Referring Video Object Segmentation via Language-aligned Track Selection

Seongchan Kim[1*] Woojeong Jin[1*] Sangbeom Lim[2*] Heeji Yoon[1*]
Hyunwook Choi[2] Seungryong Kim[1†]
[1]KAIST [2]Korea University

Abstract

*Referring video object segmentation (RVOS) requires tracking and segmenting an object throughout a video according to a given natural language expression, demanding both complex motion understanding and the alignment of visual representations with language descriptions. Given these challenges, the recently proposed Segment Anything Model 2 (SAM2) emerges as a potential candidate due to its ability to generate coherent segmentation mask tracks across video frames, and provide an inherent spatio-temporal objectness in its object token representations. In this paper, we introduce **SOLA** (**S**election by **O**bject **L**anguage **A**lignment), a novel framework that leverages SAM2 object tokens as compact video-level object representations, which are aligned with language features through a lightweight track selection module. To effectively facilitate this alignment, we propose an IoU-based pseudo-labeling strategy, which bridges the modality gap between SAM2 representations with language features. Extensive experiments show that SOLA achieves state-of-the-art performance on the MeViS dataset and demonstrate that SOLA offers an effective solution for RVOS. Our project page is available at:* https://github.com/cvlab-kaist/SOLA.

1. Introduction

Referring video object segmentation (RVOS) [4, 7, 14, 25] aims to identify and segment a specific object throughout a video sequence based on a natural language expression. RVOS has recently attracted significant research interest due to its broad applicability in various fields, including interactive video editing and human-robot interaction. However, RVOS presents several challenges, as the model must integrate both natural language comprehension and visual understanding at both the scene and object levels.

Recently, segment anything model (SAM) [15] has emerged as a powerful models in the field of segmentation, demonstrating remarkable performance across various tasks. In particular, SAM2 [24] excels in generating segmentation masks across video frames, an essential capability for accurate object tracking and motion modeling in dynamic environments. Since generation of segmentation mask tracks inherently requires maintaining object identity over time, SAM2 implicitly captures temporal-aware object information at the video level. This characteristic makes SAM2 highly relevant for addressing the challenges of RVOS. Despite this potential, certain challenges arise in directly applying SAM2 to RVOS, as SAM2 is designed solely for segmentation and lacks an understanding of natural language. Moreover, while SAM2 effectively encodes object information, how to exploit this inherent knowledge remains an important question.

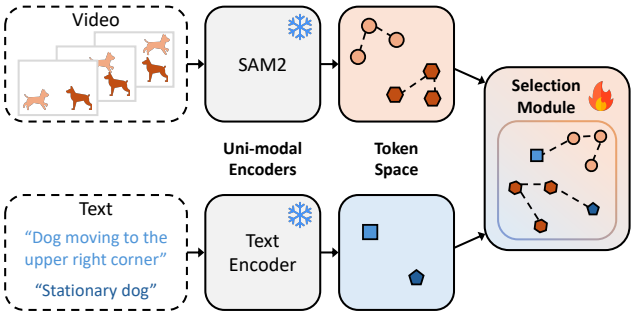

Figure 1. **Teaser.** Our method effectively bridges the modality gap by aligning the features obtained from fully frozen uni-modal encoders: the video segmentation model such as SAM2 [24] and the text encoder such as RoBERTa [20]. By directly leveraging the token representations, our approach achieves lightweight multi-modal alignment while significantly reducing the number of trainable parameters.

Based on the observation that SAM2's object token representations inherently capture temporally consistent object regions, our **SOLA** (**S**election by **O**bject **L**anguage **A**lignment) framework leverages these tokens as compact video-level representations, enabling robust motion understanding—crucial for RVOS. To associate objects with language, we introduce a lightweight *language-aligned track selection module*, effectively bridges the modality gap between SAM2's object token representations and language features, as illustrated in Figure 1. Notably, SOLA uti-

lizes solely precomputed object tokens, enabling efficient training on a single GPU while preserving SAM2's robustness and generalizability. Additionally, we introduce a novel training strategy that leverages IoU (Intersection over Union)-based pseudo-labels to supervise a simple binary classification objective, complemented by a contrastive loss designed to highlight detailed motion patterns. To utilize precomputed object tokens during training, we generate pseudo-labels based on the IoU between the mask track corresponding to the precomputed object tokens and the ground truth mask track associated with the given referring expression.

In our experiments, we evaluate our method on standard RVOS benchmarks, including MeViS [4], Ref-YouTube-VOS [16], and Ref-DAVIS [13]. Our framework achieves state-of-the-art performance on MeViS, demonstrating its effectiveness in tracking object sequences guided by complex language expressions. Additionally, our method exhibits strong generalizability and robustness across diverse settings, including zero-shot and combined dataset evaluation. Our method achieves both high-quality tracking and effective multi-modal alignment, excelling in both quantitative and qualitative evaluations.

Our main contributions are as follows:

- We propose SOLA, a novel framework that, for the first time, utilizes SAM2's object token representations for RVOS. We hypothesize that these tokens inherently encode temporal-aware objectness, enabling effective motion modeling of an object.
- By introducing a lightweight language-aligned track selection module that relies exclusively on object tokens. This approach allows for the use of precomputed tokens, enabling efficient training on a single GPU.
- We adopt a novel training strategy that utilizes IoU-based pseudo-labels for object tokens, enabling our language-aligned track selection module to effectively bridge the modality gap between SAM2's object representations and language features.
- Our method achieves new state-of-the-art results on the MeViS dataset [4] and demonstrates strong generalization on Ref-YouTube-VOS [16] and Ref-DAVIS [13], excelling in both quantitative and qualitative evaluations.

2. Related Work

Referring video object segmentation. RVOS requires segmenting objects by capturing both action and appearance from video sequences based on a given expression. RVOS was first introduced by Gavrilyuk et al. [7] with the A2D-Sentences benchmark. Since then, RVOS has garnered significant attention, leading to the development of benchmarks such as Ref-YouTube-VOS [16], Ref-DAVIS [13], and MeViS [4].

Recently, query-based models [2, 4, 9, 21, 28] have achieved impressive performance by leveraging object query tokens. These tokens are expected to capture spatial properties, appearance, and temporal dynamics while maintaining temporally consistent object mask tracks. Other approaches [8, 17] enhance language alignment by employing object tokens pre-aligned with language features. Thereby, solving RVOS demands a model that ensures temporal consistency while effectively linking textual descriptions with visual representations containing various object information.

Segment anything model. SAM [15] is known as a breakthrough in foundation models for image segmentation, with a unique ability to segment any object within an image using interactive prompts. Building on SAM, SAM2 [24] extends its capabilities to video segmentation through a memory-based transformer. SAM2's memory stores information about target objects and past interactions, enabling it to perform segmentation more accurately and efficiently while maintaining strong generalization performance.

There are previous approaches [11, 18] that utilizes SAM or SAM2 in RVOS task. However, these approaches predominantly use them only at the prompting level, treating them merely as powerful mask generation tools without tapping into their rich internal representations for more advanced video-level object understanding. Ref-SAM [18] processes textual inputs by projecting them into sparse and dense prompts, but these prompts mainly tied to image-level, propagating from a selected object through an implicit tracking module. As a result, it struggles to handle complex motion across an entire video sequence or to differentiate among objects of similar classes. Similarly, AL-RefSAM 2 [11] assigns the spatio-temporal reasoning capability on GPT-4 [1] and Grounding DINO [19]. They select pivot frames via GPT, detect objects using Grounding DINO, and then pick specific bounding boxes that best match the given language expression with GPT again. Consequently, these methods struggle to leverage video-level context and capture the object-level details necessary for understanding complex motion and inter-object distinctions.

3. Method

3.1. Overview

For given T frames of video clip $\mathcal{V} = \{I^t\}_{t=1}^{T}$, each frame $I^t \in \mathbb{R}^{C\times H\times W}$ has height H, width W, and C channels. In RVOS, a language expression is provided as additional input, and the text encoder tokenizes it into text tokens $\mathcal{E} \in \mathbb{R}^{N_w\times D}$, where N_w denotes the number of tokenized words. The objective of RVOS is to generate binary mask tracks $\mathcal{B} = \{B^t\}_{t=1}^{T}$, where each mask $B^t \in \{0,1\}^{H\times W}$ corresponds to the referred objects at time t.

