OpenVIS: Open-vocabulary Video Instance Segmentation

Pinxue Guo[1,2*], **Hao Huang**[2], **Peiyang He**[2], **Xuefeng Liu**[2], **Tianjun Xiao**[2], **Wenqiang Zhang**[1,3†]

[1]Academy for Engineering and Technology, Fudan University
[2]Amazon Web Services
[3] School of Computer Science, Fudan University
pxguo21@m.fudan.edu.cn, tonyhh@aws.com, wqzhang@fudan.edu.cn

Abstract

Open-vocabulary Video Instance Segmentation (OpenVIS) can simultaneously detect, segment, and track arbitrary object categories in a video, without being constrained to categories seen during training. In this work, we propose InstFormer, a carefully designed framework for the OpenVIS task that achieves powerful open-vocabulary capabilities through lightweight fine-tuning with limited-category data. InstFormer begins with the open-world mask proposal network, encouraged to propose all potential instance class-agnostic masks by the contrastive instance margin loss. Next, we introduce InstCLIP, adapted from pre-trained CLIP with Instance Guidance Attention, which encodes open-vocabulary instance tokens efficiently. These instance tokens not only enable open-vocabulary classification but also offer strong universal tracking capabilities. Furthermore, to prevent the tracking module from being constrained by the training data with limited categories, we propose the universal rollout association, which transforms the tracking problem into predicting the next frame's instance tracking token. The experimental results demonstrate the proposed InstFormer achieve state-of-the-art capabilities on a comprehensive OpenVIS evaluation benchmark, while also achieves competitive performance in fully supervised VIS task.

Code — https://github.com/PinxueGuo/OpenVIS

Introduction

Video understanding (Bertasius, Wang, and Torresani 2021; Li et al. 2023) is a challenging yet significant computer vision task that requires specialized algorithms and techniques, surpassing the difficulty of image understanding. To achieve a more thorough understanding, Video Instance Segmentation (VIS) (Yang, Fan, and Xu 2019) has been proposed, which can simultaneously detect, segment (Hong et al. 2022; Guo et al. 2022; Hong et al. 2023; Guo et al. 2024b; Li et al. 2025; Guo et al. 2024a), and track (Zhou et al. 2024; Hong et al. 2024) instances in a given video, becoming a new research hotspot. Despite significant progress, current VIS models possess an inherent limitation. They can only segment objects within the boundaries of their training data, meaning they are unable to identify objects beyond the categories present in the training set. Consequently, their video understanding remains restricted. Moreover, identifying new categories requires retraining with additional annotated data, leading to substantial time and resource investment. To address the limitation, we investigate a novel computer vision task called Open-vocabulary Video Instance Segmentation (OpenVIS). This task focuses on detecting, segmenting, and tracking instances in videos based on the category names of target objects, regardless of whether those categories have been seen during the training stage.

Figure 1: OpenVIS simultaneously segments, detects, and tracks arbitrary objects in a video according to their corresponding text description. The proposed InstFormer can accurately identify various objects based on their respective category names in a video, irrespective of whether the category is included in the training set.

Although recent pre-trained Vision-Language Models (VLMs) (Radford et al. 2021; Yao et al. 2021) have shown promising results in zero-shot classification and provide good foundation for open-vocabulary video instance segmentation, significant challenges still remain in leveraging these static, image-level VLMs for this video, instance-level task. To eliminate this gap, we propose InstFormer, a carefully designed framework tailored for the OpenVIS task that achieves robust open-vocabulary capabilities through lightweight fine-tuning on a limited-category labeled dataset. **Firstly**, since any object, rather than fixed categories, might be selected for identification by the end

*Work done during internship at AWS.
†Corresponding author.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

user, InstFormer first performs the open-world mask proposal by incorporating a margin instance contrastive loss into a query-based mask proposal network to generates class-agnostic instance masks, with the goal of proposing as many distinct instances within a given video as possible to meet the flexible needs of open-world perception. **Secondly**, obtaining open-vocabulary representations of instances to enable classification and tracking across frames is non-trivial. Leveraging the zero-shot capabilities of pre-trained VLMs like CLIP (Radford et al. 2021) by directly inputting masked instance images is suboptimal and inefficient for video tasks due to the domain gap between pre-training on natural images and testing on masked images, as well as the need to run the VLM's vision encoder multiple times per frame. So we propose the InstCLIP, a variant of CLIP adapted with the proposed Instance Guidance Attention, which directs instance tokens to attend different instance regions simultaneously by the generated guidance according to multiple mask proposals. These instance tokens not only enable open-vocabulary classification but also offer strong universal tracking capabilities. **Thirdly**, the training process of current instance trackers with fixed-category datasets in video instance segmentation presents a significant challenge when it comes to tracking open-vocabulary instances. To address this issue, we propose Universal Rollout Association, where the rollout tracker is trained to predict instance tokens of the next frame to achieve tracking. The rollout tracker is implemented with a simple yet history-aware RNN layer, predicting instance tokens for the next frame based on previous tracking tokens. This prediction training is independent of categories, enabling the rollout tracker to handle open-vocabulary task.

