This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

UniVS: Unified and Universal Video Segmentation with Prompts as Queries

Minghan Li[1,2*], Shuai Li[1,2*], Xindong Zhang[2] and Lei Zhang[1,2†]
[1]The Hong Kong Polytechnic University [2]OPPO Research Institute
liminghan0330@gmail.com, xindongzhang@foxmail.com, {csshuaili, cslzhang}@comp.polyu.edu.hk

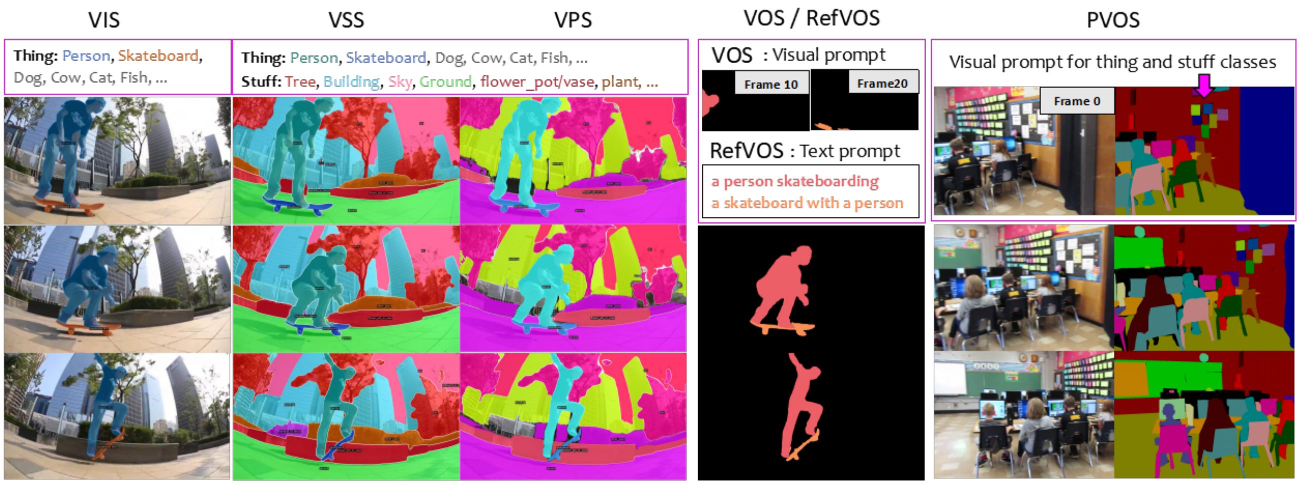

Figure 1. Illustration of different video segmentation (VS) tasks. Category-specified VS includes VIS, VSS and VPS tasks, while prompt-specified VS consists of VOS, RefVOS and PVOS tasks. Please find more video demos on our project page https://sites.google.com/view/unified-video-seg-univs.

Abstract

Despite the recent advances in unified image segmentation (IS), developing a unified video segmentation (VS) model remains a challenge. This is mainly because generic category-specified VS tasks need to detect all objects and track them across consecutive frames, while prompt-guided VS tasks require re-identifying the target with visual/text prompts throughout the entire video, making it hard to handle the different tasks with the same architecture. We make an attempt to address these issues and present a novel unified VS architecture, namely UniVS, by using prompts as queries. UniVS averages the prompt features of the target from previous frames as its initial query to explicitly decode masks, and introduces a target-wise prompt cross-attention layer in the mask decoder to integrate prompt features in the memory pool. By taking the predicted masks of entities from previous frames as their visual prompts, UniVS converts different VS tasks into prompt-guided target segmentation, eliminating the heuristic inter-frame matching process. Our framework not only unifies the different VS tasks but also naturally achieves universal training and testing, ensuring robust performance across different scenarios. UniVS shows a commendable balance between performance and universality on 10 challenging VS benchmarks, covering video instance, semantic, panoptic, object, and referring segmentation tasks. Code can be found at https://github.com/MinghanLi/UniVS.

*Equal contribution, † Corresponding author.

1. Introduction

Video segmentation (VS) partitions a video sequence into different regions or segments, facilitating many applications such as semantic-guided video restoration [35, 66, 75], video generation [54, 61], video editing and augmented reality [18, 69, 91], *etc*. VS tasks can be divided into two groups: category-specified VS and prompt-specified VS. The former focuses on segmenting and tracking entities from a predefined set of categories. Typical tasks along this line include video instance [58, 81], semantic [52] and panoptic segmentation [26, 51] (VIS/VSS/VPS), where the object category information needs to be specified. Another group focuses on identifying and segmenting specific targets throughout the video, where visual prompts or textual descriptions of the targets need to be provided. Prompt-specified VS tasks include video object segmentation (VOS) [2, 14, 56, 78], Panoptic VOS (PVOS) [79] and referring VOS (RefVOS) [63]. Each of above VS tasks has established its own protocol for dataset annotation, as well as model evaluation.

VS has a close relationship with image segmentation (IS) [12, 16, 25, 28, 43, 49, 55, 77, 94, 95], which has similar types of segmentation tasks to VS. In the past decades, the model performance on each individual IS/VS task has been significantly improved, and many well-known network architectures have been developed [3, 5, 7, 11, 17, 19, 23, 33, 47, 49, 62, 67, 74, 82, 85, 87]. Some architectures [8, 29, 39, 40, 64, 89, 90, 97] have also been proposed to deal with multiple segmentation tasks; however, these architectures require separate training and inference for different tasks because of the different annotations and evaluation protocols. Fortunately, the recent advancements in vision-language models [24, 31, 32, 42, 44, 45, 48, 60, 84, 92, 93] align and harmonize the multimodal feature representations, which bridge the labeling gaps across various IS/VS tasks. As a result, some unified segmentation models [30, 59, 98] have emerged to process multiple segmentation tasks simultaneously, which can be jointly trained on different datasets and tasks. These methods generally convert prompt-guided segmentation into the category-specified segmentation problem [8, 19, 96], which first predicts masks for all potential entities per frame and then utilizes a post-processing matching to find the target. For example, SEEM [98] captures prompt information of the target via concatenating learnable queries and prompt features as keys and values in the self-attention layer of the mask decoder, showing good versatility in various IS tasks.

Compared to IS, the additional challenge of VS lies in the temporal consistency [68] of segmentation across frames in a video sequence. Existing unified models [1, 80] for VS tasks are mostly inspired by the unified IS models. They segment the video sequence frame by frame, and then use a similarity matching step to associate common objects or find the targets for category-specified and prompt-specified VS tasks, respectively. For example, UNINEXT [80] is specially designed for object-centric IS and VS tasks. Though UNINEXT performs very well in several aspects, it becomes ineffective when segmenting 'stuff' entities, such as 'sky'. To adapt to different VS tasks, TarVIS [1] subdivides the learnable queries within the mask decoder into four groups: semantic, class-agnostic instance, object and background queries. However, TarVIS falls short in encoding linguistic information, hindering it from resolving language-guided VS tasks.

