EntitySAM: Segment Everything in Video

Mingqiao Ye[1,*] Seoung Wug Oh[2] Lei Ke[3] Joon-Young Lee[2]
[1]EPFL [2]Adobe Research [3]Carnegie Mellon University

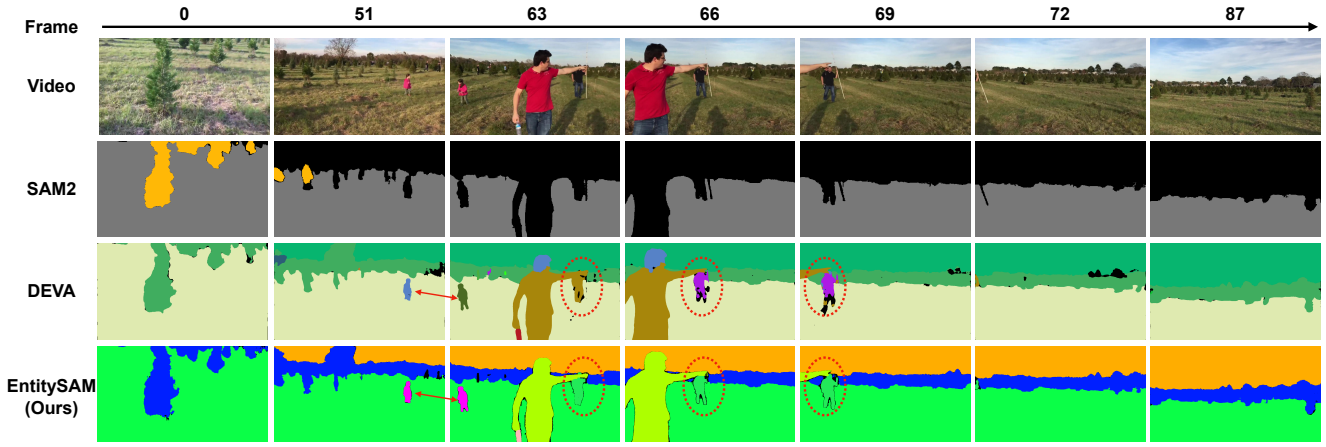


Figure 1. Zero-shot video entity segmentation performance comparison on VIPEntitySeg dataset using models trained on COCO, showing: 1) SAM 2 [39] using Mask2Former [7] mask prompts for the initial frame, 2) Mask2Former with DEVA [11] association, and 3) our proposed EntitySAM. Our **EntitySAM** enhances SAM 2 by automatically segmenting and tracking novel entities without requiring user-specified prompts, achieving superior performance compared to existing state-of-the-art zero-shot tracking methods.

Abstract

Automatically tracking and segmenting every video entity remains a significant challenge. Despite rapid advancements in video segmentation, even state-of-the-art models like SAM 2 struggle to consistently track all entities across a video—a task we refer to as Video Entity Segmentation. We propose EntitySAM, a framework for zero-shot video entity segmentation. EntitySAM extends SAM 2 by removing the need for explicit prompts, allowing automatic discovery and tracking of all entities, including those appearing in later frames. We incorporate query-based entity discovery and association into SAM 2, inspired by transformer-based object detectors. Specifically, we introduce an entity decoder to facilitate inter-object communication and an automatic prompt generator using learnable object queries. Additionally, we add a semantic encoder to enhance SAM 2's semantic awareness, improving segmentation quality. Trained on image-level mask annotations without category information from the COCO dataset, EntitySAM demonstrates strong generalization on four zero-shot video segmentation tasks: Video Entity, Panoptic, Instance, and Semantic Segmentation. Results on six popular benchmarks show that EntitySAM outperforms previous unified video segmentation methods and strong baselines, setting new standards for zero-shot video segmentation. Our code and models are at github.com/ymq2017/entitysam.

1. Introduction

Video segmentation is an essential task for various applications in video scene understanding, including video editing, AR/VR, and robotic perception. There are roughly two lines of video segmentation approaches. One line of approaches focuses on selecting semantic thing/stuff from predefined sets [27, 28, 48]. While these methods show strong performance in the benchmark, the predictions are limited to a pre-defined class set and the training data domain.

Another line of approaches works by initialing the information about the target object. The most common approach in this group is semi-supervised video object segmentation (VOS) which tracks a given mask in the first frame. The recent SAM 2 [39] also falls into this group. SAM 2 generalizes the VOS task by taking multiple prompt types, including masks, boxes, and clicks similar to image segmentation methods [19, 22], and enables interactive video segmentation. Trained on large-scale training data, SAM 2 shows

*This work was done during an internship at Adobe Research.

strong zero-shot broad-domain selection capability.

Despite recent progress, a significant challenge remains in selecting and tracking all entities from a video—a task we term **Video Entity Segmentation**, which aims to segment any distinct group with coherent semantic meaning. While SAM 2 is an effective model for tracking each individual entity, we identified two limitations in selecting all entities: (1) It initializes tracking from user prompts so it cannot track unmarked entities and new entities appearing in different frames, as shown in Figure 1. (2) It tracks each entity independently which is not only inefficient and can lead to less optimal results (e.g., mask overlapping).

In this paper, we propose **EntitySAM**, an extension of SAM 2 for zero-shot video entity segmentation. EntitySAM retains the zero-shot potential of SAM 2 while eliminating the need for explicit prompts for each entity and enabling seamless integration of new entities during tracking. Inspired by transformer-based object detectors [5, 6, 42], we sophisticatedly integrated query-based object discovery and association into SAM 2 making minimal changes to the original structure. To be specific, we replace the original mask decoder with a new entity decoder and introduce an automatic prompt generator that generates SAM 2 consumable prompts from learnable object queries. In this way, we can preserve most of the pre-trained SAM 2 weights. The proposed entity decoder differs from the original SAM 2 mask decoder in that it enables inter-entity communication to predict multiple entity masks without overlaps. Another notable addition to SAM 2 is a semantic encoder. We empirically find that SAM 2 encoder has a weak semantic awareness and supplementing semantic information using an external feature encoder greatly boosts the performance.