To facilitate research on this novel task, we propose an evaluation benchmark that utilizes readily available datasets to thoroughly assess the performance. In our benchmark, the OpenVIS model will be trained with a limited number of categories, and subsequently tested on a large number of categories. Specifically, we evaluate the proposed model on YouTube-VIS, BURST, LVVIS, and UVO datasets, encompassing a large number of novel categories, to comprehensively assess its diverse capacities. However, the training process only see the data of YouTube-VIS, which comprises only 40 categories. The experimental results demonstrate the proposed InstFormer achieves state-of-the-art capabilities in OpenVIS and competitive performance in fully supervised VIS. This indicates that InstFormer retains most of VLM's zero-shot capabilities while optimizing for specific domains, providing a sound solution for scenarios needing both extreme domain performance and generalization. Our contributions can be summarized as follows:

- We propose the InstFormer framework, which achieves open-vocabulary capabilities through lightweight fine-tuning on limited-category data, to explore the novel OpenVIS task and introduce a comprehensive evaluation benchmark.
- We introduce the contrastive instance margin loss to open-world mask proposal network to encourage the generation of distinct instance proposals.
- We present InstCLIP, designed to embed each open-vocabulary instance with an instance token. The resulting instance tokens not only enable efficient open-vocabulary classification for multiple instances but also prove effective in subsequent open-vocabulary instance tracking.
- We propose universal rollout association, which achieves tracking by training the tracker to predict instance tokens of the next frame, to overcome the limitations of trackers trained on fixed-category data that struggle to generalize to open-vocabulary instances.

Related Work

Video Instance Segmentation

There are two main paradigms for Video Instance Segmentation (Yang, Fan, and Xu 2019): offline and online approaches. Offline methods, such as VisTR (Wang et al. 2021b), Mask2Former-VIS (Cheng et al. 2021), and SeqFormer (Wu et al. 2021), process the entire video at once, using instance queries to predict instance sequences in a single step. While effective on popular datasets, their reliance on full video input limits application in long or ongoing videos. Online methods, including MaskTrack RCNN (Yang, Fan, and Xu 2019), MaskProp (Bertasius and Torresani 2020), MinVIS (Huang, Yu, and Anandkumar 2022), IDOL (Wu et al. 2022a), and DVIS (Zhang et al. 2023), process frames independently, generating mask proposals and categories, and track instances via post-processing. Our approach adopts this strategy, enabling flexibility in predicting mask proposals and leveraging InstCLIP for instance classification.

Vision-Language Models

Vision-language models (VLMs) bridge visual and textual modalities, gaining attention for their strong visual representation learning. Pre-trained VLMs like CLIP (Radford et al. 2021) and FLIP (Yao et al. 2021), leveraging large-scale datasets, exhibit impressive zero-shot object recognition. For instance, CLIP achieves 76.2% zero-shot accuracy on ImageNet after training on 400M image-text pairs, inspiring its use in tasks like classification (Radford et al. 2021; Huang, Chu, and Wei 2022), captioning (Hu et al. 2022), retrieval (Liu et al. 2021), and segmentation (Xu et al. 2021). However, applying VLMs to some tasks is challenging; for example, their performance drops with masked inputs in video instance segmentation (Liang et al. 2022), and running VLMs' vision encoder N times for N instances in a frame is computationally expensive.

Open-Vocabulary Segmentation

Open vocabulary segmentation, introduced by ZS3Net (Bucher et al. 2019), segments objects based on text descriptions, including unseen categories during training. Mainstream methods like ZSSeg (Xu et al. 2021), ZegFormer (Ding et al. 2022), and OVSeg (Liang et al. 2022) adopt a two-stage framework: extracting class-agnostic proposals and matching visual features with text descriptions to identify categories. Recent works, SAN (Xu et al. 2023) and DeOP (Han et al. 2023), address the

high computational cost of repeated VLM encoder passes, similar to InstCLIP in our framework. However, unlike these, InstCLIP requires only lightweight finetuning without modifying VLM pre-trained weights and provides instance tokens for video-level instance association.

Setting

Problem Formulation

Open-vocabulary Video Instance Segmentation (OpenVIS) aims to simultaneously segment, detect, and track open-world objects of arbitrary category based on the category name or corresponding text description in a video, regardless of whether the category has been seen during training. We are given a video consisting of T frames, denoted as $\{F_t \in \mathbb{R}^{3\times H\times W}\}_{t=1}^{T}$, where H and W represent the height and width of each frame, respectively. Additionally, we have a set of category labels denoted as $\mathcal{C}$, which represents the possible categories of objects present in the video. Our objective in OpenVIS is to accurately predict all N objects belonging to these categories within the video. Specifically, for each object i, its category label $c^i \in \mathcal{C}$ and segmentation masks across the video $\mathbf{m}^i_{p\dots q} \in \mathbb{R}^{H\times W\times(p-q)}$ need to be predicted, where $p \in [1, T]$ and $q \in [p, T]$ indicate its starting and ending frame index.

Evaluation Benchmark

To comprehensively evaluate the overall performance of the proposed OpenVIS, we introduce a novel evaluation benchmark. An ideal OpenVIS model should possess two essential properties, which form the two core focus of our evaluation: 1) open-world proposal ability to segment all possible instances within the video accurately and 2) zero-shot capability to correctly classify instances of arbitrary category. Moreover, we also further evaluate 3) the overall OpenVIS performance on both seen and unseen categories.

- **Open-world Property:** We leverage the exhaustively annotated UVO dataset (Wang et al. 2021a) to evaluate the open-world mask proposal ability. The UVO dataset provides an average of 13.52 instances annotated per video. Compared to the only 1.68 objects in YouTube-VIS, UVO is naturally a suitable dataset for measuring open-world property.
- **Zero-shot Property:** We utilize the category-rich BURST dataset (Athar et al. 2023) to evaluate the zero-shot instance classification property. The objects in BURST involve 482 categories, with 78 common categories from COCO (Lin et al. 2014) and 404 uncommon categories, which can be regard as unseen categories. The uncommon-404 categories is an ideal dataset to measure the zero-shot property. Additionally, a latest dataset LVVIS (Wang et al. 2023) novel set contains 555 unseen categories for our setting, so we also evaluate it.
- **Overall Property:** To further evaluate the overall performance on both seen and unseen categories, we also report the results on full BURST (482 categories).

