Referring Video Object Segmentation via Language-aligned Track Selection

Seongchan Kim [*1] **Woojeong Jin** [*1] **Sangbeom Lim** [2] **Heeji Yoon** [1] **HyunWook Choi** [1] **Seungryong Kim** [1]

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, 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. Code and pre-trained weights will be publicly available.

1. Introduction

Referring video object segmentation (RVOS) (Ding et al., 2023; Gavrilyuk et al., 2018; Khoreva et al., 2019; Seo et al., 2020) aims to segment a specific object throughout a video sequence based on a natural language expression. This task has garnered increasing attention for its potential applications in video editing and human-robot interaction. However, RVOS remains challenging, as it requires complex reasoning over both spatial and temporal visual cues while simultaneously aligning them with language at the object and scene level.

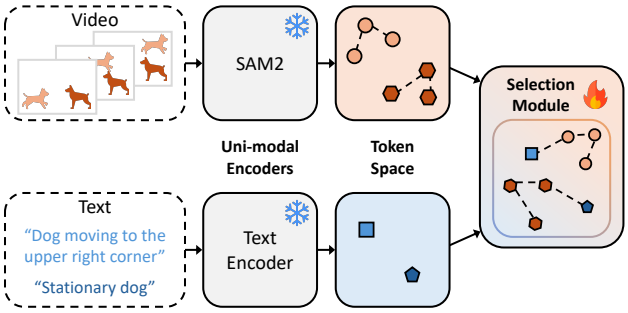

Figure 1. **Teaser.** Our method effectively bridges the modality gap by aligning the features obtained from frozen uni-modal encoders: the video segmentation model such as SAM2 (Ravi et al., 2024) and the text encoder such as RoBERTa (Liu et al., 2019). By directly leveraging the token representations, our approach achieves lightweight multi-modal alignment while significantly reducing the number of trainable parameters.

Recently, Segment Anything Model (SAM) (Kirillov et al., 2023) has emerged as a powerful models in the field of segmentation, demonstrating remarkable performance across various tasks. In particular, SAM2 (Ravi et al., 2024) extends these capabilities to video, offering strong performance in segmenting objects consistently across frames. Since mask track generation inherently involves maintaining object identity over time, SAM2 encodes temporally consistent object representations—making it a promising foundation for RVOS. However, SAM2 is not designed for language understanding, and how to effectively bridge its powerful representations with natural language remains an open challenge.

In this work, we propose **SOLA** (**S**election by **O**bject **L**anguage **A**lignment), the first framework to directly utilize SAM2's object token representations for language-guided video object segmentation. We hypothesize that these tokens encode rich, temporally coherent object-level information, and can serve as compact video-level features. To connect them with language, we introduce a lightweight *language-aligned track selection module* that efficiently aligns the frozen SAM2 object tokens with language features (Figure 1). Notably, our design leverages only precomputed object tokens, allowing the entire model to be trained efficiently on a single GPU with minimal additional parameters,

[1]KAIST AI [2]Korea University. Correspondence to: Seungryong Kim <seungryong.kim@kaist.ac.kr>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

while retaining SAM2's robustness and generalization.

To supervise this module without training SAM2, we propose a novel training strategy based on *IoU-based pseudo-labeling*. Given that RVOS datasets only provide expression-mask pairs, we assign pseudo-labels by comparing each candidate mask track to the ground-truth via mean Intersection-over-Union (mIoU). This enables the use of a binary classification objective, further enhanced by a contrastive alignment loss to emphasize discriminative motion patterns. This strategy ensures effective supervision without requiring explicit language-token annotations.

We validate our approach on standard RVOS benchmarks, including MeViS (Ding et al., 2023), Ref-YouTube-VOS (Li et al., 2018), and Ref-DAVIS (Khoreva et al., 2018). SOLA achieves state-of-the-art performance on MeViS, while also demonstrating strong generalization in zero-shot and combined dataset evaluations. These results confirm that our proposed module effectively aligning language with object motion dynamics, without any retraining of SAM2 or reliance of pre-aligned external models.