We train EntitySAM on the COCO dataset [23] without using any category annotation and evaluate it on various zero-shot video segmentation tasks. First we introduce a video entity segmentation task. Video entity segmentation is class-agnostic, focusing on the segmentation and tracking of "entities" rather than predefined object categories. For this, we re-purpose the video panoptic segmentation dataset [28] by removing category annotation. In addition to this, we evaluate our method on multiple video instance and semantic segmentation benchmarks in zero-shot setting. Experimental results show that EntitySAM greatly outperforms previous unified video segmentation models [52] and strong baselines that combine an image detector [7] and a mask propagator [11]. Moreover, our method can be easily adapted to class-specific video panoptic segmentation by integrating state-of-the-art vision-language models [1].

2. Related Works

Taxonomy of Video Segmentation Tasks Video segmentation tasks can be broadly categorized into class-specific and class-agnostic paradigms. Class-specific tasks, such as Video Instance Segmentation [36, 42, 48], Video Semantic Segmentation [30, 40], and Video Panoptic Segmentation [20, 28], operate within a closed vocabulary, segmenting "thing" or "stuff" categories based on predefined labels throughout video sequences.

In contrast, class-agnostic tasks, such as Video Object Segmentation (VOS) [34, 46], rely on prompt-based guidance. Unlike image-level prompts, video prompts typically require initial mask input in the first frame for temporal propagation. Recent work, such as SAM 2 [39], demonstrates that large-scale pre-training can generalize VOS to zero-shot applications. However, its reliance on explicit prompts becomes inefficient when tracking multiple dense objects with frequent occlusions or reappearances.

Recent advances have expanded video segmentation to open-world and open-vocabulary paradigms. The BURST dataset [3] introduces annotations across 482 diverse categories with a focus on thing-level segmentation, while UVO [13] takes a class-agnostic open-world approach, treating all video objects as a single category without predefined labels. Entity Segmentation [26, 37] has shown promising generalization to novel categories at the image level, but extending this approach to video introduces significant challenges for zero-shot evaluation. To bridge this gap, we propose Video Entity Segmentation, a new task aimed at segmenting entities across video sequences without relying on predefined categories or explicit prompts. This approach extends image-level entity understanding to maintain temporal consistency across frames, enabling generalized handling of novel categories in a zero-shot setting. By avoiding rigid class definitions, our framework provides greater flexibility, particularly beneficial for real-world applications with unknown or fluid category boundaries.

Query-based Video Segmentation Query-based video segmentation models are widely used for dense prediction tasks, including video instance [6, 14, 17, 18, 42–44] and semantic segmentation [2, 21]. Inspired by detection transformers [5], these methods use queries to represent object information and apply cross-attention with image features to generate segmentation masks, demonstrating effectiveness in video instance segmentation. Temporal association is achieved through either direct query matching [17, 45] or additional tracking modules [15]. While effective in controlled scenarios, these query-based methods without explicit memory propagation struggle with long or complex video sequences due to challenges in modeling long-term temporal dependencies [36].

Memory-based Video Segmentation Memory-based methods [9, 24, 31, 33, 47, 49, 50] are effective in Video Object Segmentation [4, 12, 46, 46], using initial ground truth masks as prompts. Memory features are stored in memory banks for segmentation in subsequent frames, excelling in long-term video tracking [8, 10]. SAM 2 [39] validates the

generalizability of memory-based designs but processes objects separately, limiting efficiency and lacking inter-object communication. Recent works have attempted to integrate query-based and memory mechanisms, such as Cutie [10] for object-level VOS and UniVS [29] for multi-task inference. However, joint training of memory and query modules is computationally challenging for long videos. Our EntitySAM addresses these issues by leveraging SAM 2's frozen memory encoder and introducing a novel query-based entity decoder for efficient joint training, combining the strengths of both approaches.

3. Method

We present EntitySAM for Video Entity Segmentation, enabling automatically segmenting and tracking all entities in the video sequence. By integrating query and memory-based video segmentation designs, EntitySAM handles multi-object tracking with inter-object communication and object-level contextual information. In Section 3.1, we first introduce the Video Entity Segmentation task, which offers improved generalization and reduced annotation complexity compared to other video segmentation tasks, such as Video Instance / Panoptic / Object Segmentation. In Section 3.2, we provide an overview of SAM 2's architecture and key characteristics. Section 3.3 details the architectural design of our EntitySAM based on SAM 2, highlighting our enhanced feature fusion and decoding mechanisms, along with comprehensive training and inference protocols.

3.1. Video Entity Segmentation

Video Entity Segmentation aims to segment and track all entities within video sequences using a class-agnostic approach. In this context, an "entity" encompasses both thing and stuff categories, representing any distinct group with coherent semantic meaning. The task requires continuous tracking and segmentation of these entities through temporal sequences, with annotations represented as tube masks that capture the spatial-temporal evolution of each entity.

Specifically, given an input video sequence $I \in \mathbb{R}^{T\times H\times W\times 3}$, both annotations and predictions consist of N non-overlapping masklets $M_{i\,i=1}^{N}$, where each masklet $M_i \in 0,1^{T\times H\times W}$. For predictions, the task additionally requires a confidence score $s_i \in [0,1]$ associated with each masklet. Under the non-overlapping constraint, these N masklets can be merged into a single video sequence $M \in \mathbb{N}^{T\times H\times W}$, where each pixel value represents the corresponding prediction ID.

