Temporal Object-Aware Vision Transformer for Few-Shot Video Object Detection

Yogesh Kumar and Anand Mishra

Indian Institute of Technology Jodhpur, India
kumar.204@iitj.ac.in, mishra@iitj.ac.in

Abstract

Few-shot Video Object Detection (FSVOD) addresses the challenge of detecting novel objects in videos with limited labeled examples, overcoming the constraints of traditional detection methods that require extensive training data. This task presents key challenges, including maintaining temporal consistency across frames affected by occlusion and appearance variations, and achieving novel object generalization without relying on complex region proposals. Our novel object-aware temporal modeling approach addresses these challenges by incorporating a filtering mechanism that selectively propagates high-confidence object features across frames. This enables efficient feature progression, reduces noise accumulation, and enhances detection accuracy in a few-shot setting. By utilizing few-shot trained detection and classification heads with focused feature propagation, we achieve robust temporal consistency without depending on explicit object tube proposals. Our approach achieves performance gains, with AP improvements of 3.7% (FSVOD-500), 5.3% (FSYTV-40), 4.3% (VidOR), and 4.5% (VidVRD) in the 5-shot setting. Further results demonstrate improvements in 1-shot, 3-shot, and 10-shot configurations. We make the code public at: https://github.com/yogesh-iitj/fs-video-vit.

Introduction

Object detection in videos has witnessed significant progress with the advent of deep learning-based methods in the last few years (Zhu et al. 2017; Wu et al. 2019; Yu et al. 2021; Fan, Tang, and Tai 2022a). However, most traditional methods focus on closed-set detection, where models are trained on predefined object categories with extensive labeled samples per category. This makes them impractical for many real-world applications that require the detection of novel object categories with limited training samples. Few-Shot Video Object Detection (FSVOD) (Fan, Tang, and Tai 2022a) addresses this challenge by enabling models to detect novel objects in videos using only a few examples, as shown in Figure 1.

Unlike widely-studied few-shot object detection in images (Dong et al. 2022; Wang et al. 2020; Kang et al. 2019; Wu et al. 2020; Sun et al. 2021), FSVOD must account for temporal consistency, motion blur, occlusion, and variations in object appearance between frames, making the task substantially more difficult (Fan, Tang, and Tai 2022a). Existing approaches for FSVOD (Fan et al. 2020; Zhou, Koltun, and Krähenbühl 2020; Han et al. 2023a; Kumar et al. 2024) face several limitations, such as they often process frames independently without leveraging temporal information (Fan et al. 2020), struggle with false positives due to weak temporal matching (Fan, Tang, and Tai 2022b), use region or tube proposal networks not optimized for few-shot settings (Fan et al. 2020; Fan, Tang, and Tai 2022b; Han et al. 2023b), and have difficulties in discriminating with visually-similar objects and occlusions (Kumar et al. 2024).

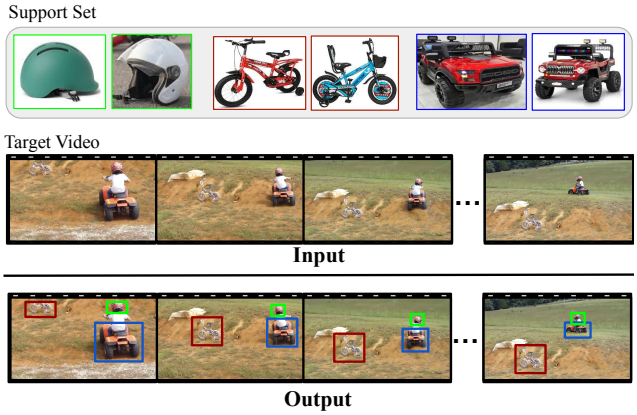

Figure 1: Given a support set containing novel objects and a target video, the goal of FSVOD (Fan, Tang, and Tai 2022b) is to detect all the instances of novel objects on the target video. We propose an object-aware temporally consistent few-shot object detection framework that significantly improves the state of the art for this task.

Recent advances in vision-language and video grounding models (Heigold et al. 2023a; Jung et al. 2025; Kumar and Mishra 2023; Kumar et al. 2025b,a) have shown strong capabilities in open-set video object detection and grounding. Video OWL-ViT (Heigold et al. 2023a), built on CLIP's vision-language pretraining, can be naturally adapted for few-shot video object detection by replacing text prompts with visual prompts from support images, enabling detection of novel objects from just a few examples. However, this adaptation lacks object-aware temporal modeling, as it pro-

cesses frames independently of the detected objects, leading to inconsistent detection across time.

Our approach leverages a vision-language pretrained frame encoder to transfer semantic knowledge from image-text data, enabling recognition of novel object categories. Furthermore, contrastively-trained vision encoders are more effective in distinguishing between visually similar objects across different categories and handle partially visible or occluded objects compared to CNN backbones (like ResNet (He et al. 2016)) used in prior methods (Fan, Tang, and Tai 2022b; Han et al. 2023b; Kumar et al. 2024). This robust feature representation serves as the foundation for our temporal consistency mechanism. To enhance this consistency, we fuse representations of previously detected objects with current frame detections, preserving relevant temporal information and adapting to appearance changes over time. Unlike methods that process frames independently, our approach maintains memory of past observations for better temporal reasoning. Our detection is directly optimized to condition on few-shot visual examples, providing better results compared to traditional region proposal networks. Further, unlike conventional tracking-by-detection pipelines (Fan, Tang, and Tai 2022b; Han et al. 2023b), our end-to-end learning of video-specific object representations leads to superior generalization.