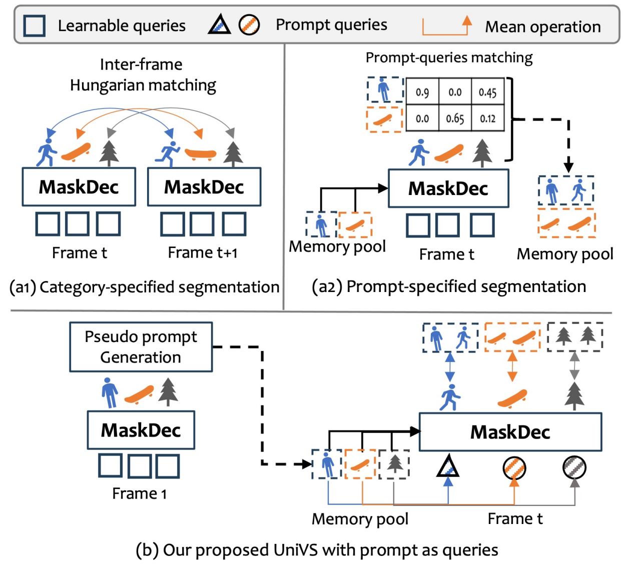

Figure 2. Comparison between the existing unified segmentation methods and ours. In existing methods for category-specified segmentation tasks (see (a1)), entities need to be first detected per frame and then matched across frames, while in methods for prompt-specified segmentation tasks (see (a2)), targets need to be identified from the predicted masks. In contrast, our proposed UniVS (see (b)) uses predicted masks as pseudo visual prompts and averages prompt features to decode masks across videos, avoiding the heuristic post-processing process.

From the above discussions, we can see that it remains a challenge to develop a unified VS framework to effectively handle all VS tasks. This is mainly because category-specified and prompt-specified VS tasks attribute to different focuses. As shown in Figs. 2(a1) and 2(a2), category-specified segmentation prioritizes the precise detection per frame and the inter-frame association for common objects, while prompt-specified segmentation concentrates on accurately tracking the target with text/visual prompts in video sequences, where the target can be an uncommon object or a part of an object. The different focuses of the two types of VS tasks make it challenging to integrate them within a single framework while achieving satisfactory results.

To alleviate the above issues, we propose a novel unified VS architecture, namely **UniVS**, by using prompts as queries. For each target of interest, UniVS averages the prompt features from previous frames as its initial query. A target-wise prompt cross-attention (ProCA) layer is introduced in the mask decoder to integrate comprehensive prompt features stored in the memory pool. The initial query and the ProCA layer play a crucial role in the explicit and accurate decoding of masks. On the other hand, by taking the predicted masks of entities from previous frames as their visual prompts, UniVS can convert different VS tasks into the task of prompt-guided target segmentation task, eliminating the heuristic inter-frame matching. The overall process of UniVS is depicted in Fig. 2(b). UniVS not only unifies the different VS tasks (see Fig. 1) but also naturally achieves universal training and testing, resulting in robust performance across different scenarios. It shows a commendable balance between performance and universality on 10 challenging VS benchmarks, covering VIS, VSS, VPS, VOS, RefVOS and PVOS tasks. To the best of our knowledge, ***UniVS is the first work which can unify all the existing VS tasks successfully in a single model***.

2. Related Work

We first briefly introduces the representative models designed for category-specified and prompt-specified VS tasks, and then introduce the existing unified and universal models for processing multiple VS tasks simultaneously.

Category-specified VS tasks includes video instance [81], semantic [52] and panoptic segmentation [51] tasks, which aim to correctly partition the regions and label them with specific categories or open-world vocabulary[17, 37, 86]. Building on representative query-based image detection and segmentation methods [4, 8, 36, 96], recent approaches [21–23, 27, 33, 34, 39, 40, 67, 71, 74] focus on efficiently encoding both short-term and medium-term temporal information. Some methods [23, 34, 41, 71, 74] utilize the consistency of the relative relationship between objects in a short period of time to associate entities across frames, while the latest state-of-the-arts [21, 22, 39, 40, 89] design learnable trackers between multiple video clips to better learn entity motion information over short to medium time periods, thus maintaining more temporally consistent segmentation results. This continual learning strategy also bridges the gap between training and inference.

Prompt-specified VS tasks consist of video object segmentation (VOS) [2, 14, 57, 78], panoptic VOS [79] and referring VOS [63], which aim to re-identify and segment the target objects with visual/text prompts in a video. Recent offline models [9, 11, 38, 50, 72, 83, 87] focus more on designing long-term information propagation modules to transfer previous image features and corresponding masks to the target frame to predict masks. This helps more precisely identify and track the movement trajectory of objects throughout the entire video sequence (more than 200 frames), but also makes the network architecture heavy [73, 83]. Due to the slow speed of offline models, online models [70, 73, 88] come back into focus, aiming to keep the right balance between speed and performance.

Unified VS models [1, 6, 10, 39, 40, 64, 73, 80, 89] have emerged recently. TubeLink [39] and DVIS [89] employ a single framework to accomplish different category-specified VS tasks, but they still need to be trained separately for each task and cannot be used for prompt-specified VS tasks. UniRef [73] uses a model to unify prompt-specified VS tasks, but it cannot handle category-specified VS tasks. To handle more types of VS tasks, TarVIS [1] decouples learnable queries of mask decoder into four groups. But, it fails to solve RefVOS due to the lack of language encoding capability. UNINEXT [80] is a unified object-centric segmentation model on images and videos, demonstrating universal applicability in several aspects. Unfortunately, it cannot handle entities with 'stuff' categories.

By far, there still lacks a unified model that can accommodate all VS tasks simultaneously. This is mainly because different VS tasks have different focuses, which can be contradictory. To address this difficulty, we use prompts as queries to unify different tasks into a unified framework.

3. Methodology

UniVS contains of three main modules: the Image Encode, the Prompt Encoder and the Unified Video Mask Decoder, as depicted in Fig. 3. The Image Encoder transforms RGB images to the feature tokens, while the Prompt Encoder translates raw visual/text prompts into the prompt embeddings. The Unified Video Mask Decoder explicitly decodes masks for any entity or prompt-guided target in the video.