For evaluation, we first review traditional Video Panoptic Quality (VPQ), which computes metrics per category:

$$VPQ = \frac{1}{|C|}\sum_{c\in C}\frac{\sum_t IoU_{c,t}}{TP_c + 0.5\times FP_c + 0.5\times FN_c} \quad (1)$$

To better evaluate video entity segmentation, we propose Video Entity Quality (VEQ), a class-agnostic metric that assesses performance across temporal sequences:

Table 1. Comparison of Different Video Segmentation Tasks

Task	Category		Inference		
	Things	Stuff	Automatic Prompt	Cross-Dataset	Zero-shot
Video Instance Seg.	✓		✓	✓	
Video Semantic Seg.		✓	✓	✓	
Video Object Seg.	✓			✓	✓
Video Panoptic Seg.	✓	✓	✓	✓	
Video Entity Seg.	✓	✓	✓	✓	✓

$$\text{VEQ} = \frac{\sum_t IoU_t}{TP + 0.5\times FP + 0.5\times FN} \quad (2)$$

$$\text{VEQ}^{SQ} = \frac{\sum_t IoU_t}{TP} \quad (3)$$

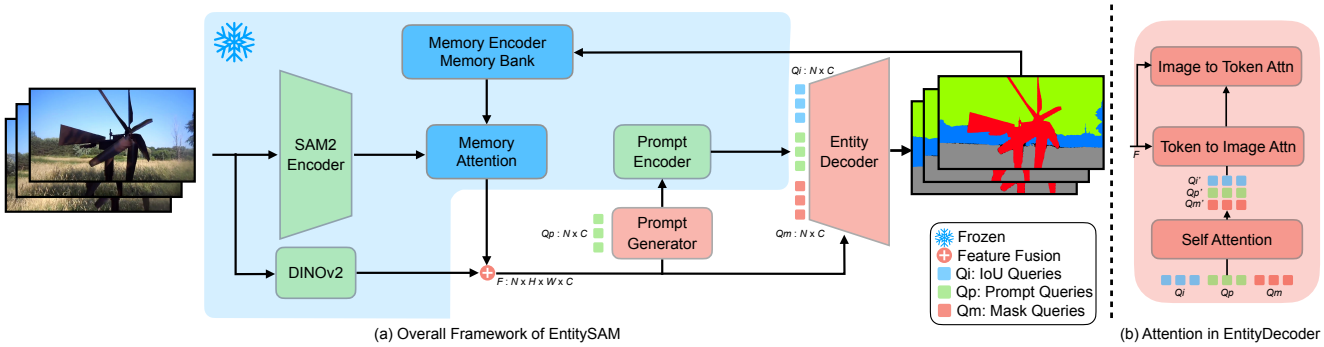
$$\text{VEQ}^{RQ} = \frac{\sum_t TP}{TP + 0.5\times FP + 0.5\times FN} \quad (4)$$

where IoU_t represents the temporal intersection over union across frames, measuring both spatial accuracy and temporal consistency. Here, TP, FP, and FN are computed globally across all entities without category differentiation. For consistency with Video Panoptic Segmentation evaluation, we also implement a class-agnostic variant of the STQ (Segmentation and Tracking Quality) metric, denoted as STQ^{EN}, which emphasizes tracking consistency across frames. The formulation of VEQ provides a more direct assessment of video-level segmentation and tracking performance while eliminating the complexity of category-specific evaluation. Moreover, this class-agnostic approach reduces annotation requirements, enables large-scale training, and enhances model generalizability across datasets, particularly for open-vocabulary and zero-shot scenarios.

As illustrated in Table 1, Video Entity Segmentation presents distinct advantages compared to existing video segmentation paradigms: Unlike Video Object Segmentation, it operates automatically and accommodates multiple dynamic objects. In contrast to Video Instance Segmentation, it extends beyond foreground objectness to include background entities. Compared to Video Semantic Segmentation, it provides instance-level annotations. Unlike Video Panoptic Segmentation, it adopts a class-agnostic approach that eliminates the constraint of pre-defined category vocabularies. This unified entity representation allows both thing and stuff categories to be learned jointly and subsequently classified using external Vision Language Models, enabling advanced open-vocabulary evaluation and reasoning-based segmentation tasks.

3.2. Method Background: SAM 2

SAM 2 [39] extends the Segment Anything model [22] to encompass both image and video domains. Through an iterative process of data annotation and model training, SAM 2 leverages both manual and automated datasets generated by its data engine. For video segmentation functionality,



(a) Overall Framework of EntitySAM (b) Attention in EntityDecoder

Figure 2. (a) Overview of the EntitySAM framework. EntitySAM utilizes the frozen encoder and memory parameters from SAM 2, incorporating a dual encoder design for enhanced semantic features (Section 3.3.1). The PromptGenerator (Section 3.3.2) automatically generates prompts from Prompt Queries. The enhanced features and distinct query groups are processed by the EntityDecoder (Section 3.3.3) to produce video mask outputs. (b) Self-attention and cross-attention mechanisms in EntityDecoder layers.

SAM 2 incorporates a classical memory mechanism to facilitate object tracking.

The architecture processes video input through multiple stages: initially, a multi-scale Image encoder extracts frame-level features, which then undergo cross-attention with previous video-level memory features to derive per-mask temporal information. Concurrently, input prompts are processed through a dedicated prompt encoder to generate prompt encodings. These various features and prompts are integrated through cross-attention mechanisms and dynamic convolution operations within the mask decoder to generate mask outputs. The model maintains temporal coherence by encoding and storing mask sequences in a memory bank, which updates frame-by-frame and feeds into memory attention for processing subsequent frames.

Through extensive training on 35M manual and automated mask annotations, SAM 2 shows exceptional zero-shot performance across various video object segmentation benchmarks, including DAVIS [35], YTVOS [46], LVOS [16], and MOSE [12]. However, SAM 2 exhibits notable limitations: it requires initial mask/prompt specification at the sequence start and cannot effectively accommodate new objects introduced during tracking. Moreover, the model's efficiency deteriorates when handling large numbers of concurrent objects, indicating challenges in scaling to more complex scenarios with multiple dynamic objects.