In summary, we make the following contributions: (i) We revisit the relatively under-explored task of Few-Shot Video Object Detection (FSVOD) and introduce simple yet effective task-specific innovations. Our key contribution is a novel object-aware temporal modeling approach, that selectively propagates high-confidence object features across frames. This enables efficient feature progression, reducing noise accumulation, and improving detection accuracy in few-shot video scenarios. (ii) For the first time, we adapt large-scale pre-trained vision-language models for FSVOD, aligning with recent trends in open-world detection. This integration enables novel object detection by leveraging broad visual-linguistic knowledge, facilitating efficient adaptation and strong generalization to unseen categories. (iii) We perform extensive evaluations against baselines and state-of-the-art methods on four public FSVOD benchmarks. Our approach achieves significant performance gains, with AP improvements of 3.7% (FSVOD-500), 5.3% (FSYTV-40), 4.3% (VidOR), and 4.5% (VidVRD) in the 5-shot setting. Further, we observe consistent improvements across 1-shot, 3-shot, and 10-shot setups, demonstrating robust generalization across different few-shot scenarios.

Related Work

Few-Shot Object Detection: Recent advances in Few-Shot Object Detection (FSOD) have been driven by attention mechanisms, transformers, and meta-learning techniques. Approaches like CoAE (Hsieh et al. 2019) and OS2D (Osokin, Sumin, and Lomakin 2020) have pushed the boundaries of one-shot object detection in images through co-attention mechanisms and versatile one-stage approach, respectively. For traditional FSOD, methods such as feature reweighting (Kang et al. 2019), contrastive proposal encoding (Sun et al. 2021), and hierarchical learning (She et al. 2022) have demonstrated improved performance across various benchmarks. In addition, several DETR variants (Dong et al. 2022; Dai et al. 2021) have emerged and shown promising results for few-shot object detection in images. Although FSOD in images provides the foundational techniques and can be trivially extended to videos by performing few-shot detection on frames independently. Our experiments suggest that they often fall short. The inherent temporal nature of videos introduces additional complexities that require specialized approaches to maintain consistency across frames while preserving the few-shot learning paradigm. We introduce object-aware temporal consistency to overcome the inherent challenges specific to video.

While few-shot object detection in images has been extensively studied, its video counterpart remains relatively underexplored. Qi et al. (Fan, Tang, and Tai 2022b) have formally introduced a few-shot video object detection task and accompanying datasets. They proposed a proposal-based approach, where object detection trajectories are first proposed and later refined using the matching network. Kumar et al. (Kumar et al. 2024) have specifically studied this problem for one-shot settings using a novel query-guided variant of DETR. More recently, Wentano et al. (Han et al. 2023b) enhanced temporal information by context fusion. Our work addresses key limitations of existing FSVOD methods by utilizing a transformer-based backbone (OWL-ViT) pre-trained on large-scale image-text pairs, replacing the CNN-based backbones used in prior approaches and improves semantic knowledge transfer. In addition, unlike proposal-based networks not optimized for few-shot scenarios, our proposal-free method conditions detection directly on few-shot-trained heads and improves generalization to novel object categories.

Open-World Object Detection: There is a fundamental shift in computer vision literature toward detecting novel visual concepts driven by models pre-trained on large-scale vision and language data (Han and Lim 2024; Kaul, Xie, and Zisserman 2023). These large models help in detecting novel objects by leveraging their broad visual and linguistic knowledge, efficient adaptation, and improved generalization to unseen categories. Recent works like OWL-ViT (Minderer et al. 2022b) and Video OWL-ViT (Heigold et al. 2023b) leverage vision-language models pretrained on large-scale data to detect open-vocabulary concepts. Methods like GLIP (Li et al. 2022) and OpenSeed (Zhang et al. 2023) further advance this paradigm by incorporating grounded pretraining and structured knowledge distillation.

This growing trend highlights the importance of flexible, prompt-driven detection systems that can adapt to diverse visual inputs. Open-world object detectors can be extended for image-conditioned detection, enabling users to detect objects by providing a few visual examples. In this work, we leverage OWL-ViT encoders for image-conditioned detection in a few-shot setup and enhance the existing Video OWL-ViT (Heigold et al. 2023a) by introducing an object-aware decoder that significantly improves temporal consistency across video frames, and thereby the few-shot video object detection performance.

Method

Few-shot video object detection (FSVOD) aims to detect novel object categories that were not encountered during training. The task involves a *support set* containing a few images of novel object classes and a *target video* consisting of multiple frames. The goal is to identify and localize all instances of the objects from the support set across every frame of the target video, as shown in Figure 1.

Problem Formulation

In the FSVOD task, we aim to locate and classify all instances of objects belonging to novel classes in a target video using only a few support examples. Formally, given an N-way K-shot support set $\mathcal{S} = \{(\mathbf{I}_{i,j}, c_i)|i = 1, \cdot, N; j = 1, \cdot, K\}$ where $\mathbf{I}_{i,j}$ represents the j-th support image of class c_i, and a target video $\mathcal{V} = \{\mathbf{F}_t|t = 1, ..., T\}$ with T frames, our goal is to detect all instances of the N novel classes across all frames. For each frame $\mathbf{F}_t$, the model predicts a set of object bounding boxes along with their class labels.

Design Motivation and Architectural Overview

Our proposed approach is designed to address the following unique challenges associated with FSVOD task: (i) the limited supervision requiring efficient knowledge transfer, (ii) the necessity for temporal consistency across sequential frames despite appearance variations due to motion, occlusion, and viewpoint shifts, and (iii) the open-set recognition capability to distinguish between novel objects and background elements without explicit training.

To address these challenges, we propose an object-aware, temporally consistent few-shot video object detection framework. Our approach consists of the following main components: a language-aligned vision encoder that provides semantically rich visual representations, a temporal fusion decoder that selectively propagates high-confidence object features across frames and, few-shot matching and detections heads that aligns target frame features with support examples. Figures 2,3 illustrates our overall framework.