3.1. Image and Prompt Encoders

Image encoder contains a backbone and a pixel decoder [8]. The backbone maps the RGB image $X \in R^{3\times H\times W}$ into multi-scale features, and the pixel decoder further fuses features across scales to enhance the representation: $\{F_s\}_{s=1}^4 = \text{ImEnc}(X)$, where $F_s \in R^{C\times H_s\times W_s}$ is the s-th scale image embeddings, H_s, W_s are its height and width and C is the number of channels. The resolutions of multi-scale feature maps are 1/32, 1/16, 1/8 and 1/4 of that of the input image, respectively.

Prompt encoder converts the input visual/text prompts into the prompt embeddings. The visual prompts can be clicked points, boxes, masks and scribbles, *etc*. For the i-th target, we use $X_i^* \in R^{l_i\times 3}$ and $y_i^* \in R^{l_i}$ to represent its prompt-specified pixels and corresponding segment IDs, where l_i is the total number of prompt-specified pixels. To convert the visual prompts into image embeddings, we adopt the Visual Sampler strategy proposed in SEEM [98]. It samples l^* points from the prompt-specified pixels for each target, and extracts point features from the 3rd scale image embeddings F_3 as its visual prompt embeddings:

$$P_i^* = \text{VisualSampler}(X_i^*,\ F_3), \tag{1}$$

where the shape of P_i^* is $l^* \times C$.

The language prompt can be category names, such as 'person', and textual expressions, such as 'a person is skateboarding'. Following [13, 60], we feed a category name or an expression into a tokenizer to get its string tokens $L_i^* \in R^{l^*\times C^t}$, which are input into the CLIP text encoder

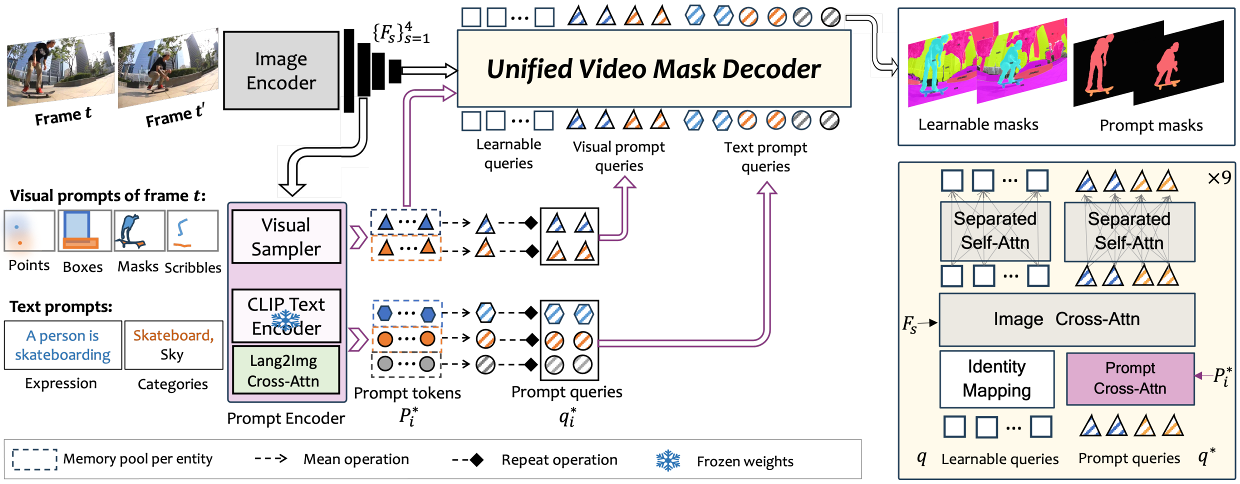

Figure 3. Training process of our unified video segmentation (UniVS) framework. UniVS contains three main modules: the Image Encoder (grey rectangle), the Prompt Encoder (purple rectangle) and the Unified Video Mask Decoder (yellow rectangle). The Image Encoder transforms the input RGB images to the feature space and outputs image embeddings. Meanwhile, the Prompt Encoder translates the raw visual/text prompts into prompt embeddings. The Unified Video Mask Decoder explicitly decodes the masks for any entity or prompt-guided target in the input video by using prompts as queries (striped triangles, hexagons and circles).

to obtain the text embeddings. We then introduce a single cross-attention layer to achieve language-image embeddings interaction, where the query is text embeddings, the keys and values are flattened multi-scale image embeddings $F \in R^{C\times(\sum_{s=1}^{3} H_s W_s)}$. This process can be formulated as:

$$P_i^* = \text{Lang2Img-CA}(\text{CLIPTextEnc}(L_i^*)\cdot W^{t2v},\ F), \quad (2)$$

where the shape of text embeddings P_i^* is $l^* \times C$, and l^* is the length of string tokens. The matrix $W^{t2v} \in R^{C^t\times C}$ maps the text embddings of dimension C^t to the visual space of dimension C. Note that we freeze the weights of CLIP text encoder to take advantage of the strong open vocabulary capabilities of CLIP.

3.2. Unified Video Mask Decoder

The unified video mask decoder aims to decode the masks for prompt-specified targets, which can be described as:

$$M_i^t = \text{MaskDec}(\{F_s^t\}_{s=1}^4,\ P_i^*,\ y_i^*), \forall i \in [1, N_p], t \in [1, V]$$

where P_i^* and y_i^* are prompt features and associated segment IDs for the i-th target, and M_i^t is its predicted masks in the t-th frame. N_p and V are the total numbers of provided targets and frames in the video, respectively. To achieve our goal, we adapt the mask decoder of Mask2Former [6, 8], which was initially designed for generic segmentation tasks with a set of learnable queries, by introducing a side stream that takes the mean of prompt features as input queries. As illustrated in the right yellow area of Fig. 3, our proposed unified video mask decoder comprises four key components: target-wise prompt cross-attention layer, an image cross-attention layer, a separated self-attention layer and a feed-forward network (FFN, which is omitted in Fig. 3).

Initial prompt query (P→Q). The visual/text prompt embeddings of the i-th target $P_i^* \in R^{l^*\times C}$ consist of l^* prompt tokens, which are point features from the visual prompt or string tokens of category names and expressions. We compute the average of all prompt tokens associated with it as the initial query for the target: $q_i^* = \sum_{l=1}^{l^*} P_{i,l}^*$. If the input contains a video clip with T frames, the initial query will be repeated T times to generate a clip-level initial query $q_i^* \in R^{T\times C}$. Using the mean of prompt features as the initial query provides an informative and stable starting point for the unified video mask decoder.

