Decoupled Motion Expression Video Segmentation

Hao Fang[1,2] Runmin Cong[1,2*] Xiankai Lu[3] Xiaofei Zhou[4]
Sam Kwong[5] Wei Zhang[1,2]
[1]School of Control Science and Engineering, Shandong University
[2]Key Laboratory of Machine Intelligence and System Control, Ministry of Education
[3]School of Software, Shandong University
[4]School of Automation, Hangzhou Dianzi University
[5]School of Data Science, Lingnan University

Abstract

Motion expression video segmentation aims to segment objects based on input motion descriptions. Compared with traditional referring video object segmentation, it focuses on motion and multi-object expressions and is more challenging. Previous works achieved it by simply injecting text information into the video instance segmentation (VIS) model. However, this requires retraining the entire model and optimization is difficult. In this work, we propose DMVS, a simple framework constructed on the existing query-based VIS model, emphasizing decoupling the task into video instance segmentation and motion expression understanding. Firstly, we use a frozen video instance segmenter to extract object-specific contexts and convert them into frame-level and video-level queries. Secondly, we interact two levels of queries with static and motion cues, respectively, to further encode visually enhanced motion expressions. Furthermore, we propose a novel query initialization strategy that uses video queries guided by classification priors to initialize motion queries, greatly reducing the difficulty of optimization. Without bells and whistles, DMVS achieves state-of-the-art performance on the MeViS dataset at a lower training cost. Extensive experiments verify the effectiveness and efficiency of our framework.

1. Introduction

Referring Video Object Segmentation (RVOS) is a multi-modal video task, which aims to segment the target object that is specified by a provided language description across the entire video. Existing RVOS datasets [15, 21, 38] typically consist of videos where the objects are salient and possess distinct and unchanging features. Therefore, referring image segmentation methods [1, 9, 21, 25] can achieve good results on these datasets. To alleviate this limitation, Motion Expression Video Segmentation (MEVS) is recently proposed [10], which highlights the importance of the temporal motion characteristics of videos. The new dataset for the task, MeViS [10], contains a large number of motion expressions to indicate objects in intricate environments, and one expression may refer to multiple target objects. The current methods [2, 11, 38, 44] designed for traditional RVOS datasets encounter significant performance degradation.

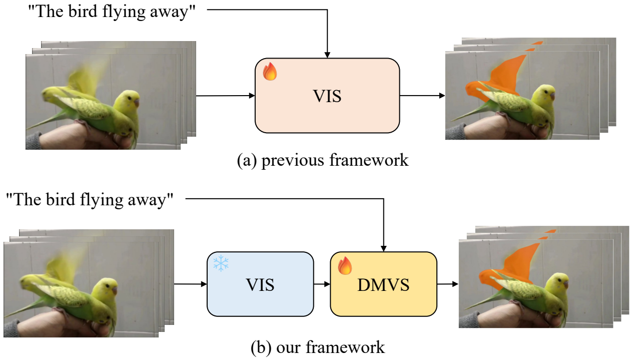


Figure 1. Comparisons of previous framework and our DMVS framework. (a) Previous methods inject textual information into the video instance segmenter, which required retraining the overall framework. (b) Our DMVS decouples the task into video instance segmentation and motion expression understanding, video instance segmenter freezes during the training process.

Due to the multi-object and inter-frame motion correlation features of the dataset, researchers attempted to inject textual information into the video instance segmentation (VIS) model to address this new task, as shown in Fig. 1 (a). The VIS task aims to simultaneously detect, segment, and track all object instances in videos [48]. A major challenge in MEVS is the precise capture and alignment of cross-

[*] Corresponding author

temporal motion, therefore offline VIS methods [7, 19, 41] are mainly favored. LMPM [10] use language queries instead of conventional object queries and selects matching object trajectories through threshold, simply converting the VIS model VITA [19] to the MEVS model. Also based on VITA, DsHmp [18] employs Mask2Former [8] to segment possible objects according to static cues, and then designs modules to use motion cues to segment specified objects. However, this framework has several intrinsic issues. Mask2Fomer is expected to use static cues to segment as many candidates as possible, but we do not have the ground truth for all objects specified by static cues. For example, DsHmp trains Mask2Former to first find all the birds, but in reality, only the bird that flies away is supervised. This is inconsistent with the author's motivation to decouple referring video segmentation into static perception and motion perception. In addition, training the entire framework from scratch also leads to higher computational costs. As a result, the question arises: How to truly decouple motion expression video segmentation?

In this work, we introduce Decoupled Motion Expression Video Segmentation (DMVS), a simple framework constructed on the existing query-based VIS model, emphasizing decoupling the task into video instance segmentation and motion expression understanding. As shown in Fig. 1 (b), it first uses frozen video instance segmenter to track and segment all candidates, and then recognizes the specified object based on the motion expression, requiring only training the DMVS module with fewer parameters. Firstly, we fully utilize object representation output by the VIS model to design our DMVS module. As with previous works [10, 18], we choose VITA [19] as our video instance segmenter, which accomplishes video-level understanding by associating frame-level object queries. Through VITA, we obtain frame queries that independently represent the object information of each frame and video queries that uniformly represent the entire video. Therefore, we conduct motion expression understanding by associating video queries and frame queries without using spatio-temporal backbone features.

To further encode visually enhanced text information, we propose motion expression encoder module. Inspired by DsHmp [18], we decouple motion expressions to static cues and motion cues. The object categories of frame queries mainly come from the category set defined by the VIS dataset. Frame queries interact with static cues to focus on potential candidates of a specified category. Video queries uniformly represent the target object of the entire video clip, including the overall contextual information of the video scene. Video queries interact with motion cues to focus on the target of a specific motion in a category independent manner. The fusion of sentence embedding with frame queries for static perception and video queries for motion perception completes the final encoding of text information.

