Toward Open-Vocabulary Video Object Detection

Zhihong Zheng(AI Class)[1], Hantao Qi(AI Class)[1], Yanglan Fan(AI Class)[2]

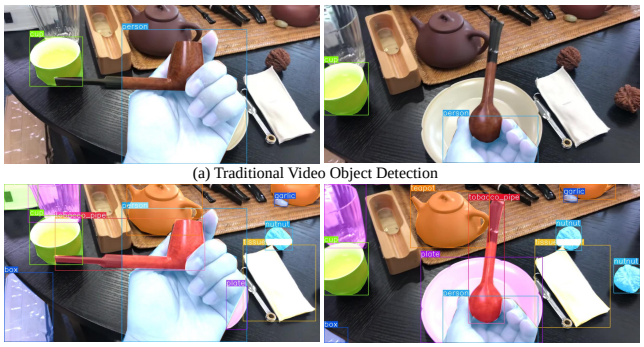
[1]Institute of Artificial Intelligence, Xiamen University
{36920241153287,36920240157319}@stu.xmu.edu.cn
[2]School of Information, Xiamen University
23020241154389@stu.xmu.edu.cn

Abstract

Traditional Video Object Detection typically involves detecting and classifying objects within a video from a closed set of predefined categories, which limits its ability to generalize to new, unseen categories that may appear in real-world video data. To address this limitation, we make three key contributions. First, we introduce the novel task of *Open-Vocabulary Video Object Detection*, which aims to detect objects in videos from open-set categories, including those that were never seen during training. Second, drawing inspiration from the fields of open-vocabulary multi-object tracking and open-vocabulary video instance segmentation, we establish a training and evaluation framework for open-vocabulary video object detection models, which can serve as a reference for future work. Third, to fully leverage the temporal information in videos, we introduce pixel-level and instance-level temporal attention mechanisms into existing open-vocabulary object detection methods. We propose a baseline model, OV-VOD, which aggregates pixel-level and instance-level features from adjacent frames into key frames, aiming to effectively enhance detection performance. Although our extensive experiments on the LV-VIS dataset have not yet validated the effectiveness of our approach, we remain confident that aggregating features from adjacent frames is a promising direction for improving performance. This has been demonstrated by numerous experiments in the traditional video object detection field. We will continue exploring this direction until a suitable method for enhancing detection performance is found.

Introduction

Although traditional video object detection methods have been well-developed, they are fundamentally limited by the need to detect and classify objects from a closed set of training categories, which restricts the ability of these methods to generalize to new concepts. In real-world scenarios, this closed-set vocabulary paradigm has limited practical value, as new categories frequently emerge in actual applications. This is one of the key reasons why traditional video object detection methods struggle to be effectively applied in practice. In contrast, recent open-vocabulary object detection methods aim to detect and classify all objects within an image. However, these methods have been primarily applied to static images. When applied to videos, such static detection approaches fail to leverage the rich temporal information inherent in video data. Moreover, video-specific challenges, such as object occlusion, unusual poses, and motion blur, are typically absent in images, making it difficult to directly transfer open-vocabulary object detection methods from image-based tasks to videos without sacrificing performance. To address this gap, we introduce the task of *Open-Vocabulary Video Object Detection*, which aims to detect and classify any object within an open set of categories across video frames,as illustrated in Fig.1.



(a) Traditional Video Object Detection

(b) Open-Vocabulary Video Object Detection

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

Accurate benchmarking of open-vocabulary video object detection methods requires a video dataset that encompasses a large and diverse set of object categories. However, existing datasets in the video object detection domain, such as ImageNet VID and EPIC-KITCHENS-55(Damen et al. 2018), are insufficient in terms of category diversity, see Table 1. This limitation hinders the further development of open-vocabulary approaches in the video domain due to the lack of suitable datasets tailored to open-vocabulary video tasks.

Dataset	VID	EPIC-55	TAO	BURST	LV-VIS
Videos	4417	272	1488	2914	4828
Instances	1298k	174k	51378	16089	25588
Objects	2005k	326k	168k	600k	544k
Categories	30	351	363	482	1196

Table 1: Comparison of Key Statistics Between the Datasets Used in This Paper and Traditional Video Object Detection Datasets.

To address this issue, we draw inspiration from the fields of open-vocabulary tracking and open-vocabulary video instance segmentation, transferring open-vocabulary video datasets from these domains to the detection domain. Since the labels used for tracking and segmentation tasks can be readily converted to bounding box annotations required for detection tasks, this cross-domain adaptation is highly feasible. Specifically, we leverage two open-vocabulary video instance segmentation datasets, LV-VIS(Wang et al. 2024), and one open-vocabulary tracking dataset, BURST(Athar et al. 2023) and TAO(Dave et al. 2020).