Prompt cross-attention (ProCA). The initial query may not be sufficient to provide a distinct representation for targets, particularly for those with similar characteristics, such as 'person' and 'black T-shirt' in Fig. 3. To enhance the uniqueness in representation, we introduce an entity-wise prompt cross-attention layer to learn prompt information to better differentiate between targets:

$$\text{ProCA}(q_i^*, P_i^*) = \text{Softmax}(\frac{q_i^* W^Q\ (P_i^* W^K)^T}{\sqrt{d_k}})\ P_i^* W^V, \quad (3)$$

where the query is q_i^*, and the keys and values are the prompt tokens P_i^*. W^Q, W^K and W^V represent the projection weights. The ProCA layer is placed in the front of image cross-attention layer to avoid forgetting prompt information as the decoder layer goes deep.

Image cross-attention and separated self-attention. The ProCA layer facilitates the incorporation of prompt information, while the image cross-attention layer focuses on extracting entity details from the input frames. We only compute the image cross-attention between each frame's query and the corresponding image features to reduce the memory overhead. Furthermore, the separated self-attention (Sep-SA) layer serves for two purposes. On one hand, it isolates the interactions between learnable queries and prompt queries, minimizing unnecessary negative impacts. On the other hand, by flattening learnable/prompt queries in the time dimension, it facilitates content interactions of the target of interest across spatial and time domains. The Sep-SA layer can be formulated as:

$$\mathrm{SepSA}(\mathbf{q}, \mathbf{q}^*) = \mathrm{SA}(\mathbf{q}, \mathbf{q}) \ \& \ \mathrm{SA}(\mathbf{q}^*, \mathbf{q}^*), \tag{4}$$

where $\mathbf{q} \in R^{NT \times C}$ and $\mathbf{q}^* \in R^{N_p T \times C}$ represent flattened learnable and flattened prompt queries, respectively, and N and N_p are the numbers of them.

Overall architecture. In addition to the ProCA, image cross-attention and SepSA layers, the FFN further allows the mask decoder to learn non-linear relationships from data. These four key components constitute a transformer layer, and our unified video mask decoder is composed of nine such transformer layers. In addition, there are two mask decoding streams, which share the same set of weights, to decode learnable queries and prompt queries, respectively. Note that the ProCA layer is omitted for learnable queries to simplify the representation.

To obtain the predicted masks for the t-th frame, we linearly combine mask coefficients with the finest-scale feature map $F_4^t \in R^{C \times H/4 \times W/4}$, *i.e.*, $[M^t, M^{*t}] = f_{\text{mask}}([\mathbf{q}^t, \mathbf{q}^{*t}]) \cdot F_4^t$, where mask coefficients are generated by passing the output queries through a multi-layer perception, denoted as f_{mask}.

4. Training and Inference

4.1. Training Stages

Training Losses consist of three terms: pixel-wise mask supervision loss, classification loss, and ReID loss. The training process of UniVS consists of three stages: image-level training, video-level training and long video fine-tuning. In the first stage, UniVS is trained on multiple image segmentation datasets, pretraining the model with image-level annotations for a good visual representation. In the second stage, we feed a short clip of three frames to the pretrained model, and fine-tune it on video segmentation datasets to perceive entity changes over a short period of time. In the third stage, we employ long video sequences of more than five frames to further fine-tune the unified video mask decoder, encouraging it to learn more discriminative features and trajectory information over a longer time period. To optimize memory usage, we freeze the backbone weights in the last two stages and further freeze the pixel decoder in the final stage. In each iteration, all samples within a batch come from the same dataset. We found that this sampling

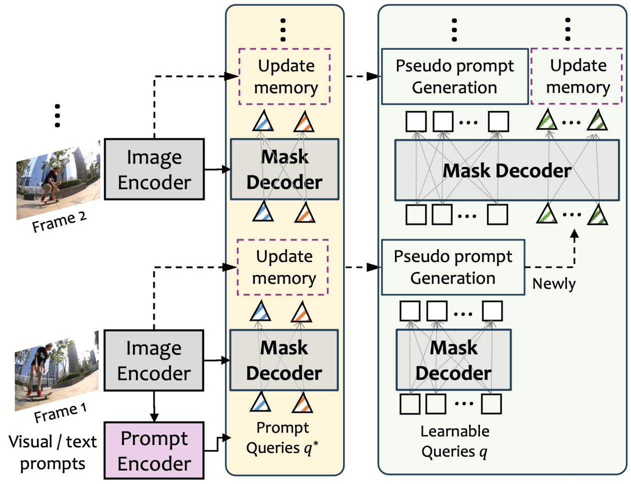

Figure 4. Inference process of our UniVS on prompt-specified and category-specified video segmentation tasks, respectively.

strategy can make the training more stable compared with the mixed sampling from different datasets.

More detailed information about the training losses and stages can be found in the **Supplementary Material**.

4.2. Unified Streaming Inference Process

In UniVS, the model input can be a single frame or a clip of multiple frames. In this subsection, we take a single frame as input to elucidate the unified inference process for generic category-specified and prompt-specified VS tasks.

For **prompt-specified VS tasks**, UniVS takes video frames and visual/text prompts as input, and the inference process is illustrated in the yellow boxes of Fig. 4. UniVS can process multiple targets simultaneously. First, the image encoder transforms the first frame into multi-scale image embeddings. Consequently, the prompt encoder converts visual/text prompts of the target into prompt tokens. In our design, each target has its dedicated memory pool to store associated prompt tokens, and its prompt query is obtained by averaging the tokens in the memory pool. These queries are used by the mask decoder to predict the masks of targets in the current frame, which are then used as visual prompts of targets and fed back to the prompt encoder, thereby updating the target's memory pool with new prompt information. In short, UniVS utilizes the prompt information of target objects stored in the memory pool to identify and segment the targets in subsequent frames, eliminating the cumbersome post-matching step in other unified models like SEEM [98] and UNINEXT [80], where targets need to be filtered out from all predicted entities.

For **category-specified VS tasks**, UniVS adopts a periodic object detection strategy and transforms the segmentation into a prompt-guided target segmentation problem. The detailed process is depicted in the light green box in

Fig. 4. First, UniVS employs learnable queries to identify all entity masks presented in the first frame, then employs non-maximum suppression (NMS) and classification thresholding to filter out redundant masks and those with low classification confidence. The remaining target objects also serve as their visual prompts, with which UniVS employs the prompt-guided target segmentation stream to directly predict their masks in the following frames, eliminating the need of cross-frame entity matching in previous methods. Furthermore, to identify newly appearing objects in subsequent frames, UniVS performs target detection for every a few frames using a learnable query and compares them to previously detected objects stored in a memory pool. We use the bi-Softmax approach [53] to distinguish between old and new objects in the video.