Due to the lack of additional information, the object queries are mainly initialized randomly in the VIS model. In RVOS, ReferFormer [44] first regards the language information as a set of queries, so subsequent RVOS methods also follow this paradigm. However, regardless of the method, there is a certain degree of optimization difficulty because there is a significant gap from the final decoded output queries. In an ideal scenario, RVOS essentially selects VIS output based on textual information. Naturally, we propose using video queries to initialize motion queries. We select object queries based on the classification score Top K to eliminate the interference of background queries. Finally, motion queries are interactively decoded with enhanced text features to generate masks and classification results.

The contributions can be summarized as follows:

- We propose DMVS, a simple yet effective framework constructed on the existing query-based VIS model, emphasizing decoupling motion expression video segmentation into video instance segmentation and motion expression understanding.
- We suggest interacting two levels of queries with static and motion cues, respectively, to further encode visually enhanced motion expressions. We propose a novel query initialization strategy that uses video queries guided by classification priors to initialize motion queries, greatly reducing the difficulty of optimization.
- We conduct experiments on the recently released MeViS dataset, and our methods demonstrate significant superiority over the existing methods. Meanwhile, extensive ablation validates the methods effectiveness and efficiency.

2. Related Work

Video Instance Segmentation. The VIS task is designed to detect, segment, and track all object instances within videos concurrently [48]. Generally speaking, the existing VIS techniques can be categorized into two main types: online and offline methods. Online methods [13, 20, 24, 45, 48, 50] handle video instance segmentation frame by frame. They then implement post-processing procedures to explicitly establish the association of instances across different frames. As an illustration, MinVIS [20] and IDOL [45] utilize distinctive instance queries to achieve matching between frames. On the contrary, offline methods [7, 14, 19, 41, 43, 53, 54] accept a video clip as the input and produce a sequence of instances in an end-to-end manner. For example, Seqformer [43] and VITA [19] first localize the instances in every single frame and subsequently learn a highly effective representation of video-level instance queries.

Referring Video Object Segmentation. The RVOS task aims to segment the target object that is specified by a provided language description across the entire video [15]. The current methods can be divided into two types: multi-

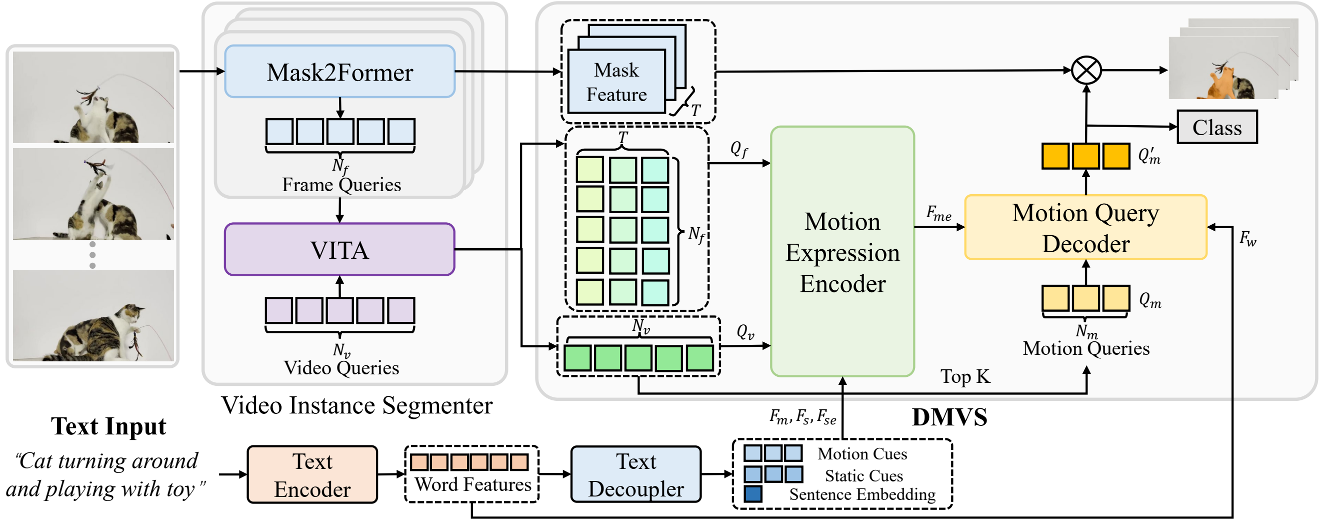


Figure 2. The overall framework of DMVS. We decouple motion expression video segmentation into video instance segmentation and motion expression understanding. We first use frozen VITA as a video instance segmenter to identify, segment, and track all objects. Then based on the frame queries Q_f and video queries Q_v generated by VITA, a Motion Expression Encoder is employed to interact with motion cues F_m and static cues F_s to generate visual information enhanced motion expression features F_{me}. Finally, we use video queries guided by classification priors to initialize motion queries Q_m, and then use Motion Query Decoder to decode layer by layer for target object recognition and mask predictions.

stage and one-stage methods. The multi-stage methods [1, 21, 23, 38] handle each frame of the video clip independently using an image-level model [4–6]. In URVOS [38], the process starts with an initial mask prediction carried out by an image-level model. Subsequently, the mask is propagated across frames via semi-supervised VOS methods [12, 16, 30, 35]. Inspired by DETR [3], one-stage methods [2, 31, 44, 47] of Transformer [40] structure have been extensively proposed. Compared to rely on complicated pipelines, MTTR[2] and Referformer[44] initially formulate the task as a sequence prediction problem by an end-to-end framework, which significantly streamlines the overall pipeline. SOC[31] and MUTR[47] attain remarkable performance by effectively aggregating both the information within a single frame and between different frames.