To address this, we propose a novel framework SOLA, which, for the first time, leverages SAM2 object token to ef-

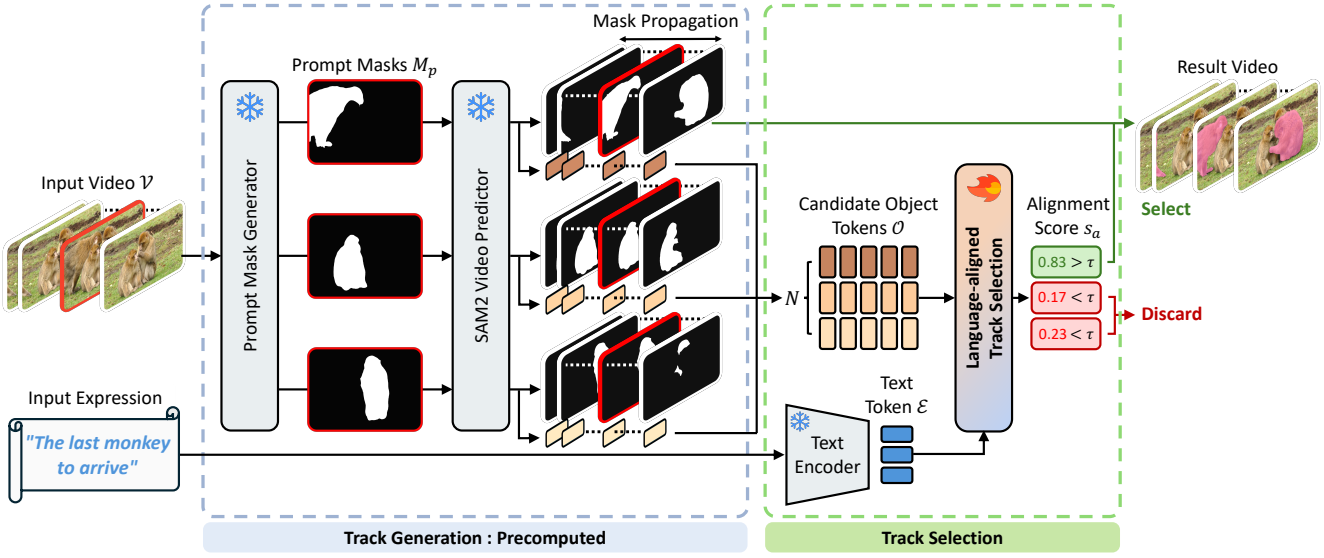

Figure 2. **Overall pipeline of the proposed SOLA framework.** Our method selects the correct object mask track among candidates via a language-aligned track selection module. We first generate candidate mask tracks and corresponding object tokens from the fully frozen SAM2. These tokens are then aligned with language expressions, producing alignment scores that indicate selection probabilities. Mask tracks with scores above a predefined threshold are selected and merged into the final binary segmentation mask. By leveraging precomputed object tokens from SAM2, our approach minimizes trainable parameters, enabling efficient training on a single GPU.

ficiently bridge SAM2's knowledge with language features, as illustrated in Figure 2. Specifically, we first generate N candidate mask tracks $\mathcal{M} \in \{0,1\}^{N\times T\times H\times W}$ and their corresponding object tokens $\mathcal{O} \in \mathbb{R}^{N\times T\times D}$ using SAM2, where D denotes the feature dimension. Next, we select the N_v valid mask tracks $\mathcal{M}_v \in \mathbb{R}^{N_v\times T\times H\times W}$ that align with the given expression from the candidates $\mathcal{M}$ through our lightweight *language-aligned track selection module*. This module efficiently bridges the modality gap between the object token representations of SAM2 and language features. Through this process, our framework obtain high-quality object mask tracks that are precisely aligned with complex natural language expressions.

3.2. Preliminary - SAM2

In this section, we provide an overview of the Segment anything model 2 (SAM2) [24], which is a promptable video segmentation model that consists of an image encoder, a prompt encoder, a mask decoder, and a memory encoder.

Image encoder. The image encoder extracts high-resolution image embeddings from individual frames. These spatial features retains detailed object and scene information. The embeddings are later conditioned on user prompts and past memory for mask generation.

Prompt encoder. The prompt encoder supports three types of user inputs: points $\mathcal{P}_g$, bounding boxes $\mathcal{P}_b$, and masks $\mathcal{P}_m$. It generates prompt tokens representing user inputs that specify the target object for segmentation.

Mask decoder. The mask decoder takes memory-conditioned image embeddings from the memory attention layer and prompt tokens from the prompt encoder as inputs. It generates three mask predictions, each paired with a predicted Intersection over Union (IoU) score and an output mask token. These mask tokens serve as memory values. The final mask is selected based on the highest IoU score, and its associated token is converted into an object pointer $\mathcal{O}^{i,t} \in \mathbb{R}^D$ at time t to update the memory for $i = 1, \ldots, N$ and $t = 1, \ldots, T$.

Memory module. SAM2 incorporates a memory module that conditions the features of the current frame on both previous frames and user-provided prompts. Each memory entry consists of two elements: the spatial embedding fused with the predicted mask and a corresponding mask token. By cross-attending to this memory, the model effectively captures fine-grained correspondences and spatial information, ensuring temporal consistency across frames.

3.3. Track generation

As our method select the valid mask tracks among the candidates, we first prompt SAM2 to ensure it generates all the objects exist in a video. Since some objects only appear momentarily, we adopt a strategy of selecting frames at predefined frame intervals as a prompt frame I_p for mask generation.

Prompt mask generation. We use two types of input prompts: grid points $\mathcal{P}_g$, and bounding boxes $\mathcal{P}_b$, along with frame I_p. The bounding boxes are obtained from external object detection models, only for inference to efficiently

capture potential objects. These prompts are used to generate N binary masks $M_p \in \{0,1\}^{N\times H\times W}$, as

$$M_p = \text{SAM2}_{\text{Image}}(I_p; \{\mathcal{P}_g, \mathcal{P}_b\}), \tag{1}$$

where $\text{SAM2}_{\text{Image}}(\cdot)$ denotes the SAM2 image predictor.

Mask track propagation. The generated masks M_p are propagated both forward and backward across the entire video $\mathcal{V}$ by the SAM2 video predictor $\text{SAM2}_{\text{Video}}(\cdot)$, to obtain mask tracks $\mathcal{M}$ and the corresponding object tokens $\mathcal{O}$:

$$\mathcal{O}, \mathcal{M} = \text{SAM2}_{\text{Video}}(\mathcal{V}; M_p). \tag{2}$$

Notably, grid point prompts $\mathcal{P}_g$ cover both the foreground objects and the surrounding background, as the points are evenly distributed across the frame.

3.4. Object representation

Revisiting the architecture of SAM2 [24], the object pointer obtained from the spatiotemporal-aware memory bank serves as an auxiliary high-level representation of the objects to be segmented. Specifically, each object pointer $\mathcal{O}^{i,t}$, corresponding to a mask $\mathcal{M}^{i,t}$ within a frame, is hypothesized to encode certain object-level information at that timestep t. Consequently, the sequence of these object pointers accumulated over time can be considered as temporal-aware object information, which inherently captures *object motion*. Based on this intuition, as we generate N candidate mask tracks $\mathcal{M}$ using SAM2, we simultaneously extract T object pointers $\{\mathcal{O}^{i,t}\}_{t=1}^{T}$ per track i and concatenate them along the temporal dimension. We define the resulting representation as the *object token*, denoted as $\mathcal{O}^i$, which encapsulates both spatial and motion characteristics of the object over time. Motivated by these considerations, we utilizes these object tokens to model complex motions of objects.

3.5. Track selection

Once we successfully gather N candidate mask tracks and their consistent feature representations, $\mathcal{O}$, we can address RVOS by selecting tracks that semantically match the given expression. To determine this, we introduce a lightweight *language-aligned track selection module*, which aligns SAM2's token representations and language features, thus outputs scores reflecting correspondence between each mask track and given language expression. We define these scores as alignment score s_a, representing the probability of selection. Thus, the module takes object tokens $\mathcal{O}$ and text tokens $\mathcal{E}$ as input, and produces $s_a \in \mathbb{R}^N$ along with an alignment token $\mathcal{O}_a \in \mathbb{R}^{N\times D}$.

$$\mathcal{O}_a, s_a = \text{TS}(\mathcal{O}; \mathcal{E}), \tag{3}$$

where $\text{TS}(\cdot)$ denotes the track selection module. As depicted in Figure 3, the track selection module is composed of short-term motion encoder followed by object-language alignment layers and language aligned motion aggregation module.