Remarks. Existing VS methods mostly assume smooth object motion within a short clip to associate entities across frames. However, for videos containing complex trajectories or large scene changes, this assumption does not hold, resulting in a decline of tracking accuracy. In contrast, our proposed UniVS overcomes this limitation by using prompts as queries to achieve explicit mask decoding.

5. Experiments

5.1. Experimental Settings

Datasets. The VIS datasets include YouTube-VIS 2019/2021 (YT19/21) [81] and OVIS [58]. The VSS and VPS datasets respectively include VSPW [52] and VIPSeg. The VOS datasets include DAVIS [57], YouTube-VOS 2018 (YT18) [78], MOSE [14] and BURST [2]. RefVOS datasets include RefDAVIS [63] and Ref-YouTube-VOS (RefYT) [63], while PVOS uses VIPOSeg [79] as dataset. The 2017 version of DAVIS and RefDAVIS are used [57]. Note that VSPW, VIPSeg and VIPOseg share the same original videos but have different protocols and guidelines for annotation. Due to the high cost of annotating datasets, the video scenes contained in each dataset lack diversity, being mostly sports scenes or animal scenes.

Implementation Details. We use Detectron2 [76] and follow Mask2Former [8] baseline settings. For data augmentation, we use the large-scale jittering (LSJ) augmentation [15] with a random scale sampled from range 0.5 to 4.0, followed by fixed-size cropping and padding to 1024×1024. We use the distributed training framework with 16 V100/A100 GPUs. Each mini-batch has 3 images or 1 video clip (2-7 frames) per GPU. ResNet50 [20] and SwinT/B/L [46] are adopted as the backbone networks, while we use the CLIP Text Encoder, whose corresponding visual encoder is ResNet50x4. We sample 32 points per visual prompt for visual prompt-guided targets. Considering the difference in the number of objects on various videos, we increase the number of learnable queries from 100 to 200 in the unified video mask decoder.

More detailed information about experimental settings can be found in the **Supplementary Material**.

5.2. Video Benchmark Results

We compare the results of recent high-performance VS models. The results of individual models trained on a single task/dataset are shown in Table 1, and the results of unified models trained individually or jointly on different tasks are shown in Table 2. To assess the generalization capability of competing models, we present the quantitative performance comparison on 10 benchmarks of six VS tasks, including VIS, VSS, VPS, VOS, RefVOS and PVOS. The detailed quantitative comparison on more image/video benchmarks can be found in the **Supplementary Material**.

Individual models are specifically designed based on the characteristics of each task, resulting in high performance on each benchmark, as shown in Table 1. For VIS task, GenVIS-off [21] with sequential learning achieves impressive performance of 51.3, 46.3 and 34.5 mAP on YT19, YT21 and OVIS, respectively. For the recently proposed VSS/VPS tasks, the baselines [26, 52] only obtain decent performance. For VOS task, XMem[11] and DEAOT [83] employ long-term information propagation modules, achieving state-of-the-art performance of 86.2 on DAVIS and 86.0 on YT18, respectively. For RefVOS task, SgMg [50] achieves the best performance of 62.0 on RefYT by using a spectrum-guided multi-granularity approach. For the newly proposed PVOS task, PAOT [79] extends the typical VOS method DEAOT, obtaining 75.4 on VIPOSeg.

Unified models trained individually on different tasks include VideoK-Net [40], Video-kMax [64], Tube-Link [39], DVIS [89] and TubeFormer [27], as listed in Table 2, which tend to handle both thing and stuff categories within a single framework. Video-kMax with clip k-means mask transformer achieves high performance on VSS task, reaching 44.3 mIoU on VSPW. By using the strong Mask2Former[8] as baseline, Tube-Link and DVIS inherit its ability to handle both thing and stuff classes, achieving remarkable performance on VIS and VPS tasks, as evidenced by the 33.8 mAP on OVIS and 43.2 VPQ on VIPSeg. It should be noted that the above unified models are trained individually on each dataset so that they can achieve excellent performance on each benchmark; however, this makes them lack the generalization ability to other datasets. Additionally, they cannot handle prompt-specified VS tasks, including VOS, RefVOS and PVOS.

Unified models trained jointly on different tasks aim to accommodate as many VS tasks as possible within a single model, where all tasks can be accomplished using the same set of trained weights. As shown in the top rows of Table 2, UniRef [73] unifies all prompt-specified VS tasks, but it fails to deal with category-specified VS tasks, such as

Video Tasks		VIS			VSS		VPS	
Method	Back-bone	YT19 mAP	YT21 mAP	OVIS mAP	VSPW mIoU	VSPW mVC_8	VIPSeg VPQ	VIPSeg STQ
IDOL[74]	R50	49.5	43.9	30.2	-	-	-	-
MinVIS[23]	R50	47.4	44.2	25.0	-	-	-	-
MDQE[34]	R50	*	44.5	29.2	-	-	-	-
GenVIS-off[21]	R50	**51.3**	**46.3**	**34.5**	-	-	-	-
TCB[52]	R101	-	-	-	**37.8**	**87.8**	-	-
ClipPanoFCN[51]	R50	-	-	-	-	-	**22.9**	**31.5**
IDOL[74]	SwinL	49.5	43.9	42.6	-	-	-	-
MinVIS[23]	SwinL	-	55.3	41.6	-	-	-	-
GenVIS-off[21]	SwinL	**63.8**	**60.1**	**45.4**	-	-	-	-
CFFM[65]	MiTB5	-	-	-	**49.3**	**90.8**	-	-

(a) Category-specified VS models

Video Tasks		VOS		RefVOS		PVOS
Method	Back-bone	DAVIS G^{th}	YT18 G^{th}	RefDAVIS J&F	RefYT J&F	VIPOSeg $G^{th\&sf}$
XMem[11]	R50	**86.2**	85.7	-	-	-
DeAOT[83]	R50	85.2	**86.0**	-	-	-
ReferFormer[72]	R50	-	-	58.5	55.6	-
OnlineRefer[70]	R50	-	-	59.3	57.3	-
SgMg[50]	SwinT	-	-	**61.9**	**62.0**	-
PAOT[79]	R50	-	-	-	-	**75.4**
DeAOT[83]	Swin-B	**86.2**	**86.2**	-	-	-
OnlineRefer[70]	Swin-L	-	-	**64.8**	63.0	-
SgMg[50]	SwinB	-	-	63.3	**65.7**	-
PAOT[79]	SwinB	-	-	-	-	**75.3**

(b) Prompt-specified VS models

Table 1. Quantitative performance comparison on individual VS models, which are specifically designed based on the characteristics of each task. '-' means the model lacks this capability. The best results are shown in **bold**.