In this work, we evaluate open-vocabulary video object detection models using the three aforementioned video datasets. Among them, the LV-VIS dataset is particularly well-suited for assessing the generalization capability of open-vocabulary video object detection methods on novel categories due to its large number of categories, most of which are distinct from those annotated in commonly used datasets such as MS-COCO(Lin et al. 2014) and LVIS(Gupta, Dollar, and Girshick 2019). As a result, model performance is primarily benchmarked based on results obtained on the LV-VIS dataset.

One intuitive approach to addressing open-vocabulary video object detection is to treat each video frame as an independent image and apply existing open-vocabulary object detection methods to detect objects frame by frame. However, this image-based approach neglects the temporal information unique to videos and fails to leverage the inter-frame correlations. Video-specific challenges, such as object occlusion, unusual poses, and motion blur, further exacerbate the limitations of this naive approach, leading to suboptimal performance.

In this paper, we aim to propose the first *Open-Vocabulary Video Object Detection model*, OV-VOD, which aims to improve the detection performance of target frames by aggregating proposal-level and pixel-level features from supporting frames. Specifically, memory attention is utilized to aggregate proposal-level features from supporting frames into the proposal-level features of the target frame, addressing challenges such as appearance degradation caused by blur, occlusion, and abnormal poses in videos.Although the performance of our model has not yet reached a satisfactory level, we firmly believe that aggregating features from supporting frames is an effective approach for improving video detection methods. Therefore, we plan to further refine our aggregation methods in future work while exploring ways to enhance model performance from the perspective of open-vocabulary detection.

Related Work

Open-Vocabulary Object Detection has rapidly advanced with the emergence of various large models. ViLD(Gu et al. 2021) was the first to distill knowledge from Visual Language Models (VLMs) into lightweight detectors through a knowledge distillation approach. ViLD utilizes the principle of open-vocabulary classification, where visual features are aligned with extensive textual features during the pre-training phase of the VLM. It introduces both image and text branches, distilling the VLM knowledge into the image branch to facilitate open-vocabulary detection. DetPro(Du et al. 2022) later extended the work of ViLD by incorporating specific contextual learning for downstream tasks, resulting in improved detection performance. It modifies the static text prompts in ViLD by adapting them to the task-specific context, thereby enhancing detection performance.

On the other hand, RegionCLIP(Zhong et al. 2022) learns visual region representations by aligning image regions with corresponding region-level textual descriptions. It generates pseudo-labels for region-text pairs using CLIP(Radford et al. 2021), followed by fine-tuning on manually annotated detection datasets. The GLIP(Li et al. 2022; Zhang et al. 2022) series unifies object detection and phrase-based grounding for pre-training. Notably, it leverages a self-training mechanism to generate anchor boxes from a large corpus of image-text pairs, enabling powerful detection and grounding performance.

Furthermore, Detic(Zhou et al. 2022) adopts a joint training approach to balance the dataset, as new categories often suffer from limited sample sizes. By incorporating image-level supervision, Detic improves long-tail detection performance.OV-DETR(Zang et al. 2022) employs region-text alignment, replacing ground truth (GT) bounding boxes with language as direct supervisory signals. This approach allows for better alignment of visual and textual features during base class training. It replaces the standard binary matching mechanism with a conditional binary matching strategy, enabling an end-to-end open-vocabulary detection approach.

Recently, BARON(Wu et al. 2023) proposed aligning embeddings of different regions into a bag, rather than just aligning a single region. Specifically, it groups contextually relevant regions into a ”bag” and treats each region as a ”word” in a sentence. The bag of regions is then passed through a text encoder to obtain bag-of-regions embeddings, which are subsequently aligned with the embeddings of cropped regions in the VLM's image encoder.

Although these Open-Vocabulary object detection methods achieve strong performance on images, their direct application to videos is limited by the inability to fully exploit the temporal information inherent in video data, which may lead to suboptimal results.

Video Object Detection usually explores rich temporal information to improve detection performance. According to the usage of temporal information, off-the-shelf video object detection methods can be typically divided into two categories: pos-processing and feature aggregation methods.

Post-processing methods first obtain bounding boxes from multiple frames using a detector, then apply linking or tracking algorithms to associate these boxes into tubelets.

For example, Seq-NMS(Wu et al. 2019a) constructs high-scoring bounding box sequences from consecutive frames and leverages them to boost the confidence of weaker detections. T-CNN(Kang et al. 2018) propagates bounding boxes across frames using optical flow, and combines tracking algorithms to generate long sequences of tubelets. However, these post-processing methods heavily rely on the quality of the detector and fail to fully utilize the temporal information inherent in videos, resulting in suboptimal performance.