Short-term motion encoder. Since RVOS deals with video data, target objects are not limited to appearance cues; they are often defined by key motion cues. Thus, vision-language alignment in RVOS requires not only frame-level object features but also temporal encoding. The initial short term motion encoder is to encode the momentary motions of objects, by implementing 1D convolutional network along temporal dimension of each object token. The output object token is $\mathcal{O}^i \in \mathbb{R}^{T'\times D}$, where T' denotes the reduced temporal dimension.

Object-language alignment layer. The object-language alignment layer, repeats L times, sequentially performs three types of attention layers: inter-object attention, motion attention, and object-to-language attention.

Understanding an object's motion implies both its interactions with the surrounding environment and its internal dynamicity. We address these temporal and spatial contexts using motion attention and inter-object attention, respectively. Both attentions are standard self-attention [26], but each operate along a different dimension for distinct pursuit.

Inter-object attention is applied to all object tokens $\mathcal{O}^t \in \mathbb{R}^{N\times D}$ within the same frame t. As we aforementioned in Section 3.3, using grid point prompts allows us to obtain mask tracks correspond to both foreground and background regions. Thus inter-object attention between all these tokens captures both object relations and interactions between objects and surroundings, leading to a comprehensive understanding of the global context. On the other hand, motion attention aims to aggregate long-term motion information for each object throughout the video, operating along the temporal dimension of each object token $\mathcal{O}^i \in \mathbb{R}^{T'\times D}$.

Subsequently, we employ object-to-language cross-attention to align visual object tokens $\mathcal{O}$ with language features $\mathcal{E}$, generating language-aligned object token $\mathcal{O}' \in \mathbb{R}^{N\times T'\times D}$. Finally, these alignment tokens $\mathcal{O}'$ serves a input to the language-aligned object aggregation block for further processing.

Language-aligned object aggregation. The language-aligned object aggregation block takes the language-aligned object token $\mathcal{O}'$ as input and outputs s_a along with $\mathcal{O}_a$ which serves as the object representative. We define $\mathcal{O}_a \in \mathbb{R}^{N\times D}$ as the weighted sum of object tokens, computed using the frame weighting matrix $w_a \in \mathbb{R}^{N\times T'}$, given by:

$$w_a = \text{softmax}(\text{Avg}_{N_w}(\mathcal{O}'\mathcal{E}^{\text{T}})), \tag{4}$$

where $\text{Avg}_{N_w}(\cdot)$ represents the mean along the N_w dimension. Using such operation, we can obtain $\mathcal{O}_a$ and s_a as follows:

$$\begin{aligned} \mathcal{O}_a &= \text{Avg}_T(w_a \otimes \mathcal{O}'), \\ s_a &= \text{sigmoid}(\text{Avg}_T(\text{Avg}_{N_w}(\mathcal{O}'\mathcal{E}^{\text{T}}))), \end{aligned} \tag{5}$$

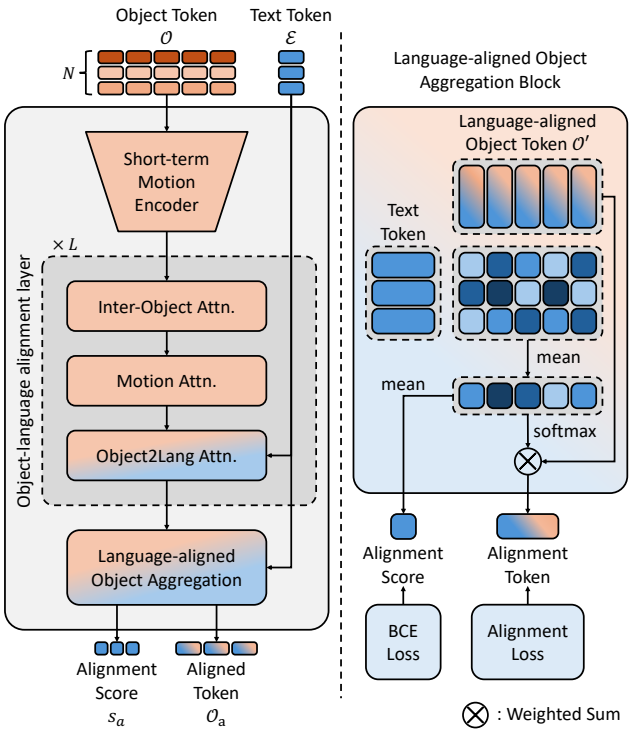

Figure 3. **Architecture of the language-aligned track selection module**, which takes object tokens and text tokens as inputs, aligning these representations to effectively capture object dynamics.

where $\otimes$ denotes element-wise multiplication operation, and $\text{Avg}_T(\cdot)$ represents the mean along the T dimension.

Here, $\mathcal{O}_a$, is not only aligned with the language expression but has also incorporated temporally aggregated motion information, making it a rich video-level object representation. Then, the alignment score of each object s_a^i is mapped to the $[0, 1]$ range, following a sigmoid activation. The i-th mask track is then selected or discarded based on whether s_a^i exceeds threshold τ. The selected mask tracks $\mathcal{M}_v$ are merged to form the final output binary mask track $\mathcal{B}$ for the given expression.

3.6. Training objective

Pseudo labeling. As we utilize SAM2 object token representations in a fully frozen state, our goal is to select the correct object tokens, representing SAM2 generated mask tracks, that matches to the referred object in a given expression. However, since RVOS datasets provide only language expressions paired with ground-truth mask tracks, there is no explicit label for each generated mask tracks and corresponding object token. This means that direct supervision for learning alignment between SAM2 object tokens and language features is unavailable in current setting. To address this issue, we introduce a novel *IoU-based pseudo-labeling* strategy that enables our model to directly identify which generated mask tracks correspond to a given expression. Specifically, we compute the mean Intersection over Union (mIoU) between each candidate mask track and the ground-truth track associated with that expression. Candidate object tokens exceeding predefined mIoU threshold τ are labeled as positive samples, while the rest are labeled as negative samples. The core motivation of this approach is to create a reliable supervision signal that bridges the gap between the frozen SAM2 object tokens and the language expressions. This mIoU-based pseudo-labeling strategy not only provides a clear supervision but also simplifies training to a straightforward binary classification objective.

Loss functions. The total loss $\mathcal{L}$ is a combination of Binary Cross-Entropy (BCE) loss $\mathcal{L}_{\text{BCE}}$ and alignment loss $\mathcal{L}_{\text{align}}$: $\mathcal{L} = \lambda_1 \mathcal{L}_{\text{BCE}} + \lambda_2 \mathcal{L}_{\text{align}}$.

The BCE loss $\mathcal{L}_{\text{BCE}}$ is applied to enforce alignment between the object features and the language, as follows:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\Big(y^i \log(s_a^i) + (1-y^i)\log(1-s_a^i)\Big), \quad (6)$$

where y^i represents the pseudo binary classification label. Alignment score s_a is defined based on whether the mask track $\mathcal{M}^i$ of $\mathcal{O}^i$ corresponds to the target object designated by the $\mathcal{E}$.

The alignment loss $\mathcal{L}_{\text{align}}$ is a modified form of contrastive loss, designed to encourage each alignment token $\mathcal{O}_a^i$ to push mismatched sentences away in semantic space, and vice versa. We define the positive anchor $\mathcal{A}_p \in \mathbb{R}^D$ as the mean vector of the text tokens $\mathcal{E}$, ensuring semantic closeness between corresponding tokens. In contrast, the negative anchors $\mathcal{A}_n \in \mathbb{R}^{N_{\text{neg}} \times D}$ consists of N_{neg} learnable embeddings, which are trained to represent a distinct negative latent space, forcing the tokens to be pushed farther apart. $\mathcal{L}_{\text{align}}$ is defined as follows:

$$\mathcal{L}_{\text{align}} = -\frac{1}{N}\sum_{i=1}^{N}\left(y^i \mathcal{L}_{\text{pos}}(\mathcal{O}_a^i) + (1-y^i)\mathcal{L}_{\text{neg}}(\mathcal{O}_a^i)\right), \quad (7)$$

where

$$\mathcal{L}_{\text{pos}} = d(\mathcal{O}_a^i, \mathcal{A}_p) - \sum_{j=1}^{N_{\text{neg}}} d(\mathcal{O}_a^i, \mathcal{A}_n^j),$$

$$\mathcal{L}_{\text{neg}} = d(\mathcal{O}_a^i, \mathcal{A}_n^{k^*}) - d(\mathcal{O}_a^i, \mathcal{A}_p) - \sum_{j=1, j\neq k^*}^{N_{\text{neg}}} d(\mathcal{O}_a^i, \mathcal{A}_n^j). \quad (8)$$

Here, the distance function is computed as $d(\mathbf{x}, \mathbf{y}) = 1 - \cos(\mathbf{x}, \mathbf{y})$, where $\cos(\mathbf{x}, \mathbf{y})$ is the cosine similarity between vectors $\mathbf{x}$ and $\mathbf{y}$. The index k^* represents the closest negative anchor to the alignment token.