Video Tasks				VIS			VSS		VPS		VOS		RefVOS		PVOS
Method	Backbone	Joint Training	Universal	YT19 mAP	YT21 mAP	OVIS mAP	VSPW mIoU	VSPW mVC_8	VIPSeg VPQ	VIPSeg STQ	DAVIS G^{th}	YT18 G^{th}	DAVIS J&F	RefYT J&F	VIPOSeg $G^{th\&sf}$
VideoK-Net[40]	ResNet50	✗	✗	40.5	*	*	*	26.1	33.1	-	-	-	-	-	
Video-kMax[64]	ResNet50	✗	✗	*	*	*	44.3	86.0	38.2	39.9	-	-	-	-	-
Tube-Link[39]	ResNet50	✗	✗	52.8	47.9	29.5	42.3	86.8	39.2	39.5	-	-	-	-	-
DVIS-off[89]	ResNet50	✗	✗	52.6	47.4	33.8	*	*	**43.2**	42.8	-	-	-	-	-
UniRef[73]	ResNet50	✓	✗	-	-	-	-	-	-	-	*	**81.4**	63.5	60.6	*
TarVIS[1]	ResNet50	✓	✗	*	**48.3**	31.1	*	*	33.5	43.1	**82.6**	*	-	-	-
UNINEXT[80]	ResNet50	✓	✗	**53.0**	*	**34.0**	-	-	-	-	74.5	77.0	**63.9**	**61.2**	-
UniVS	ResNet50	✓	✓	47.4	46.6	30.8	**48.2**	**88.5**	38.6	**45.8**	70.5	69.2	57.9	56.2	**60.2**
UniVS	SwinT	✓	✓	52.4	51.6	33.0	51.3	89.4	38.9	51.7	71.7	70.3	58.5	56.2	62.3
TubeFormer[27]	A-R50×64	✗	✗	47.5	41.2	*	63.2	92.1	*	*	-	-	-	-	-
VideoK-Net+[40]	SwinB	✗	✗	51.4	*	*	*	*	39.8	46.3	-	-	-	-	-
Video-kMax[64]	ConvNeXtL	✗	✗	*	*	*	**63.6**	91.8	51.9	51.7	-	-	-	-	-
Tube-Link[39]	SwinL	✗	✗	64.6	58.4	*	59.7	90.3	*	*	-	-	-	-	-
DVIS-off[89]	SwinL	✗	✗	**64.9**	60.1	**49.9**	*	*	**57.6**	55.3	-	-	-	-	-
UniRef[73]	SwinL	✓	✗	-	-	-	-	-	-	-	*	**82.6**	66.3	**67.4**	*
TarVIS[1]	SwinL	✓	✗	*	**60.2**	43.2	*	*	48.0	52.9	**85.2**	*	-	-	-
UNINEXT[80]	ConvNeXtL	✓	✗	64.3	*	41.1	-	-	-	-	77.2	78.1	**66.7**	66.2	-
UNINEXT[80]	ViT-H	✓	✗	66.9	*	49.0	-	-	-	-	81.8	78.6	72.5	70.1	-
UniVS	SwinB	✓	✓	57.8	56.5	39.0	59.4	90.4	46.7	56.1	75.0	70.9	58.6	57.4	68.2
UniVS	SwinL	✓	✓	60.0	57.9	41.7	59.8	**92.3**	49.3	**58.2**	76.2	71.5	59.4	58.0	**68.6**

Table 2. Overall quantitative performance comparison of unified VS models. '-' means the model lacks this capability and '*' means the result is not reported. The results of UniVS with SwinT backbone and UNINEXT with ViT-H backbone are also listed in gray color. However, due to use of different backbones, they are not considered in the performance comparison. The best results are shown in **bold**.

VIS, VSS and VPS. TarVIS [1] can handle most VS tasks except for RefVOS due to the lack of language encoding capabilities. UNINEXT [80] focuses solely on object-centric segmentation without considering entity segmentation of stuff categories, thereby being incapable of handling VSS, VPS and PVOS tasks. In contrast, our UniVS is the only approach that accommodates all VS tasks within a single model, demonstrating the highest generalization capability in universal segmentation.

Specifically, UniRef achieves outstanding performance on VOS and RefVOS tasks, reaching 81.4 on YT18. TarVIS achieves state-of-the-art performance on datasets with simple scenes, such as 48.3 mAP on YT21 of VIS task and 82.6 on DAVIS of VOS task. UNINEXT utilizes a large amount of image/video data to jointly train the model, allowing it to achieve state-of-the-art performance in instance-level segmentation benchmarks, such as 53.0/34.0 mAP on YT19/OVIS for VIS task and 63.9/61.2 on DAVIS and RefYT for RefVOS task. Our UniVS achieves comparable performance on VIS task, top-ranked performance on VSS/VPS tasks and slightly lower performance in prompt-specified VS tasks. For VSS and VPS tasks, UniVS brings 2∼4% performance improvement, reaching 48.2 mIoU and 88.5 mVC_8 on VSPW and 45.8 STQ on VIPSeg.

Compared with UniRef, TarVIS and UNINEXT, our UniVS can handle category-specified VS tasks, text prompt-guided segmentation and entity segmentation for stuff categories simultaneously. Pursuing a more universal segmentation capability, however, may sacrifice performance in individual tasks due to task conflicts and limited video data. Compared to UniRef, which utilizes heavy temporal propagation modules to pass reference frames and

Video Tasks		VIS		VPS		VOS	RefVOS
P→Q	ProCA	YT21	OVIS	VIPSeg		YT18	RefYT
		mAP	mAP	VPQ	STQ	G^{th}	J&F
✗	✗	45.9	17.2	40.1	37.9	-	-
✓	✗	39.9	10.0	40.6	37.3	58.6	42.1
✓	✓	52.7	21.7	35.4	49.2	67.4	54.9

(a) Prompt as queries (P→Q) and prompt cross-attention (ProCA)

Video Tasks		VIS		VPS		VOS	RefVOS
Stage 3	Tracker	YT21	OVIS	VIPSeg		YT18	RefYT
		mAP	mAP	VPQ	STQ	G^{th}	J&F
✗	Similarity	50.7	15.0	35.9	44.5	60.8	-
✗	VS→P	52.7	21.7	35.4	49.2	67.4	54.9
✓	VS→P	54.7	23.6	38.6	45.8	69.2	56.2

(b) Unified training and inference, where VS→P refers to the transformation from all VS tasks to prompt-guided target segmentation.

