MeViS: A Large-scale Benchmark for Video Segmentation with Motion Expressions

Henghui Ding† Chang Liu‡ Shuting He‡ Xudong Jiang‡ Chen Change Loy†
†S-Lab, Nanyang Technological University ‡School of EEE, Nanyang Technological University
https://henghuiding.github.io/MeViS

Figure 1. Examples of video clips from Motion expressions Video Segmentation (MeViS) are provided to illustrate the dataset’s nature and complexity. The expressions in MeViS primarily focus on motion attributes and the referred target object cannot be identified by examining a single frame solely. For instance, the first example features three parrots with similar appearances, and the target object is identified as “The bird flying away”. This object can only be recognized by capturing its motion throughout the video.

Abstract

This paper strives for motion expressions guided video segmentation, which focuses on segmenting objects in video content based on a sentence describing the motion of the objects. Existing referring video object datasets typically focus on salient objects and use language expressions that contain excessive static attributes that could potentially enable the target object to be identified in a single frame. These datasets downplay the importance of motion in video content for language-guided video object segmentation. To investigate the feasibility of using motion expressions to ground and segment objects in videos, we propose a large-scale dataset called MeViS, which contains numerous motion expressions to indicate target objects in complex environments. We benchmarked 5 existing referring video object segmentation (RVOS) methods and conducted a comprehensive comparison on the MeViS dataset. The results show that current RVOS methods cannot effectively address motion expression-guided video segmentation. We further analyze the challenges and propose a baseline approach for the proposed MeViS dataset. The goal of our benchmark is to provide a platform that enables the development of effective language-guided video segmentation algorithms that leverage motion expressions as a primary cue for object segmentation in complex video scenes. The proposed MeViS dataset has been released at https://henghuiding.github.io/MeViS.

1. Introduction

Language-guided video segmentation is an emerging field that involves segmenting and tracking target objects using natural language expressions. This field has traditionally been a sub-branch of semi-supervised video object segmentation, where referring expressions are used to describe the target object. Existing referring video object datasets, such as [13,21,44], commonly feature videos with isolated and salient objects that have obvious static features. The corresponding expressions often contain static attributes such as object color, which can be observed in a single frame. As a result, motion properties of videos are often
given less emphasis, and referring image segmentation methods can be used for referring video segmentation, achieving good results [1, 10, 21, 31].

In this paper, we wish to highlight the significance of temporal motion properties of videos and explore the potential of using motion expressions to segment objects in videos. To this end, we propose a new large-scale dataset called Motion expressions Video Segmentation (MeViS) to aid our investigation. The MeViS dataset comprises 2,006 videos with a total of 8,171 objects, and 28,570 motion expressions are provided to refer to these objects.

We take several steps to ensure that the MeViS dataset places emphasis on the temporal motions of videos. First, we carefully select video content that contains multiple objects that coexist with motion and exclude videos with isolated objects that can be easily described by static attributes. Second, we prioritize language expressions that do not contain static clues, such as category names or object colors, in cases where target objects can be unambiguously described by motion words alone. This is distinct from previous datasets, such as [13, 21, 44], which include obvious static clues in their expressions. Additionally, MeViS differentiates itself from referring image segmentation datasets, such as [20, 36, 53, 62], which do not account for the temporal properties of video content. Moreover, unlike existing referring video object segmentation datasets that focus on single-target expressions, where one expression refers to only one target object, MeViS expands this task to include multi-object expressions that refer to multiple target objects. This feature enables expressions to refer to an unlimited number of target objects, making the proposed MeViS more challenging and reflective of real-world scenarios.

The proposed MeViS dataset poses notable challenges in capturing and understanding motions in both video and language. The language expressions may describe motion that spans a random number of frames, requiring the capture of fleeting movements and long-term actions that occur throughout the entire video. This poses significant challenges for both understanding motion in the video content and in the accompanying language expressions. Capturing fleeting movements requires attention on each individual frame, while understanding long and complex movements that span across many frames demands temporal context across the entire video. With the proposed dataset, we benchmark 5 existing referring video object segmentation (RVOS) methods [2, 10, 11, 44, 55] and conduct a comprehensive comparison. The experimental results demonstrate that MeViS presents more challenges than existing datasets, and current RVOS methods are unable to effectively address motion expression-guided video segmentation.