Feature aggregation methods 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(Zhu et al. 2017), use optical flow networks to guide the aggregation of features across frames. The DFF(Zhu et al. 2016) method applies a large-parameter network to sparse keyframes and propagates deep features to other frames through the flow field, thereby improving inference speed. While these methods allow for end-to-end training by performing early-stage feature aggregation, the performance gains are often limited.

In contrast, proposal-level aggregation methods aggregate features at the proposal stage. For example, SELSA(Wu et al. 2019b) aggregates semantic features from the entire sequence rather than just adjacent frames, while MEGA(Chen et al. 2020) takes into account both global and local temporal information, utilizing a memory mechanism to aggregate features at the proposal level. These methods achieve better performance by fully leveraging the temporal relationships between proposal-level features. However, since they are built upon the base models, their performance is often constrained by the limitations of the underlying models, leading to suboptimal results.

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 $\mathcal{C}_{train}$, traditional VOD aims to train a model $f_\theta(\cdot)$. This model is designed to be evaluated on a test dataset $\mathcal{D}_{test} = \{V_i\}_{i=1}^{N}$, where $V_i \in \mathbb{R}^{T_i \times H_i \times W_i \times 3}$ 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 $\{\mathbf{b}_t\}_{t=1}^{T_i} \in \mathbb{R}^{T_i \times H_i \times W_i \times 3}$ and corresponding class labels $c \in \mathcal{C}_{train}(\mathcal{C}_{base})$ 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 $\mathcal{D}_{test}$. Specifically, during inference, given a test video sequence $V_i \in \mathbb{R}^{T_i \times H_i \times W_i \times 3}$, the trained model is expected to predict all object bounding boxes $\{\mathbf{b}_t\}_{t=1}^{T_i} \in \mathbb{R}^{T_i \times H_i \times W_i \times 3}$ and the category label $c \in (\mathcal{C}_{train} \cup \mathcal{C}_{novel})$ and each object $\mathbf{b}$ in V_i:

$$f_\theta(V_i) = \{\{\hat{b}_k, c_k\}_{k=1}^{K_t}\}_{t=1}^{T_i}, \tag{1}$$

Where K_t represents the total number of objects in the t-th frame, and category c^k belongs to the intersection of the training categories and novel categories. Additionally, $\hat{b}_k = \{x, y, w, h\}$ 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 and use 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.

Structure of OV-VOD

After defining the Open-Vocabulary VOD task, this section introduces our proposed Open-Vocabulary VOD method, OV-VOD, as illustrated in Fig.2. Overall, our model introduces two key improvements at the video level compared to existing methods,ViLD. (i)an Instance-Level Memory Attention Module, This module employs attention mechanisms at the instance level to aggregate instance features from the key frame and supporting frames in memory, thereby enhancing the proposal-level features of the key frame; (ii)a Pixel-Level Memory Attention Module, This module utilizes attention mechanisms at the pixel level to aggregate pixel-level features from the key frame and supporting frames in memory, improving the pixel-level representation of the key frame.It is worth noting that the second module is a conceptual proposal and has not been experimentally validated due to time constraints.We detail the architecture in the following sections.

Instance-Level Memory Attention

The primary difference between videos and images is that videos are continuous sequences of images, where the objects across the frames generally exhibit high temporal consistency. Therefore, an intuitive approach to improving the performance of Open-Vocabulary object detectors is to fully exploit the temporal information in videos, similar to traditional video object detection methods. Specifically, when detecting on a key frame I_t, we aggregate a set of supporting frames $\{I_s\}_{s=1}^{T}$ from the same video.where T denotes the number of supporting frames used during inference for the target frame. Inspired by SAM2(Ravi et al. 2024), we extend the ViLD framework by aggregating instance-level features.

Let F_t denote the proposal-level features of the target frame, and $\{F_s\}_{s=1}^{T}$ denote the proposal-level features of the supporting frames. First, the proposal-level features of the target frame F_t are passed through a self-attention layer to obtain spatially enhanced features F_t' allowing for the extraction of spatial relationships between different proposals