Table 3. Ablation studies on (a) prompt as queries and (b) unified training and inference on VS tasks. '-' means that the model lacks this capability. For VIS task, the results are evaluated in the development set (1/10 of the training set, excluded during training).

masks to the current frame, UniVS adopts a parameter-free Visual Sampler to extract prompt information, resulting in a slight decrease in performance. Compared to UNINEXT, which uses 900 object queries, we train UniVS using only 200 object queries with fewer video data. Additionally, UniVS needs to accommodate stuff categories in object segmentation. These two factors explain the slightly lower performance of our approach in instance-level segmentation benchmarks compared to UNINEXT. In future work, more diverse training video data and long-term information propagation modules will be explored to improve the performance of our unified segmentation models.

Our results with stronger backbones further validate the aforementioned conclusion, as shown in the bottom rows of Tables 1 and 2. Our UniVS is the only model that is capable of handling all the six VS tasks within a single framework, setting new state-of-the-art performance on VSS and VPS tasks, and achieving 92.3 mVC_8 on VSPW and 58.2 STQ on VIPSeg. Overall, our UniVS can obtain an appropriate balance between performance and universality capability.

5.3. Ablation Studies

We analyze UniVS through a series of ablation studies using the ResNet50 backbone [20]. To test the generality of the proposed components for universal video segmentation, ablation studies are performed on various VS tasks.

Prompts as Queries and ProCA. We validate the importance of each component by adding them one at a time. As shown in Table 3a, by introducing the prompt as queries (P→Q), the baseline can handle prompt-specific VS tasks but achieve relatively lower performance on all VS tasks. Moreover, by introducing the prompt cross-attention (ProCA) to further extract comprehensive prompt features, the performance on most VS tasks is significantly improved, especially on OVIS for VIS task and YT18 for VOS task, whose performance is increased by $\sim$10%.

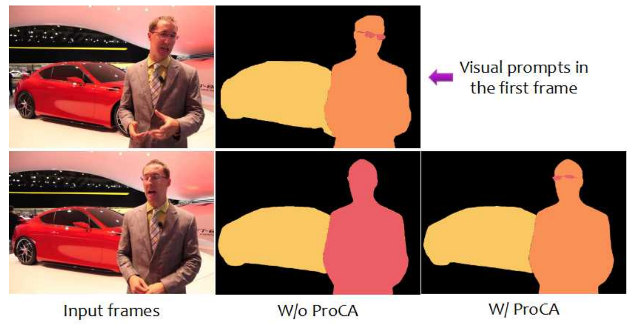

Figure 5. Qualitative results of UniVS w/o and w/ ProCA on VOS task.

The predicted masks of UniVS without and with the ProCA layer are shown in Fig. 5. It is evident that, for the uncommon object 'glasses', the ProCA layer could provide precise prompt features to prevent the mask from spilling over to other objects like 'person'.

Unified Training and Inference. In Table 3b, we study the performance when using different ways to associate entities across video clips. Using query embedding similarity as a tracker fails to differentiate entity trajectories in complex scenarios, leading to poor performance. By adopting our unified streaming inference process in Sec. 4.2, which transfers all VS tasks to the prompt-guided target segmentation, the performance of all VS tasks improves significantly, particularly in complex scenes (with an increase of $\sim$6%). Additionally, introducing the third training stage improves the model's ability to capture medium-term object motion trajectories, bringing $\sim$2% performance improvement.

6. Conclusion

In this paper, we attempted to accommodate all video segmentation tasks within a single model, and proposed a novel unified architecture, namely UniVS, by using prompts as queries. We averaged the prompt features stored in the memory pool as the initial query of the prompt-guided target, and introduced a target-wise prompt cross-attention layer to integrate comprehensive prompt features, and converted different VS tasks into prompt-guided target segmentation during inference. Extensive experimental results demonstrated that UniVS achieved competitive or even better performance on category-specified VS tasks, while achieved slightly lower performance on prompt-specified VS tasks. Overall, by using a single model with the same set of trained model parameters, UniVS resulted in a commendable balance between performance and universality. Its potentials can be further released with more training video-language data.

References

[1] Ali Athar, Alexander Hermans, Jonathon Luiten, Deva Ramanan, and Bastian Leibe. Tarvis: A unified approach for target-based video segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18738–18748, 2023. 2, 3, 7

[2] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1674–1683, 2023. 2, 3, 6

[3] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Int. Conf. Comput. Vis.*, pages 9157–9166, 2019. 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, pages 213–229. Springer, 2020. 3

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2017. 2

[6] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 3, 4

[7] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12475–12485, 2020. 2

[8] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 2, 3, 4, 6

[9] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. *arXiv preprint arXiv:2310.12982*, 2023. 3

[10] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *Int. Conf. Comput. Vis.*, 2023. 3

[11] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *Eur. Conf. Comput. Vis.*, pages 640–658. Springer, 2022. 2, 3, 6, 7

[12] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3213–3223, 2016. 2

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[14] 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. *arXiv preprint arXiv:2302.01872*, 2023. 2, 3, 6

[15] Xianzhi Du, Barret Zoph, Wei-Chih Hung, and Tsung-Yi Lin. Simple training strategies and model scaling for object detection. *arXiv preprint arXiv:2107.00057*, 2021. 6

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.*, 88:303–338, 2010. 2

[17] Pinxue Guo, Tony Huang, Peiyang He, Xuefeng Liu, Tianjun Xiao, Zhaoyu Chen, and Wenqiang Zhang. Openvis: Open-vocabulary video instance segmentation. *arXiv preprint arXiv:2305.16835*, 2023. 2, 3

[18] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. 2023. 1

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Int. Conf. Comput. Vis.*, pages 2961–2969, 2017. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 6, 8

[21] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 3, 6, 7

[22] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. 2022. 3

[23] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. *Adv. Neural Inform. Process. Syst.*, 2022. 2, 3, 7

[24] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3D-language pre-training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10984–10994, 2023. 2

[25] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 2

[26] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9859–9868, 2020. 1, 6

[27] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13914–13924, 2022. 3, 6, 7

[28] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 9404–9413, 2019. 2

[29] Feng Li, Hao Zhang, Shilong Liu, Lei Zhang, Lionel M Ni, Heung-Yeung Shum, et al. Mask dino: Towards a unified

transformer-based framework for object detection and segmentation. *arXiv preprint arXiv:2206.02777*, 2022. 2

[30] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023. 2

[31] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. pages 12888–12900. PMLR, 2022. 2

[32] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2

[33] Minghan Li, Shuai Li, Lida Li, and Lei Zhang. Spatial feature calibration and temporal fusion for effective one-stage video instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11215–11224, 2021. 2, 3

[34] Minghan Li, Shuai Li, Wangmeng Xiang, and Lei Zhang. Mdqe: Mining discriminative query embeddings to segment occluded instances on challenging videos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10524–10533, 2023. 3, 7