In addition to proposing the MeViS dataset, we present a baseline approach, named Language-guided Motion Perception and Matching (LMPM), to address the challenges posed by the dataset. Our approach generates language-conditional queries to detect potential target objects in the video and represents them using object embeddings, which are more robust and computationally efficient than object feature maps [15]. We then perform Motion Perception on the object embeddings to capture the temporal context and obtain a global view of the video, enabling the model to understand both fleeting and long-term motions. Next, we use a Transformer decoder to decode language-related information from the motion-aggregated object embeddings and predict object trajectories. Finally, we perform similarity matching between the language features and the predicted object trajectories to identify the target object(s).

Our contributions provide a foundation for developing more advanced language-guided video segmentation algorithms that leverage motion expressions as a primary cue for object segmentation and identification in complex video scenes. In particular, we propose a new language-guided video segmentation dataset, MeViS, and conduct comprehensive evaluations of state-of-the-art referring video object segmentation methods on the MeViS dataset, providing a reference for future works. We also develop a simple baseline approach, LMPM, which points to potential solutions to some of the challenges and future research directions.

2. Related Work

Referring Image Segmentation. Referring image segmentation [5, 9, 10, 27, 28], also known as referring expression segmentation, involves grounding the target object in images based on natural language expressions that describe its properties and generating a corresponding segmentation mask. This task requires both language and image understanding and is one of the most fundamental yet challenging tasks in computer vision. Referring image segmentation was first introduced by Hu et al. [16] in 2016 and has received considerable attention since then. In the pre-Transformer era, mainstream methods typically employed Fully Convolutional Networks (FCN) [6, 7, 34] and Recurrent Neural Networks (RNN) to extract image features and language features, respectively, and then fused the multimodal features via some specially designed modules [12, 23, 30, 37]. For example, Liu et al. [30] introduced a Recurrent Multimodal Interaction (RMI) module to recurrently fuse the feature of each word into the image features. Li et al. [23] proposed a Recurrent Refinement Network (RRN) that progressively refines the segmentation mask based on pyramid features in FCN.

In addition to one-stage methods that fuse multi-modal features and conduct segmentation, some methods decouple referring image segmentation into instance segmentation and language-object matching [19, 29, 61]. For instance, Yu et al. use the off-the-shelf instance segmentation model Mask R-CNN [14] to detect all instances first and then
select the one that best matches the language as output. A holistic understanding of language and vision information is crucial for referring image segmentation, and many works have explored this direction [18, 58, 60]. For instance, Ye et al. introduce a Cross-Modal Self-Attention (CMSA) model [60] to select the most meaningful words in the expression and pixels in the image to achieve better contextual understanding. Recently, the success of Transformer [47] in vision tasks has inspired many studies in referring image segmentation. Ding et al. [9, 10] first introduced Transformer into referring segmentation and proposed a Vision-Language Transformer (VLT). Following Ding et al. [9, 10], more Transformer-based methods have been proposed [22, 52, 59]. For example, Wang et al. [52] employ the Vision-Language Decoder to deal with visual and text tokens extracted by CLIP [43]. Yang et al. [59] focus on multi-modal feature fusion and propose a Language-Aware Vision Transformer (LAVT).

**Referring Video Segmentation.** Referring video object segmentation is an emerging area [3, 17, 31, 38, 39, 45, 46, 49, 50, 54, 57, 63] that aims to segment the target object indicated by a given expression across the entire video clip. It was first introduced in 2018 by A2D [13] and DAVIS17-RVOS [21], where A2D [13] seeks to segment actors according to descriptions of their actions in video content, and DAVIS17-RVOS [21] replaces masks with language as the reference for the target object in video object segmentation. Later, Seo et al. [44] built the Refer-YouTube-VOS based on the YouTube-VOS-2019 dataset [56]. These datasets typically provide an expression for a single object, and the expression usually describes the static attributes of the target object, such as its color and shape.

Existing methods typically treat referring video segmentation as a form of semi-supervised video object segmentation [41] by replacing mask reference with language reference. For instance, Khoreva et al. [21] employ the referring image segmentation method MAttNet [61] to achieve frame-level segmentation and then perform post-processing for temporal consistency. URVOS [44] employs cross-modal attention to perform per-frame segmentation and propagate the mask across clips with a memory attention module. RefVOS [1] independently segments each frame based on the fused features of language and image/frame, without utilizing temporal information. Liang et al. [26] introduces a top-down approach that first detects all object tracklets and then selects the target object by matching between language and tracklet features. Most recently, ReferFormer [55] and MTTR [2] employ Transformer [47] to address referring video object segmentation.