Object-aware Temporally consistent FSVOD framework

Language-Aligned Vision Encoder: Traditional CNN backbones like ResNet (He et al. 2016), trained solely on visual classification tasks, often struggle to generalize to novel classes with limited examples. To overcome this limitation, the recent trend (Han and Lim 2024; Heigold et al. 2023a) is to use a pretrained vision encoder that has been aligned with language semantics through large-scale vision-language pretraining. We adopt this by utilizing an OWL-ViT encoder pretrained on large-scale image-text pairs. This encoder transforms each input image $\mathbf{I}$ into a grid of patch embeddings $\mathbf{E} \in \mathbb{R}^{P \times D}$, where P is the number of patches and D is the embedding dimension.

Support Set Encoding: **(i) Support Feature Extraction.** Given a support set $\mathcal{S}$ containing K examples for each of the N novel classes, we extract patch-level features from each support image. For each support image $\mathbf{I}_{i,j}$, the encoder

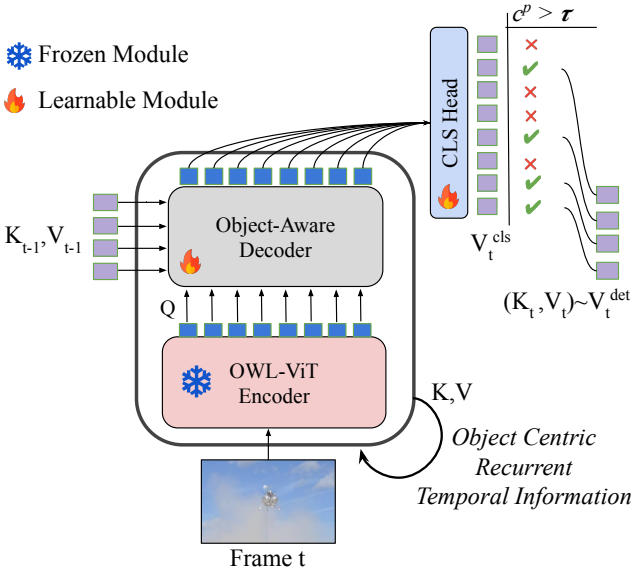

Figure 2: Object-aware Temporal Consistency. We utilize the OWL-ViT encoder output tokens as queries but, critically, only forward the matched object tokens ($c^p > \tau$) as key-value pairs to the next frame's decoder. This mechanism selectively propagates high-confidence object features across frames. This enables efficient feature progression, reducing noise accumulation. We empirically found a threshold (τ) and compared it with class probability (c^p) scores for selective propagation. This selective propagation mechanism significantly reduces noise accumulation across frames and maintains focused, consistent visual representations of detected objects throughout the video sequence.

generates both patch embeddings, $\mathbf{E}_{i,j} \in \mathbb{R}^{P \times D}$, and their corresponding objectness scores for each patch, $\mathbf{s}_{i,j} \in \mathbb{R}^P$. A higher objectness score indicates a higher probability of an object being present in that particular patch. To obtain an object-centric representation, we select the patch with the highest objectness score.

$$p^* = \arg\max_p s^p_{i,j}, \quad \mathbf{z}_{i,j} = \mathbf{e}^{p^*}_{i,j}, \tag{1}$$

where $\mathbf{z}_{i,j}$ is the selected patch embedding that best represents the target object in the support image. **(ii) Support Prototype Generation.** After extracting object-centric features ($\mathbf{z}_{i,j}$) from each support image, we aggregate information from the K support examples to form a class prototype. For each class c_i, the final class prototype $\mathbf{z}_i$ is computed by averaging across the K selected object-centric features. This aggregation creates a robust representation for each novel class that captures essential visual characteristics across multiple support examples, allowing the model to handle intra-class variations more effectively.

Object-Aware Temporal Frame Decoder: **(i) Target Frame Feature Extraction.** For each frame $\mathbf{F}_t$ in the target video, we extract patch embeddings using the OWL-ViT image encoder, $\mathbf{F}_t \in \mathbb{R}^{P \times D}$. **(ii) Temporal Fusion.** To ensure temporal consistency in object detection across frames,

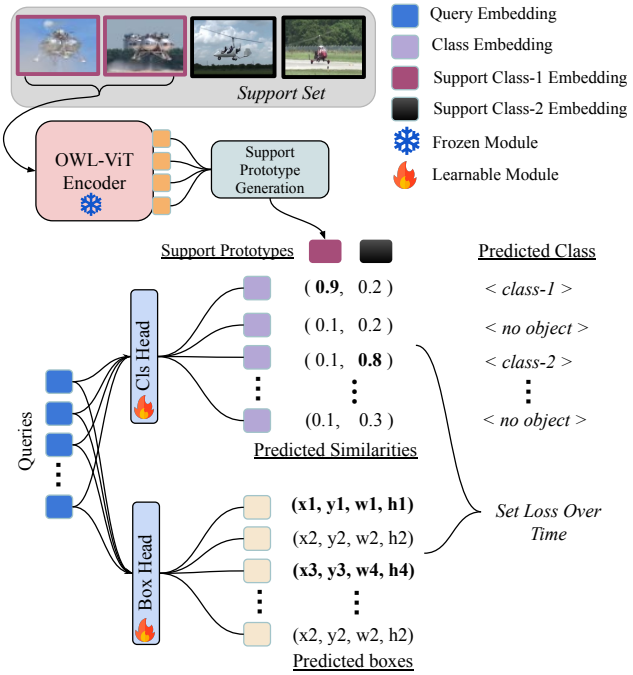

Figure 3: Few-shot Classification and Detection Heads. Our architecture processes object queries through parallel projection heads for classification and localization. Classification embeddings are compared against support embeddings via cosine similarity, while the detection head predicts bounding box coordinates. The temporal consistency is maintained by propagating matched object queries across frames.

we propose a filtering mechanism that selectively propagates high-confidence object features from one frame to the next. This approach facilitates efficient feature progression and minimizes noise accumulation by leveraging the embeddings of successfully detected objects from previous frames. By doing so, it enhances temporal consistency and improves the overall accuracy of object detection in video sequences.