[35] Minghan Li, Qi Xie, Qian Zhao, Wei Wei, Shuhang Gu, Jing Tao, and Deyu Meng. Video rain streak removal by multiscale convolutional sparse coding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6644–6653, 2018. 1

[36] Shuai Li, Minghan Li, Ruihuang Li, Chenhang He, and Lei Zhang. One-to-few label assignment for end-to-end dense detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7350–7359, 2023. 3

[37] Shuai Li, Minghan Li, Pengfei Wang, and Lei Zhang. Opensd: Unified open-vocabulary segmentation and detection. *arXiv preprint arXiv:2312.06703*, 2023. 3

[38] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Yan Lu, and Bhiksha Raj. Rˆ 2vos: Robust referring video object segmentation via relational multimodal cycle consistency. *arXiv preprint arXiv:2207.01203*, 2022. 3

[39] Xiangtai Li, Haobo Yuan, Wenwei Zhang, Guangliang Cheng, Jiangmiao Pang, and Chen Change Loy. Tube-link: A flexible cross tube baseline for universal video segmentation. *arXiv preprint arXiv:2303.12782*, 2023. 2, 3, 6, 7

[40] Xiangtai Li, Wenwei Zhang, Jiangmiao Pang, Kai Chen, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. Video k-net: A simple, strong, and unified baseline for video segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18847–18857, 2022. 2, 3, 6, 7

[41] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 214–223, 2021. 3

[42] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 2

[43] 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 *Eur. Conf. Comput. Vis.*, pages 740–755. Springer, 2014. 2

[44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2

[45] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with ringattention. *arXiv preprint*, 2024. 2

[46] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Int. Conf. Comput. Vis.*, pages 10012–10022, October 2021. 6

[47] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3431–3440, 2015. 2

[48] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10437–10446, 2020. 2

[49] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 616–625, 2018. 2

[50] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *Int. Conf. Comput. Vis.*, pages 920–930, October 2023. 3, 6, 7

[51] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21033–21043, 2022. 1, 3, 7

[52] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4133–4143, 2021. 1, 3, 6, 7

[53] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 164–173, 2021. 6

[54] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 1

[55] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2594–2605, 2022. 2

[56] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 2

[57] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 3, 6

[58] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip Torr, and Song Bai. Occluded video instance segmentation: Dataset and challenge. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*,

2021. 1, 6

[59] Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, et al. Freeseg: Unified, universal and open-vocabulary image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 19446–19455, 2023. 2

[60] 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. pages 8748–8763. PMLR, 2021. 2, 3

[61] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

[62] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Comput. and Computer-Assisted Intervent.*, pages 234–241. Springer, 2015. 2

[63] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Eur. Conf. Comput. Vis.*, pages 208–223. Springer, 2020. 2, 3, 6

[64] Inkyu Shin, Dahun Kim, Qihang Yu, Jun Xie, Hong-Seok Kim, Bradley Green, In So Kweon, Kuk-Jin Yoon, and Liang-Chieh Chen. Video-kmax: A simple unified approach for online and near-online video panoptic segmentation. *arXiv preprint arXiv:2304.04694*, 2023. 2, 3, 6, 7

[65] Guolei Sun, Yun Liu, Henghui Ding, Thomas Probst, and Luc Van Gool. Coarse-to-fine feature mining for video semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3126–3137, June 2022. 7

[66] Lingchen Sun, Rongyuan Wu, Zhengqiang Zhang, Hongwei Yong, and Lei Zhang. Improving the stability of diffusion models for content consistent super-resolution. *arXiv preprint arXiv:2401.00877*, 2023. 1

[67] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2, 3

[68] Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, and Luca Bertinetto. Do different tracking tasks require different appearance models? *Thirty-Fifth Conference on Neural Infromation Processing Systems*, 2021. 2

[69] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Int. Conf. Comput. Vis.*, pages 15943–15953, 2023. 1

[70] Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. Onlinerefer: A simple online baseline for referring video object segmentation. In *Int. Conf. Comput. Vis.*, pages 2761–2770, 2023. 3, 7

[71] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *Eur. Conf. Comput. Vis.*, 2022. 3

[72] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4974–4984, 2022. 3, 7

[73] Jiannan Wu, Yi Jiang, Bin Yan, Huchuan Lu, Zehuan Yuan, and Ping Luo. Segment every reference object in spatial and temporal spaces. In *Int. Conf. Comput. Vis.*, pages 2538–2550, October 2023. 3, 6, 7

[74] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *Eur. Conf. Comput. Vis.*, 2022. 2, 3, 7

[75] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 1

[76] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 6

[77] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 373–381, 2016. 2

[78] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Eur. Conf. Comput. Vis.*, pages 585–601, 2018. 2, 3, 6

[79] Yuanyou Xu, Zongxin Yang, and Yi Yang. Video object segmentation in panoptic wild scenes. *arXiv preprint arXiv:2305.04470*, 2023. 2, 3, 6, 7

[80] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15325–15336, 2023. 2, 3, 5, 7

[81] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Int. Conf. Comput. Vis.*, pages 5188–5197, 2019. 1, 3, 6

[82] Tien-Ju Yang, Maxwell D Collins, Yukun Zhu, Jyh-Jing Hwang, Ting Liu, Xiao Zhang, Vivienne Sze, George Papandreou, and Liang-Chieh Chen. Deeperlab: Single-shot image parser. *arXiv preprint arXiv:1902.05093*, 2019. 2

[83] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. In *Adv. Neural Inform. Process. Syst.*, 2022. 3, 6, 7

[84] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. MAGVIT: Masked generative video transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2023. 2

[85] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 2

[86] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianfeng Gao, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. *arXiv preprint arXiv:2303.08131*, 2023. 3

[87] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4158–4166, 2018. 2, 3

[88] Jiaming Zhang, Yutao Cui, Gangshan Wu, and Limin Wang.

Joint modeling of feature, correspondence, and a compressed memory for video object segmentation. *arXiv preprint arXiv:2308.13505*, 2023. 3

[89] Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. Dvis: Decoupled video instance segmentation framework. *arXiv preprint arXiv:2306.03413*, 2023. 2, 3, 6, 7

[90] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021. 2

[91] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 1

[92] Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 2

[93] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16793–16803, 2022. 2

[94] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 633–641, 2017. 2

[95] Minghao Zhou, Hong Wang, Qian Zhao, Yuexiang Li, Yawen Huang, Deyu Meng, and Yefeng Zheng. Interactive segmentation as gaussion process classification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 19488–19497, 2023. 2

[96] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2, 3

[97] Xueyan Zou*, Zi-Yi Dou*, Jianwei Yang*, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee*, and Jianfeng Gao*. Generalized decoding for pixel, image and language. 2022. 2

[98] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023. 2, 3, 5