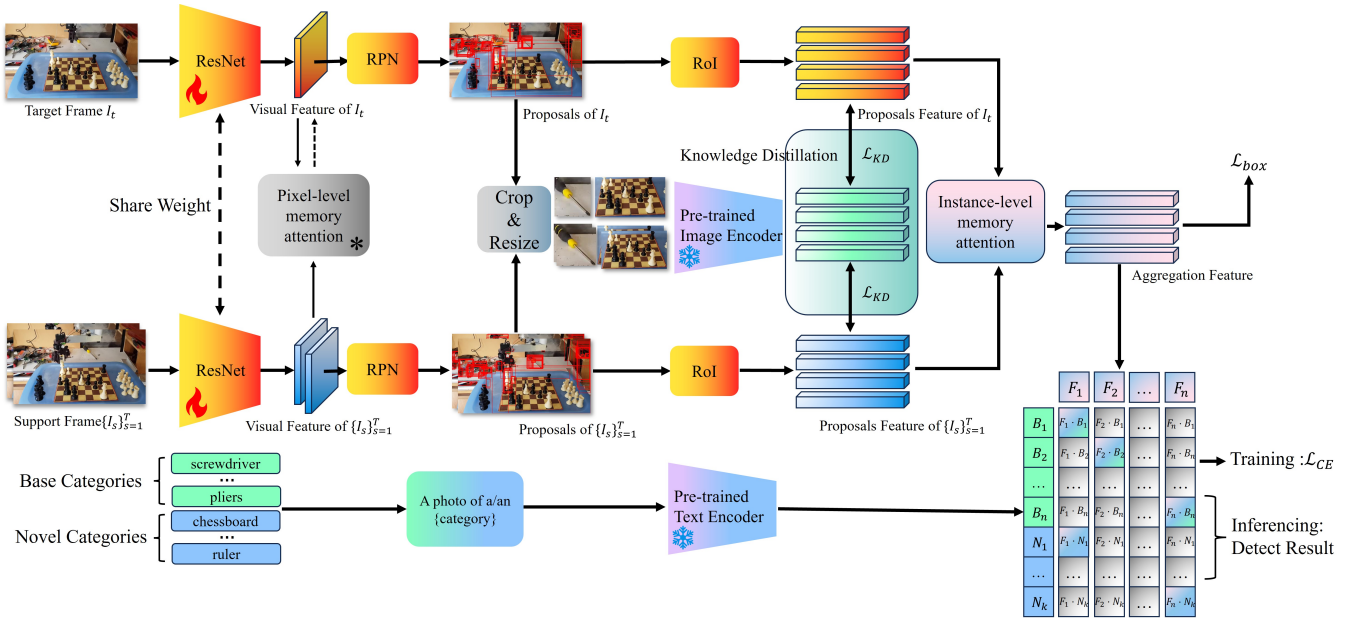


Figure 2: Overview of OV-VOD.∗Indicates a conceptual module that has not yet undergone rigorous experimental validation.

within the same frame. Next, the enhanced features F_t' of the target frame and the features of the supporting frames $\{F_s\}_{s=1}^T$ are processed through a cross-attention layer, followed by a projection layer, to obtain the aggregated features F_t^*.Through the cross-attention mechanism, the proposal-level features of the key frame can effectively learn temporal information across different frames, enhancing the proposal-level features in the target frame by incorporating similar object representations from different frames.

The attention layers are stacked n times, with n as a hyperparameter. By learning through multiple layers of self-attention and cross-attention, the proposal-level features of the target frame effectively integrate spatial information from different instances and temporal information across different frames, leading to improved final classification performance. The entire process is illustrated in Fig.3

Pixel-Level Memory Attention*

Due to time constraints, we were unable to conduct a detailed design and experimental analysis of the Pixel-Level Memory Attention module. However, for the sake of completeness, we have decided to present our preliminary idea in this paper. Our initial concept is largely aligned with the Instance-Level Memory Attention approach, where the features of the target frame are enhanced by aggregating features from supporting frames in memory. Although we cannot provide a detailed description of the internal structure of this module, we can present an abstract representation of the process.

Let Φ denote the feature extraction network, then the pixel-level features of the key frame f_t can be expressed as:

$$f_t = \Phi(FPN(I_t)), \tag{2}$$

where FPN represents the feature pyramid network. Similarly, the pixel-level features of the supporting frames can be denoted as $\{f_s\}_{s=1}^T$. The aggregated pixel-level features of the target frame, f_t^*, can then be expressed as:

$$f_t^* = PLMA(f_t, \{f_s\}_{s=1}^T), \tag{3}$$

where PLMA denotes Pixel-Level Memory Attention.

In future work, we plan to refine the internal design of the Pixel-Level Memory Attention module and provide rigorous experimental validation to demonstrate its effectiveness.

Training and Loss

Due to the memory attention module we introduced requiring the learning of temporal information, we were inspired by OV-Track(Li et al. 2023) and used image generation algorithms to generate a support frame for each static image on the LVIS dataset, so that each image can be viewed as a video segment with only two frames.