### 3. MeViS Dataset
In this section, we introduce the newly built large-scale dataset MeViS by first presenting the video collection and annotation process in Section 3.1 and then providing the dataset statistics and analysis in Section 3.2.

#### 3.1. Motion Expression Annotation

**Video Collection.** We gather and choose videos from publicly available video segmentation datasets with high-quality mask annotations [8, 42, 48, 51], and select the ones that meet our criteria for motion and object complexity. Our selection process involves the following rules:

- **R1.** We only include videos that have multiple objects within the frame in MeViS; videos with only one or two salient objects are not considered. We specifically look for videos that depict many objects with similar appearances, such as the first example video in Figure 1 which shows three yellow parrots.

- **R2.** We select videos that contain objects that demonstrate substantial motion and movement. Videos depicting objects that have little or no motion are excluded.

After reviewing over 4,000 potential candidates, we carefully selected the most appropriate and suitable videos that meet our rigorous standards for both visual and linguistic content. Ultimately, by prioritizing quality over quantity, we chose 2,006 videos to create a benchmark that is diverse and representative of a wide range of real-world video scenarios. The basic language annotation methodology and procedure for MeViS follow the ReferIt [20], which is an interactive game-like approach that involves two players taking turns to annotate and validate. The following section will introduce the process of language expression annotation and validation in more detail.

**Language Expression Annotation.** We developed a web-based annotation system for annotating language expressions. The system randomly selects a video from the MeViS dataset and displays all object masks of the selected video on the web system. The annotator needs to choose one or several objects from the video and write the corresponding referring expression according to the guidelines for annotating language expressions. To ensure that the language expressions in our dataset align with our focus on motion-based video segmentation, we established several guidelines for annotating the language expressions:

- **A1.** Target objects must exhibit significant motion. Objects that remain stationary or only demonstrate minimal motion should be disregarded.

- **A2.** If an object can be unambiguously described by its motion or action, static attributes such as color should not be included in the expression.

- **A3.** If multiple objects cannot be differentiated based solely on their motion or action, they can be described together if their motion or action can unambiguously identify them, such as “The two lions fighting and running amidst a group of lions.”
A.4. If it is not possible to differentiate single or multiple objects based solely on their motion or action, limited static attributes can be included in the expression.

**Language Expression Validation.** Upon receiving annotated “video-object-expression” samples from the annotators, the validation process begins by displaying the video and expression and prompting the validator to select and submit the objects referred to in the expression. The validator must find the targets independently and submit their selection. The system then compares the targets chosen by the validator with the annotations submitted by the annotator. A sample is considered valid if the validator and annotator independently selected the same target object(s) using the same expression. If the targets selected by the validator do not match the annotation submitted by the annotator, the sample will be forwarded to another validator for a second opinion. If the second validator also fails to identify the correct targets, the sample will be considered invalid and excluded from the dataset. Validators have the authority to reject samples that are deemed inappropriate or fall short of quality standards. Moreover, we stress the importance of the following validation criterion:

**V1.** The corresponding sentence will be removed from the dataset when the target object described by a sentence can be identified through a single frame without the need for motion information.

By establishing these validation criteria, we aim to ensure that the language sentences in our dataset accurately express motion and are of high quality, while also increasing the level of difficulty in the language-guided video segmentation task, thereby enabling a more robust evaluation of the performance of different models and methods.

### 3.2. Dataset Analysis and Statistics

In Table 1, we present a statistical analysis of the newly proposed MeViS dataset, using 5 previous referring video object segmentation datasets as references, including A2D Sentence [13], J-HMDB Sentence [13], DAVIS16-RVOS [21], DAVIS17-RVOS [21], and Refer-Youtube-VOS [44]. As shown in Table 1, MeViS contains 2,006 videos and 8,171 objects. Compared to Refer-Youtube-VOS [44], which is based on the existing VOS dataset [56], MeViS has more objects (8,171 vs. 7,451), more expressions (28,570 vs. 15,009), and more annotation masks (443k vs. 131k). In the following, we discuss how the proposed dataset MeViS intentionally increases the complexities of language-guided video segmentation by considering the challenges of both linguistic and visual modalities.

