In *Proceedings of the European Conference on Computer Vision*. 350–368.

[63] Wenqi Zhu, Jiale Cao, Jin Xie, Shuangming Yang, and Yanwei Pang. 2025. CLIP-VIS: Adapting CLIP for Open-Vocabulary Video Instance Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* 35, 2 (2025), 1098–1110. doi:10.1109/TCSVT.2024.3474698

[64] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. 2017. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 408–417.

[65] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. 2017. Deep feature flow for video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2349–2358.