4. Experiments

4.1. Datasets and evaluation metrics

Dataset. We evaluate our method on three video datasets: MeViS [4], Ref-YouTube-VOS [25], and Ref-DAVIS [13]. MeViS, a newly established dataset focused on motion information analysis, comprises 2,006 videos and 28,570 sentences, which are divided into three subsets: the training set with 1,712 videos, the validation set with 140 videos, and the testing set with 154 videos. Ref-YouTube-VOS is the largest RVOS dataset, containing 3,978 videos with approximately 13,000 annotations. Ref-DAVIS builds upon DAVIS17 [23] by incorporating linguistic annotations for a variety of objects, featuring a total of 90 videos.

Evaluation metrics. Following [4, 9, 21], we evaluate our method on the MeViS dataset using the commonly used $\mathcal{J}\&\mathcal{F}$ metrics. The $\mathcal{J}$ metric, or region similarity, calculates the Intersection over Union (IoU) between predicted and ground-truth masks to assess segmentation quality, while the $\mathcal{F}$-measure evaluates contour accuracy. To provide an overall effectiveness score for our method, we report the average of these two metrics, referred to as $\mathcal{J}\&\mathcal{F}$.

4.2. Implementation details

Precomputing SAM2 object tokens. Since we utilize SAM2 in a fully frozen state, training focuses exclusively on the language-aligned selection module. Inspired by FuseMix [27], we precompute SAM2 mask tracks on the RVOS dataset, eliminating the need for on-the-fly inference during training. This approach enables efficient training, taking approximately 7 hours on a single RTX 3090 GPU using the MeViS [4] dataset.

Track generation. We generate mask tracks using SAM2-L [24], prompted by grid points and bounding boxes obtained from Grounding DINO-T [19] every fourth frame. To avoid generating redundant tracks, we apply IoU-based filtering, similar to Non-Maximum Suppression (NMS) [22], propagating only distinct prompt masks.

Language-aligned track selection module. We employ pre-trained RoBERTa [20] as the text encoder. Hyperparameters are set as follows: $N_{\text{neg}} = 32$ for number of negative anchors, and loss weights of $\lambda_1 = 1.0, \lambda_2 = 0.3$ and $\tau = 0.5$ for selection thresholding.

4.3. Quantitative results

Main results. Table 1 presents the quantitative results of our method on the MeViS [4] dataset, which is widely regarded as the most challenging benchmark in the RVOS field. Our method achieves state-of-the-art performance, underscoring its effectiveness. Additionally, compared to previous methods, SOLA significantly reduces the number of trainable parameters to 32.9M while achieving the highest $\mathcal{J}\&\mathcal{F}$ score of 48.6. This low number of trainable parameters is achieved through our design, which relies solely exclusively on object tokens.

Methods	# of trainable parameters	Metrics		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
URVOS [25]	-	27.8	25.7	29.9
LBDT [5]	95.6 M	29.3	27.8	30.8
MTTR [2]	-	30.0	28.8	31.2
ReferFormer [28]	112.9 M	31.0	29.8	32.2
VLT+TC [3]	<u>38.3 M</u>	35.5	33.6	37.3
LMPM [4]	66.4 M	37.2	34.2	40.2
HTR [21]	-	42.7	39.9	45.5
DsHmp [9]	92.4 M	<u>46.4</u>	<u>43.0</u>	<u>49.8</u>
SOLA	**32.9 M**	**48.6**	**45.2**	**52.1**

Table 1. **Quantitative comparison on MeViS.** The best results are highlighted in **bold**, and the second-best results are <u>underlined</u>.

Methods	Ref-YouTube-VOS			Ref-DAVIS		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
ReferFormer [12]	35.0	34.2	35.8	40.5	36.8	44.2
LMPM [4]	31.5	30.0	32.9	39.9	36.7	43.2
DsHmp [9]	45.8	43.7	47.9	42.6	37.8	47.3
SOLA	**47.9**	**44.3**	**51.5**	**45.4**	**43.0**	**47.7**

Table 2. **Zero-shot quantitative comparison on Ref-YouTube-VOS and Ref-DAVIS.** The best results are in **bold**. The models are trained on the training set of MeViS and evaluated on Ref-YouTube-VOS and Ref-DAVIS.

Methods	MeViS			Ref-YouTube-VOS		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
ReferFormer [12]	36.6	34.1	39.1	46.8	46.2	47.5
LMPM [4]	40.5	37.8	43.2	37.6	36.0	39.2
DsHmp [9]	42.5	37.5	47.4	51.4	48.5	54.3
SOLA	**48.9**	**45.2**	**52.6**	**55.4**	**52.0**	**58.8**

Table 3. **Quantitative comparison on combined dataset.** The best results are in **bold**. The models are jointly trained on the training sets of MeViS and Ref-YouTube-VOS and evaluated separately on their respective evaluation datasets.

Zero-shot evaluation. Since our method utilizes object tokens obtained from SAM2 in a fully frozen state, we conducted a zero-shot experiment to evaluate its generalization capability. We trained our model on the MeViS [4] dataset and evaluated it on the Ref-YouTube-VOS [25] and Ref-DAVIS [13] datasets. As shown in Table 2, SOLA achieved superior performance, surpassing the previous state-of-the-art methods. This demonstrates that our approach not only effectively bridges the modality gap between SAM2 token features and language features but also inherits the intrinsic robustness of SAM2 representations.

Combined dataset evaluation. Table 3 presents the quantitative results obtained by training on a naively combined

Methods	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
w/o selection module	36.9	30.0	43.8
w/ selection module	**48.6**	**45.2**	**52.1**

Table 4. **Ablation study on our proposed selection method.**

Methods	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
w/o $\mathcal{L}_{\text{align}}$	44.5	41.4	47.6
w/ $\mathcal{L}_{\text{align}}$	**48.6**	**45.2**	**52.1**

(a) Different loss functions.

Inter-object attn.	Motion attn.	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
✗	✓	44.3	41.6	47.0
✓	✗	44.9	42.2	47.0
✓	✓	**48.6**	**45.2**	**52.1**

(b) Effects of employing different types of attention layers.

Table 5. **Ablation studies on various settings of our method.**

dataset of MeViS [4] and Ref-YouTube-VOS [25], followed by individual evaluations on each dataset. The results highlight the robustness of our method, as it maintains strong performance across different datasets without requiring dataset-specific tuning. Furthermore, the scalability of our approach is evident, as it effectively leverages multiple datasets without performance degradation, suggesting its potential for broader generalization in RVOS.

4.4. Ablation studies

We conduct our ablation studies on the MeViS [4] dataset to examine the effectiveness of our proposed language-aligned selection module and its components.

Effect of the proposed selection method. The quantitative results in Table 4 demonstrate that our language-aligned track selection module effectively interprets complex language expressions. *w/o selection module* refers to an baseline approach that relies solely on Grounding DINO [19], which detects objects at the frame level by understanding the correspondence between text and objects in an image. However, this approach does not incorporate temporal information, limiting its ability to associate objects with motion patterns or temporal events described in the text. Consequently, it struggles with understanding complex expressions that require video-level reasoning. In contrast, *w/ selection module* represents our framework, SOLA, which selects the referred object tracks from the candidates by leveraging language-aligned object tokens. By considering both spatial and temporal information, our selection module enables a more comprehensive understanding of complex expressions, leading to improved RVOS.

Ablation on losses. In Table 5a, we evaluate the model's performance under different loss configurations. When using only BCE loss (w/o $\mathcal{L}_{\text{align}}$), we observe a performance reduction of 4.1 $\mathcal{J}\&\mathcal{F}$ compared to the combined setting of BCE and alignment loss (w/ $\mathcal{L}_{\text{align}}$). This result indicates that alignment loss enhances the model's discriminative ability, improving its understanding complex motions and enabling more precise alignment with given expression.