3.3. EntitySAM

Coupled with the proposed Video Entity Segmentation task, we present EntitySAM, a novel model that achieves strong zero-shot video entity segmentation performance while requiring only image-level class-agnostic entity annotations. As illustrated in Figure 2, our model preserves the memory-based architecture of the original SAM by freezing both the pretrained image encoder and memory-related modules, thus retaining its advantages in long-term tracking of objects. Meanwhile, EntitySAM comprises three learnable components: Dual Visual Encoder, EntityPromptGenerator, and EntityDecoder. The following sections will detail these architectural components.

3.3.1 Dual Visual Encoder

Our EntitySAM leverages the frozen Segment Anything Model 2 (SAM 2) image encoder as its backbone, as well as its frozen memory encoder and memory attention layers. The frozen design preserves SAM 2's zero-shot feature extraction capabilities, which were developed through training on extensive video object tracking datasets. Additionally, freezing the memory encoder substantially reduces GPU memory requirements during the training of multiple dense entity tracking in video sequences.

Although SAM 2 exhibits impressive performance across multi-granularity class-agnostic segmentation tasks, we reveal its limitations in maintaining consistent entity-level representations as in Table 9. Specifically, SAM 2 struggles to generate reliable confidence scores for entity classification, where an 'entity' refers to a semantically coherent group. Thus in addition to the original SAM 2's backbone feature, we integrate the backbone features from DINOv2 [32], which produces consistent feature representations for semantically similar regions by large-scale self-supervised training.

As illustrated in Figure 2, EntitySAM incorporates DINOv2 through a linear projection layer, aligning its feature dimensionality with that of SAM 2. The aligned features are then concatenated and fed into the decoder to form a fused representation. This enhances the encoded visual representation by preserving the zero-shot capabilities for video entity segmentation while effectively mitigating SAM 2's limitations in entity-level understanding.

3.3.2 Automatic Prompt Generator

In SAM 2's architecture, the prompt encoder plays a crucial role in image-level segmentation, where input prompts, such as points and bounding boxes, are processed to generate prompt embeddings for decoder segmentation. However, in video-level operations, SAM 2 bypasses the prompt

encoder, focusing solely on object tracking using historical memory features. To extend the model's capability for new entity tracking in video sequences, EntitySAM incorporates prompt embeddings during video training.

Our EntitySAM implements a prompt generator utilizing learnable queries that perform cross-attention with enhanced visual features to predict input prompts for the prompt encoder. The resulting prompt embeddings are subsequently fed into the decoder at each frame to localize both existing and novel entities within the video sequence. Concurrently, this module predicts an entity confidence score, denoted as $S(\text{entity})$, which performs binary classification between foreground entities and background regions.

More formally, let $\mathcal{Q} \in \mathbb{R}^{N \times D}$ represent the learnable queries and $\mathcal{F} \in \mathbb{R}^{HW \times D}$ denote the enhanced visual features, where N is the number of queries, D is the feature dimension, and H, W are spatial dimensions. The prompt generation and scoring process can be expressed as: $\mathcal{A} = \text{CrossAttention}(\mathcal{Q}, \mathcal{F})$, $\mathcal{P} = \text{FC}_{\text{prompt}}(A)$, and $\mathcal{S} = \text{FC}_{\text{score}}(A)$, where $A \in \mathbb{R}^{N \times D}$ is the cross-attention output, $\mathcal{P} \in \mathbb{R}^{N \times D'}$ represents the generated prompts with D' matching the expected prompt dimension required by the prompt encoder, and $\mathcal{S} \in \mathbb{R}^{N \times 1}$ denotes the entity confidence scores.

3.3.3 Entity Mask Decoder

The Entity Mask Decoder predicts output masks based on enhanced features in Section 3.3.1 and the learnable prompts through a multi-stage processing pipeline in Section 3.3.2. It employs two groups of batched queries: mask queries $\mathcal{Q}_m \in \mathbb{R}^{N \times 1 \times C}$ and IoU queries $\mathcal{Q}_{iou} \in \mathbb{R}^{N \times 1 \times C}$, where N denotes the number of entities and C represents the feature dimension. These queries are concatenated with input prompts for self-attention to facilitate inter-entity communication. These queries are subsequently reshaped for cross-attention operations between token and image representations. In the final layer, mask queries undergo dynamic convolution with upsampled mask features to generate output masks, while IoU queries are processed through a dedicated prediction head to estimate conditioned IoU scores. The computational flow can be formalized as:

$$\begin{aligned}[\mathcal{Q}'_m, \mathcal{Q}'_{iou}] &= \text{SelfAttention}(\text{concat}(\mathcal{Q}_m; \mathcal{Q}_{iou})) \\ &\quad \text{where concat}(\mathcal{Q}_m; \mathcal{Q}_{iou}) \in \mathbb{R}^{1 \times 2N \times C} \\ F_{out} &= \text{CrossAttention}(\text{reshape}(\mathcal{Q}'_m, \mathcal{Q}'_{iou}), F) \\ &\quad \text{where reshape}(\mathcal{Q}'_m, \mathcal{Q}'_{iou}) \in \mathbb{R}^{N \times 2 \times C}.\end{aligned} \tag{5}$$

The supervision mechanism integrates outputs from both the prompt generator (entity scores) and mask decoder (masks and conditioned IoU predictions) through Hungarian matching with ground truth annotations. The expected IoU for query q can be expressed as:

$$\begin{aligned}E(\text{IoU}_q) &= E(\text{IoU}_q|\text{Ent}_q)S(\text{Ent}_q) + E(\text{IoU}_q|\neg\text{Ent}_q)S(\neg\text{Ent}_q) \\ &= E(\text{IoU}_q|\text{Ent}_q)S(\text{Ent}_q),\end{aligned} \tag{6}$$

where $S(\text{Ent}_q)$ represents the entity confidence score from PromptGenerator in Section 3.3.2 and $E(\text{IoU}_q|\neg\text{Ent}_q) = 0$.

The resulting output masks are processed through the frozen memory encoder, memory bank, and memory attention mechanisms, with gradient computation detached for memory-related components. This enables efficient training of the queries-based tracking component while maintaining frozen memory modules. Consequently, our approach achieves scalable training for over 50 entities in 8-frame video clips, significantly surpassing SAM 2's original capacity of handling only 3 video objects during training.