Following (Yang, Fan, and Xu 2019), we utilize the Average Precision (i.e., AP) and Average Recall (i.e., AR) at the video level as the main metrics. Additionally, our OpenVIS model is only trained on YouTube-VIS (a widely-used VIS dataset comprising 40 categories). This ensures that the categories present in the training data are small-scale subsets of those found in the test data. More discussion and analysis of the evaluation benchmark can be found in *Supplementary*.

Method

In this section, we detail how we bridge the gap between static, image-level VLMs and the video, instance-level demands of the OpenVIS task, leading to the InstFormer, a carefully designed framework tailored for OpenVIS that achieves open-vocabulary capabilities through lightweight fine-tuning on limited-category labeled data.

Open-world Mask Proposal

An open-world proposed mask proposal network needs to propose as many distinct instances as possible to meet the flexible needs of open-world perception, as each instance has the possibility of being selected to identify by end user. To achieve this goal, we first adopted a query-based image segmentation model Mask2Former (Cheng et al. 2022) as the mask proposal network, predicting N class-agnostic masks $M_t = \{m_t^i\}_{i=1}^{N} \in [0,1]^{N\times H\times W}$ and their corresponding instance queries $Q_t = \{q_t^i\}_{i=1}^{N} \in \mathbb{R}^{N\times C}$ for each frame $F_t \in \mathbb{R}^{3\times H\times W}$ of a video:

$$M_t, Q_t = \Psi(\Phi(F_t), Q^0), \tag{1}$$

where Φ and Ψ indicate the backbone and transformer decoder of the mask proposal network respectively. The $Q^0 \in \mathbb{R}^{N\times C}$ denotes the N learnable initial query embeddings. Despite the mask proposal network mentioned above can generate category-agnostic masks for all candidate instance, its training process on a dataset with a limited number of objects results in the redundant assignment of instance queries to the same instance. To ensure that these initial queries can perceive as many distinct instances as possible on given video, we introduce a contrastive instance margin loss to the open-world mask proposal network:

$$\mathcal{L}_{SC} = \sum_{i=0}^{N}\sum_{j=0}^{N} \max(0, \cos(Q_t^i, Q_t^j) - \alpha), \tag{2}$$

where $\cos(\cdot,\cdot)$ refers to the cosine similarity ranging $[-1, 1]$, and α is the margin that determines how similar tokens should be penalized. This loss function will penalize instances that are excessively similar, thereby promoting diverse assignments of queries to distinct instances.

Open-vocabulary Instance Representation

Leveraging pre-trained VLMs like CLIP (Radford et al. 2021) for zero-shot capabilities by directly inputting masked instance images is suboptimal and inefficient for real-time video tasks due to the domain gap between pre-training on natural images and testing on masked images, as well as the need to run CLIP vision encoder N times per frame. The proposed InstCLIP efficiently represents each instance with an instance token, enabling open-vocabulary instance