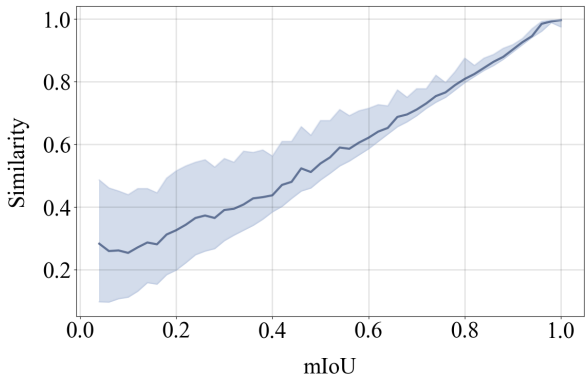

Figure 4. **Spatial and motion information in object tokens.** The bold line represents the mean similarity, while the shaded region indicates the variance. The results show a certain correlation: as the mIoU between mask tracks increases, the similarity between their associated tokens also rises nearly proportionally. This tendency suggests that object tokens inherently capture spatial information, implicitly encoding object motions over time.

Ablation on different types of attention. Table 5b shows the model's performance with different attention configurations. Using only motion attention allows the model to aggregate long-term temporal information across frames, improving motion modeling but neglecting object relationships and scene-level context within each frame. Conversely, using only inter-object attention encodes spatial relationships among objects including surrounding backgrounds, but lacks temporal awareness. Combining both attention types, our method effectively captures temporal object dynamics as well as spatial interactions, resulting in comprehensive global context understanding.

4.5. Analysis on object token of SAM2

As our method solely relies on the object tokens of SAM2, it is important to investigate whether they contain sufficient information to model diverse aspects of corresponding objects. To this end, we conducted two experiments to explore whether these tokens capture motion information from their corresponding masks and whether they possess a minimally sufficient level of semantic knowledge to align with language expressions.

First, we analyzed the relationship between object tokens and their corresponding masks by comparing the cosine similarity of object tokens based on their mIoU (mean Intersection over Union) of their masks. As illustrated in Figure 4, the similarity between tokens tends to increase nearly proportionally to the mIoU. This suggests a correla-

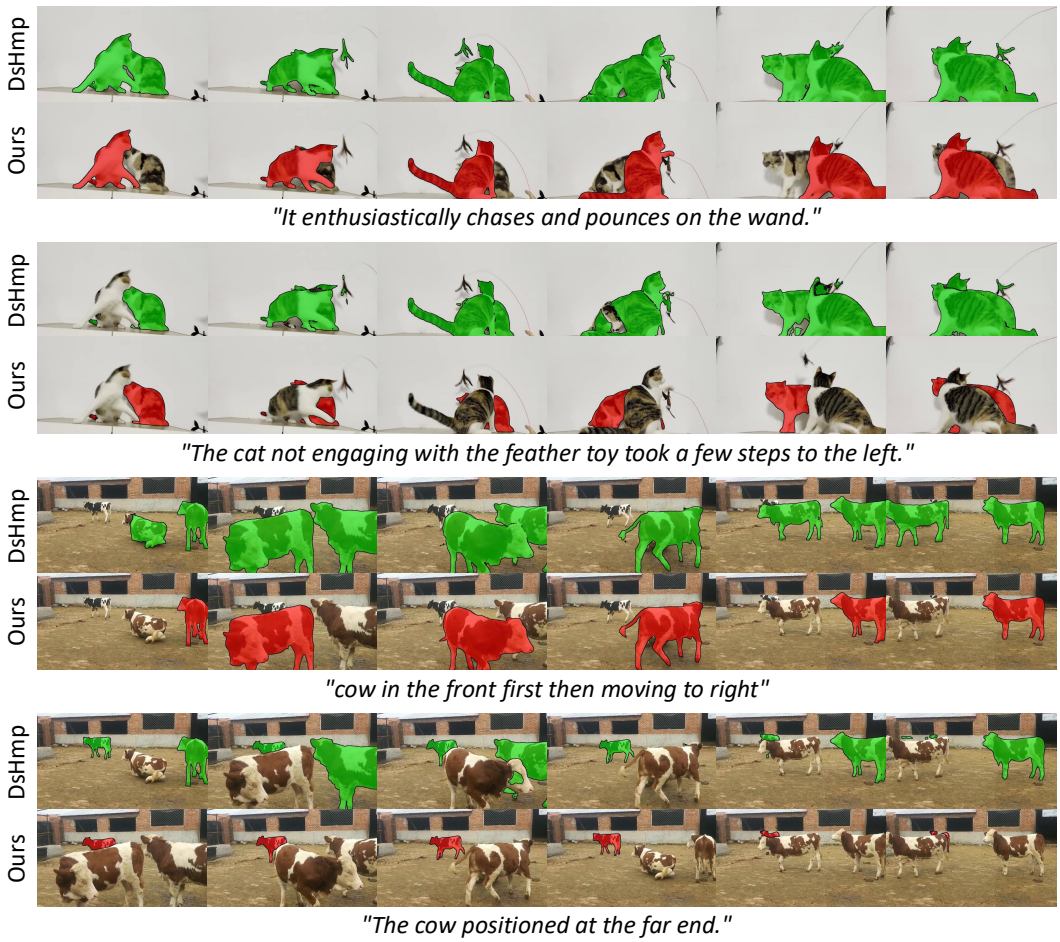

Figure 5. **Qualitative results of our model on MeViS.** SOLA shows its ability to understand complex language expressions.

tion between the spatial proximity of masks and the similarity of their object tokens. It provides a reasonable indication that object tokens implicitly encode spatial information, which can be extended to spatial trajectories when temporally connected across frames.

Second, to assess the degree of semantic content encoded in the object tokens, we conducted a simple classification task using the PASCAL-VOC dataset [6], an image dataset containing 20 object categories with pixel-level segmentation and class annotations per mask. Specifically, we added a linear classification head on top of the object tokens and trained it to predict object categories. The classifier achieved a maximum accuracy of 85.3%, indicating that the tokens may possess at least a basic ability to differentiate between object classes, and thus contain meaningful semantic information.

These findings suggest that SAM2 object tokens obtain intrinsic properties such as object motion and semantic information. Their potential to encode such *objectness* makes them a promising candidate for aligning with language expressions, offering a lightweight module design.

4.6. Qualitative results

In Figure 5, our proposed method demonstrates its ability to understand complex language expressions. The model captures both appearance cues—such as *"The cat"* and *"The cow"* attributes—and complex motion cues, including *"moving to right"*. SOLA can select the referred object even when the expression relies solely on motion (e.g., *"chases"*, *"pounces"*).

5. Conclusion and discussion

We proposed SOLA, a novel framework that leverages SAM2 object tokens as compact video-level object representation. We align these object tokens with language features using a lightweight track selection module with only 32.9M trainable parameters. Additionally, we employ an IoU-based pseudo-labeling strategy to effectively bridge the modality gap between SAM2 representations and language features. Our experiments demonstrate that SOLA achieves state-of-the-art results on the MeViS dataset. This validates the effectiveness of SOLA in addressing the challenges of complex motion understanding and multi-modal alignment in RVOS.

References

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

[2] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4975–4985. IEEE Computer Society, 2022. 2, 6

[3] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 6

[4] 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, 6, 7

[5] 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. 6

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 8

[7] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5958–5966. IEEE Computer Society, 2018. 1, 2

[8] 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. 2

[9] 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, 6, 1, 3, 4, 5

[10] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1

[11] Shaofei Huang, Rui Ling, Hongyu Li, Tianrui Hui, Zongheng Tang, Xiaoming Wei, Jizhong Han, and Si Liu. Unleashing the temporal-spatial reasoning capacity of gpt for training-free audio and language referenced video object segmentation. *arXiv preprint arXiv:2408.15876*, 2024. 2

[12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 6, 1

[13] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV*, 2018. 2, 6

[14] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *14th Asian Conference on Computer Vision*, pages 123–141. Springer, 2019. 1

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 2

[16] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2018. 2

[17] 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. 2

[18] Yonglin Li, Jing Zhang, Xiao Teng, Long Lan, and Xinwang Liu. Refsam: Efficiently adapting segmenting anything model for referring video object segmentation. *arXiv preprint arXiv:2307.00997*, 2023. 2

[19] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 6, 7

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. 1, 6, 2

[21] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, Mubarak Shah, and Ajmal Mian. Temporally consistent referring video object segmentation with hybrid memory. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2, 6

[22] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th international conference on pattern recognition (ICPR'06)*, pages 850–855. IEEE, 2006. 6

[23] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 724–732. IEEE, 2016. 6

[24] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1, 2, 3, 4, 6

[25] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European Conference on Computer Vision*, pages 208–223, 2020. 1, 6, 7

[26] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4

[27] Noël Vouitsis, Zhaoyan Liu, Satya Krishna Gorti, Valentin Villecroze, Jesse C Cresswell, Guangwei Yu, Gabriel Loaiza-Ganem, and Maksims Volkovs. Data-efficient multimodal fusion on a single gpu. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27239–27251, 2024. 6, 1

[28] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4964–4974. IEEE, 2022. 2, 6

Referring Video Object Segmentation via Language-aligned Track Selection

Supplementary Material

A. Additional qualitative results

Qualitative results on MeViS. Figure A.1 presents the qualitative results on MeViS [4], comparing the performance of DsHmp [9] with SOLA. Our approach consistently demonstrates superior capability in accurately selecting the target object as specified by the referring expression. Specifically, Figure A.2 illustrates scenarios involving a single video with two distinct expressions. SOLA accurately identifies the precise object corresponding to each expression, whereas DsHmp demonstrates limitations in distinguishing between objects described by different expressions. Figure A.3 illustrates a scenario where the given expression exclusively describes motion-related information (e.g., *"Going right."*). Our language-aligned track selection module can establish correspondence with the expression using motion cues from the language alone, independently of appearance-based features.