3.3.4 Training Loss and Open-Vocabulary Inference

We employ an end-to-end training objective for EntitySAM, formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{Mask}} + \lambda_1 \mathcal{L}_{\text{Entity}} + \lambda_2 \mathcal{L}_{\text{IoU}}, \tag{7}$$

where $\mathcal{L}$ supervises both mask prediction and binary entity classification, following the Mask2former framework [7]. The hyperparameters λ_1 and λ_2 balance the respective loss terms and are empirically set to 1.0 and 1.0.

During inference, EntitySAM accommodates both class-agnostic and class-specific evaluation protocols. For class-agnostic scenarios, entity masks are directly generated and merged based on confidence scores and mask overlap metrics. In class-specific evaluation, we select the frame with the highest entity confidence score for each masklet prediction. The predicted entity masks are utilized to crop corresponding regions of interest, which, together with carefully designed prompts containing candidate category names, are processed through GPT-4o [1] for final classification. We demonstrate in the supplementary materials that this mask-guided cropping approach achieves superior classification performance compared to alternative methods.

4. Experiments

4.1. Experimental Setup

Datasets Video entity segmentation typically requires extensive class-agnostic dense entity video segmentation annotations. Although it eliminates the need for class-specific annotations, creating frame-by-frame annotations remains labor-intensive. Inspired by [51, 52], our training pipeline leverages COCO [23] panoptic segmentation data without category label annotation (denoted as COCO*) and extends it to pseudo-video sequences. Our approach utilizes only image-level annotations, eschewing video-level labels, to achieve effective video entity segmentation. The pipeline

also enables zero-shot cross-dataset evaluation across multiple benchmarks. We demonstrate our model's versatility through comprehensive zero-shot evaluations on various tasks: video panoptic segmentation on VIPSeg [28], video semantic segmentation on VSPW [27], video instance segmentation on Youtube-VIS 2019 [48], Youtube-VIS 2021 [48], and OVIS [36], and video entity segmentation on our proposed VIPEntitySeg dataset—a class-agnostic variant of VIPSeg. Notably, all zero-shot evaluations are performed using a single trained model, demonstrating its robust zero-shot generalization capabilities across different video segmentation tasks and datasets.

Implementation details In our experimental setup, we employ two backbone architectures: ViT-Small and ViT-Large. The models are trained using the AdamW optimizer with a learning rate of 5×10^{-5} and weight decay of 0.05. Input images are cropped to 1024×1024 pixels for compatibility with SAM 2 ViT/16 and 896×896 pixels for DINOv2 ViT/14. Training is conducted with a batch size of 8. The loss function combines multiple components: cross-entropy and dice losses for mask prediction, L1 loss for IoU prediction, and focal loss for binary entity classification. Our training strategy consists of two phases: initial training on static COCO images for 40K iterations, followed by training on COCO pseudo-video sequences with clip length 8 for 10K iterations. Both stages are class-agnostic. During inference, we utilize a single entity category output mode with confidence scoring. Video entity quality (VEQ) is evaluated in a class-agnostic manner. For class-specific evaluation tasks, we implement a multi-step process: first selecting the highest-confidence mask from the masklet for each prediction, then extracting the corresponding frame region within the mask, and finally employing GPT-4o for classification.

4.2. Zero-shot Comparison

Zero-shot Video Entity Segmentation In Table 2 (left), we evaluate our approach on VIPEntitySeg using class-agnostic metrics against several baseline methods. For SAM 2, we adopt two prompting manners, a grid point prompt sampling initialization and a COCO-trained Mask2Former [7] mask prompts initialization. We additionally compare against state-of-the-art zero-shot segmentation models, including: 1) FC-CLIP [51] integrated with MinVIS [17] association, 2) DEVA [11] coupled with COCO-trained Mask2Former [7] for per-frame segmentation, and 3) OV-DVIS++ [52] in both online and offline modes. All these SOTA models were trained on COCO with category annotations, with DEVA utilizing additional external datasets for its association model. We compare with these methods on our class-agnostic VIPEntitySeg benchmark. Our EntitySAM achieves significant performance improvements over all baselines and SOTA models with both ViT-Small and ViT-Large backbones, despite training without category annotations. This superior performance demonstrates the effectiveness of our class-agnostic training approach and proposed modules. Notably, we observe substantial improvements in VEQ^{RQ}, indicating enhanced cross-frame entity identification and reduced false positives. Figure 3 further illustrates our model's robustness by evaluating VEQ across varying frame intervals, demonstrating that EntitySAM maintains consistent performance over extended temporal sequences.

Regarding the class-agnostic STQ^{EN} metric, which emphasizes tracking performance, DEVA achieves higher scores through its tracking association model trained on external video object segmentation datasets. However, DEVA relies on a two-stage pipeline of detection followed by association, lacking end-to-end integration. While our approach achieves lower scores, it delivers comparable performance through a streamlined process that requires only COCO training data and single-pass inference.

Zero-shot Video Panoptic Segmentation In Table 2 (right), we extend our evaluation to VIPSeg for class-specific video panoptic segmentation. Several baseline models, including Mask2Former and DEVA, are constrained to VIPSeg evaluation due to their closed-vocabulary nature. When compared to open-vocabulary models such as FC-CLIP and OV-DVIS++, our approach demonstrates substantial improvements, particularly with the ViT-Small backbone. These results validate our approach of "post-classification with multimodal models", indicating that it surpasses traditional open-vocabulary segmentation training pipelines despite operating without category annotations. Our EntitySAM achieves competitive performance on both class-agnostic and class-specific tasks, suggesting that category annotations may not be essential for effective open-vocabulary video segmentation.

Zero-shot Video Semantic Segmentation Table 3 presents our results on video semantic segmentation, obtained by merging per-pixel mask predictions. Our method achieves improvements in both mean Intersection over Union (mIoU) and Video Consistency (VC) metrics. The VSPW evaluation dataset, which contains the same video content but with extended frame sequences, validates our model's capability for effective long-term tracking.