#### Video Content

As shown in Table 1, MeViS has an average of 4.28 objects per video, which is higher than previous datasets. Furthermore, as depicted in Figure 2, MeViS contains longer videos, with an average duration of 13.16 seconds, which is significantly longer than the Refer-Youtube-VOS dataset. These intentional design choices make MeViS more complex and challenging for language-guided video segmentation. This is in contrast to existing datasets such as A2D Sentence [13] and DAVIS16-RVOS [21], where only one or two salient objects per category are present, and the model can choose the most prominent object as the target or identify the target object based on the category name. For example, in Figure 3(b), there is only one person in the foreground, and the model can simply identify the target by the term “a person” while ignoring “skateboarding”.

The proposed MeViS dataset addresses this limitation by selecting videos with more objects that have diverse and dynamic motions. Moreover, MeViS features many videos with objects of the same...
category, such as a group of tigers or rabbits. For instance,
in Figure 3(a), there are three giraffes with highly similar
appearances, and the most salient/foreground one is not
the target object in this sample, making it challenging to
identify the target object(s) through saliency or category
information alone. By including more challenging videos,
MeViS better simulates real-world scenarios, making it a
valuable resource for studying motion expression-guided
video understanding in complex environments.

Target Object(s). As we have included longer videos in our
MeViS dataset, we have also observed a significant increase
in the duration of target objects, ensuring adequate object
motions. As shown in Figure 2(b), the object durations in
our dataset have an average of 10.88 seconds, which is
more than two times longer than the average duration of
Refer-Youtube-VOS. Compared to previous datasets, such as
A2D Sentence [13] and J-HMDB Sentence [13], which
focus on salient actions of a few categories [44], our MeViS
dataset includes more categories from open-world [8, 42,
48, 51], presenting improved difficulties in the diversity
of target objects. Besides, as shown in Table 1, previous
datasets usually have one sentence referring to one single
object, which means that finding multiple objects requires
multiple expressions, and each object must be searched for
individually*. In contrast, we add a more natural way of
selecting target objects, where one expression may refer to
several objects, denoted as “multi-object expression”. An
example of multi-object expression is shown in Figure 3(a),
where “Giraffes that turns around” refers to two giraffes.

*Although some datasets like DAVIS16-RVOS [21] contain samples
of multiple instances as targets, the mask annotations provided by
DAVIS16 [40] do not distinguish instances, and some target masks cover
several instances. Thus, each sample in DAVIS16-RVOS can be considered
as a single-sentence single-object pair.

As shown in Table 1, on average, each expression in MeViS
refers to 1.59 objects, which is larger than existing datasets
where the average is only 1 object per expression.

Language Expression. One of the key distinguishing as-
pects of the MeViS dataset is its emphasis on describing ob-
ject motions in language expressions. The previous largest
RVOS dataset Refer-Youtube-VOS [44] provides two types
of language annotations: full-video expression and first-
frame expression. The first-frame expression is based solely
on static attributes of the first frame image, whereas the full-
video expression considers the entire video. However, in
many cases, even the full-video expressions contain static
attributes that could potentially enable the target object to
be identified in a single frame, for example, “A person
on the right dressed in blue black...”. In contrast, to
explore the practicality of employing motion expressions
for object localization and segmentation in videos, MeViS
is intentionally designed to include a range of diverse and
dynamic object motions, making it more challenging to
identify the target object based on static attributes alone. In
MeViS, there are significantly more motion expressions that
explicitly identify the target object based on its distinctive
actions or movements. The language expressions in the
proposed MeViS contain more motion attributes, such as
object position moving through the video and actions that
span several frames. The word cloud of the newly proposed
MeViS is visualized in Figure 4. From the word cloud
figure, we can observe that MeViS dataset has a large
number of words that describe motions, like “walking”,
“moving”, “playing”, and many relative directions that are
related to motions, such as “left”, “right”, etc.