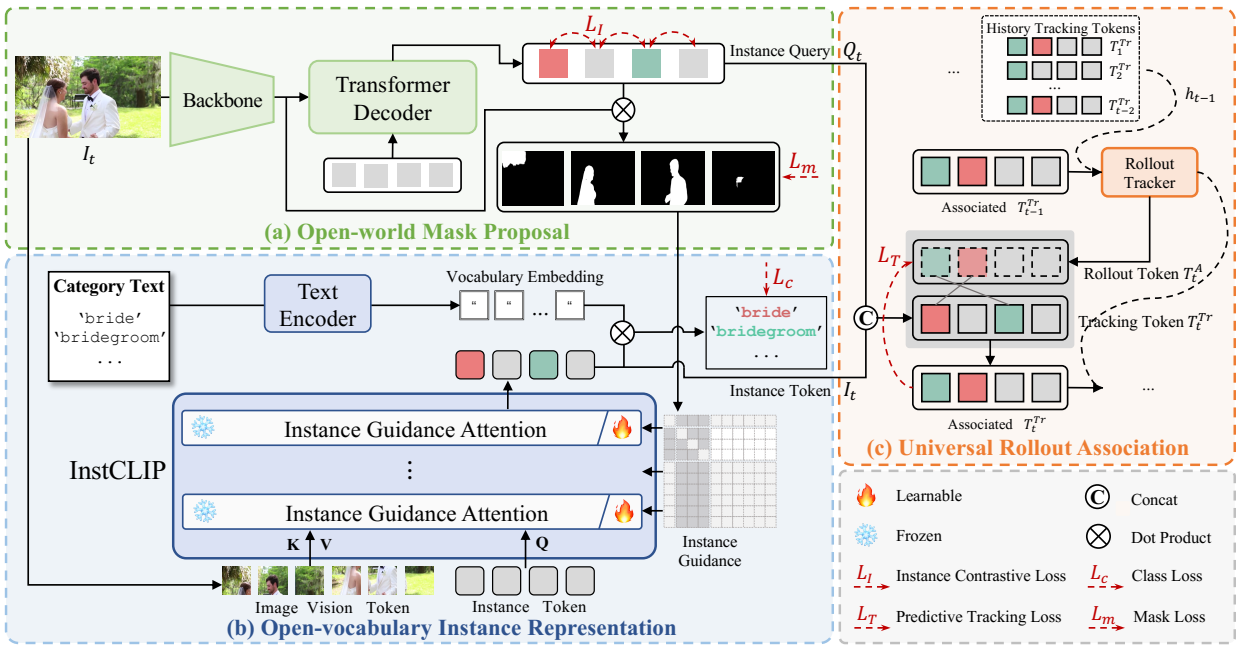


Figure 2: Overview of the proposed InstFormer framework for OpenVIS. (a) **Open-world Mask Proposal**: Generate class-agnostic instance masks with a query-based transformer, which is encouraged to propose all potential object instances. (b) **Open-vocabulary Instance Representation**: InstCLIP embeds open-vocabulary instance tokens using Instance Guidance Attention efficiently. These tokens enable open-vocabulary instance classification and provide robust open-vocabulary tracking capabilities. (c) **Universal Rollout Association**: Associate instances of any category across frames with the proposed universal rollout tracker, which is trained to predict the instance tracking tokens of the next frame, termed the rollout token.

classification and providing strong open-vocabulary tracking capabilities. Specifically, InstCLIP is a Vision Transformer (Dosovitskiy et al. 2020) adapted from the pre-trained CLIP vision encoder, consisting of L Instance Guidance Attention layers. We generate attention masks from mask proposals for Instance Guidance Attention to guide N instance tokens to embed N instances in a single forward pass through the encoder. Instance Guidance Attention takes as input the concatenated tokens $X_t^{l-1} \in \mathbb{R}^{1+N+P}$ from the previous attention layer and the guidance attention mask $\mathcal{M} \in \mathbb{R}^{(1+N+P)\times(1+N+P)}$:

$$\begin{aligned} X_t^l &= \text{InstAttn}(X_t^{l-1}, \mathcal{M}) \\ &= \text{softmax}(W^q X_t^{l-1} \cdot W^k X_t^{l-1} + \mathcal{M}) \cdot W^v X_t^{l-1}, \end{aligned} \tag{3}$$

where W^q, W^k, W^v are weights of query, key, and value projection layer, respectively. X_t^{l-1} consists of vision tokens $V_t \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times C}$ from image patch embedding, N initial instance tokens $I^l \in \mathbb{R}^{N\times C}$, and a register token $R^l \in \mathbb{R}^{1\times C}$. P is the number of vision tokens. The register token, inspired by (Darcet et al. 2023), is a token permitted to attend to all vision tokens. It plays the role of collecting low-informative feature, which helps obtain cleaner attention maps from instance tokens to vision tokens.

The initial instance tokens I^0 and register token R^0 are learnable embeddings. The instance guidance $\mathcal{M}$, generated as illustrated in Fig. 3, directs instance tokens to attend different instance regions by acting as the attention mask in self-attention layers of the vision transformer. Instance tokens are independently guided to enhance attention to specific regions while suppressing attention to other regions based on the logits value of the instance masks. After L instance guidance attention layers, these N instance tokens aggregate CLIP features of N instance. So classification can be directly calculated by comparing them with vocabulary embeddings extracted by the CLIP text encoder:

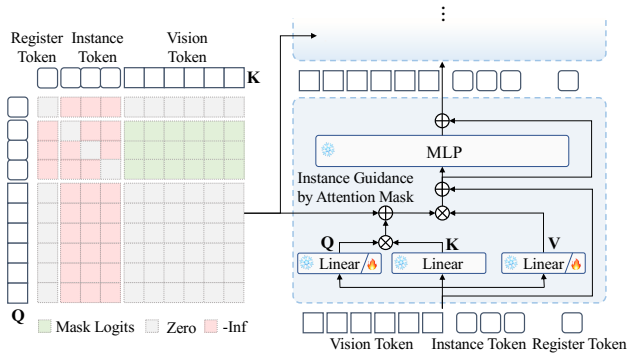
$$C_t = \text{argmax}(\text{softmax}(I_t^L \cdot E^\top)) \in \mathbb{K}^N, \tag{4}$$

where $E \in \mathbb{R}^{K\times C}$ is the vocabulary embeddings of K categories. InstCLIP is designed with the principle of minimizing modifications from CLIP, to fully unleash the zero-shot capability of the pre-trained CLIP. Only the linear projections for the query and value of the attention layer are adjusted using the parameter-efficient fine-tuning approach LoRA (Hu et al. 2021) during training, while almost parameters of CLIP remain frozen.

Universal Rollout Association

To prevent trackers optimized on closed-set data from failing to generalize to open-vocabulary instance tracking, the proposed Universal Rollout Association fully leverages the open-vocabulary characteristics of instance tokens and transforms the tracking problem into predicting the next frame's instance tracking token for training.

Instance Tracking Tokens. We form universal instance tracking tokens T_t^{Tr} by combining the instance tokens I_t



(a) Instance Guidance Generation (b) Details of Instance Guidance Attention Layer

Figure 3: The architecture of InstCLIP and the generation of the corresponding instance guidance mask.

from InstCLIP with the instance queries Q_t from the proposal network to handle open-vocabulary tracking: $T_t^{Tr} = \text{Concat}(I_t, Q_t) \in \mathbb{R}^{N\times 2C}$. These tokens and queries are naturally aligned in a token/query-based architecture. The former, which leverages CLIP features with zero-shot capabilities, proves particularly effective for tracking open-vocabulary instances in subsequent experiments. The latter, characterized by its ability to generate class-agnostic mask proposals, has also been demonstrated in MinVIS (Huang, Yu, and Anandkumar 2022) to distinguish instances between frames.