Motion Expressions Video Segmentation. Compared to traditional RVOS task, this task focuses on segmenting objects according to motion description of the objects in video. Most existing RVOS benchmarks, including Ref-YouTube-VOS [38] and Ref-DAVIS 2017 [21], primarily focus on single salient objects and static attributes. MeViS [10] is a new large-scale video dataset, ontains a large number of motion expressions to indicate objects in intricate environments, and one expression may refer to multiple target objects. The current RVOS methods have encountered significant difficulties, and researchers are trying to draw on the VIS model to address this new task. LMPM [10] replaces randomly initialized queries with language-conditional queries and selects matching object trajectories through threshold, simply converting the VIS model VITA [19] to RVOS model. Based on LMPM [10], DsHmp [18] decouples static and hierarchical motion perception and employs contrastive learning to differentiate the motions of objects that appear visually alike.

3. Method

In this section, we first give a brief overview of VITA [19], a video instance segmenter for DMVS. Then, we introduce the architecture of our proposed DMVS, which is built on top of VITA, as shown in Fig. 2.

3.1. Video Instance Segmenter

In this paper, we adopt VITA [19] for the video instance segmenter, which is an offline VIS method that is constructed upon an existing Transformer-based image instance segmentation model. VITA achieves a comprehensive understanding at the video level by establishing associations among the object tokens at the frame level. Given an input video of T frames of resolution $H \times W$, VITA first uses Mask2Former [8] to process each frame in a complete frame-independent manner, generating 1) frame queries $Q_f \in \mathbb{R}^{T \times N_f \times C}$ which hold object-centric information, where N_f is the number of object queries and C is the number of channels; and 2) mask features $F_{mask} \in \mathbb{R}^{T \times \frac{H}{S} \times \frac{W}{S} \times C}$ from the pixel decoder, where S is the stride of the feature map. Then, Object Encoder builds temporal communication on frame queries by employing self-attention along the temporal axis. Finally, Object Decoder aggregates information from frame queries Q_f to video queries $Q_v \in \mathbb{R}^{N_v \times C}$, which are eventually used for predicting categories and masks of objects in videos at once.

3.2. DMVS

We now propose the novel motion expression video segmentation method DMVS, which can be largely divided into three phases (Fig. 2). We first use frozen VITA as a video instance segmenter to identify, segment, and track all objects. Then based on the frame queries Q_f and video queries Q_v generated by VITA, a Motion Expression Encoder is employed to interact with motion cues F_m and static cues F_s to generate visual information enhanced motion expression features F_{me}. We use video queries guided by classification priors to initialize motion queries Q_m, and then use Motion Query Decoder to decode motion queries layer by layer. Finally, the predicted masks are obtained by multiplying motion queries Q_m and mask features F_{mask}, and these with class scores higher than a threshold are selected as output.

Motion Expression Encoder. Extensive referring works [26, 46] have confirmed that the interaction between text and visual features before query decoding can help enhance the information of each mode. We believe that frame queries and video queries can provide sufficient object-specific visual information without interacting with dense backbone features, as RVOS methods did before. Therefore, the challenge now is how to integrate object queries of the two levels with the text information.

Frame queries independently represent all the objects of each frame, and static information is required to determine the object category. Video queries integrate global context information, which is more suitable for interaction with motion information across time domain. Inspired by DsHmp [18], we introduce a decoupling of the given expression into static and motion words, serve as cues for static perception of frame queries and motion perception of video queries.

As shown in Fig. 2, given the sentence Cat turning around and playing with toy, we first use text encoder [27] to extract word features $F_w \in \mathbb{R}^{K_w \times C}$, where K_w represents the longest word number of sentences in the dataset. We employ text decoupler [37] to detect adjectives, prepositions and nouns within the sentence, obtaining static cues such as cat, toy. At the same time, we draw out adverbs and verbs, obtaining motion cues like turning around, playing with. Consequently, we obtain static word features as $F_s \in \mathbb{R}^{K_s \times C}$ and motion word features as $F_m \in \mathbb{R}^{K_m \times C}$, where K_s and K_m denote the lengths of static words and motion words respectively. And we also acquire the sentence-level feature $F_{se} \in \mathbb{R}^{C}$ through the operation of pooling the features of every single word.

We first use decoupled static cues and motion cues to enhance the object queries respectively. Specifically, we employ cross-attention to inject static cues into frame queries:

$$Q'_f = Q_f + \mathrm{softmax}\left(\frac{Q_f F_s^T}{\sqrt{C}}\right) F_s, \tag{1}$$

where $Q'_f \in \mathbb{R}^{T \times N_f \times C}$ is frame queries enhanced by static cues. Similarly, we employ cross-attention to inject motion cues into video queries:

$$Q'_v = Q_v + \mathrm{softmax}\left(\frac{Q_v F_m^T}{\sqrt{C}}\right) F_m, \tag{2}$$

where $Q'_v \in \mathbb{R}^{N_v \times C}$ is video queries enhanced by motion cues. Dshmp [18] only allows static cues and motion cues to interact with object queries when they are initialized at the beginning. The randomly initialized queries do not contain any object cues and cannot extract valid text information. In essence, it relies on the combination of all text features in the decoding stage to establish modal associations. We use frame and video queries that already contain all object information to fuse with static and motion cues, respectively, fully leveraging the important role of decoupled text features.