4. Experiment

Evaluation Metrics. Similar to previous studies such as
[21, 44], we employ two widely used metrics, \( \mathcal{J} \) and \( \mathcal{F} \), to
assess the performance of methods on the newly proposed
MeViS dataset. The region similarity metric \( \mathcal{J} \) computes
the Intersection over Union (IoU) of the predicted and

![Figure 3](https://via.placeholder.com/150)

(a) “Giraffes turning around”

(b) “A person skateboarding”

![Figure 4](https://via.placeholder.com/150)

Figure 4. Word cloud of the top 100 words in the MeViS dataset. MeViS has a large number of words that describe motions, like “walking”, “moving”, “playing”, and many position words that are related to motions, such as “left”, “right”.

Figure 3. (a) Example from MeViS. (b) Example from Refer-
Youtube-VOS [44]. Compared to Refer-Youtube-VOS: ● Videos
in MeViS contain more objects in complex environments, making
it impossible to identify the target object via saliency or category
information alone. ● The number of target objects indicated by
language expression in MeViS is arbitrary, from 1 to many.
reflects the contour accuracy and gather global temporal context ("TC" in Table 2). Incorporating an attention module at the head to perceive long-term motions in the MeViS dataset. In order to further outperform VLT by a large margin of 3.2% in terms of selected frames from the video, we find that ReferFormer [55], which is trained using five randomly selected frames, outperforms VLT [10].

Furthermore, by comparing the results of VLT [10] with motion, despite their success on other benchmark datasets. This presents significant challenges for recognizing motions in the video content and the corresponding language expressions. Detecting fleeting actions requires meticulous perceiving of every frame while comprehending complex and extended motion spanning multiple frames requires contextual understanding across the entire duration of the video. Current state-of-the-art methods, such as [2, 11, 55], rely on random sampling of a few frames, which may miss frames containing crucial information described by the given expression. Furthermore, these methods fail to effectively extract temporal contextual information and performance gain of approximately 5% \( J \& F \), underscoring the significance of temporal context for MeViS. However, it is worth noting that longer temporal information does not necessarily lead to better performance on DAVIS_{17}-RVOS and Refer-Youtube-VOS.

We also conduct a cross-dataset experiment by training on referring image segmentation datasets and testing on referring video segmentation datasets. The results in Table 3 show that both the image-based method VLT [10] and video-based method ReferFormer [55] achieve competitive results on Refer-Youtube-VOS [44] and DAVIS_{17}-RVOS [21] when trained on image datasets Ref-COCO, Ref-COCO+, and Ref-COCOg. These results suggest that the expressions in Refer-Youtube-VOS [44] and DAVIS_{17}-RVOS [21] provide static clues like in the image domain, and many target objects can be identified by examining a single frame solely. In contrast, when trained on referring image segmentation datasets and tested on MeViS, both VLT [10] and ReferFormer [55] perform worse, indicating that there is a significant expression-gap (e.g., static vs. motion) between MeViS and these image domain datasets.

### 4.2. LMPM: A Simple Baseline Approach

The MeViS dataset introduces unique challenges in detecting and understanding object motions in both video and language contexts. The motions described by language expressions can occur over a random number of frames, making it necessary to capture fleeting actions and movements that occur throughout the entire video.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Type</th>
<th>Temporal</th>
<th>DAVIS_{17}-RVOS</th>
<th>RYV</th>
<th>MeViS</th>
</tr>
</thead>
<tbody>
<tr>
<td>VLT [10]</td>
<td>Image</td>
<td>1 frame</td>
<td>60.4</td>
<td>63.1</td>
<td>27.8</td>
</tr>
<tr>
<td>RFormer [55]</td>
<td>Video</td>
<td>5 rand. frames</td>
<td>60.2</td>
<td>62.8</td>
<td>31.0</td>
</tr>
<tr>
<td>VLT+TC</td>
<td>Video</td>
<td>All frames</td>
<td>60.3</td>
<td>62.7</td>
<td>35.5</td>
</tr>
<tr>
<td>RFormer+TC</td>
<td>Video</td>
<td>All frames</td>
<td>59.9</td>
<td>63.0</td>
<td>36.3</td>
</tr>
</tbody>
</table>

For module details, please refer to “Motion Perception” in Section 4.2. Adding temporal context via this module results in both VLT and ReferFormer achieving a performance gain of approximately 5% \( J \& F \), underscoring the significance of temporal context for MeViS. However, it is worth noting that longer temporal information does not necessarily lead to better performance on DAVIS_{17}-RVOS and Refer-Youtube-VOS.