Zero-shot Video Instance Segmentation As shown in Table 4, we evaluate EntitySAM on three popular video instance segmentation datasets. Our approach on extending video entity segmentation to video instance segmentation requires no model retraining—we simply add a background/stuff class during inference and filter out predictions with that class to obtain instance results. Our model demonstrates significant performance improvements over state-of-the-art open-vocabulary video instance segmentation models. In particular, on the challenging OVIS dataset, which features frequent occlusions, our model maintains robust

Table 2. Comparison of **Zero-shot** Video Entity Segmentation (left) and Video Panoptic Segmentation (right) on COCO → VIPSeg evaluation. EntitySAM is trained on the COCO dataset without category annotations. In VIPSeg, – means the results of open-vocabulary evaluation for closed-vocabulary models are not available. * denotes that DEVA is trained with multiple external video object segmentation datasets for the temporal association model.

Method	Backbone	Class Agnostic Training	VIPEntitySeg VEQ	VEQ^{SQ}	VEQ^{RQ}	STQ^{EN}	VIPSeg VPQ	VPQ^{Th}	VPQ^{St}	STQ
SAM 2.1 [39] (Everything Init)	ViT-S	✓	26.9	82.5	32.6	38.7		-		
SAM 2.1 [39] (Mask2former Init)	ViT-S + ResNet50		46.2	83.6	55.2	34.5		-		
DEVA* [11]	ResNet-50		49.0	83.6	58.6	**42.4**		-		
FC-CLIP [51]	ResNet-50		41.5	84.0	49.4	29.7	22.3	25.5	19.1	19.7
OV-DVIS++(Online) [52]	ResNet-50		45.9	83.7	54.8	33.6	24.4	26.8	22.4	22.0
OV-DVIS++(Offline) [52]	ResNet-50		46.7	82.8	56.3	35.6	23.8	26.4	21.4	24.4
Ours	ViT-S	✓	**51.8**	**84.7**	**61.1**	41.0	**28.7**	**32.9**	**25.1**	**31.4**
SAM 2.1 [39] (Everything Init)	ViT-L	✓	30.5	83.1	36.7	41.7		-		
SAM 2.1 [39] (Mask2former Init)	ViT-L + Swin-L		50.6	84.4	60.0	39.1		-		
DEVA* [11]	Swin-L		50.8	83.8	60.6	**45.3**		-		
FC-CLIP [51]	ConvNext-L		43.4	84.8	51.2	32.4	27.9	30.9	25.0	24.2
OV-DVIS++(Online) [52]	ConvNext-L		49.4	83.5	59.2	37.9	28.9	31.3	26.8	28.4
OV-DVIS++(Offline) [52]	ConvNext-L		49.0	83.1	59.0	40.0	30.4	31.9	29.1	32.2
Ours	ViT-L	✓	**54.6**	**84.7**	**64.5**	43.3	**31.4**	**37.5**	26.0	**33.5**

Table 3. **Zero-Shot** Open-vocabulary Video Semantic Segmentation on COCO → VSPW. EntitySAM is trained on the COCO dataset without category annotations.

Method	Backbone	Class Agnostic Training	VSPW mVC_8	mVC_{16}	mIOU
FC-CLIP [51]	ResNet-50		84.9	82.7	24.3
OV-DVIS++(Online) [52]	ResNet-50		92.7	91.5	27.6
OV-DVIS++(Offline) [52]	ResNet-50		92.4	91.3	28.4
Ours	ViT-S	✓	**94.0**	**93.0**	**34.6**
FC-CLIP [51]	ConvNext-L		89.9	88.4	28.9
OV-DVIS++(Online) [52]	ConvNext-L		94.2	93.3	34.3
OV-DVIS++(Offline) [52]	ConvNext-L		93.9	93.0	34.1
Ours	ViT-L	✓	**94.6**	**93.7**	**35.5**

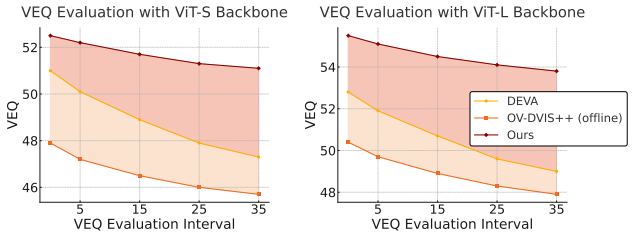


Figure 3. VEQ comparison with different evaluation time intervals. EntitySAM improves over long intervals.

performance, benefiting from the entity learning strategy employed during training.

4.3. Ablation Experiments

We perform ablation studies using the ViT-Small backbone on the COCO → VIPEntitySeg video entity segmentation task with zero-shot protocol to evaluate the effectiveness of our proposed modules.

Ablation on the Proposed Modules Table 5 presents ablation studies of our proposed modules. Our baseline, SAM 2 with memory-based tracking modules, cannot generate predictions automatically without appropriate prompting. Using grid-point sampling leads to numerous false positive predictions, achieving a VEQ of only 26.9. Removing the memory design and implementing solely a query-based EntityDecoder with query association for tracking improves performance to 41.3, though this remains suboptimal. Combining memory-based and query-based approaches significantly enhances performance to 50.2, surpassing both DEVA and OV-DVIS++. Finally, incorporating a complementary DINOv2 dual feature encoder achieves the best result (51.8) for the ViT-Small backbone.

Ablation on Query Design As illustrated in Figure 2, our design incorporates three query types: two groups in EntityDecoder and one group in PromptGenerator. Our ablation experiment of query design in Table 6 reveals that using a single query group to predict all outputs (mask, IoU, and confidence score) yields suboptimal performance with a VEQ of 49.2. Adding a second group of prompt queries improves performance to 49.7. Further separation into a third query group for distinct mask and IoU prediction achieves the best performance of 51.8. These results suggest that different query groups learn specialized distributions for their respective prediction tasks, departing from traditional detection transformers' shared-query architecture.