Rollout Association. And to prevent the tracker from being constrained by fixed-category object data, we reframe the tracking problem by training the tracker to predict the instance tracking token for the next frame. When associating instances in frame-t with those from previous frames, the rollout tracker predicts the instance tracking token for frame-t based on the tokens from history frames. This predicted instance tracking token, referred to as the rollout association token, is denoted as $T_t^A \in \mathbb{R}^{N\times C}$:

$$T_t^A = \mathcal{R}(T_{1\to t-1}^{Tr}) = \text{RNN}(T_{t-1}^{Tr}, h_{t-1}), \tag{5}$$

where $\mathcal{R}$ denotes the concise yet effective rollout tracker implemented by a single RNN layer and h_{t-1} is the hidden state of RNN remaining instance temporal information. Finally, by comparing the rollout association token for frame t with the actual instance tracking token using Hungarian matching on the similarity score $S_{ij} = \cos(T_t^A, T_t^{Tr})$, the instance association for frame-t can be completed. The rollout tracker is trained with the loss:

$$\mathcal{L}_T = \sum_{i=1}^{N}\sum_{j=1}^{N} \text{CE}(\cos(T_t^A(i), T_t^{Tr}(j)), \mathbf{1}_{i=j}), \tag{6}$$

where $\mathbf{1}_{i=j} \in \{0, 1\}$ is an indicator function evaluating to 1 if $i = j$. This process is independent of categories, allowing the rollout tracker to handle open-vocabulary instances. Meanwhile, the incorporation of historical information in this history-aware tracker enhances robustness compared to tracking based solely on the previous frame.

Experiment

Implementation Details

Model Architecture. We regard a COCO (Lin et al. 2014)-pretrained Mask2Former (Cheng et al. 2022) as our mask proposal network. By default, the transformer decoder has 100 queries, with a dimension of 256 for the query embedding and transformer decoder. For InstCLIP, we select a ViT-B/32 of CLIP (Radford et al. 2021) as foundation vision transformer. The number of instance tokens of InstCLIP is also set to 100, aligning with the 100 instance queries of the mask proposal network. We initialize the instance tokens and register token using CLIP's learned class tokens. The text encoder is a 12-layer transformer, the same as that in CLIP. For input prompts, we ensemble 14 prompts (e.g., "a photo of a {category name}") from (Liang et al. 2022) to boost zero-shot classification ability.

Training. InstFormer is trained using a two-stage approach and CLIP weights are frozen during the entire training. In first stage, the open-world mask proposal network and InstCLIP (LoRA adapter) are trained for 6k iterations with $\mathcal{L}_I$ and instance segmentation loss. Subsequently, we train the rollout tracker in second stage, with all other weights frozen, using $\mathcal{L}_T$ for an additional 600 iterations. The whole training is done on 8 V100 GPUs for 3 hours.

Baselines. To better assess the performance of the proposed InstFormer framework, we introduce several baselines for comparison, as shown in Tab. 1. For fully-supervised methods, we provide STCN Tracker (Athar et al. 2023), Box Tracker (Athar et al. 2023) and MinVIS (Huang, Yu, and Anandkumar 2022). Both the first two methods ultilize the Mask-RCNN (He et al. 2017) but with different tracking strategies as in (Athar et al. 2023). MinVIS (Huang, Yu, and Anandkumar 2022) is an advanced VIS model in fully-supervised VIS task. All of them are trained on full BURST dataset with all 482 categories. For open-vocabulary methods, we employ three approaches: Detic-SORT, Detic-OWTB, and OV2Seg (Wang et al. 2023). The first two methods utilize the open-vocabulary detector Detic (Zhou et al. 2022), paired with the classical multi-object tracker SORT (Bewley et al. 2016) and the state-of-the-art open-world tracker OWTB (Liu et al. 2022), respectively. OV2Seg introduces the CLIP text encoder and a momentum-updated query for tracking, also achieving open-vocabulary video instance segmentation. These baselines may be trained with different datasets. We provide their corresponding training datasets and category numbers in Tab. 1. L-1203 represents the entire LVIS (Gupta, Dollar, and Girshick 2019) dataset with all 1203 categories. L-866 indicates the LVIS subset with 866 frequent categories. B and Y denote BURST (Athar et al. 2023) with 482 categories and YouTube-VIS (Yang, Fan, and Xu 2019) with 40 categories, respectively. C is the COCO (Lin et al. 2014) dataset with 80 categories, which may be used to pretrain some modules. And for clarity, we provide the proportion of novel categories during inference for each approach and setting in the table in gray% (higher means more challenging).

Method	OV	Training Categories	BURST			LVVIS$_{novel}$
			All	*Common*	*Uncommon*	
Fully-supervised						
MRCNN (He et al. 2017)-BoxTracker	×	L-1203	1.4 0%	3.0 0%	0.9 0%	-
MRCNN (He et al. 2017)-STCNTracker	×	L-1203	0.9 0%	0.7 0%	0.6 0%	-
MinVIS (Huang, Yu, and Anandkumar 2022)	×	CB-482	1.4 0%	5.5 0%	0.5 0%	-
Open-vocabulary						
Detic (Zhou et al. 2022)-SORT (Bewley et al. 2016)	✓	L-866	1.9 15%	1.8 0%	2.5 18%	3.4 100%
Detic (Zhou et al. 2022)-OWTB (Liu et al. 2022)	✓	L-866	2.7 15%	2.8 0%	1.8 18%	4.2 100%
OV2Seg (Wang et al. 2023)	✓	L-866	3.7 15%	3.9 0%	2.4 18%	11.9 100%
InstFormer (Ours)	✓	CY-103	**4.2** 84%	**7.4** 0%	**3.5** 96%	**12.2** 100%

Table 1: Overall OpenVIS performance and zero-shot property comparison with baselines on BURST and LVVIS with AP metric. OV indicates whether the method has the ability to handle the open-vocabulary setting. The Training Categories column shows the training dataset and the number of categories involved. The gray% represents the proportion of novel categories during inference for each approach and setting (higher means more challenging).