Next, we use object queries containing specific cues to enhance sentence-level feature. Specifically, we use serial cross-attention to gradually encode sentence embedding:

$$F'_{se} = F_{se} + \mathrm{softmax}\left(\frac{F_{se} {Q'_f}^T}{\sqrt{C}}\right) Q'_f, \tag{3}$$

$$F_{me} = F'_{se} + \mathrm{softmax}\left(\frac{F'_{se} {Q'_v}^T}{\sqrt{C}}\right) Q'_v, \tag{4}$$

where $F'_{se} \in \mathbb{R}^{C}$ is a sentence embedding that integrates frame-level static object information. $F_{me} \in \mathbb{R}^{C}$ is a sentence embedding that integrates video-level motion object information, and it is also the final encoded motion expression used in the subsequent decoding process.

Motion Query Decoder. The premise of Motion Query Decoder is to initialize the motion query. Due to the lack of additional information, the object queries are mainly initialized randomly in the VIS model:

$$Q_m \sim \mathcal{N}(0, 1), \tag{5}$$

where $Q_m \in \mathbb{R}^{N_m \times C}$ is motion queries sampled from a normal distribution with a mean of 0 and a variance of 1, N_m is the number of motion queries.

In RVOS methods, ReferFormer [44] first views the language embedding as queries, so subsequent RVOS methods also follow this paradigm. For example, LMPM [10] repeates sentence embedding to initialize object queries:

$$Q_m = \mathrm{repeat}(F_{se}, N_m), \tag{6}$$

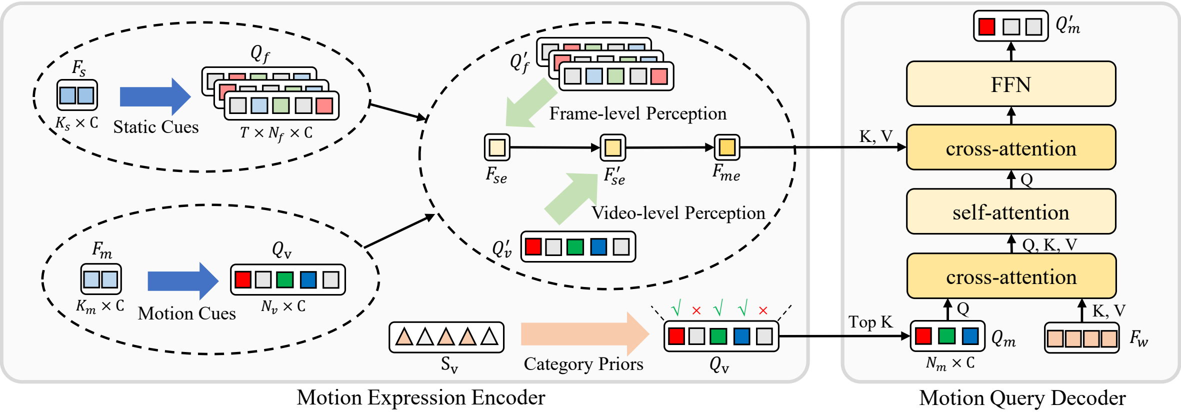


Figure 3. Architecture of the proposed Motion Expression Encoder and Motion Query Decoder. We decouple motion expressions to static cues and motion cues and interact them separately with frame queries and video queries. The fusion of sentence embedding with frame queries for static perception and video queries for motion perception completes the encoding of text information. Motion queries are obtained through video queries TopK and interact with encoded text information to obtain the final decoded query embeddings.

where $Q_m \in \mathbb{R}^{N_v \times C}$ is motion queries initialized by repeating sentence embedding.

However, regardless of the method, there is a certain degree of optimization difficulty because there is a significant gap from the final decoded output queries. In an ideal scenario, RVOS essentially selects VIS output based on textual information. Naturally, we propose using video queries to initialize motion queries. The default number of queries in VIS is relatively large, we select object queries based on the classification score Top K to eliminate the interference of background queries:

$$Q_m = \mathrm{TopK}(Q_v, S_v, N_m), \tag{7}$$

where $S_v \in \mathbb{R}^{N_v}$ is the classification score corresponding to each video query output by VIS model, TopK represents selecting N_m queries with the highest classification scores in Q_v. Through this initialization method, motion queries almost contain queries corresponding to all instances, and the decoding process changes from identifying and segmenting objects from scratch to matching the most suitable objects from all instances and refining them. This allows the optimization process to focus on referring subtask, reducing optimization difficulty and training costs.

Next, we suggest Motion Query Decoder which extracts information from motion expression features, not frame queries or video queries. Specifically, motion queries use cross-attention to interact with word features, extract the most primitive word-level information, and determine the importance of each word. Then use cross-attention to interact with visually enhanced sentence embedding to extract the overall motion expression representation:

$$Q'_m = \mathrm{Decoder}(Q_m, F_w, F_{me}), \tag{8}$$

where $Q'_m \in \mathbb{R}^{N_m \times C}$ is the decoded motion queries used for final classification and mask prediction. One decoder consists of two cross-attention layers, a self-attention layer and a ffn layer. Motion Query Decoder effectively captures video contexts and aggregates motion expression information into the motion queries. As a result, Motion Query Decoder shows fast convergence speed while achieving high accuracy, and significantly reduces training memory compared to previous RVOS methods.

3.3. Training and Inference

Finally, the output motion queries Q'_m from Motion Query Decoder are passed into the classification head $\mathcal{H}_c$ and the mask generation head $\mathcal{H}_m$:

$$\mathcal{S} = \mathcal{H}_c(Q'_m), \tag{9}$$

where $\mathcal{S} \in \mathbb{R}^{N_m}$ is a binary classification score, $\mathcal{H}_c$ is a single linear layer.

$$\mathcal{M} = F_{mask} \cdot \mathcal{H}_m(Q'_m), \tag{10}$$

where $\mathcal{M} \in \mathbb{R}^{T \times N_m \times \frac{H}{S} \times \frac{W}{S}}$ is mask predictions, $\mathcal{H}_m$ is three MLP layers to generate mask embeddings, $\cdot$ denotes dot product operation.