### 4.1. Dataset Necessity and Challenges

To show the necessity and validity of MeViS in motion expression understanding, we compare the results of state-of-the-art referring image segmentation method VLT [10] and referring video segmentation method ReferFormer [55] on DAVIS_{17}-RVOS [21], Refer-Youtube-VOS [44], and MeViS, as shown in Table 2. When trained on referring video segmentation dataset, such as Refer-Youtube-VOS [44] and testing on itself, the image-based method VLT [10] that does not use any temporal design can achieve exception results of 60.4% \( J \& F \) and 63.1% \( J \& F \) on video datasets DAVIS_{17}-RVOS [21] and Refer-Youtube-VOS [44], respectively, which are even better than video method ReferFormer [55]. The results suggest that for DAVIS_{17}-RVOS [21] and Refer-Youtube-VOS [44], the temporal context is not essential, and image-based methods that use static clues can achieve good performance on these two datasets. However, on the proposed MeViS, VLT [10] only achieves a score of 27.8% \( J \& F \), suggesting that referring image segmentation methods without temporal designs struggle to address the unique challenges presented by videos in our dataset, particularly in handling motion, despite their success on other benchmark datasets. Furthermore, by comparing the results of VLT [10] with ReferFormer [55], which is trained using five randomly selected frames from the video, we find that ReferFormer outperforms VLT by a large margin of 3.2% in terms of \( J \& F \). This further highlights the importance of analyzing long-term motions in the MeViS dataset. In order to further prove this point, we enhance VLT and ReferFormer by incorporating an attention module at the head to perceive and gather global temporal context ("TC" in Table 2).
Figure 5. The overview architecture of the proposed baseline approach Language-guided Motion Perception and Matching (LMPM).

We first detect all possible target objects in each frame of the video and use object embeddings to represent them through Language-Guided Extractor. Then, Motion Perception is conducted on all the object embeddings of the video to grasp the global temporal context. By leveraging language queries and object embeddings with motion information, we generate object trajectories through a Transformer Decoder. Finally, we match the language features with the predicted object trajectories to identify the target object(s).

Table 4. Ablation study of the baseline approach LMPM.

<table>
<thead>
<tr>
<th>ID</th>
<th>Language Query</th>
<th>Motion Perception</th>
<th>Matching</th>
<th>$J &amp; F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>✓</td>
<td></td>
<td></td>
<td>31.0</td>
</tr>
<tr>
<td>ii</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>36.3</td>
</tr>
<tr>
<td>iii</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>37.2</td>
</tr>
</tbody>
</table>

Implementation Details. We set all the hyper-parameters related to the Language-Guided Extractor to the default settings of Mask2Former [4], including the backbone, Transformer decoder. We train 150,000 iterations using AdamW optimizer [35] with a learning rate of 0.00005. Tiny Swin Transformer [33] is employed as our backbone in all the experiments. The input frames are resized to have a minimum size of 448 pixels on the shorter side during inference. Motion Perception consists of six layers, and the Transformer decoder employs three layers. For the hyper-parameter settings, we set $\sigma$, $N_1$, and $N_2$ to 0.8, 20, and 10, respectively. We use RoBERTa [32] as a text encoder that is consistent with the ReferFormer and is frozen all the time.

Ablation study of LMPM. In Table 4, we present an ablation study of the baseline approach LMPM. We perform the following three experiments: (i) First, we use language queries to detect potential target object trajectories and output the best trajectory, similar to ReferFormer [55]. This variant achieves a $J & F$ score of 31.0%. It relies solely on language information with 5 randomly sampled frames and neglects global temporal context across the video, making it unable to effectively process long-term motions. (ii) With the help of Motion Perception, the $J & F$ score significantly improves by 5.3%, as it captures temporal contextual information and a global view of the entire video, which are critical for MeViS. (iii) Since MeViS contains multi-objects expressions, outputting only the object with
the highest score is insufficient. We introduce a matching mechanism to identify the target object(s), enabling our method to handle not only single-object expressions but also multi-object expressions. This variant outperforms (ii) by 0.9% in terms of $J \& F$ score.