Ablation on EntityDecoder Depth Table 7 studies the influence of the entity decoder's depth layer. For the ViT-Small backbone, a decoder depth of 4 achieves optimal performance with a VEQ of 51.8 on COCO → VIPSeg setting. Further increasing decoder depth shows no additional performance benefits.

Ablation on Training Stages As detailed in Section 4.1, our training process consists of two stages: initial training on single images (treated as zero-length videos) followed by training on pseudo video clips of length 8. During the first stage, the memory component remains unused while the EntityDecoder learns to identify new entities in each image. Table 8 shows that this image-only training achieves limited video entity segmentation performance with a VEQ of 31.4. However, this stage exposes the model to diverse en-

Table 4. **Zero-Shot** Open-vocabulary Video Instance Segmentation on COCO → Validation sets of YouTube-VIS 2019, 2021, and OVIS. EntitySAM is trained on the COCO dataset without category annotations.

Method	Backbone	Class Agnostic Training	Youtube-VIS 2019					Youtube-VIS 2021					OVIS				
			AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
Detic[53]-OWTB[25]	ResNet-50		17.9	-	-	-	-	16.7	-	-	-	-	9.0	-	-	-	-
MindVLT[41]	ResNet-50		23.1	-	-	-	-	20.9	-	-	-	-	11.4	-	-	-	-
FC-CLIP [51]	ResNet-50		28.9	43.9	29.9	32.4	41.2	25.5	40.0	26.8	28.0	37.0	11.8	25.1	10.5	8.5	16.4
OV-DVIS++(Online) [52]	ResNet-50		34.5	50.6	39.2	39.5	49.5	30.9	46.7	34.8	34.4	45.8	14.8	31.2	13.1	10.5	24.7
OV-DVIS++(Offline) [52]	ResNet-50		34.4	51.8	39.2	40.0	50.6	31.0	48.5	34.3	34.6	46.8	13.0	28.4	10.9	9.8	25.1
Ours	ViT-S	✓	**44.9**	**60.7**	**49.5**	**43.6**	**62.4**	**38.8**	**52.5**	**43.1**	**36.2**	**56.5**	**21.0**	**35.8**	**20.8**	**12.3**	**36.0**

Table 5. Ablation Study on the Proposed Modules.

Model	**Memory**	**EntityDecoder**	**Dual Visual Encoder**	**VEQ**
SAM 2 [39]	✓			26.9
Ours		✓		41.3
	✓	✓		50.2
	✓	✓	✓	**51.8**

Table 6. Ablation Study on Query Design.

EntityDecoder	PromptGenerator	COCO → VIPSeg		
		VEQ	VEQ^{SQ}	VEQ^{RQ}
Shared Queries		49.2	84.0	58.6
Shared Queries	✓	49.7	84.2	59.0
Separate Queries	✓	**51.8**	**84.7**	**61.1**

Table 7. Ablation Study on EntityDecoder Depth.

Decoder	Decoder depth	COCO → VIPSeg		
		VEQ	VEQ^{SQ}	VEQ^{RQ}
EntityDecoder	2	51.3	84.7	60.5
	4	**51.8**	**84.7**	**61.1**
	6	51.6	84.6	61.0
	8	51.1	84.4	60.6

tities, and when combined with the subsequent video training stage—which aligns EntityDecoder with memory encoders—performance improves substantially to 51.8 VEQ. This two-stage approach outperforms training exclusively on pseudo video sequences.

Ablation on Dual Encoder In Section 3.3.1, we introduced encoder feature fusion designs for dual encoders. As shown in Table 9, our ablation studies reveal that incorporating features from either CLIP [38] or SAM [22] alongside the SAM 2 encoder features yields no improvement in zero-shot performance. This can be attributed to CLIP's vision encoder being optimized for classification rather than segmentation tasks, resulting in representations less suitable for segmentation. Similarly, SAM, being trained with an approach analogous to SAM 2, provides features that largely overlap with SAM 2's capabilities rather than offering complementary information. In contrast, DINOv2, trained through self-supervision, demonstrates significant benefits when combined with SAM 2. Its self-supervised training enables DINOv2 to learn consistent image representations that effectively complement SAM 2's supervised features, ultimately enhancing the overall performance.

Table 8. Ablation Study on Training stages.

Model	Image	Pseudo Video	COCO → VIPSeg		
			VEQ	VEQ^{SQ}	VEQ^{RQ}
Ours	✓		31.4	79.4	39.5
		✓	49.2	84.2	58.4
	✓	✓	**51.8**	**84.7**	**61.1**

Table 9. Ablation Study on Dual Encoder Feature.

Encoder Feature	Dual Encoder	COCO → VIPSeg		
		VEQ	VEQ^{SQ}	VEQ^{RQ}
SAM 2	-	50.2	84.5	59.4
	SAM [22]	47.6	84.2	56.5
	CLIP [38]	49.2	84.0	58.6
	DINOv2 [32]	**51.8**	**84.7**	**61.1**

4.4. Qualitative experiments and visualization

Figure 4 demonstrates our model's performance on zero-shot video sequences, showcasing consistent entity tracking and segmentation across frames. Additional qualitative results are provided in the supplementary materials.

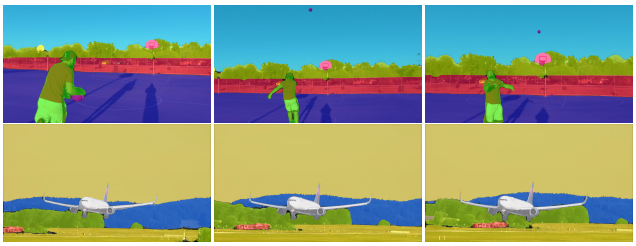


Figure 4. Visualization of zero-shot video entity segmentation.