Our contributions are as follows:

- We propose SOLA, the first framework to utilize SAM2's object token representations for RVOS, based on the hypothesis that these tokens inherently encode temporal-aware objectness.
- We design a lightweight language-aligned track selection module that relies solely on precomputed SAM2 tokens, enabling efficient training with minimal parameters on a single GPU.
- We introduce a novel training strategy based on IoU-derived pseudo-labels, allowing effective alignment of frozen object tokens and language features through a simple classification objective and contrastive supervision.
- We achieve new state-of-the-art results on MeViS and demonstrate strong generalization to Ref-YouTube-VOS and Ref-DAVIS, 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. (Gavrilyuk et al., 2018) with the A2D-Sentences benchmark. Since then, RVOS has garnered significant attention, leading to the development of benchmarks such as Ref-YouTube-VOS (Li et al., 2018), Ref-DAVIS (Khoreva et al., 2018), and MeViS (Ding et al., 2023).

Recently, query-based models (Wu et al., 2022; Botach et al., 2022; Ding et al., 2023; He & Ding, 2024; Miao et al., 2024) 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 (Han et al., 2023; Li et al., 2023a) 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 (Kirillov et al., 2023) 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 (Ravi et al., 2024) 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 (Li et al., 2023b; Huang et al., 2024) 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 (Li et al., 2023b) 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 (Huang et al., 2024) assigns the spatio-temporal reasoning capability on GPT-4 (Achiam et al., 2023) and Grounding DINO (Liu et al., 2023). 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

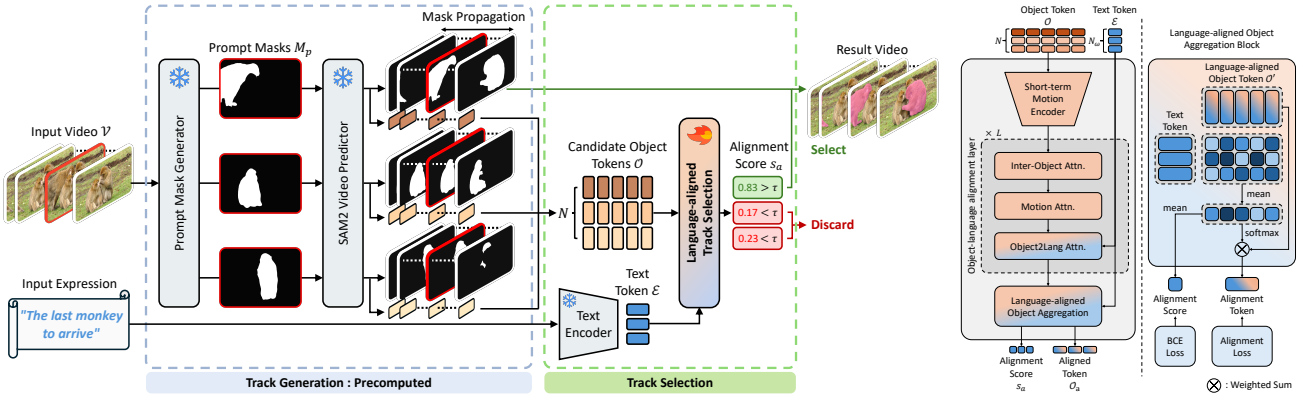

Figure 2. **Overall pipeline of the proposed SOLA framework and the language-aligned track selection module. (Left)** Our method selects the correct object mask track among candidates via a language-aligned track selection module. We first generate candidate mask tracks and corresponding object tokens 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. **(Right)** Architecture of the language-aligned track selection module, which takes object tokens and text tokens as inputs and aligns these representations to effectively capture object dynamics and query-relevant semantics. By leveraging precomputed object tokens from SAM2, our approach minimizes trainable parameters, enabling efficient training on a single GPU.

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 efficiently 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*, which efficiently bridges the modality gap between SAM2 object tokens and language features.

3.2. Preliminary - SAM2

SAM2 (Ravi et al., 2024) is a promptable video segmentation model composed of an image encoder, a prompt encoder, a mask decoder, and a memory module. Given user prompts—points $\mathcal{P}_g$, bounding boxes $\mathcal{P}_b$, or masks $\mathcal{P}_m$—the prompt encoder produces prompt tokens that, together with memory-conditioned image embeddings, are decoded into segmentation masks. For each predicted mask, the decoder also outputs a mask token, which is converted into an *object pointer* $\mathcal{O}^{i,t} \in \mathbb{R}^D$ at time t for object i $(i = 1,\dots,N, t = 1,\dots,T)$. These pointers, together with spatial embeddings fused with predicted masks, are stored in the memory module and cross-attended by subsequent frames, ensuring temporal consistency across the video.

We hypothesize that each object pointer encodes high-level information of the object associated with mask $\mathcal{M}^{i,t}$, and that its temporal sequence inherently captures *object motion*. Our framework exploits this property by repurposing object pointers as compact object representations for language-aligned track selection.