We attach the proposed module DMVS on top of video instance segmenter, and the entire model get trained end-to-end. Note that frame-level outputs of Mask2Former [8] and video-level outputs of VITA [19] are not used for loss computation. Only the loss of video-level outputs of DMVS is considered. The total loss for model learning is as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{mask} + \lambda_{cls}\mathcal{L}_{cls}, \tag{11}$$

where $\mathcal{L}_{mask}$ is mask loss, consisting of the binary cross-entropy loss and the dice loss [34]. $\mathcal{L}_{cls}$ is the classification loss.

Table 1. Comparison with state-of-the-art models on MeViS val and val^u datasets.

Methods	Reference	Val			Val^u		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
URVOS [38]	ECCV'2020	27.8	25.7	29.9	-	-	-
LBDT [11]	CVPR'2022	29.3	27.8	30.8	-	-	-
MTTR [2]	CVPR'2022	30.0	28.8	31.2	-	-	-
ReferFormer [44]	CVPR'2022	31.0	29.8	32.2	-	-	-
VLT+TC [9]	TPAMI'2022	35.5	33.6	37.3	-	-	-
LMPM [10]	ICCV'2023	37.2	34.2	40.2	40.2	36.5	43.9
DsHmp [18]	CVPR'2024	46.4	43.0	49.8	55.3	51.0	60.4
DMVS (Ours)	CVPR'2025	**48.6**	**44.2**	**52.9**	**58.3**	**52.6**	**63.9**

We use the entire video as input in the inference process, and DMVS learns motion queries that represent referring instances of the entire video. For traditional RVOS datasets with single-object expressions [21, 38], we select the mask with the highest prediction score as the final prediction result. For MeViS dataset [10] with multi-object expressions, we select the masks with prediction confidence scores greater than the confidence threshold σ as the final prediction results.

4. Experiments

4.1. Datasets and Metrics

Dataset. We conduct experiments on motion expression video segmentation dataset MeViS [10] and traditional RVOS datasets: Ref-YouTube-VOS [38] and Ref-DAVIS17 [21]. MeViS [10] is a novel dataset designed for motion description analysis, featuring 2,006 video clips and 443,000 high-quality object segmentation masks. It includes 28,570 descriptive sentences referencing 8,171 objects in complex scenarios. The dataset is partitioned into 1,662 training videos, 190 validation videos, and 154 test videos. Ref-YouTube-VOS [38] contains 3,471 training videos annotated with 12,913 expressions and 507 validation videos with 2,096 expressions. Ref-DAVIS17 [21] offers 90 videos accompanied by 1,544 expressions.

Evaluation Metrics. we utilize region similarity $\mathcal{J}$ (average IoU), contour accuracy $\mathcal{F}$ (mean boundary similarity), and their average $\mathcal{J}\&\mathcal{F}$ as our evaluation indicators.

4.2. Implementation Details

MeViS. We use frozen VITA [19] as the video instance segmenter, and the default weight of VITA is trained on the OVIS [36] dataset using Swin Transformer [28] as the image backbone. The training process is conducted over 40,000 iterations, utilizing the AdamW optimizer [29] with a learning rate set to 5e-5. During training, we randomly extract clips of T = 10 frames per video, resizing the shorter side of each frame to 360 pixels and the longer side to 640 pixels. Motion Query Decoder employs 6 layers, the number of queries N_f, N_v and N_m are set to 100, 100 and 20. The coefficient for loss is set as λ_{cls} = 2.0. Threshold σ is set to 0.7. We use RoBERTa [27] as the text encoder that is frozen all the time.

Ref-Youtube-VOS and Ref-DAVIS17. Follow VITA [19], we process images from Ref-COCO/+/g [32, 51] to generate pseudo-videos for joint training with Ref-Youtube-VOS. The implementation details are consistent with MeViS. In addition, we directly use the weights trained on Ref-Youtube-VOS to test Ref-DAVIS17.

4.3. Main Results

MeViS. In Tab. 1, we evaluate the proposed approach DMVS on MeViS [10] dataset. DMVS achieves superior performance compared to other state-of-the-art methods. DMVS achieves 48.6% $\mathcal{J}\&\mathcal{F}$, 44.2% $\mathcal{J}$ and 52.9% $\mathcal{F}$ on MeViS val set, outperforming the leading method DsHmp [18] by 2.2% $\mathcal{J}\&\mathcal{F}$, 1.2% $\mathcal{J}$ and 3.1% $\mathcal{F}$, respectively. On the MeViS val^u set, DMVS surpasses DsHmp by a remarkable 3% $\mathcal{J}\&\mathcal{F}$. These results demonstrate the effectiveness of decoupling motion expression video segmentation into video instance segmentation and motion expression understanding.

Ref-Youtube-VOS and Ref-DAVIS17. We compare our method to previous models on Ref-YouTube-VOS [38] and Ref-DAVIS17 [21] datasets in Tab. 2. On Ref-YouTube-VOS, DMVS achieves 64.3 % $\mathcal{J}\&\mathcal{F}$, which is 0.7 % higher than the previous state-of-the-art DsHmp [18]. On Ref-DAVIS17, DMVS achieves 65.2 % $\mathcal{J}\&\mathcal{F}$ and surpasses DsHmp [18] by 1.2 %. The improvement on Ref-YouTube-VOS is not significant, mainly because the dataset mainly focuses on static expressions, lacks motion expressions, and has less dependence on temporal information.