5. Conclusion

We present EntitySAM, a framework that integrates query and memory mechanisms to enhance SAM 2 for zero-shot video entity segmentation. Our approach eliminates the need for user-defined prompt inputs and enables automatic tracking of new entities across frames. Trained exclusively on COCO image-level mask annotations without category information, EntitySAM achieves superior performance on the Video Entity Segmentation task. The model demonstrates strong generalization capabilities across various zero-shot evaluations, including Video Panoptic / Semantic / Instance Segmentation. These results provide valuable insights into class-agnostic training approaches and their effectiveness in zero-shot scenarios.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 5

[2] Zhaochong An, Guolei Sun, Zongwei Wu, Hao Tang, and Luc Van Gool. Temporal-aware hierarchical mask classification for video semantic segmentation. *arXiv preprint arXiv:2309.08020*, 2023. 2

[3] Ali Athar, Jonathon Luiten, Paul Voigtlaender, Tarasha Khurana, Achal Dave, Bastian Leibe, and Deva Ramanan. Burst: A benchmark for unifying object recognition, segmentation and tracking in video. In *WACV*, 2023. 2

[4] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. In *arXiv preprint arXiv:1905.00737*, 2019. 2

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[6] Bowen Cheng, Anwesa Choudhuri andd Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G. Schwing. Mask2former for video instance segmentation. In *https://arxiv.org/abs/2112.10764*, 2021. 2

[7] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 1, 2, 5, 6

[8] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022. 2

[9] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *NeurIPS*, 2021. 2

[10] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *arXiv*, 2023. 2, 3

[11] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *ICCV*, 2023. 1, 2, 6, 7

[12] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *ICCV*, 2023. 2, 4

[13] Yuming Du, Wen Guo, Yang Xiao, and Vincent Lepetit. Uvo challenge on video-based open-world segmentation 2021: 1st place solution. *ICCV Workshop*, 2021. 2

[14] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. *Advances in Neural Information Processing Systems*, 35:23109–23120, 2022. 2

[15] Miran Heo, Sukjun Hwang, Jeongseok Hyun, Hanjung Kim, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. A generalized framework for video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14623–14632, 2023. 2

[16] Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. Lvos: A benchmark for long-term video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13480–13492, 2023. 4

[17] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. In *NeurIPS*, 2022. 2, 6

[18] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *NeurIPS*, 2021. 2

[19] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023. 1

[20] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *CVPR*, 2020. 2

[21] Dahun Kim, Jun Xie, Huiyu Wang, Siyuan Qiao, Qihang Yu, Hong-Seok Kim, Hartwig Adam, In So Kweon, and Liang-Chieh Chen. Tubeformer-deeplab: Video mask transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13914–13924, 2022. 2

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 1, 3, 8

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 5

[24] Yong Liu, Ran Yu, Fei Yin, Xinyuan Zhao, Wei Zhao, Weihao Xia, and Yujiu Yang. Learning quality-aware dynamic memory for video object segmentation. In *ECCV*, 2022. 2

[25] Yang Liu, Idil Esen Zulfikar, Jonathon Luiten, Achal Dave, Deva Ramanan, Bastian Leibe, Aljoša Ošep, and Laura Leal-Taixé. Opening up open world tracking. In *CVPR*, 2022. 8

[26] Qi Lu, Jason Kuen, Shen Tiancheng, Gu Jiuxiang, Guo Weidong, Jia Jiaya, Lin Zhe, and Yang Ming-Hsuan. High-quality entity segmentation. In *ICCV*, 2023. 2

[27] Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4133–4143, 2021. 1, 6

[28] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *CVPR*, 2022. 1, 2, 6

[29] Xindong Zhang Minghan Li, Shuai Li and Lei Zhang. Univs: Unified and universal video segmentation with prompts as queries, 2024. 3

[30] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6819–6828, 2018. 2

[31] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 2

[32] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 4, 8

[33] Kwanyong Park, Sanghyun Woo, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Per-clip video object segmentation. In *CVPR*, 2022. 2

[34] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 2016. 2

[35] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 4

[36] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation. *IJCV*, 2022. 2, 6

[37] Lu Qi, Jason Kuen, Yi Wang, Jiuxiang Gu, Hengshuang Zhao, Zhe Lin, Philip Torr, and Jiaya Jia. Open-world entity segmentation. In *arXiv preprint arXiv:2107.14228*, 2021. 2

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 8

[39] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1, 2, 3, 7, 8

[40] Guolei Sun, Yun Liu, Henghui Ding, Thomas Probst, and Luc Van Gool. Coarse-to-fine feature mining for video semantic segmentation. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3126–3137, 2022. 2

[41] Haochen Wang, Cilin Yan, Shuai Wang, Xiaolong Jiang, Xu Tang, Yao Hu, Weidi Xie, and Efstratios Gavves. Towards open-vocabulary video instance segmentation. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 4057–4066, 2023. 8

[42] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *CVPR*, 2021. 2

[43] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *European Conference on Computer Vision*, pages 553–569. Springer, 2022.

[44] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, 2022. 2

[45] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *European Conference on Computer Vision*, pages 588–605. Springer, 2022. 2

[46] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. In *ECCV*, 2018. 2, 4

[47] Xiaohao Xu, Jinglu Wang, Xiao Li, and Yan Lu. Reliable propagation-correction modulation for video object segmentation. In *AAAI*, 2022. 2

[48] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 1, 2, 6

[49] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In *NeurIPS*, 2021. 2

[50] Zongxin Yang, Jiaxu Miao, Yunchao Wei, Wenguan Wang, Xiaohan Wang, and Yi Yang. Scalable video object segmentation with identification mechanism. *TPAMI*, 2024. 2

[51] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In *NeurIPS*, 2023. 5, 6, 7, 8

[52] Tao Zhang, Xingye Tian, Yikang Zhou, Shunping Ji, Xuebo Wang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, and Yu Wu. Dvis++: Improved decoupled framework for universal video segmentation. *arXiv preprint arXiv:2312.13305*, 2023. 2, 5, 6, 7, 8

[53] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 8