	Training Data	AP	AP$_{c1}$	AR_{100}
MTRCNN (Yang, Fan, and Xu 2019)	YouTubeVIS	7.6	-	9.3
MTRCNN (Yang, Fan, and Xu 2019)	UVO	11.2	-	17.4
TAM (Yang et al. 2023)	SA-1B	-	1.7	24.1
SAM-PT (Rajič et al. 2023)	SA-1B	-	6.7	**28.8**
InstFormer (Ours)	YouTubeVIS	**16.7**	**7.2**	24.7

Table 2: Comparison of open-world instance proposal property on UVO. AP$_{c1}$ indicates the class-agnostic AP.

Main Results

Overall Performance. We evaluate the overall performance of the InstFormer on the BURST validation set. Since the mask proposal network is pre-trained on COCO and InstFormer is trained on YouTube-VIS, so there are 103 categories have been seen during training. As illustrated in Tab. 1, our proposed InstFormer framework outperforming fully-supervised baselines by a large margin (AP from 1.4 to 4.2). And despite InstFormer seen fewer categories compared to other open-vocabulary baselines, it still achieved state-of-the-art OpenVIS performance (BURST 4.2 AP, LVVIS$_{novel}$ 12.2 AP), demonstrating InstFormer achieves obvious advantages over other methods. Qualitative results can be found in Fig. 1 and *Supplementary*.

Zero-shot Instance Classification. To measure zero-shot instance classification property, we report the results of BURST-uncommon with 404 categories and LVVIS-novel with 555 categories in Tab. 1. Specifically, for BURST-uncommon, where 96% categories are novel to us, we achieve a 45% (AP from 2.4 to 3.5) improvement over the OV2Seg. For LVVIS-novel, we also achieve the best performance even InstFomer only seen 103 categories, which the compared methods like OV2Seg have seen 866 categories. This demonstrates that InstCLIP successfully maintains the zero-shot capability of the pre-trained CLIP model.

Open-world Instance Proposal. In this section, we evaluate the performance of the open-world mask proposal, which is a critical component for achieving OpenVIS, using the extensively annotated UVO dataset. As reported in Tab. 2, our mask proposal network, trained solely on the YouTube-VIS dataset with contrastive instance margin loss, outperforming the baseline method (Yang, Fan, and Xu 2019) traind on YouTube-VIS and even on UVO itself. Compared with the most advanced mask proposal approaches empowered by the Segment Anything Model(SAM) (Kirillov et al. 2023) trained with the extensive dataset SA-1B (Kirillov et al. 2023), our mask proposal network can also achieve comparable performance.

	OV	AP	AP$_{50}$	AP$_{75}$	AR$_{10}$
Fully-supervised					
MaskTrack (Yang, Fan, and Xu 2019)	×	30.3	51.1	32.6	35.5
SipMask (Cao et al. 2020)	×	33.7	54.1	35.8	40.1
CrossVIS (Yang et al. 2021)	×	36.3	56.4	38.9	40.7
VISOLO (Han et al. 2022)	×	38.6	56.3	43.7	42.5
MinVIS (Huang, Yu, and Anandkumar 2022)	×	47.4	69.0	52.1	55.7
IDOL (Wu et al. 2022b)	×	49.5	74.0	52.9	58.1
GenVIS (Heo et al. 2023)	×	50.0	71.5	54.6	59.7
DVIS (Zhang et al. 2023)	×	51.2	73.8	57.1	59.3
MinVIS-CLIP	✓	30.6	51.2	32.0	40.7
InstFormer (Ours)	✓	**51.8**	**75.6**	**57.2**	**60.0**
Open-vocabulary					
Detic-SORT	✓	14.6	-	-	-
Detic-OWTB	✓	17.9	-	-	-
Ov2seg	✓	27.2	-	-	-

Table 3: Performance comparison in the fully-supervised VIS on YouTube-VIS.

Fully-supervised VIS. An ideal OpenVIS model should excel in both open-set and closed-set scenarios. We evaluate the proposed InstFormer on YouTube-VIS, as shown in Tab. 3, where it achieves top-tier performance in fully-supervised VIS, outperforming both OpenVIS baselines and fully-supervised methods. Since InstFormer is trained on YouTube-VIS while other baselines are not, we introduce an open-vocabulary baseline, MinVIS-CLIP, for fair com-

	BURST	YouTube-VIS	UVO
1 MinVIS-CLIP	2.1	30.6	9.0
2 + InstCLIP	3.3	48.6	13.2
3 + $\mathcal{L}_{SC}$	3.5	48.5	15.8
4 + InstCLIP Token	3.9	50.2	16.1
5 + Rollout Tracker	**4.2**	**51.8**	**16.7**

Table 4: Ablation study of InstFormer on diverse datasets.

	AP_{All}	AP_{Com}	AP_{Uncom}	Once
1 N times CLIP	2.11	3.58	1.81	×
2 + Instance Token	1.09	0.87	1.14	✓
3 + Binary Guidance	1.70	2.13	1.62	✓
4 + Designed Guidace	3.28	6.88	2.52	✓
5 + Register Token	**3.87**	**7.01**	**3.22**	✓

Table 5: Ablation study of InstCLIP on BURST.

parison. MinVIS-CLIP modifies MinVIS by replacing its closed-set classification head with a frozen CLIP for open-vocabulary capability. Serving as the starting-point baseline for our framework, experiments show InstFormer surpasses MinVIS-CLIP by 21.2 AP, proving the proposed open-vocabulary methods benefit fully-supervised VIS tasks as well.