4.4. Ablation Study

Since the main focus of this paper on decoupling motion expression video segmentation, we conduct ablation study on MeViS [10] val set to demonstrate the effectiveness of

Table 2. Comparison with state-of-the-art models on Ref-Youtube-VOS and Ref-DAVIS17 datasets.

Methods	Reference	Ref-Youtube-VOS			Ref-DAVIS17		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
MTTR [2]	CVPR'2022	55.3	54.0	56.6	-	-	-
ReferFormer [44]	CVPR'2022	59.4	58.0	60.9	59.6	56.5	62.7
OnlineRefer [42]	ICCV'2023	62.9	61.0	64.7	62.4	59.1	65.6
HTML [17]	ICCV'2023	61.2	59.5	63.0	-	-	-
R2VOS [22]	ICCV'2023	61.3	59.6	63.1	-	-	-
SgMg [33]	ICCV'2023	62.0	60.4	63.5	61.9	59.0	64.8
TempCD [39]	ICCV'2023	62.3	60.5	64.0	62.2	59.3	65.0
SOC [31]	NeurIPS'2023	62.4	61.1	63.7	63.5	60.2	66.7
LoSh [52]	CVPR'2024	63.7	62.0	65.4	62.9	60.1	65.7
DsHmp [18]	CVPR'2024	63.6	61.8	65.4	64.0	60.8	67.2
DMVS (Ours)	CVPR'2025	**64.3**	**62.4**	**66.2**	**65.2**	**62.2**	**68.2**

Table 3. Ablation study of main components of DMVS.

Model	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
VITA-ROVS	39.7	36.6	42.8
Baseline	44.2	40.7	47.7
Baseline+MQI	46.3	42.4	50.2
Baseline+MQI+MEE	**48.6**	**44.2**	**52.9**

Table 4. Ablation study of motion expression encoder.

$Q_{f/v}$	Q'_f	Q'_v	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
✗	✗	✗	46.3	42.4	50.2
✓	✗	✗	47.1	42.9	51.3
✓	✓	✗	47.6	43.3	51.9
✓	✗	✓	47.8	43.5	52.1
✓	✓	✓	**48.6**	**44.2**	**52.9**

Table 5. Ablation study of motion query initialization method.

Initialization	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Random	46.5	41.8	51.2
Text	46.8	42.6	51.0
Video	**48.6**	**44.2**	**52.9**

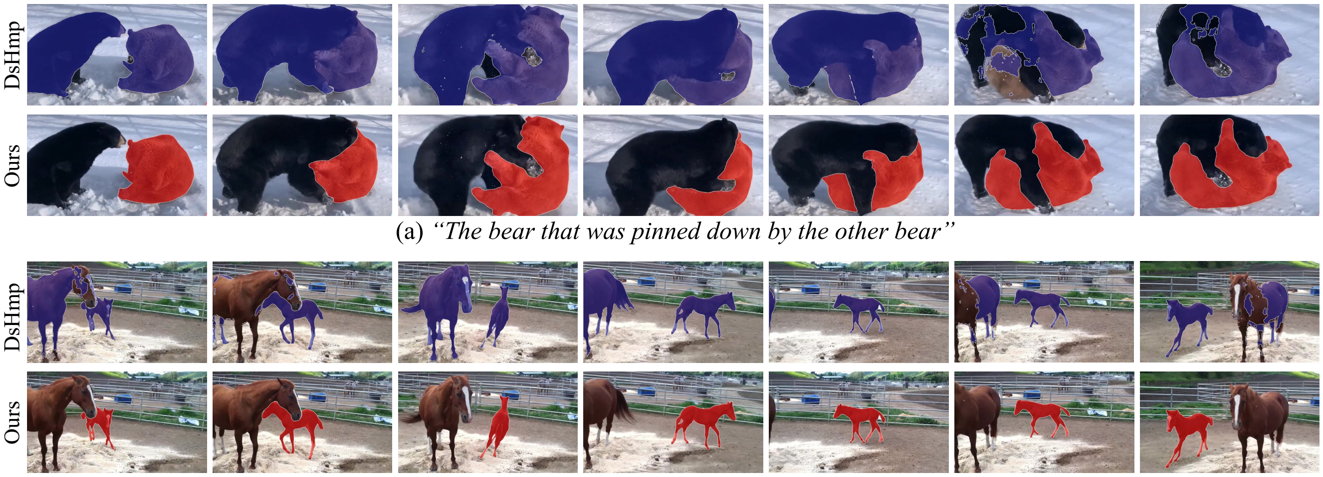
each component and the impact of different configurations.

Effectiveness of main components. In Tab. 3, we conduct experiments to verify the effectiveness of each key component in of our framework. VITA-ROVS is essentially a reproduction of LMPM [10], which is the fundamental model for using VITA for RVOS tasks. Baseline is a simple implementation of DMVS that uses randomly initialized motion queries to interact with text features and video queries. Baseline achieves 44.2 % $\mathcal{J}\&\mathcal{F}$, an improvement of 4.5 % $\mathcal{J}\&\mathcal{F}$ compared to VITA-RVOS, which fully demonstrates the effectiveness of decoupling motion expression video segmentation into video instance segmentation and motion expression understanding. MQI is Motion Query Initialization through video queries with classification priors. MQI improves the performance by 2.1% $\mathcal{J}\&\mathcal{F}$, emphasizing that the query initialization method has a significant impact on performance. Next, we present Motion Expression Encoder (MEE) to integrate video contextual object information and text features. Utilizing MEE improves the $\mathcal{J}\&\mathcal{F}$ by 2.3%.