For the training loss, we adapted the loss from ViLD. First, $\mathcal{L}_{text}$ initially adopts the CLIP(Radford et al. 2021) approach to replace the original classification loss, assuming f_t^* is the feature used for classification in the target frame, and t_i represents the text embedding of the i-th category obtained through a pre-trained text encoder. Then, $\mathcal{L}_{text}$ can be expressed as:

$$\mathbf{z}(t) = \left[sim(\boldsymbol{f}_t^*, \mathbf{e}_{bg}), \cdots, sim(\boldsymbol{f}_t^*, \mathbf{t}_{|C_B|})\right], \tag{4}$$

$$\mathcal{L}_{text} = \frac{1}{N} \sum_{r \in P} \mathcal{L}_{CE}(softmax(\mathbf{z}(r)/\tau), y_r), \tag{5}$$

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

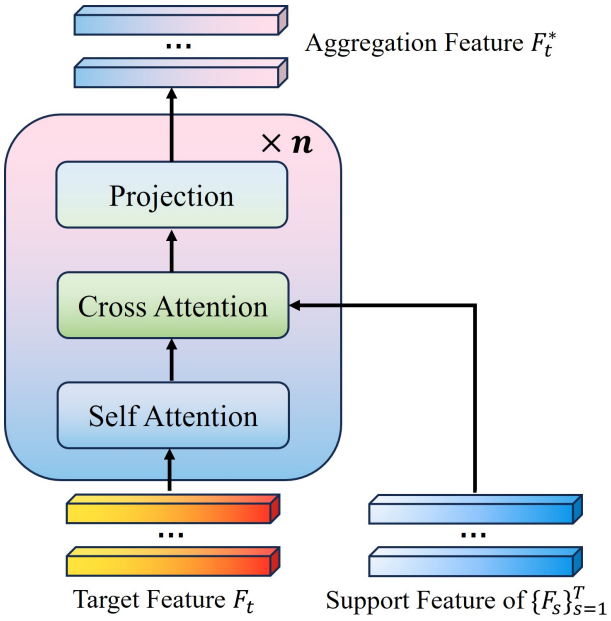


Figure 3: The architecture of Instance-Level Memory Attention

For the knowledge distilling loss. Let $\mathcal{V}$ denotes Pretrained image encoder. Specifically, we align region embedding $\mathcal{R}(\phi(I), \tilde{r})$ to image embedding $\mathcal{V}(crop(I, \tilde{r}))$.And to make training more efficient, we extract M proposals $\tilde{r} \in \tilde{P}$ offline for each training image, and precomputed the M image embeddings. These proposals can contain objects in both $\mathcal{C}_B$ and $\mathcal{C}_N$, as the network can generalize. Thus, $\mathcal{L}_{KD}$ can be expressed as:

$$\mathcal{L}_{\text{KD}} = \frac{1}{M} \sum_{\tilde{r} \in \tilde{P}} \|\mathcal{V}(\text{crop}(I, \tilde{r})) - \mathcal{R}(\phi(I), \tilde{r})\|_1, \quad (6)$$

The total training loss is simply a weighted sum of both objectives:

$$\mathcal{L} = \mathcal{L}_{text} + w \cdot \mathcal{L}_{\text{KD}} + \mathcal{L}_{box}. \quad (7)$$

where w is a hyperparameter weight of distilling the image embedding.

Experiments

We first introduce the relevant datasets and evaluation metrics. Subsequently, we provide additional details about the proposed method. Finally, we present our experimental results along with an analysis of those results.

Datasets and Metrics

We trained OV-VOD on the LVIS dataset. Since the introduced memory attention module requires learning temporal information from supporting frames, but the LVIS dataset is an open-vocabulary image dataset where images are entirely independent, we addressed this issue by adopting the approach of OV-Track. Specifically, we used an image generation algorithm to generate one adjacent frame for each image in the training set, converting each image in LVIS into a short video clip consisting of two frames.

We then evaluated our model's performance on the large-scale open-vocabulary video instance segmentation dataset, LV-VIS. It is worth noting that the original annotation format of LV-VIS only includes mask labels. However, due to the similarity between instance segmentation and object detection tasks, the mask annotations can be easily converted into bounding box annotations.

In future work, we plan to further evaluate our model on the BURST and TAO datasets. These datasets are primarily used for evaluating open-vocabulary tracking methods, but their annotations can also be converted into bbox format, making them suitable for assessing the performance of Open-Vocabulary VOD methods.

Notably, all methods presented in the experiments were not fine-tuned on the respective evaluation datasets. For a fair comparison, all models were trained on the LVIS dataset and subsequently evaluated in a zero-shot manner on the corresponding evaluation datasets.

LVIS is a widely used image open-vocabulary detection dataset, which contains a large set of 1203 categories. Following ViLD setting, we take frequent and common categories as the base categories and set 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 video detection methods.

TAO is a dataset designed for evaluating open-vocabulary multi-object tracking methods. Since the annotations for multi-object tracking tasks are similar to those for detection tasks, we believe that despite the potential for incomplete object annotations, the dataset's category diversity makes it suitable for evaluating open-vocabulary video detection methods. TAO contains a total of 363 categories, following the LVIS setting, with 73 novel categories and 290 base categories.