4.3. MeViS Benchmark Results

Quantitative results. We performed a comprehensive evaluation of the MeViS dataset to assess the performance of existing methods in the more challenging motion-expression scenarios. We evaluated 1 modified image-based method VLT [10] and 4 recent state-of-the-art video-based methods, including URVOS [44], LBDT [11], MTTR [2], and ReferFormer [55], on the validation set† of MeViS. The evaluation results, presented in Table 5, indicate that the current state-of-the-art methods could only achieve performance ranging from 27.8% $J \& F$ to 31.0% $J \& F$ on the validation set of MeViS, while their results on other benchmarks like Refer-Youtube-VOS [44] and DAVIS17-RVOS [21] are usually above 60% $J \& F$. Our experiments demonstrate that while notable progress has been made in language-guided video object segmentation on existing benchmarks, the challenges presented by MeViS underline the need for further exploration of motion expression-guided video segmentation in complex scenarios. These challenges can arise from various factors, including both linguistic and visual modalities, such as the use of motion expressions and highly dynamic objects or fast-paced motions in videos, which can impact the overall performance of algorithms.

Visualizations. Figure 6 displays some of the success and failure cases of the baseline approach LMPM. Example (a) and (b) depict successful cases where LMPM effectively processes expressions of long-term motions such as “moving to the front ...” and “goes out of the screen”. In contrast, example (c) and (d) are failure cases. In example (c), the expression involves the target object disappearing and reappearing, which poses a significant challenge for the model’s global understanding of the video. In this case, our model becomes disoriented after the target object reappears. Example (d) shows a sentence describing a long-term motion while involving multiple target objects. Our method successfully identifies the multiple targets, i.e., the “two goats walking from the distance”, at the beginning of the video. However, one of the targets is lost during the later stages of the video when the motions of objects became complex and tangled. These two failure cases demonstrate the complexity and challenges of MeViS, emphasizing the importance of a strong ability to comprehend the global temporal context of the entire video and to understand the motion expression for models working on MeViS.

5. Conclusion and Discussion

The ability to effectively understand and leverage motion expressions as a primary cue for object segmentation in videos remains an unresolved challenge that requires attention in future research. The proposed large-scale benchmark MeViS provides a foundation for developing more advanced language-guided video segmentation algorithms.

Future Directions. There are many interesting research directions and remaining challenges to be addressed with the MeViS dataset. These include but are not limited to: (i) exploring new techniques for better motion understanding and modeling in both visual and linguistic modalities, (ii) designing more elegant and robust models that can effectively handle diverse motion types spanning across a range of frames, including long-term/short-term and complex motions, (iii) developing advanced models that can handle complex scenes with various types of objects and expressions, (iv) creating more efficient models that can effectively reduce the number of redundant detected objects, (v) designing effective cross-modal fusion methods to better leverage the complementary information between language and visual signals, (vi) investigating the potential of transfer learning and domain adaptation in language-guided video segmentation, and (vii) developing methods that can better handle the open-world concepts in both the visual and linguistic domain. These challenges require significant research efforts to advance the state-of-the-art in language-guided video segmentation.

---

Table 5. MeViS Benchmark Results.

<table>
<thead>
<tr>
<th>Methods</th>
<th>$J &amp; F$</th>
<th>$J$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>URVOS [44]</td>
<td>27.8%</td>
<td>25.7%</td>
<td>29.9%</td>
</tr>
<tr>
<td>LBDT [11]</td>
<td>29.3%</td>
<td>27.8%</td>
<td>30.8%</td>
</tr>
<tr>
<td>MTTR [2]</td>
<td>30.0%</td>
<td>28.8%</td>
<td>31.2%</td>
</tr>
<tr>
<td>ReferFormer [55]</td>
<td>31.0%</td>
<td>29.8%</td>
<td>32.2%</td>
</tr>
<tr>
<td>VLT+TC [10]</td>
<td>35.5%</td>
<td>33.6%</td>
<td>37.3%</td>
</tr>
<tr>
<td>LMPM (ours)</td>
<td>37.2%</td>
<td>34.2%</td>
<td>40.2%</td>
</tr>
</tbody>
</table>

†The test set is utilized for evaluation during the competition periods.
Acknowledgement. This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). This study is also supported by the research grant of NTU Presidential Postdoctoral Fellowship. The data of MeViS is released for non-commercial research purposes only.

References