Qualitative results on Ref-YouTube-VOS. Figure A.4 presents the qualitative results on the Ref-YouTube-VOS [25] dataset in a zero-shot setting, where the model has been trained on MeViS dataset. The results highlights our model's remarkable capability to generalize across diverse videos and expressions, despite not having seen the dataset during training. This generalization underscores the strength of our approach in leveraging the intrinsic robustness of SAM2 representations.

B. Results on corrupted setting

To demonstrate the robustness of our method, we evaluated it on a perturbed dataset with ImageNet-C [10] derived corruption. we intentionally corrupted all video frames with gaussian noise or motion blur, simulating common distortions in real-world scenarios such as low-light environments or rapid camera movements. Since these perturbations represent data types not originally present in the dataset, our method's ability to effectively handle them shows its robustness inherited from SAM2 and highlights its suitability for practical applications. Table A.1 presents the quantitative results, showing that our proposed method outperforms previous approaches [4, 9, 12] even under corruption scenarios.

Qualitative results on MeViS with image corruption. Figures A.5 and A.6 visualize the results presented in Table A.1. These results demonstrate that SOLA consistently retains its ability to select the correct object even in corrupted environments.

Methods	Algorithm	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
ReferFormer [12]	Motion blur	26.3	25.4	27.1
LMPM [4]		33.3	31.2	35.4
DsHmp [9]		38.0	35.0	41.1
SOLA		**39.8**	**36.6**	**43.0**
ReferFormer [12]	Gaussian noise	26.9	24.0	29.9
LMPM [4]		36.0	33.4	38.6
DsHmp [9]		43.4	39.5	47.2
SOLA		**44.4**	**40.5**	**48.3**

Table A.1. **Quantitative result on a corrupted version of MeViS.** The best results are in **bold**. The models are trained on the original training set and evaluated on the corrupted version of the validation set. The image corruption algorithms are derived from ImageNet-C [10], with corruption severity 5.

C. Additional ablation studies on MeViS

Existence of background object tokens. The quantitative results presented in Table A.2a underscore the effectiveness of incorporating background object tokens during both training and inference. During training, background object tokens refer to object tokens corresponding to mask tracks that have low IoU with the ground-truth mask track, while during inference, they are derived from mask tracks obtained using grid point prompts. Given that the inter-object attention is designed to capture object relationships and scene-level understanding, the inclusion of background object tokens in both training and inference significantly enhances performance. This comprehensive interactions between foreground and background objects proves its effectiveness, enabling a more enhanced video-level understanding of language.

Ablation on the number of object-language alignment layers. Table A.2b shows the results of using different numbers of attention block layers. Our method achieves the highest performance when two layers are adopted, compared to the settings with one or three layers.

D. Detailed implementation details

Precomputing SAM2 object tokens. Since our method operates with a fully frozen SAM2 and trains only the language-aligned selection module using object tokens, we adopt a highly efficient training strategy similar to FuseMix [27]. Specifically, we first perform SAM2 mask propagation on the given RVOS dataset to generate can-

Train	Inference	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
✗	✗	45.7	42.4	48.9
✓	✗	47.5	43.9	51.1
✓	✓	**48.6**	**45.2**	**52.1**

(a) Effects of including background object tokens.

# of Alignment Layers	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
1	42.5	40.0	45.1
2	**48.6**	**45.2**	**52.1**
3	48.2	44.8	51.5

(b) Effects of the number of object-language alignment layers.

Table A.2. **Additional ablation studies on various settings of our method.**

didate mask tracks and their corresponding object tokens in advance. By precomputing these tokens beforehand, we eliminate the need for SAM2 inference during training phase, allowing us to focus solely on optimizing the language-aligned track selection module. The entire training process, using the MeViS [4] training dataset, takes approximately 7 hours on a single RTX 3090 GPU.

Track generation. We employ grid points and bounding boxes from the object detection model, Grounding DINO (GDINO)-T [19] every fourth frame to generate prompt masks, which serve as input for the SAM2-L [24] video predictor. To reduce redundant mask track generation, we filter out similar prompt masks based on their Intersection over Union (IoU) scores. Specifically, we first propagate the mask track sequence starting from the largest prompt mask. Then, for each subsequent prompt mask, we filter it out if its IoU with the previously generated mask tracks at the corresponding frame exceeds 0.7, ensuring that only distinct prompt masks propagate new tracks.

Language-aligned track selection module. We employ pre-trained RoBERTa [20] as the text encoder. Training is conducted over 13 epochs, with an initial learning rate of 5e-6 that gradually decreases throughout training. We set the hyperparameter values for λ_1, λ_2, N_{neg}, τ to 1.0, 0.3, 32, and 0.5, respectively.

E. Limitations and future works

While our approach effectively solve RVOS, certain aspects remain beyond the scope of our work. The training objectives of the text encoder and the RVOS model differ: the text encoder is trained to identify the best matching words from the vocabulary, while the RVOS model focuses on extracting key cues from sentences essential for locating the corresponding objects. In our future work, we aim to explore tuning the text encoder to capture features that are particularly beneficial for the RVOS task.