For the first frame ($t = 0$), we process the frame embeddings directly. For subsequent frames ($t > 0$), we incorporate information from previous frame detections via a cross-frame interaction that selectively integrates object-level features to enhance temporal consistency. Specifically, we select embeddings from the few-shot classification head ($\mathbf{V}_t^{cls}$) of the previous frame where detection confidence exceeds a threshold τ (shown in Figure 2):

$$\mathbf{V}_{t-1}^{\text{det}} = \{\mathbf{v}_{t-1}^{p} | \hat{c}_{t-1,i}^{p} > \tau\}, \tag{2}$$

where $\hat{c}_{t-1,i}^{p}$ represents the classification probability for patch p and class i in frame $t-1$. The current frame embeddings $\mathbf{F}_t$ serve as queries, while the embeddings from Cls Head with high confidence score responsible for object detection (Figure 2), $\mathbf{V}_{t-1}^{\text{det}}$ from the previous frame serve as both keys and values. The cross-frame interaction operation is computed as:

$$\mathbf{A}_t = \text{Softmax}\left(\frac{\mathbf{F}_t(\mathbf{V}_{t-1}^{\text{det}})^T}{\sqrt{D}}\right). \tag{3}$$

The resulting attention-weighted features are used to update the current frame representations:

$$\hat{\mathbf{F}}_t = \mathbf{F}_t + \mathbf{A}_t \mathbf{V}_{t-1}^{\text{det}}. \tag{4}$$

These temporally enhanced features $\hat{\mathbf{F}}_t$ are then fed to the classification and localization heads for object detection in the current frame. By using attention between current frame features and previously detected objects representations, the model maintains better temporal consistency while detecting objects in the current frame.

Few-Shot Detection Heads: The temporally enhanced frame embeddings $\hat{\mathbf{F}}_t$ are processed through parallel classification and localization heads to produce the final detection results.

Classification Head. The classification head compares frame embeddings with support class prototypes to determine object categories. It projects frame embeddings to the classification space:

$$\mathbf{V}_t^{cls} = \mathbf{W}_{\text{cls}} \hat{\mathbf{F}}_t, \tag{5}$$

where $\mathbf{W}_{\text{cls}}$ is a learnable projection matrix.
For each patch embedding $\mathbf{v}_t^p$ and class prototype $\mathbf{z}_i$, we compute cosine similarity:

$$s_{p,i} = \frac{\mathbf{v}_t^p \cdot \mathbf{z}_i}{||\mathbf{v}_t^p|| \cdot ||\mathbf{z}_i||}. \tag{6}$$

Each patch is assigned a probability distribution over the N classes and background using softmax normalization of similarity scores, translating feature similarity into class probabilities.

Localization Head. The localization head predicts bounding box coordinates for detected objects through a multi-layer perceptron:

$$\hat{\mathbf{B}}_t = \text{MLP}_{\text{box}}(\hat{\mathbf{F}}_t), \tag{7}$$

where each row $\hat{\mathbf{b}}_t^p \in \mathbb{R}^4$ represents the predicted bounding box parameters (x, y, w, h) for each corresponding patch p. For each patch p, if its maximum classification score exceeds threshold κ, we consider it as a valid detection.

Training Objective: Our model is trained end-to-end by optimizing a combined loss function across all video frames:

$$\mathcal{L} = \sum_{t=1}^{T} \Big[\sum_{m=1}^{M_t} \lambda_{\text{cls}} \cdot \mathcal{L}_{\text{cls}}(\hat{\mathbf{c}}_t^{\sigma_t(m)}, c_t^m) + \sum_{m=1}^{M_t} \lambda_{\text{box}} \cdot \mathcal{L}_{\text{box}}(\hat{\mathbf{b}}_t^{\sigma_t(m)}, \mathbf{b}_t^m) \Big].$$

The classification loss ($\mathcal{L}_{\text{cls}}$) uses cross-entropy to measure discrepancy between predicted and ground-truth class labels, while the localization loss ($\mathcal{L}_{\text{box}}$) combines L1 loss for coordinate distances and generalized IoU loss for handling varying box sizes. Following DETR (Carion et al. 2020), $\sigma_t(m)$ denotes the index of the prediction matched to ground-truth object m, and M_t is the number of objects in frame t. The hyperparameters λ_{cls} and λ_{box} control the weighting of the classification and localization losses.

	Method	FSVOD-500			FSYTV-40			VidOR			VidVRD		
		AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Non-temporal	FR-CNN	18.2	26.4	19.6	9.3	15.4	9.6	21.6	29.1	23.2	14.3	18.6	14.2
	FSOD	21.1	31.3	22.6	12.5	20.9	13.0	-	-	-	-	-	-
	RDN	18.2	27.9	19.7	8.1	13.4	8.6	-	-	-	-	-	-
	CoAE	18.4	29.3	18.9	9.8	18.4	10.1	35.8	43.7	36.1	30.2	38.1	31.8
	Retrieval-Based	1.3	2.4	1.6	0.6	1.1	0.8	1.4	2.6	1.3	1.8	2.5	2.0
	Owl-ViT	14.8	25.3	16.7	10.6	19.8	11.2	30.1	37.3	31.2	23.6	32.1	23.9
	UP-DETR	20.1	27.6	21.2	11.8	19.4	12.1	39.7	47.5	40.0	11.6	18.3	12.5
Temporal	DeepSort+YOLOv8 + CLIP	13.2	21.4	14.4	5.3	9.4	6.2	23.2	30.1	23.6	23.4	33.1	24.1
	TrackFormer + CLIP	14.2	22.6	15.3	6.2	10.5	6.8	25.3	32.7	26.2	25.4	33.8	27.3
	ByteTrack + CLIP	14.7	23.9	15.8	7.9	15.6	8.1	24.9	31.4	25.3	25.6	36.2	26.2
	MEGA	16.8	26.4	17.7	7.8	13.0	8.3	-	-	-	-	-	-
	CTracker	20.1	30.6	21.0	8.9	14.4	9.1	-	-	-	-	-	-
	FairMOT	20.3	31.0	21.2	9.6	16.0	9.5	-	-	-	-	-	-
	CenterTrack	20.6	30.5	21.9	9.5	15.6	9.7	-	-	-	-	-	-
	TACF	26.9	39.2	27.9	15.9	24.2	17.9	-	-	-	-	-	-
	Video Owl-ViT	25.8	36.5	26.1	15.7	26.3	15.9	44.3	52.9	45.7	44.2	53.7	45.4
	FSVOD	25.1	36.8	26.2	14.6	21.9	16.1	45.1	54.3	46.2	40.7	53.5	43.7
	QDETRv	26.1	33.8	23.7	14.1	22.8	15.3	43.4	51.8	44.6	42.8	50.8	41.7
	Ours	**30.6**	**42.9**	**32.1**	**21.2**	**29.8**	**23.5**	**49.4**	**57.3**	**50.2**	**48.7**	**57.8**	**49.2**