Ablation Study

In this section, we conduct ablation study on key designs of our framework to demonstrate their effectiveness. In the experiments of non-tracker components, to avoid performance changes caused by trackers, we default to using Hungarian matching with the instance query for association.

Effectiveness of InstCLIP. In Tab. 4, comparing Line 1 and Line 2, InstCLIP demonstrates significant improvements over the MinVIS-CLIP baseline, where the masked image is directly input into CLIP for zero-shot classification. Specifically, on BURST, we observe an increase in AP from 2.1 to 3.3 (57% improvement). Similarly, on UVO, the AP rises from 9.0 to 13.1 (46% increase). Notably, for YouTube-VIS, there is a remarkable gain of 18 AP (58% improvement). This shows that InstFormer retains most of CLIP's zero-shot capabilities while optimizing for YouTubeVIS domain, offering an effective solution for scenarios requiring both extreme domain performance and generalization.

Instance Tokens for Association. InstCLIP's instance tokens not only achieve open-vocabulary classification effectively and efficiently, as illustrated in Tab. 4 Line 4 and Tab. 6, but also aid in tracking instances of any vocabulary.

Key design of InstCLIP. In this part, we ablate the key design of InstCLIP including instance guidance mask, instance tokens, and register tokens. In Tab. 5, directly introducing N instance tokens into CLIP to enable CLIP to classify N instances in a single-forward doesn't work well, as instance representations cannot aggregate into tokens without specific guidance (Line 2). Masking the background region for each instance with the binary instance mask from mask proposal network allows instance token to know what should attend (Line 3). Line 4 reveals that the effectiveness of InstCLIP hinges on the generated Instance Attention Mask. The register token, specifically designed for collecting low-informative features, indeed assists InstCLIP in obtaining superior instance tokens and vision tokens (Line 5).

Instance Token	Rollout Tracker	$\mathcal{L}_{TC}$	BURST	YouTube-VIS
×	×	×	3.5	48.5
✓	×	×	3.9	50.2
×	RNN	✓	3.6	49.9
✓	RNN	✓	**4.2**	**51.8**
✓	Linear	✓	3.7	49.1
✓	MLP	✓	3.4	49.7

Table 6: Ablation study of Rollout Association in both OpenVIS (on BURST) and fully-supervised VIS (on YouTube-VIS).

Contrastive Instance Margin Loss. We study the effect of the contrastive margin loss to open-world mask proposal on UVO. As shown in Tab. 4 Line 3 and Tab. **??** of Supp, introducing the contrastive instance margin loss encourages the mask proposal network provide more distinct instances, thereby improving both AP and AR. More details see Supp.

Effectiveness of Rollout Association. Tab. 6 ablates the key components of the rollout association. Given the instance token provides a richer open-vocabulary tracking feature, the predictive tracking loss-driven rollout tracker achieves a 10.3% improvement in AP performance in the OpenVIS setting. Additionally, it provides a 3.2% boost in fully-supervised VIS tasks, demonstrating the effectiveness of rollout association in normal fully-supervised tracking instances. RNN offers historical information offered by the hidden state, aiding in handling object occlusion and reappearance issues. We also replace the RNN layer with a linear layer or a two-layer MLP with the larger capacity.

Conclusion

We propose InstFormer, a framework for the OpenVIS task that achieves strong open-vocabulary capabilities through lightweight fine-tuning on limited-category data, bridging the gap between image-level VLMs and video, instance-level requirements. InstFormer introduces a mask proposal network with margin instance contrastive loss to detect all potential object instances. It includes InstCLIP for efficient open-vocabulary instance embedding, enabling classification and robust tracking. Additionally, a universal rollout association trains a tracker to predict instance tokens for the next frame, ensuring universal tracking. We also present a comprehensive evaluation benchmark to advance research in this field. Experiments show InstFormer achieves state-of-the-art OpenVIS performance and competitive fully supervised VIS results.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No.62072112), Scientific and Technological innovation action plan of Shanghai Science and Technology Committee (No.22511102202, No.22511101502, and No.21DZ2203300).

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.

Bertasius, G.; and Torresani, L. 2020. Classifying, segmenting, and tracking object instances in video with mask propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9739–9748.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 813–824. PMLR.

Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; and Upcroft, B. 2016. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, 3464–3468. IEEE.

Bucher, M.; Vu, T.-H.; Cord, M.; and Pérez, P. 2019. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32.

Cao, J.; Anwer, R. M.; Cholakkal, H.; Khan, F. S.; Pang, Y.; and Shao, L. 2020. Sipmask: Spatial information preservation for fast image and video instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 1–18. Springer.

Cheng, B.; Choudhuri, A.; Misra, I.; Kirillov, A.; Girdhar, R.; and Schwing, A. G. 2021. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1290–1299.

Darcet, T.; Oquab, M.; Mairal, J.; and Bojanowski, P. 2023. Vision Transformers Need Registers. *arXiv preprint arXiv:2309.16588*.

Ding, J.; Xue, N.; Xia, G.-S.; and Dai, D. 2022. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11583–11592.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Guo, P.; Hong, L.; Zhou, X.; Gao, S.; Li, W.; Li, J.; Chen, Z.; Li, X.; Zhang, W.; and Zhang, W. 2024a. ClickVOS: Click Video Object Segmentation. *arXiv preprint arXiv:2403.06130*.