3.3. Track generation

As our method selects valid mask tracks among candidates, we first prompt SAM2 to ensure it generates all the objects existing 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. Notably, grid point prompts $\mathcal{P}_g$ cover both foreground objects and the surrounding background, as the points are evenly distributed across the frame.

Mask track propagation. The generated masks M_p are propagated 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 pointers:

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

For each candidate track i, we concatenate the pointer sequence $\{\mathcal{O}^{i,t}\}_{t=1}^{T}$ over time to form an *object token* $\mathcal{O}^i \in \mathbb{R}^{T\times D}$, which serves as a compact representation

that encodes both spatial and temporal cues for modeling complex object motion.

3.4. Track selection

Given N candidate mask tracks and their corresponding object tokens $\mathcal{O}$, we address RVOS by selecting tracks that semantically match the input expression. To this end, we propose a lightweight *language-aligned track selection module* that aligns SAM2 object tokens with language embeddings $\mathcal{E}$ to produce alignment scores $s_a \in \mathbb{R}^N$ and corresponding alignment tokens $\mathcal{O}_a \in \mathbb{R}^{N\times D}$:

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

where $\text{TS}(\cdot)$ is the track selection module. As shown in Figure 2 (Right), it consists of a short-term motion encoder, object-language alignment layers, and a language-guided motion aggregation module.

Short-term motion encoder. Since objects in RVOS are often distinguished by motion as well as appearance, we apply a 1D convolution along the temporal axis of each object token to capture short-term dynamics. This produces an updated representation $\mathcal{O}^i \in \mathbb{R}^{T'\times D}$, where T' is the reduced temporal resolution.

Object-language alignment layer. Repeated L times, the object-language alignment layer sequentially applies three types of attention: inter-object attention, motion attention, and object-to-language cross-attention. The first two are standard self-attention (Vaswani, 2017) applied along different dimensions to capture inter-object interactions and internal dynamics, respectively. Inter-object attention operates on object tokens $\mathcal{O}^t \in \mathbb{R}^{N\times D}$ at each frame t, allowing the model to reason about both object-object and object-background interactions—made possible by the grid-point prompting that covers foreground and background alike. Motion attention is applied along the temporal axis of each object token $\mathcal{O}^i \in \mathbb{R}^{T'\times D}$, capturing object-specific motion across time.

Finally, object-to-language cross-attention aligns visual object tokens $\mathcal{O}$ with language features $\mathcal{E}$, yielding language-aware object tokens $\mathcal{O}' \in \mathbb{R}^{N\times T'\times D}$, which are passed to the subsequent module for final selection.

Language-aligned object aggregation. The language-aligned object aggregation block takes $\mathcal{O}' \in \mathbb{R}^{N\times T'\times D}$ as input and produces an alignment score $s_a \in \mathbb{R}^N$ and an object-level representation $\mathcal{O}_a \in \mathbb{R}^{N\times D}$. We compute $\mathcal{O}_a$ as a weighted sum of temporally aligned object tokens, using a frame-wise weighting matrix $w_a \in \mathbb{R}^{N\times T'}$ defined as:

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

where $\text{Avg}_{N_w}(\cdot)$ denotes averaging along the object dimension.

The final representation and alignment score are computed as:

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

where $\otimes$ indicates element-wise multiplication, and $\text{Avg}_T(\cdot)$ averages along the temporal dimension.

The resulting $\mathcal{O}_a$ encodes both motion and semantic features aligned with the language expression, yielding a compact video-level object representation. The alignment scores s_a^i are computed via sigmoid function and thresholded by τ to select the relevant mask tracks. The selected tracks $\mathcal{M}_v$ are then merged to form the final binary segmentation output $\mathcal{B}$ for the given expression.

3.5. Training objective

IoU-based pseudo labeling. Since SAM2 remains frozen, our goal is to identify which SAM2-generated tracks match the referred target. However, RVOS datasets provide ground-truth masks only for the referred object, without labels for individual candidate tracks. To address this, we adopt an IoU-based pseudo-labeling strategy: for each expression, we compute the mIoU between every candidate mask track and the ground-truth mask, labeling candidates above a threshold τ as positive and the rest as negative. This reduces training to a simple binary classification task with minimal annotation overhead.