Table 1: Performance comparison on FSVOD-500, FSYTV-40, VidOR, and VidVRd in 5-shot set-up. Our Object-aware Video OWL-ViT consistently surpasses both state-of-the-art and baseline methods across all datasets and metrics.

Split	Vid-OR	VidVRD	FSVOD	FSYTV
#Train Videos	6,164	758	2553	1627
#Train Object Queries	57,599	4,395	-	-
#Train Object Categories	75	30	320	30
#Test Videos	32	42	949	608
#Test Object Queries	407	112	-	-
#Test Object Categories	9	5	100	10

Table 2: Overview of the datasets used in this work.

Experiments and Results

Datasets

In this work, we utilized the FSVOD-500, FSYTV-40, VidOR, and VidVRD datasets, as proposed and repurposed in (Fan, Tang, and Tai 2022b; Kumar et al. 2024). These datasets are specifically designed for few-shot video object detection, providing a structured evaluation framework with disjoint training, validation, and testing classes. Detailed statistics of the datasets are provided in Table 2.

Baselines

We compare our method against temporal and non-temporal few-shot object detectors, covering diverse architectural designs and learning strategies.

Non-temporal Methods. We establish strong baselines using image-based few-shot detectors compared in (Fan, Tang, and Tai 2022b). Traditional detectors like Faster R-CNN (Ren et al. 2015) validate core detection performance, while FSOD (Fan et al. 2020) and RDN (Deng et al. 2019) from (Fan, Tang, and Tai 2022b) highlight the impact of specialized few-shot designs. We adapt CoAE (Hsieh et al. 2019), originally for one-shot detection, by selecting high-confidence predictions for few-shot extension. CoAE employs meta-learning and attention for data efficiency. To assess large-scale pre-training, we include OWL-ViT (Minderer et al. 2022a) and UP-DETR (Wang et al. 2021). Retrieval-based detection (Radenović, Tolias, and Chum 2016) is also evaluated as a proposal selection strategy.

Temporal Methods. We evaluate video-based detectors across three categories. First, we adopt temporal baselines from (Fan, Tang, and Tai 2022b), including MEGA (Chen et al. 2020), CTracker (Peng et al. 2020), FairMOT (Zhang et al. 2021), and CenterTrack (Zhou, Koltun, and Krähenbühl 2020). Second, we assess tracking-by-detection methods, DeepSort+YOLOv8 (Wojke, Bewley, and Paulus 2017), TrackFormer (Meinhardt et al. 2022), and ByteTrack (Zhang et al. 2022) combined with CLIP (Radford et al. 2021) for tube-support similarity. Models are trained on support examples to generate spatio-temporal tubes, and tube-support similarity is computed via average pooling. Third, we compare with SoTA one/few-shot video detectors, including FSVOD (Fan, Tang, and Tai 2022b), TACF (Han et al. 2023b), and QDETRv (Kumar et al. 2024). Lastly, we adapt Video OWL-ViT (Heigold et al. 2023a) by replacing its text-encoder with OWL-ViT's image-encoder, integrating our few-shot detection and classification heads.

Performance Measures

Following (Fan, Tang, and Tai 2022b), we adopt Average Precision (AP) as an evaluation metric, computed as the area under the precision-recall curve. AP_{50} and AP_{75} denote performance at IoU thresholds of 50% and 75%, respectively.

Implementation Details

We implement our model using pretrained OWL ViT-L/16 for the vision encoder. The temporal fusion module employs a 4-head cross-attention mechanism with 1024-dimensional hidden states. The classification and localization heads use