Effectiveness of motion expression encoder. In Tab. 4, we conduct ablation experiments to evaluate the effectiveness of motion expression encoder. The sentence embedding directly interacts with the frame and video queries output by VITA, resulting in 0.8% $\mathcal{J}\&\mathcal{F}$ improvement. After enhancing frame queries with static clues and video queries with motion clues, model continues to increase by 0.5% and 0.7% $\mathcal{J}\&\mathcal{F}$, respectively. The experimental results demonstrate that by decoupling motion perception and interacting with object queries of different level, understanding of motion expression is effectively improved.

Motion query initialization method. Tab. 5 shows the impact of different motion query initialization methods. Random is Eq. (5), Text is Eq. (6), both motion queries require interaction with text features and video queries in the motion query decoder. Video is Eq. (7), the video queries TopK guided by classification prior brings rich object information to motion queries, resulting in 2% $\mathcal{J}\&\mathcal{F}$ improvement.

Number of motion queries N_m. Tab. 6 shows results obtained with varying numbers of motion queries N_m. The number of video queries N_v = 100, so the maximum number of motion queries N_m = 100. When N_m = 10, the performance reduces by 0.8% $\mathcal{J}\&\mathcal{F}$ due to missing queries with instances. When N_m = 50 or 100, there is a significant decrease in performance due to the introduction of too many background queries, which increases the difficulty of



(a) *"The bear that was pinned down by the other bear"*

(b) *"Baby horse moving around"*

Figure 4. Qualitative comparison of our method to the main counterpart DsHmp [18] on the MeViS [10] val set.

Table 6. Ablation study of number of motion queries N_m.

N_m	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
10	47.8	43.2	52.4
20	**48.6**	**44.2**	**52.9**
50	47.3	42.9	51.7
100	46.2	41.6	50.8

Table 7. Ablation study of VITA weights pretrained on different VIS datasets.

Dataset	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
YouTube-VIS 2019 [48]	45.4	42.2	48.6
YouTube-VIS 2021 [49]	46.7	43.1	50.3
OVIS [36]	**48.6**	**44.2**	**52.9**

Table 8. Efficiency comparison with the state-of-the-art method.

Method	Learnable params	GPU memory	Training time
DsHmp [18]	102.3M	22G	17 hours
DMVS (Ours)	**12.7M**	**6G**	**7 hours**

optimization.

Pretrained weights of VITA. As shown in Tab. 7, we demonstrate the impact of different pretrained weights of VITA [19]. MeViS is a long video dataset, and OVIS [36] is also a long video instance segmentation dataset, so the pretrained weight of OVIS achieves the highest performance. YouTube-VIS 2019 [48] mainly focuses on short videos, while YouTube-VIS 2021 [49] improves its performance to some extent due to the addition of some long videos in the training set.

Efficiency of DMVS. In Tab. 8, we compare the training efficiency with the previous state-of-the-art DsHmp [18]. Thanks to freezing the entire video instance segmenter, our model's learnable parameters are only 12.4% of DsHmp. When training a video clip on each GPU (video length T = 8), our model only occupies 6G memory, which provides the possibility for high batchsize training with fewer GPUs. Our model also has a significant advantage in training time compared to DsHmp.

4.5. Qualitative Results

As shown in Fig. 4, we present some visualization results on MeViS [10]. DsHmp [18] does not truly understand motion expressions and tends to segment all instances in the video. On the contrary, DMVS can understand the motion expressions ”was pinned down by” and ”moving around”, correctly segmenting the specified objects ”the bear” and ”baby horse”. Moreover, the segmentation results have good temporal consistency, without errors due to the movement of the objects. The qualitative results further demonstrate the effectiveness of DMVS.

5. Conclusion

In this paper, we propose DMVS, a simple and efficient referring video segmentation framework, emphasizing decoupling motion expression video segmentation into video instance segmentation and motion expression understanding. The frozen video instance segmenter extracts high-quality frame-level and video-level object queries. Additionally, our motion expression encoder interacts two levels of queries with static and motion cues, respectively, to further encode visually enhanced motion expressions. Furthermore, we use video queries guided by classification priors to initialize motion queries, greatly reducing optimization difficulty and training costs. Extensive experiments demonstrate the effectiveness and efficiency of our framework.

Acknowledgements

This work was supported in part by the Taishan Scholar Project of Shandong Province under Grant tsqn202306079, in part by the the National Natural Science Foundation of China Grant 62471278, Grant 61991411, and Grant 62271180, and in part by the Research Grants Council of the Hong Kong Special Administrative Region, China under Grant STG5/E-103/24-R.

References

[1] Miriam Bellver, Carles Ventura, Carina Silberer, Ioannis Kazakos, Jordi Torres, and Xavier Giro-i Nieto. A closer look at referring expressions for video object segmentation. *Multimedia Tools and Applications*, 82(3):4419–4438, 2023. 1, 3

[2] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4985–4995, 2022. 1, 3, 6, 7

[3] 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. 3

[4] Jinpeng Chen, Runmin Cong, Yuxuan Luo, Horace Ip, and Sam Kwong. Saving 100x storage: Prototype replay for reconstructing training sample distribution in class-incremental semantic segmentation. *Advances in Neural Information Processing Systems*, 36:35988–35999, 2023. 3

[5] Jinpeng Chen, Runmin Cong, Yuxuan Luo, Horace Ho Shing Ip, and Sam Kwong. Strike a balance in continual panoptic segmentation. In *European Conference on Computer Vision*, pages 126–142. Springer, 2024.