Loss functions. The total loss combines a Binary Cross-Entropy loss $\mathcal{L}_{\text{BCE}}$, applied between the alignment scores s_a^i and pseudo-labels y^i, with an alignment loss $\mathcal{L}_{\text{align}}$:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{BCE}} + \lambda_2 \mathcal{L}_{\text{align}}. \tag{6}$$

The alignment loss is a modified contrastive loss with a positive anchor $\mathcal{A}_p \in \mathbb{R}^D$ defined as the mean of text tokens $\mathcal{E}$, and learnable negative anchors $\mathcal{A}_n \in \mathbb{R}^{N_{\text{neg}}\times D}$ representing unrelated concepts:

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

with

$$\begin{aligned} \mathcal{L}_{\text{pos}} &= d(\mathcal{O}_a^i, \mathcal{A}_p) - \sum_{j=1}^{N_{\text{neg}}} d(\mathcal{O}_a^i, \mathcal{A}_n^j), \\ \mathcal{L}_{\text{neg}} &= d(\mathcal{O}_a^i, \mathcal{A}_n^{k^*}) - d(\mathcal{O}_a^i, \mathcal{A}_p) - \sum_{j\neq k^*} d(\mathcal{O}_a^i, \mathcal{A}_n^j), \end{aligned} \tag{8}$$

where $d(\mathbf{x}, \mathbf{y}) = 1 - \cos(\mathbf{x}, \mathbf{y})$ denotes cosine distance and k^* is the index of the closest negative anchor to $\mathcal{O}_a^i$.

4. Experiments

4.1. Datasets and evaluation metrics

Dataset. We evaluate our method on three video datasets: MeViS (Ding et al., 2023), Ref-YouTube-VOS (Seo et al., 2020), and Ref-DAVIS (Khoreva et al., 2018). 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 (Perazzi et al., 2016) by incorporating linguistic annotations for a variety of objects, featuring a total of 90 videos.

Evaluation metrics. Following (Ding et al., 2023; Miao et al., 2024; He & Ding, 2024), 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 (Vouitsis et al., 2024), 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 (Ding et al., 2023) dataset.

Track generation. We generate mask tracks using SAM2-L (Ravi et al., 2024), prompted by grid points and bounding boxes obtained from Grounding DINO-T (Liu et al., 2023) every fourth frame. To avoid generating redundant tracks, we apply IoU-based filtering, similar to Non-Maximum Suppression (NMS) (Neubeck & Van Gool, 2006), propagating only distinct prompt masks.

Language-aligned track selection module. We employ pre-trained RoBERTa (Liu et al., 2019) 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 (Ding et al., 2023) 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	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
URVOS (Seo et al., 2020)	-	27.8	25.7	29.9
LBDT (Ding et al., 2022)	95.6 M	29.3	27.8	30.8
MTTR (Botach et al., 2022)	-	30.0	28.8	31.2
ReferFormer (Wu et al., 2022)	112.9 M	31.0	29.8	32.2
VLT+TC (Ding et al., 2021)	<u>38.3 M</u>	35.5	33.6	37.3
LMPM (Ding et al., 2023)	66.4 M	37.2	34.2	40.2
HTR (Miao et al., 2024)	-	42.7	39.9	45.5
DsHmp (He & Ding, 2024)	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	35.0	34.2	35.8	40.5	36.8	44.2
LMPM	31.5	30.0	32.9	39.9	36.7	43.2
DsHmp	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	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
w/o selection module	36.9	30.0	43.8
w/ selection module	**48.6**	**45.2**	**52.1**

Table 3. **Ablation study on our selection module.**

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 (Ding et al., 2023) dataset and evaluated it on the Ref-YouTube-VOS (Seo et al., 2020) and Ref-DAVIS (Khoreva et al., 2018) 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.

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**

Table 4. **Ablation study on the effect of the alignment loss.**

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**

Table 5. **Ablation study on the effects of different attention modules.**

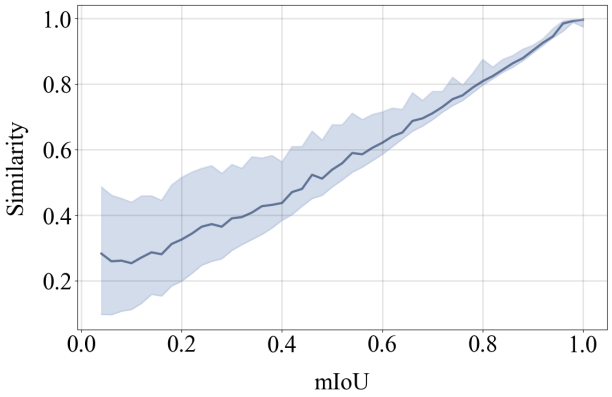

Figure 3. **Spatial and motion information in SAM2's object tokens.** The bold line represents the mean similarity, while the shaded region indicates the variance. The results show a certain correlation: as the mIoU between mask tracks increases, the similarity between their associated tokens also rises nearly proportionally. This tendency suggests that object tokens inherently capture spatial information, implicitly encoding object motions over time.