BURST is a recently introduced video dataset extending TAO. BURST contains 425 base categories and 57 novel categories following the partitions in LVIS.

Implementation Details

Baseline Models. We selected several existing two-stage open-vocabulary object detection methods as our baseline models, specifically ViLD and DetPro, due to their methodological similarities and incremental improvements. We first chose to validate the effectiveness of our approach on ViLD.

OV-VOD. To ensure a fair comparison with the baselines, we used the same backbone, ResNet-50, for our experiments. The number of heads in the memory attention module was set to 16, and the memory attention layer was stacked for a total of 4 layers. Following the settings in ViLD, the temperature coefficient τ was set to 0.1 during training and

0.007 during inference. The weight coefficient w for the loss $\mathcal{L}_{\text{KD}}$ was set to 0.5.

Training Details. We train OV-VOD on LVIS for 6 epochs with a batch size of 1, Each GPU processes a mini-batch, with each mini-batch containing one target frame and its supporting frames.We adopt SGD optimizer. Due to limited computational resources, we initialized the model with ViLD's pre-trained parameters on LVIS and froze these parameters during the training of the memory attention module. The base learning rate was set to 0.1 and decayed to one-tenth of its value at the 3rd and 5th epochs. The momentum was set to 0.9, and the weight decay was set to 0.0001. A warmup strategy was applied during the first 1,000 iterations.To ensure fairness in the experiments, we adopted the same data augmentation strategy as DetPro. Training was conducted on 4 4090 GPUs for approximately 11 hours, and all inference was performed on a single 4090 GPU.

Note that since the experiments are not yet fully completed, the above experimental parameter settings reflect the best results obtained so far and may not represent the final parameter configuration of the model.

Results on LV-VIS dataset

Due to time constraints, we have not yet achieved satisfactory experimental results. However, we still present many of the results obtained during our experiments, even though they are far from ideal. We also provide a detailed analysis of the reasons behind their suboptimal performance. After analyzing the results of early experiments and making corresponding improvements, performance has shown some improvement. Nevertheless, the final model's performance remains unsatisfactory. In subsequent sections, we will analyze the reasons for the model's suboptimal performance under the current best settings and propose directions for improvement in future work.All relevant experimental processes and the results of the baselines on the LV-VIS validation set are shown in Table 2.

We first conducted a series of experiments on ViLD, where expi represents the experiment index. Due to the long experimental period, the data for experiments exp2 and exp3 were not saved because their model performance was very poor and similar to that of exp1. We also report the performance results of DetPro. From an overall performance comparison, DetPro demonstrates the best results so far. However, we acknowledge that due to time constraints, many SOTA methods have not yet been thoroughly tested, and completing these experiments will be a focus of future work.In the following sections, we will present the detailed ablation results for each experiment and provide related analysis.

Ablation Study on LV-VIS dataset

Due to time constraints, we were unable to achieve satisfactory model performance. Therefore, we present some failed experimental results along with the best performance achieved so far. The detailed ablation results for each experiment are shown in Table 3.

Experiment (a) was our initial attempt at implementing the memory attention module. In this experiment, we did not freeze any model parameters; instead, we simply loaded the pre-trained ViLD parameters and randomly initialized the parameters of the memory attention module. A uniform learning rate and weight decay strategy were applied to all model parameters.As shown in Table 3, the performance of Experiment (a) was very poor. The most critical metric, AP_n, decreased by 4.9 compared to the baseline. Furthermore, we observed that the model's recall rate AR also dropped significantly, with a reduction of 30.1 compared to the baseline. This directly indicates that a large portion of the model parameters deviated substantially. Specifically, while we expected the model to learn the parameters of the memory attention module during fine-tuning, the overly large learning rate caused the model to forget much of the previously acquired knowledge.We concluded that the primary reason for the failure of Experiment (a) was the catastrophic forgetting of previously learned knowledge during the fine-tuning process.

Experiment (b) was conducted after identifying the reasons for the failure of Experiment (a). Specifically, in Experiment (b), we froze all parameters initialized with pre-trained weights to ensure that the pre-trained knowledge would not be forgotten during fine-tuning.As shown in Table 3, the performance of Experiment (b) improved significantly compared to Experiment (a), with most metrics approaching the baseline. However, the performance still fell short of the baseline, with the critical AP_n metric remaining 0.9 lower than the baseline. Nonetheless, the substantial performance improvement from Experiment (a) to Experiment (b) highlights the issues in the fine-tuning training strategy used in Experiment (a).In future work, we will explore assigning different learning rates and weight decay values to different parameters, which will be discussed in detail in the results analysis and future outlook sections.