[6] Jinpeng Chen, Runmin Cong, Yuxuan Luo, Horace Ho Shing Ip, and Sam Kwong. Replay without saving: Prototype derivation and distribution rebalance for class-incremental semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

[7] 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. 2

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2, 3, 5

[9] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7900–7916, 2022. 1, 6

[10] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2694–2703, 2023. 1, 2, 3, 4, 6, 7, 8

[11] Zihan Ding, Tianrui Hui, Junshi Huang, Xiaoming Wei, Jizhong Han, and Si Liu. Language-bridged spatial-temporal interaction for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4964–4973, 2022. 1, 6

[12] Hao Fang, Feiyu Pan, Xiankai Lu, Wei Zhang, and Runmin Cong. Uninext-cutie: The 1st solution for lsvos challenge rvos track. *arXiv preprint arXiv:2408.10129*, 2024. 3

[13] Hao Fang, Peng Wu, Yawei Li, Xinxin Zhang, and Xiankai Lu. Unified embedding alignment for open-vocabulary video instance segmentation. In *European Conference on Computer Vision*, pages 225–241. Springer, 2024. 2

[14] Hao Fang, Tong Zhang, Xiaofei Zhou, and Xinxin Zhang. Learning better video query with sam for video instance segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[15] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5958–5966, 2018. 1, 2

[16] Qingfeng Guan, Hao Fang, Chenchen Han, Zhicheng Wang, Ruiheng Zhang, Yitian Zhang, and Xiankai Lu. Structural transformer with region strip attention for video object segmentation. *Neurocomputing*, 596:128076, 2024. 3

[17] Mingfei Han, Yali Wang, Zhihui Li, Lina Yao, Xiaojun Chang, and Yu Qiao. Html: Hybrid temporal-scale multimodal learning framework for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13414–13423, 2023. 7

[18] Shuting He and Henghui Ding. Decoupling static and hierarchical motion perception for referring video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13332–13341, 2024. 2, 3, 4, 6, 7, 8

[19] 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, 3, 5, 6, 8

[20] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. *Advances in Neural Information Processing Systems*, 35:31265–31277, 2022. 2

[21] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*, pages 123–141. Springer, 2019. 1, 3, 6

[22] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Bhiksha Raj, and Yan Lu. Robust referring video object segmentation with cyclic structural consensus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22236–22245, 2023. 7

[23] Chen Liang, Yu Wu, Tianfei Zhou, Wenguan Wang, Zongxin Yang, Yunchao Wei, and Yi Yang. Rethinking cross-modal

interaction from a top-down perspective for referring video object segmentation. *arXiv preprint arXiv:2106.01061*, 2021. 3

[24] Huaijia Lin, Ruizheng Wu, Shu Liu, Jiangbo Lu, and Jiaya Jia. Video instance segmentation with a propose-reduce paradigm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1739–1748, 2021. 2

[25] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4761–4775, 2021. 1

[26] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 4

[27] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 4, 6

[28] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[29] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[30] Sihan Luo, Xia Yuan, and Yongshun Liang. Multiframe spatiotemporal attention-guided semisupervised video segmentation. *Journal of Image and Graphics*, 29(05):1233–1251, 2024. 3

[31] Zhuoyan Luo, Yicheng Xiao, Yong Liu, Shuyan Li, Yitong Wang, Yansong Tang, Xiu Li, and Yujiu Yang. Soc: Semantic-assisted object cluster for referring video object segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 7

[32] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 6

[33] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 920–930, 2023. 7

[34] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 5

[35] Feiyu Pan, Hao Fang, Runmin Cong, Wei Zhang, and Xiankai Lu. Video object segmentation via sam 2: The 4th solution for lsvos challenge vos track. *arXiv preprint arXiv:2408.10125*, 2024. 3

[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: A benchmark. *International Journal of Computer Vision*, 130(8): 2022–2039, 2022. 6, 8

[37] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015. 4

[38] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 208–223. Springer, 2020. 1, 3, 6

[39] Jiajin Tang, Ge Zheng, and Sibei Yang. Temporal collection and distribution for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15466–15476, 2023. 7

[40] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[41] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8741–8750, 2021. 2

[42] Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. Onlinerefer: A simple online baseline for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2761–2770, 2023. 7

[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. 2

[44] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 1, 2, 3, 4, 6, 7

[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] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15325–15336, 2023. 4

[47] Shilin Yan, Renrui Zhang, Ziyu Guo, Wenchao Chen, Wei Zhang, Hongyang Li, Yu Qiao, Hao Dong, Zhongjiang He, and Peng Gao. Referred by multi-modality: A unified temporal transformer for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6449–6457, 2024. 3

[48] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5188–5197, 2019. 1, 2, 8

[49] Linjie Yang, Yuchen Fan, Yang Fu, and Ning Xu. The 3rd large-scale video object segmentation challenge - video instance segmentation track, 2021. 8

[50] Kaining Ying, Qing Zhong, Weian Mao, Zhenhua Wang, Hao Chen, Lin Yuanbo Wu, Yifan Liu, Chengxiang Fan, Yunzhi Zhuge, and Chunhua Shen. Ctvis: Consistent training for online video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 899–908, 2023. 2

[51] 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. 6

[52] Linfeng Yuan, Miaojing Shi, Zijie Yue, and Qijun Chen. Losh: Long-short text joint prediction network for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14001–14010, 2024. 7

[53] Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. Dvis: Decoupled video instance segmentation framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1282–1291, 2023. 2

[54] Tong Zhang, Hao Fang, Hao Zhang, Jialin Gao, Xiankai Lu, Xiushan Nie, and Yilong Yin. Learning feature semantic matching for spatio-temporal video grounding. *IEEE Transactions on Multimedia*, 2024. 2