4.4. Ablation studies

We conduct our ablation studies on the MeViS (Ding et al., 2023) dataset to examine the effectiveness of our proposed language-aligned selection module and its components.

Effect of the proposed selection method. The results in Table 3 show that our language-aligned track selection module improves interpretation of complex expressions. The *w/o selection module* baseline uses Grounding DINO (Liu et al., 2023) for frame-level object detection based on text-image correspondence, but lacks temporal modeling, limiting its ability to handle video-level reasoning. In contrast, *w/ selection module* (our SOLA framework) selects referred object tracks using language-aligned object tokens, integrating spatial and temporal cues to better understand complex queries, thus enhancing RVOS performance.

Ablation on losses. In Table 4, 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.

Ablation on different types of attention. Table 5 compares attention configurations. Motion attention alone improves temporal modeling but lacks spatial context. Inter-object attention captures spatial relationships but misses temporal cues. Combining both enables the model to capture object dynamics and spatial interactions, leading to better global context understanding.

4.5. Analysis on object token of SAM2

As our method solely relies on SAM2 object tokens, we assess whether they carry sufficient motion and semantic information. First, we analyze their relationship to spatial masks by computing cosine similarity between tokens with respect to the mIoU of their corresponding masks. As shown in Figure 3, similarity increases with mIoU, suggesting that spatially close masks yield similar tokens—implying that spatial dynamics may be captured when these tokens are tracked temporally. Second, we evaluate the semantic content of object tokens via a classification task on PASCAL-VOC (Everingham et al., 2010). A linear head trained on the tokens achieved 85.3% accuracy across 20 categories, indicating that they embed a meaningful level of semantic information. These results suggest that SAM2 object tokens encode both motion and semantics, supporting their use as a lightweight and expressive interface for language grounding.

5. Conclusion

We introduce SOLA, the first RVOS framework to leverage object tokens precomputed from frozen SAM2 as compact, temporally consistent object representations. To align them with language, we propose a lightweight track selection module with only 32.9M trainable parameters, enabling efficient training on a single GPU. To supervise the selection of language-aligned object tokens, we adopt an IoU-based pseudo-labeling strategy designed to bridge the modality gap between vision and language. SOLA achieves state-of-the-art results on MeViS and generalizes well to Ref-YouTube-VOS and Ref-DAVIS, demonstrating strong motion modeling and multi-modal alignment with minimal computational cost.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Botach, A., Zheltonozhskii, E., and Baskin, C. End-to-end referring video object segmentation with multimodal transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4975–4985. IEEE Computer Society, 2022.

Ding, H., Liu, C., Wang, S., and Jiang, X. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

Ding, H., Liu, C., He, S., Jiang, X., and Loy, C. C. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2694–2703, 2023.

Ding, Z., Hui, T., Huang, J., Wei, X., Han, J., and Liu, S. Language-bridged spatial-temporal interaction for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4964–4973, 2022.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338, 2010.

Gavrilyuk, K., Ghodrati, A., Li, Z., and Snoek, C. G. Actor and action video segmentation from a sentence. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5958–5966. IEEE Computer Society, 2018.

Han, M., Wang, Y., Li, Z., Yao, L., Chang, X., and Qiao, Y. Html: Hybrid temporal-scale multimodal learning framework for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13414–13423, 2023.

He, S. and Ding, H. Decoupling static and hierarchical motion perception for referring video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13332–13341, 2024.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Huang, S., Ling, R., Li, H., Hui, T., Tang, Z., Wei, X., Han, J., and Liu, S. Unleashing the temporal-spatial reasoning capacity of gpt for training-free audio and language referenced video object segmentation. *arXiv preprint arXiv:2408.15876*, 2024.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Khoreva, A., Rohrbach, A., and Schiele, B. Video object segmentation with language referring expressions. In *ACCV*, 2018.