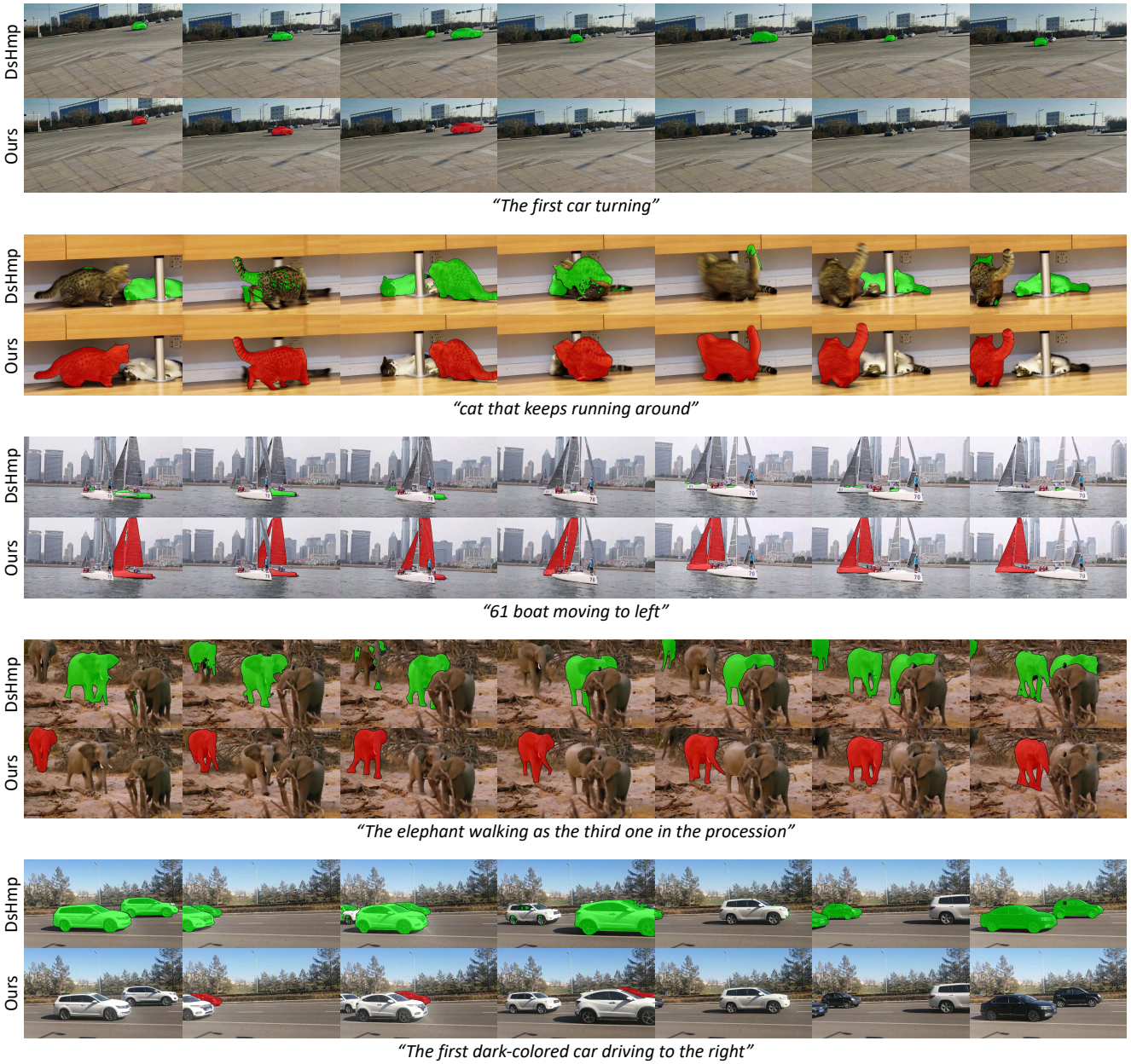

Figure A.1. **Qualitative results on MeViS.** Our proposed method outperforms previous state-of-the-art approach [9] in terms of mask quality and tracking ability, while ensuring accurate segmentation of the corresponding object based on the given expression.

Figure A.2. **Qualitative results on MeViS.** Our proposed method outperforms previous state-of-the-art approach [9] in terms of accurate selection of the corresponding object, while ensuring accurate segmentation of the corresponding object based on the given expression.

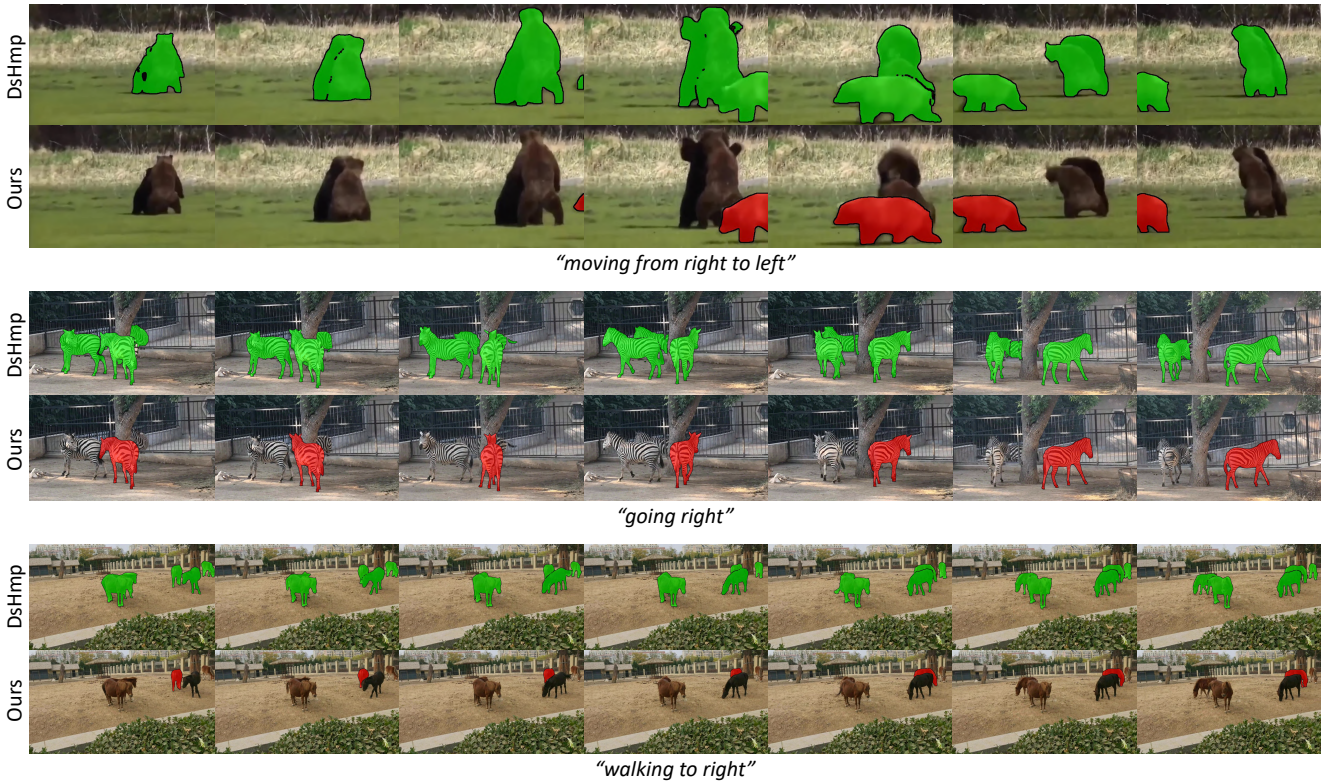

Figure A.3. **Qualitative results on MeViS.** Our proposed method outperforms previous state-of-the-art approach [9] in terms of accurate selection of the corresponding object, while ensuring accurate segmentation of the corresponding object based on the given expression. Notably, despite the given expression focusing solely on motion information, our model effectively handles the task without relying on appearance cues.

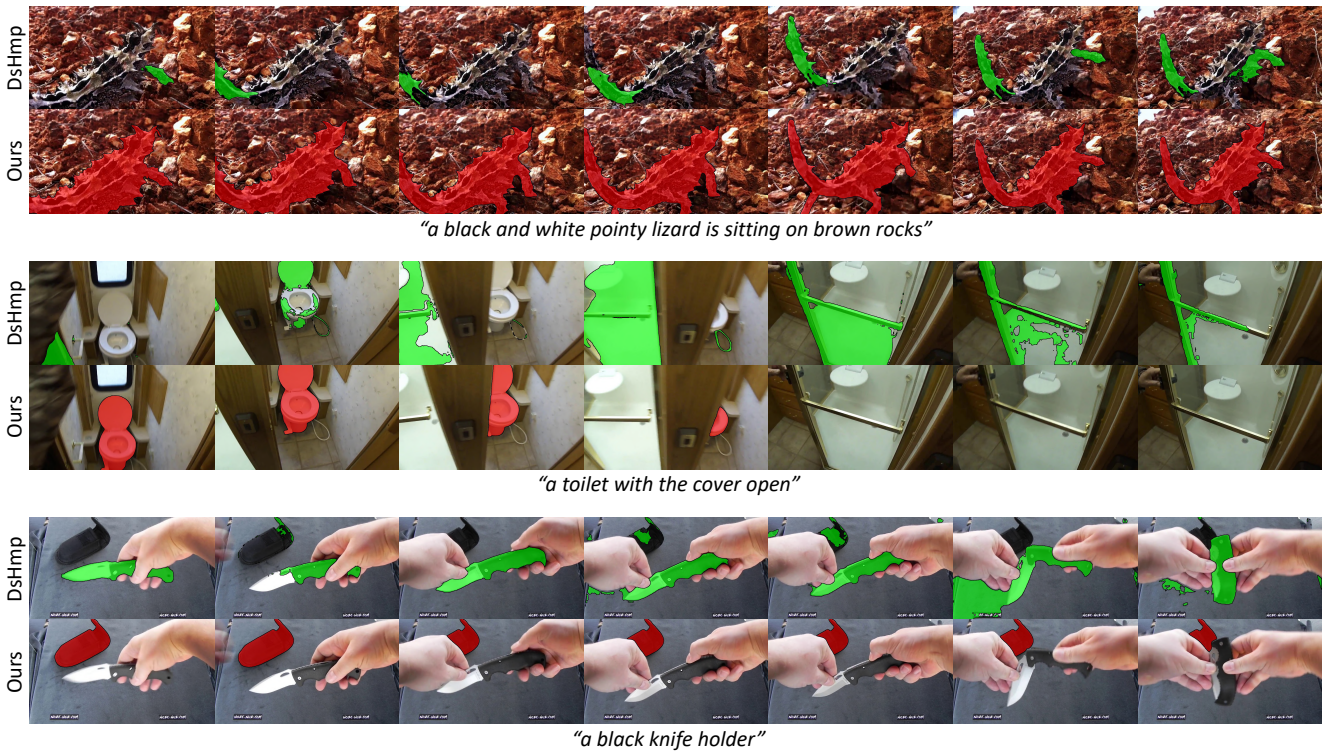

Figure A.4. **Qualitative results on Ref-YouTube-VOS.** Our proposed method outperforms previous state-of-the-art approach [9] in terms of accurate selection of the corresponding object, while ensuring accurate segmentation of the corresponding object based on the given expression.