Although Experiment (b) achieved a significant performance improvement over Experiment (a), it still fell short of the baseline. We then carefully analyzed the entire experimental pipeline to identify the reasons for the model's performance degradation.

We found that due to the design of the Mask R-CNN(Girshick et al. 2014) architecture, positive and negative sampling of image proposals was performed during training to ensure model accuracy. This design was carried over to the processing of proposals from the supporting frames, resulting in a fixed ratio of positive and negative samples during training. However, during inference, all proposals need to be classified and localized. We hypothesize that this design hinders the memory attention module from effectively learning its parameters because the distribution of positive and negative samples in the proposals differs between the training and inference stages.

To address this, we removed positive and negative sampling for supporting frame proposals during training. Additionally, to enable the cross-attention mechanism to better learn the relationships between the key frame and multiple supporting frames, we included the target frame as one of the supporting frames during training.As shown in Table 3, the results of Experiment (c) surpassed the baseline, but the performance improvement remains limited, with the critical

Method	Backbone	Detection			Instance segmentation		
		AP_n	AP_b	AP	AP_n	AP_b	AP
ViLD*	ResNet50	8.7	11.7	10.0	8.3	11.4	9.6
Detpro	ResNet50	**9.7**	**12.0**	**10.6**	**9.2**	**11.5**	**10.2**
exp1	ResNet50	3.8	5.3	4.7	3.6	5.1	4.6
...	...	...	...	...	...	...	...
exp4	ResNet50	7.8	10.2	8.8	7.3	9.8	8.4
exp5	ResNet50	8.8	11.9	10.1	8.4	11.5	9.7

Table 2: The zero-shot performance comparison on LV-VIS validation. The AP, AP_n, and AP_b mean the average precision of overall categories, novel categories, and base categories.

(a)	(b)	(c)	AP_n	AP_b	AP	AP_{50}	AR
			8.7	11.7	10.0	15.4	41.5
✓			3.8(-4.9)	5.3(-6.4)	4.7(-5.3)	6.5(-8.9)	11.4(-30.1)
	✓		7.8(-0.9)	10.2(-1.5)	8.8(-1.2)	11.2(-4.2)	36.6(-4.9)
		✓	**8.8**(+0.1)	**11.9**(+0.2)	**10.1**(+0.1)	**15.5**(+0.1)	**42.0**(+0.5)

Table 3: Ablation Study Results of Several Experiments on the LV-VIS Validation.

AP_n metric exceeding the baseline by only 0.1. Nevertheless, we believe that our model still has room for further optimization and improvement.

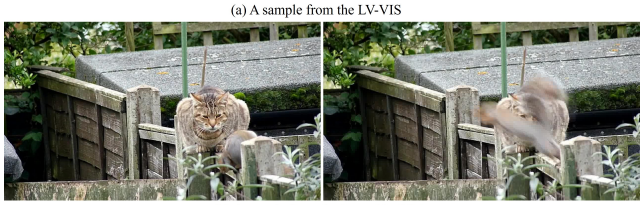
Result Analysis and outlook

In this section, we will provide a detailed analysis of the suboptimal results obtained so far and outline directions for future work.

Why was the training strategy set to fine-tuning instead of following the baseline's training strategy? Considering the limitation of computational resources, we opted for a fine-tuning strategy. Even with the more efficient training approach introduced by DetPro, ViLD still requires 20 epochs of training on the LVIS dataset. However, this could also be one of the reasons contributing to the suboptimal performance of our model. In our training strategy, all pre-trained model weights were frozen, and only the newly introduced modules were trained during the fine-tuning stage. This resulted in the model's parameters not being trained jointly, which may have caused the model to fall into a local optimum.In future work, we plan to further improve the training strategy. For example, we could follow the baseline's approach and retrain all model parameters, or adopt a strategy where smaller learning rates and larger weight decays are applied to the pre-trained parameters, while larger learning rates and smaller weight decays are applied to the new module parameters. This would allow all model parameters to be trained during the fine-tuning stage.

Why does the memory attention not work?In fact, the memory attention module did not perform as effectively as we expected. Our idea of aggregating proposal-level features from supporting frames was inspired by traditional video object detection research, where extensive experiments have shown that aggregating proposal-level features from supporting frames often significantly improves

(a) A sample from the LV-VIS



(b) A sample from the ImageNet VID

Figure 4: (a) A sample of sampled frames from a video in the LV-VIS dataset, where the objects within the frames are generally visually clear. (b) A sample of sampled frames from a video in the ImageNet VID dataset, where objects in the frames suffer from severe occlusion and motion-induced blur. These differences in data distribution between the two datasets may be the primary reason why the memory attention module does not work effectively.