Guo, P.; Li, W.; Huang, H.; Hong, L.; Zhou, X.; Chen, Z.; Li, J.; Jiang, K.; Zhang, W.; and Zhang, W. 2024b. X-Prompt: Multi-modal Visual Prompt for Video Object Segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5151–5160.

Guo, P.; Zhang, W.; Li, X.; and Zhang, W. 2022. Adaptive online mutual learning bi-decoders for video object segmentation. *IEEE Transactions on Image Processing*, 31: 7063–7077.

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.

Han, C.; Zhong, Y.; Li, D.; Han, K.; and Ma, L. 2023. Open-Vocabulary Semantic Segmentation with Decoupled One-Pass Network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1086–1096.

Han, S. H.; Hwang, S.; Oh, S. W.; Park, Y.; Kim, H.; Kim, M.-J.; and Kim, S. J. 2022. Visolo: Grid-based space-time aggregation for efficient online video instance segmentation. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2896–2905.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Heo, M.; Hwang, S.; Hyun, J.; Kim, H.; Oh, S. W.; Lee, J.-Y.; and Kim, S. J. 2023. A generalized framework for video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14623–14632.

Hong, L.; Chen, W.; Liu, Z.; Zhang, W.; Guo, P.; Chen, Z.; and Zhang, W. 2022. LVOS: A Benchmark for Long-term Video Object Segmentation. *arXiv preprint arXiv:2211.10181*.

Hong, L.; Yan, S.; Zhang, R.; Li, W.; Zhou, X.; Guo, P.; Jiang, K.; Chen, Y.; Li, J.; Chen, Z.; et al. 2024. Onetracker: Unifying visual object tracking with foundation models and efficient tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19079–19091.

Hong, L.; Zhang, W.; Gao, S.; Lu, H.; and Zhang, W. 2023. Simulflow: Simultaneously extracting feature and identifying target for unsupervised video object segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7481–7490.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; and Wang, L. 2022. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17980–17989.

Huang, D.-A.; Yu, Z.; and Anandkumar, A. 2022. Minvis: A minimal video instance segmentation framework without video-based training. *Advances in Neural Information Processing Systems*, 35: 31265–31277.

Huang, T.; Chu, J.; and Wei, F. 2022. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Li, W.; Guo, P.; Zhou, X.; Hong, L.; He, Y.; Zheng, X.; Zhang, W.; and Zhang, W. 2025. Onevos: unifying video object segmentation with all-in-one transformer framework. In *European Conference on Computer Vision*, 20–40. Springer.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2022. Open-vocabulary semantic segmentation with mask-adapted clip. *arXiv preprint arXiv:2210.04150*.

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.

Liu, Y.; Zulfikar, I. E.; Luiten, J.; Dave, A.; Ramanan, D.; Leibe, B.; Ošep, A.; and Leal-Taixé, L. 2022. Opening up open world tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19045–19055.

Liu, Z.; Rodriguez-Opazo, C.; Teney, D.; and Gould, S. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2125–2134.

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.

Rajič, F.; Ke, L.; Tai, Y.-W.; Tang, C.-K.; Danelljan, M.; and Yu, F. 2023. Segment Anything Meets Point Tracking. *arXiv preprint arXiv:2307.01197*.

Wang, H.; Yan, C.; Wang, S.; Jiang, X.; Tang, X.; Hu, Y.; Xie, W.; and Gavves, E. 2023. Towards open-vocabulary video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4057–4066.

Wang, W.; Feiszli, M.; Wang, H.; and Tran, D. 2021a. Unidentified video objects: A benchmark for dense, open-world segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10776–10785.

Wang, Y.; Xu, Z.; Wang, X.; Shen, C.; Cheng, B.; Shen, H.; and Xia, H. 2021b. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8741–8750.

Wu, J.; Jiang, Y.; Zhang, W.; Bai, X.; and Bai, S. 2021. Seqformer: a frustratingly simple model for video instance segmentation. *arXiv preprint arXiv:2112.08275*.

Wu, J.; Liu, Q.; Jiang, Y.; Bai, S.; Yuille, A.; and Bai, X. 2022a. In defense of online models for video instance segmentation. In *European Conference on Computer Vision*, 588–605. Springer.

Wu, J.; Liu, Q.; Jiang, Y.; Bai, S.; Yuille, A.; and Bai, X. 2022b. In defense of online models for video instance segmentation. In *European Conference on Computer Vision*, 588–605. Springer.

Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; and Bai, X. 2023. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2945–2954.

Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; and Bai, X. 2021. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model. *arXiv preprint arXiv:2112.14757*.

Yang, J.; Gao, M.; Li, Z.; Gao, S.; Wang, F.; and Zheng, F. 2023. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*.

Yang, L.; Fan, Y.; and Xu, N. 2019. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5188–5197.

Yang, S.; Fang, Y.; Wang, X.; Li, Y.; Fang, C.; Shan, Y.; Feng, B.; and Liu, W. 2021. Crossover learning for fast online video instance segmentation. In *proceedings of the IEEE/CVF international conference on computer vision*, 8043–8052.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. FILIP: fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Zhang, T.; Tian, X.; Wu, Y.; Ji, S.; Wang, X.; Zhang, Y.; and Wan, P. 2023. DVIS: Decoupled Video Instance Segmentation Framework. *arXiv preprint arXiv:2306.03413*.

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.

Zhou, X.; Guo, P.; Hong, L.; Li, J.; Zhang, W.; Ge, W.; and Zhang, W. 2024. Reading relevant feature from global representation memory for visual object tracking. *Advances in Neural Information Processing Systems*, 36.