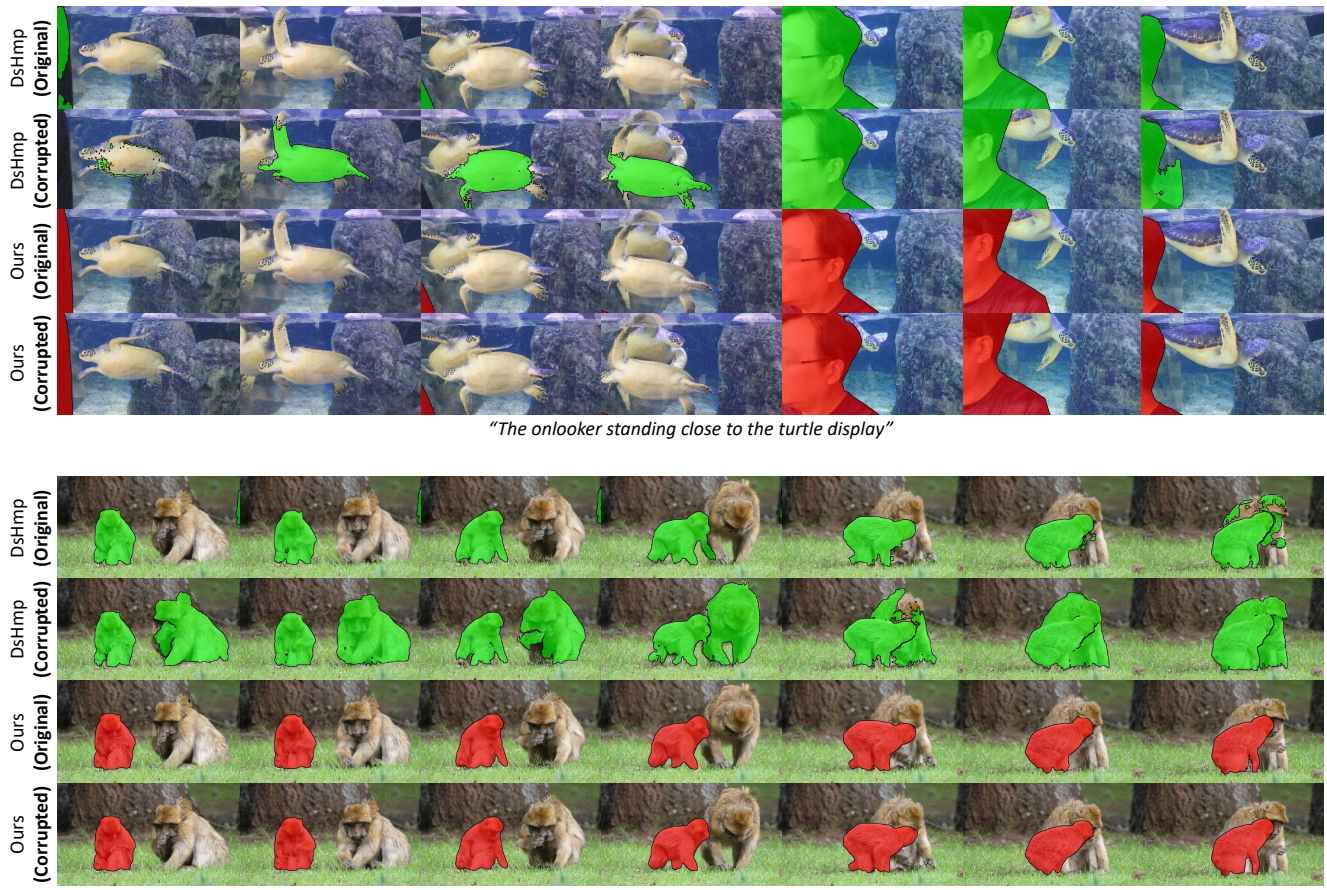

Figure A.5. **Qualitative results on corrupted version of MeViS.** Despite the *gaussian noise* distortion, our method generates high-quality outputs, demonstrating its robustness and effectiveness in handling perturbed data. Compared to previous work, our results maintain their performance even under the corrupted setting.

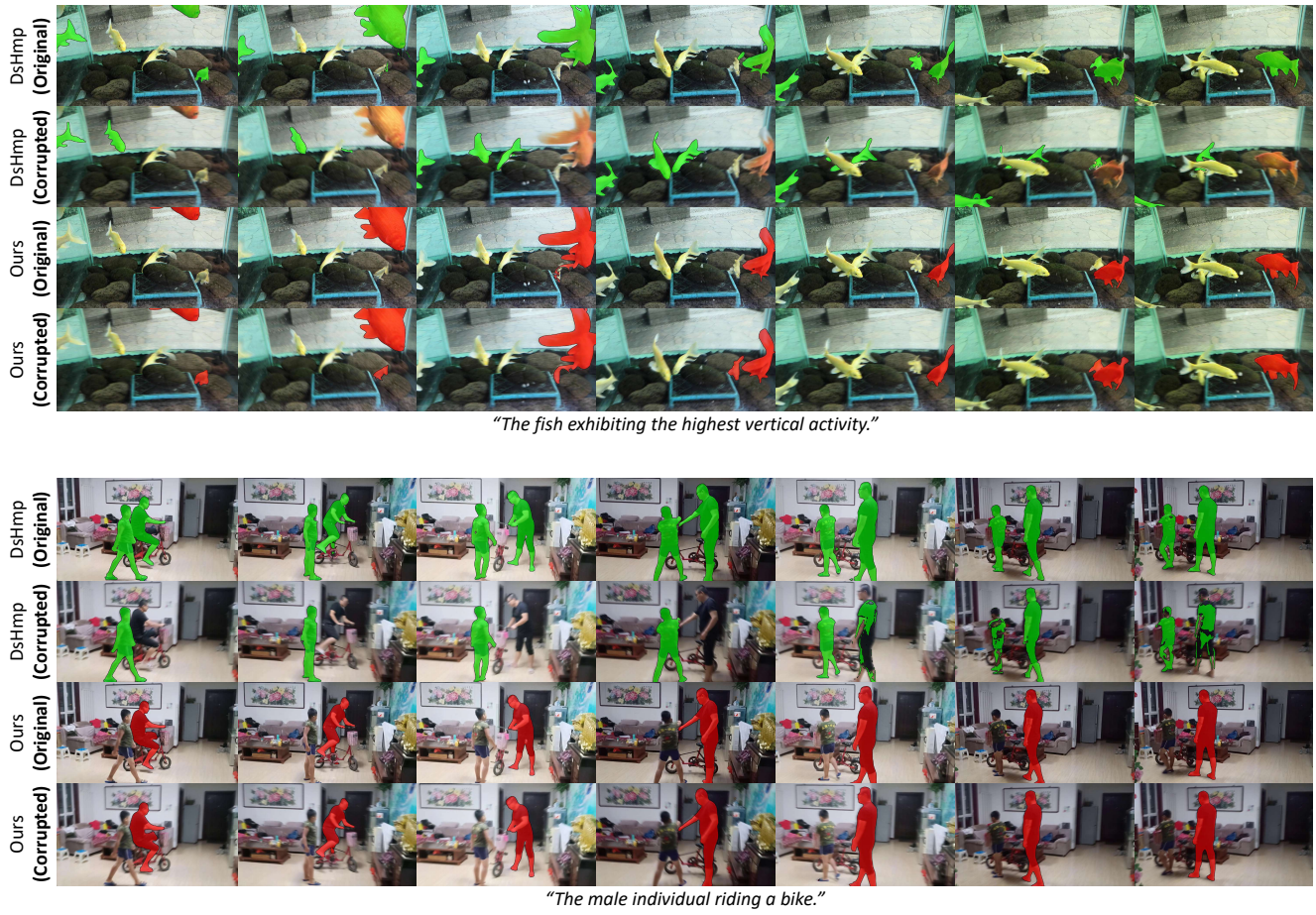

Figure A.6. **Qualitative results on corrupted version of MeViS.** Despite the *motion blur* distortion, our method generates high-quality outputs, demonstrating its robustness and effectiveness in handling perturbed data. Compared to previous work, our results maintain their performance even under the corrupted setting.