	Method	FSVOD-500			FSYTV-40			VidOR			VidVRD		
		AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
1-Shot	CoAE	17.2	27.8	17.2	6.7	17.3	9.8	33.2	42.5	35.2	29.6	37.6	31.5
	UP-DETR	17.3	24.3	18.4	9.7	16.9	9.9	35.6	45.2	37.2	8.5	15.2	9.3
	DeepSort+YOLOv8 + CLIP	12.2	20.1	12.8	4.1	7.9	4.7	21.4	28.4	22.7	21.9	31.7	22.8
	FSVOD	20.7	29.4	21.4	11.4	18.6	12.1	36.8	45.3	36.3	34.5	45.2	35.2
	QDETRv	25.2	34.9	25.8	13.6	24.1	14.2	41.8	50.3	42.2	40.3	49.7	40.1
	Video OWL-ViT	23.8	33.8	24.7	12.4	22.7	13.2	40.3	51.7	41.5	39.8	47.6	40.2
	Ours	**27.4**	**39.1**	**28.9**	**18.3**	**26.2**	**19.7**	**45.2**	**54.1**	**47.6**	**45.3**	**54.5**	**46.8**
3-Shot	CoAE	17.8	28.2	17.6	6.9	17.5	9.8	33.3	42.6	35.6	29.9	37.8	31.9
	UP-DETR	17.9	25.1	18.8	10.2	17.5	10.3	36.2	46.8	38.3	9.3	16.1	10.4
	DeepSort+YOLOv8 + CLIP	12.7	20.4	13.7	4.3	8.3	4.9	22.0	28.8	23.1	22.0	31.9	23.4
	FSVOD	21.9	30.8	22.1	12.3	19.2	12.9	37.6	46.6	37.8	35.4	46.1	36.4
	QDETRv	25.6	35.2	26.3	13.8	24.5	14.5	42.1	50.8	42.7	40.8	50.3	40.6
	Video OWL-ViT	24.1	34.6	25.2	12.7	23.2	13.8	40.6	52.5	41.8	40.4	48.3	40.7
	Ours	**28.2**	**40.5**	**31.1**	**19.6**	**27.2**	**20.9**	**46.4**	**55.6**	**48.2**	**46.5**	**55.7**	**47.2**
10-Shot	CoAE	20.1	29.9	19.7	10.2	19.8	12.3	37.2	44.9	38.2	32.3	39.7	32.7
	UP-DETR	23.2	30.2	23.7	13.7	22.3	15.2	41.8	50.2	43.1	13.7	21.2	15.4
	DeepSort+YOLOv8 + CLIP	16.6	23.7	17.2	7.8	11.5	9.4	26.3	32.9	25.8	25.1	35.6	26.6
	FSVOD	27.2	38.5	28.6	16.1	25.8	17.9	47.3	56.7	48.4	44.7	54.9	44.9
	QDETRv	26.2	33.8	23.8	14.2	23.1	15.6	43.9	52.1	45.2	43.1	53.8	44.2
	Video OWL-ViT	26.7	35.8	27.1	14.2	25.1	14.9	43.7	54.2	44.1	43.7	52.8	44.3
	Ours	**33.2**	**45.2**	**34.9**	**23.5**	**32.2**	**25.2**	**51.8**	**59.6**	**53.3**	**50.3**	**59.2**	**51.7**

Table 3: Few-shot performance comparison. Our method consistently outperforms SoTA and implemented baselines across all settings and datasets, demonstrating its effectiveness across diverse settings and datasets.

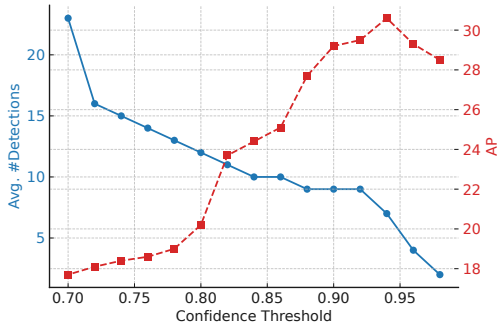

Figure 4: Threshold effect ($c^p > \tau$) during the object-aware temporal consistency.

2-layer MLPs with 512-dimensional hidden states. During training, we use AdamW optimization with a 1e-5 learning rate, 0.01 weight decay, and cosine scheduling with a linear warmup. The loss weighting parameters are set to $\lambda_{\text{cls}} = 2$, $\lambda_{\text{box}} = 5$. We set $\tau = 0.94$ and $\kappa = 0.98$ in our experiments.

Results and Discussion

Main Results. Our proposed approach achieves state-of-the-art performance across all four benchmark datasets. On FSVOD-500, we achieve AP of 30.6%, AP50 of 42.9%, and AP75 of 32.1%, surpassing TACF by margins of 3.7%, 3.7%, and 4.2%, highlighting robustness in localization and classification. Similarly, on FSYTV-40, our method yields 21.2% AP, 29.8% AP50, and 23.5% AP75, outperforming TACF by 5.3%, 5.6%, and 5.6%. Performance gains are more pronounced on VidOR and VidVRD, 49.4% AP, 57.3% AP50, and 50.2% AP75 for VidOR, and 48.7% AP, 57.8% AP50, and 49.2% AP75 for VidVRD, setting new benchmarks. These results underscore the effectiveness of our object-aware temporal modelling in capturing complex spatiotemporal dependencies with minimal supervision.

Temporal vs. Non-temporal Methods. Temporal modeling yields consistent gains across datasets. Non-temporal methods like UP-DETR (20.1% AP on FSVOD-500) underperform compared to our approach (30.6%), reflecting a 10.5% gap. On FSYTV-40, our method improves by 9.4%. These results underscore the value of capturing temporal dependencies, especially under few-shot constraints.

Comparison with adapted Temporal Methods. Despite using strong trackers (ByteTrack, DeepSort, TrackFormer) paired with CLIP's semantic features, these approaches fall short of specialized video detectors. ByteTrack+CLIP reaches only 14.7% AP on FSVOD-500, underperforming our method by 15.9%. Consistent gaps are observed across FSYTV-40 (13.3%), VidOR (24.5%), and VidVRD (23.1%), highlighting the need for dedicated few-shot modules.

Improvements over Specialized Few-Shot Methods. Our approach delivers consistent gains over prior specialized few-shot video detectors, including FSVOD, TACF, and adapted Video OWL-ViT. Compared to FSVOD, we improve by 5.5% AP on FSVOD-500, 6.6% on FSYTV-40, 4.3% on VidOR, and 8.0% on VidVRD. Similar improvements are observed over TACF and Video OWL-ViT. These results highlight our advantages in temporal aggregation via the object-aware frame encoder and superior support-query matching through few-shot tailored heads.

Advantage of Temporal Fusion (Object-Aware Decoder). In Table 4 on FSVOD-500, AP rises from 25.4% to 30.6% (+5.2%), AP_{50} from 37.1% to 42.9% (+5.8%), and AP_{75}

Figure 5: Qualitative comparison FSVOD-500 test set videos. Our method accurately detects multiple visually distinct instances of *Bactrian* camels (top) and precisely localizes the *Autogyro* (bottom), whereas FSVOD misses several camel instances.