model performance on the ImageNet VID dataset. However, this approach did not yield satisfactory results on the LV-VIS dataset.Upon closely examining the data distributions of the two datasets, a significant difference becomes apparent: many videos in ImageNet VID contain a large number of consecutive frames, while videos in LV-VIS have fewer frames, obtained through a specific frame-sampling ratio. As a result, most objects in LV-VIS are visually clear, as shown in Fig.4. Issues commonly seen in real-world videos, such as motion blur, occlusion, or abnormal poses caused by object movement, are rarely present. Aggregating proposal features from supporting frames is typically effective in addressing such issues, which are known to reduce the performance of single-frame detectors.This observation has led us to reconsider the primary challenges in the LV-VIS dataset. It appears that the challenges posed by the video dimension may not be the most significant; instead, the primary factor affecting model performance likely lies in the open-vocabulary aspect. In future work, we plan to not only en-

hance the model's performance from the video perspective but also explore ways to further improve its performance in open-vocabulary settings.Additionally, we will evaluate the memory attention module on the BURST and TAO datasets to verify whether the suboptimal performance is due to the specific video data distribution in LV-VIS.

Conclusion

In this paper, we propose a novel task, Open-Vocabulary VOD, which aims to detect and classify arbitrary objects in video frames. To better address the challenges of open-vocabulary detection in the video domain, we provide a training strategy that extends the LVIS dataset into a video dataset using image generation methods. Additionally, we introduce open-vocabulary video datasets from related domains for evaluating Open-Vocabulary VOD methods, offering valuable references for future research.Finally, we aim to design a high-performance Open-Vocabulary VOD benchmark method that mitigates the limitations of static detectors in the video domain by aggregating features from supporting frames. Although our experiments have not yet achieved satisfactory results, we remain optimistic that future work will yield promising advancements.

References

Athar, A.; Luiten, J.; Voigtlaender, P.; Khurana, T.; Dave, A.; Leibe, B.; and Ramanan, D. 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.

Chen, Y.; Cao, Y.; Hu, H.; and Wang, L. 2020. Memory Enhanced Global-Local Aggregation for Video Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Damen, D.; Doughty, H.; Farinella, G. M.; Fidler, S.; Furnari, A.; Kazakos, E.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, 720–736.

Dave, A.; Khurana, T.; Tokmakov, P.; Schmid, C.; and Ramanan, D. 2020. Tao: A large-scale benchmark for tracking any object. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, 436–454. Springer.

Du, Y.; Wei, F.; Zhang, Z.; Shi, M.; Gao, Y.; and Li, G. 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.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*.

Gu, X.; Lin, T.-Y.; Kuo, W.; and Cui, Y. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.

Gupta, A.; Dollar, P.; and Girshick, R. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5356–5364.

Kang, K.; Li, H.; Yan, J.; Zeng, X.; Yang, B.; Xiao, T.; Zhang, C.; Wang, Z.; Wang, R.; Wang, X.; and Ouyang, W. 2018. T-CNN: Tubelets With Convolutional Neural Networks for Object Detection From Videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10): 2896–2907.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10965–10975.

Li, S.; Fischer, T.; Ke, L.; Ding, H.; Danelljan, M.; and Yu, F. 2023. Ovtrack: Open-vocabulary multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5567–5577.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Wang, H.; Yan, C.; Chen, K.; Jiang, X.; Tang, X.; Hu, Y.; Kang, G.; Xie, W.; and Gavves, E. 2024. OV-VIS: Open-Vocabulary Video Instance Segmentation. *International Journal of Computer Vision*, 1–18.

Wu, H.; Chen, Y.; Wang, N.; and Zhang, Z. 2019a. Sequence Level Semantics Aggregation for Video Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Wu, H.; Chen, Y.; Wang, N.; and Zhang, Z. 2019b. Sequence Level Semantics Aggregation for Video Object Detection. *ICCV 2019*.

Wu, S.; Zhang, W.; Jin, S.; Liu, W.; and Loy, C. C. 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.

Zang, Y.; Li, W.; Zhou, K.; Huang, C.; and Loy, C. C. 2022. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, 106–122. Springer.

Zhang, H.; Zhang, P.; Hu, X.; Chen, Y.-C.; Li, L.; Dai, X.; Wang, L.; Yuan, L.; Hwang, J.-N.; and Gao, J. 2022. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35: 36067–36080.

Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L. H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; et al. 2022. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16793–16803.

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, 350–368. Springer.

Zhu, X.; Wang, Y.; Dai, J.; Yuan, L.; and Wei, Y. 2017. Flow-Guided Feature Aggregation for Video Object Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 408–417.

Zhu, X.; Xiong, Y.; Dai, J.; Yuan, L.; and Wei, Y. 2016. Deep Feature Flow for Video Recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4141–4150.