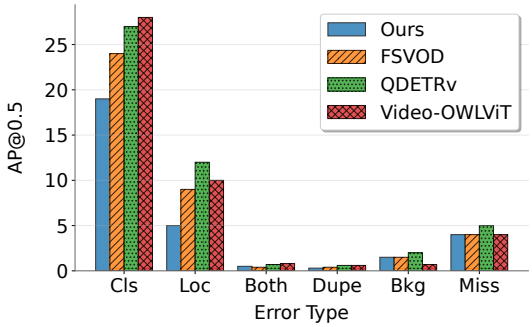

Figure 6: Error type analysis on FSVOD val set. A lower AP score is better.

from 26.9% to 32.1% (5.2%) when temporal fusion used. On FSYTV-40, AP improves from 19.2% to 21.2% (+2.0%), AP_{50} from 26.8% to 29.8% (+3.0%), and AP_{75} from 21.4% to 23.5% (+2.1%). These gains highlight the decoder's enhanced temporal reasoning detection performance.

Effect of Threshold (τ) on Temporal Propagation. Figure 4 illustrates how varying the confidence threshold τ impacts detection. As τ decreases from 0.98 to 0.70, the average number of detections rises from 2 to 23, reflecting increased object coverage. AP peaks at 30.6 when $\tau = 0.94$, then drops to 17.7 at $\tau = 0.70$, indicating precision loss due to false positives. Thus, $\tau = 0.94$ offers the best trade-off between coverage and precision.

Few-shot Performance Comparison. Table 3 shows consistent improvement of our method across baselines in 1-shot, 3-shot, and 10-shot settings. In the 1-shot case, we achieve 27.4% AP on FSVOD-500 and 18.3% on FSYTV-40, outperforming FSVOD by 6.7% and 6.9%, respectively. In the 10-shot setup, our method attains 33.2% AP on FSVOD-500 and 23.5% on FSYTV-40, with gains of 6.0% and 7.4% over FSVOD. Similar improvements across other benchmarks further validate our approach.

Qualitative Results. Figure 5 presents qualitative comparisons on two FSVOD-500 videos. In the first, our method accurately detects multiple bactrian camels missed by FSVOD. In the second, involving an autogiro, our bounding boxes are more precise.

	FSVOD-500			FSYTV-40		
Temopral Fusion	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
✗	25.4	37.1	26.9	19.2	26.8	21.4
✓	**30.6**	**42.9**	**32.1**	**21.2**	**29.8**	**23.5**

Table 4: Advantage of Object-Aware Temporal Fusion.

Error Type Analysis. Following TIDE (Bolya et al. 2020), we analyze error types in Fig. 6. Classification errors dominate all approaches (19-28 AP@0.5), with our method achieving lowest classification (19) and localization (5) errors. Video-OWLViT excels in background handling (0.7) but suffers from high classification errors (28), while QDETRv performs the poorest overall. Improving feature discrimination and localization precision are key challenges.

Efficiency Analysis. Our method achieves a strong balance between accuracy and efficiency, processing at 14 FPS with 8.1 GB of memory while attaining 30.6 AP. This is 40% faster than VideoOWL-ViT (10 FPS, 8.4 GB) and 27% faster than QDETRv (11 FPS, 7.6 GB), while providing substantially higher accuracy (+4.8 AP and +4.5 AP, respectively). Although FSVOD operates at 16 FPS with lower memory usage (5.8 GB), our approach delivers a +5.5% AP improvement with only a minor trade-off. Compared to other baselines, DeepSort runs at 9 FPS (7.9 GB), CenterTrack at 12 FPS (6.2 GB), and UP-DETR at 13 FPS (7.4 GB).

Conclusion

In this work, we introduced a novel approach for Few-Shot Video Object Detection that incorporates a language-aligned vision encoder and a filtering mechanism to selectively propagate high-confidence object features across frames. This mechanism efficiently enhances feature progression, reduces noise accumulation, and improves detection accuracy in a few-shot video setting. Through extensive experiments on four datasets, our approach consistently outperforms state-of-the-art methods, achieving superior detection performance and demonstrating its effectiveness in real-world video object detection tasks.

References

Bolya, D.; Foley, S.; Hays, J.; and Hoffman, J. 2020. TIDE: A General Toolbox for Identifying Object Detection Errors. In *ECCV*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*.

Chen, Y.; Cao, Y.; Hu, H.; and Wang, L. 2020. Memory Enhanced Global-Local Aggregation for Video Object Detection. In *CVPR*.

Dai, Z.; Cai, B.; Lin, Y.; and Chen, J. 2021. UP-DETR: Unsupervised pre-training for object detection with transformers. In *CVPR*.

Deng, J.; Pan, Y.; Yao, T.; Zhou, W.; Li, H.; and Mei, T. 2019. Relation Distillation Networks for Video Object Detection. In *ICCV*.

Dong, N.; Zhang, Y.; Ding, M.; and Lee, G. H. 2022. Incremental-DETR: Incremental Few-Shot Object Detection via Self-Supervised Learning. In *AAAI*.

Fan, Q.; Tang, C.-K.; and Tai, Y.-W. 2022a. Few-shot video object detection. In *ECCV*. Springer.

Fan, Q.; Tang, C.-K.; and Tai, Y.-W. 2022b. Few-Shot Video Object Detection. In *ECCV*.

Fan, Q.; Zhuo, W.; Tang, C.-K.; and Tai, Y.-W. 2020. Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector. In *CVPR*.

Han, G.; and Lim, S.-N. 2024. Few-Shot Object Detection with Foundation Models. In *CVPR*.

Han, W.; Lei, J.; Wang, F.; Feng, Z.; and Liang, R. 2023a. Temporal Aggregation with Context Focusing for Few-Shot Video Object Detection. In *SMC*.

Han, W.; Lei, J.; Wang, F.; Feng, Z.; and Liang, R. 2023b. Temporal Aggregation with Context Focusing for Few-Shot Video Object Detection. In *SMC*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

Heigold, G.; Minderer, M.; Gritsenko, A.; Bewley, A.; Keysers, D.; Lučić, M.; Yu, F.; and Kipf, T. 2023a. Video OWL-ViT: Temporally-consistent open-world localization in video. In *ICCV*.

Heigold, G.; Minderer, M.; Gritsenko, A.; Bewley, A.; Keysers, D.; Lučić, M.; Yu, F.; and Kipf, T. 2023b. Video OWL-ViT: Temporally-consistent open-world localization in video. In *ICCV*.

Hsieh, T.-I.; Lo, Y.-C.; Chen, H.-T.; and Liu, T.-L. 2019. One-shot object detection with co-attention and co-excitation. In *NeurIPS*.

Jung, M.; Jang, Y.; Choi, S.; Kim, J.; Kim, J.-H.; and Zhang, B.-T. 2025. Background-Aware Moment Detection for Video Moment Retrieval. In *WACV*.

Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; and Darrell, T. 2019. Few-shot object detection via feature reweighting. In *CVPR*.

Kaul, P.; Xie, W.; and Zisserman, A. 2023. Multi-Modal Classifiers for Open-Vocabulary Object Detection. In *ICML*.

Kumar, Y.; Agarwal, U.; Gupta, M.; and Mishra, A. 2025a. Aligning Moments in Time using Video Queries. In *ICCV*.

Kumar, Y.; Agarwal, U.; Gupta, M.; and Mishra, A. 2025b. Moment Alignment Transformer for Video-to-Video Moment Retrieval. *Authorea Preprints*.

Kumar, Y.; Mallick, S.; Mishra, A.; Rasipuram, S.; Maitra, A.; and Ramnani, R. 2024. QDETRv: Query-Guided DETR for One-Shot Object Localization in Videos. In *AAAI*.

Kumar, Y.; and Mishra, A. 2023. Few-Shot Referring Relationships in Videos. In *CVPR*.

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 *CVPR*.

Meinhardt, T.; Kirillov, A.; Leal-Taixe, L.; and Feichtenhofer, C. 2022. Trackformer: Multi-object tracking with transformers. In *CVPR*.

Minderer, M.; Gritsenko, A.; Stone, A.; Neumann, M.; Weissenborn, D.; Dosovitskiy, A.; Mahendran, A.; Arnab, A.; Dehghani, M.; Shen, Z.; Wang, X.; Zhai, X.; Kipf, T.; and Houlsby, N. 2022a. Simple Open-Vocabulary Object Detection. In *ECCV*.

Minderer, M.; Gritsenko, A.; Stone, A.; Neumann, M.; Weissenborn, D.; Dosovitskiy, A.; Mahendran, A.; Arnab, A.; Dehghani, M.; Shen, Z.; Wang, X.; Zhai, X.; Kipf, T.; and Houlsby, N. 2022b. Simple open-vocabulary object detection with vision transformers. In *ECCV*.

Osokin, A.; Sumin, D.; and Lomakin, V. 2020. Os2d: One-stage one-shot object detection by matching anchor features. In *ECCV*.

Peng, J.; Wang, C.; Wan, F.; Wu, Y.; Wang, Y.; Tai, Y.-W.; Wang, C.; Li, J.; Huang, F.; and Fu, Y. 2020. Chained-Tracker: Chaining Paired Attentive Regression Results for End-to-End Joint Multiple-Object Detection and Tracking. In *ECCV*.

Radenović, F.; Tolias, G.; and Chum, O. 2016. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *ECCV*.

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 *ICML*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NeurIPS*.

She, Y.; Bhat, G.; Danelljan, M.; and Yu, F. 2022. Fast Hierarchical Learning for Few-Shot Object Detection. In *IROS*.

Sun, B.; Li, B.; Cai, S.; Yuan, Y.; and Zhang, C. 2021. FSCE: Few-shot object detection via contrastive proposal encoding. In *CVPR*.

Wang, G.; Zhou, Y.; Luo, C.; Xie, W.; Zeng, W.; and Xiong, Z. 2021. Unsupervised visual representation learning by tracking patches in video. In *CVPR*.

Wang, X.; Huang, T.; Gonzalez, J.; Darrell, T.; and Yu, F. 2020. Frustratingly Simple Few-Shot Object Detection. In *ICML*.

Wojke, N.; Bewley, A.; and Paulus, D. 2017. Simple online and realtime tracking with a deep association metric. In *ICIP*.

Wu, H.; Chen, Y.; Wang, N.; and Zhang, Z. 2019. Sequence level semantics aggregation for video object detection. In *ICCV*.

Wu, J.; Liu, S.; Huang, D.; and Wang, Y. 2020. Multi-scale positive sample refinement for few-shot object detection. In *ECCV*.

Yu, Z.; Wang, G.; Chen, L.; Raschka, S.; and Luo, J. 2021. When Few-Shot Learning Meets Video Object Detection. *arXiv preprint arXiv:2103.14724*.

Zhang, J.; Zhao, H.; Wang, B.; Shi, Y.; Zhu, X.; Guo, Q.; Cai, T.; Liu, C.; Li, W.; Xu, S.; and Wang, J. 2023. Simple Open-Vocabulary Semantic Segmentation. In *CVPR*.

Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; and Wang, X. 2022. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*.

Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; and Liu, W. 2021. FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking. *IJCV*.

Zhou, X.; Koltun, V.; and Krähenbühl, P. 2020. Tracking Objects as Points. In *ECCV*.

Zhu, X.; Wang, Y.; Dai, J.; Yuan, L.; and Wei, Y. 2017. Flow-guided feature aggregation for video object detection. In *ICCV*.