Khoreva, A., Rohrbach, A., and Schiele, B. Video object segmentation with language referring expressions. In *14th Asian Conference on Computer Vision*, pp. 123–141. Springer, 2019.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Li, R., Li, K., Kuo, Y.-C., Shu, M., Qi, X., Shen, X., and Jia, J. Referring image segmentation via recurrent refinement networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753, 2018.

Li, X., Wang, J., Xu, X., Li, X., Raj, B., and Lu, Y. Robust referring video object segmentation with cyclic structural consensus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22236–22245, 2023a.

Li, Y., Zhang, J., Teng, X., Lan, L., and Liu, X. Refsam: Efficiently adapting segmenting anything model for referring video object segmentation. *arXiv preprint arXiv:2307.00997*, 2023b.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Miao, B., Bennamoun, M., Gao, Y., Shah, M., and Mian, A. Temporally consistent referring video object segmentation with hybrid memory. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

Neubeck, A. and Van Gool, L. Efficient non-maximum suppression. In *18th international conference on pattern recognition (ICPR'06)*, volume 3, pp. 850–855. IEEE, 2006.

Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., and Sorkine-Hornung, A. A benchmark dataset and evaluation methodology for video object segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 724–732. IEEE, 2016.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Seo, S., Lee, J.-Y., and Han, B. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European Conference on Computer Vision*, pp. 208–223, 2020.

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

Vouitsis, N., Liu, Z., Gorti, S. K., Villecroze, V., Cresswell, J. C., Yu, G., Loaiza-Ganem, G., and Volkovs, M. Data-efficient multimodal fusion on a single gpu. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27239–27251, 2024.

Wu, J., Jiang, Y., Sun, P., Yuan, Z., and Luo, P. Language as queries for referring video object segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4964–4974. IEEE, 2022.

Referring Video Object Segmentation via Language-aligned Track Selection – Supplementary Materials –

A. Additional qualitative results

Qualitative results on MeViS. Figure 1 presents the qualitative results on MeViS (Ding et al., 2023), comparing the performance of DsHmp (He & Ding, 2024) with SOLA. Our approach consistently demonstrates superior capability in accurately selecting the target object as specified by the referring expression. Specifically, Figure 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 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 4 presents the qualitative results on the Ref-YouTube-VOS (Seo et al., 2020) 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. Additional quantitative results

Results on corrupted setting To demonstrate the robustness of our method, we evaluated it on a perturbed dataset with ImageNet-C (Hendrycks & Dietterich, 2019) 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 1 presents the quantitative results, showing that our proposed method outperforms previous approaches (Ding et al., 2023; He & Ding, 2024; Jia et al., 2021) even under corruption scenarios.

Results on combined setting Table 2 presents the quantitative results obtained by training on a naively combined dataset of MeViS (Ding et al., 2023) and Ref-YouTube-VOS (Seo et al., 2020), 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.

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

C. Additional ablation studies on MeViS

Existence of background object tokens. The quantitative results presented in Table 3 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 4 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,

Methods	Motion blur			Gaussian noise		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
ReferFormer (Jia et al., 2021)	26.3	25.4	27.1	26.9	24.0	29.9
LMPM (Ding et al., 2023)	33.3	31.2	35.4	36.0	33.4	38.6
DsHmp (He & Ding, 2024)	38.0	35.0	41.1	43.4	39.5	47.2
SOLA	**39.8**	**36.6**	**43.0**	**44.4**	**40.5**	**48.3**

Table 1. **Quantitative results on corrupted MeViS validation set.** The table shows performance of different methods under Motion blur and Gaussian noise corruptions (severity 5). All models are trained on the original dataset. Best results are highlighted in **bold**.

Methods	MeViS			Ref-YouTube-VOS		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
ReferFormer	36.6	34.1	39.1	46.8	46.2	47.5
LMPM	40.5	37.8	43.2	37.6	36.0	39.2
DsHmp	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 2. **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.

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**

Table 3. **Effects of including background object tokens.**

we adopt a highly efficient training strategy similar to FuseMix (Vouitsis et al., 2024). Specifically, we first perform SAM2 mask propagation on the given RVOS dataset to generate candidate 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 (Ding et al., 2023) 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 (Liu et al., 2023) every fourth frame to generate prompt masks, which serve as input for the SAM2-L (Ravi et al., 2024) 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.

# 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

Table 4. **Effects of the number of object-language alignment layers.**

Language-aligned track selection module. We employ pre-trained RoBERTa (Liu et al., 2019) 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.

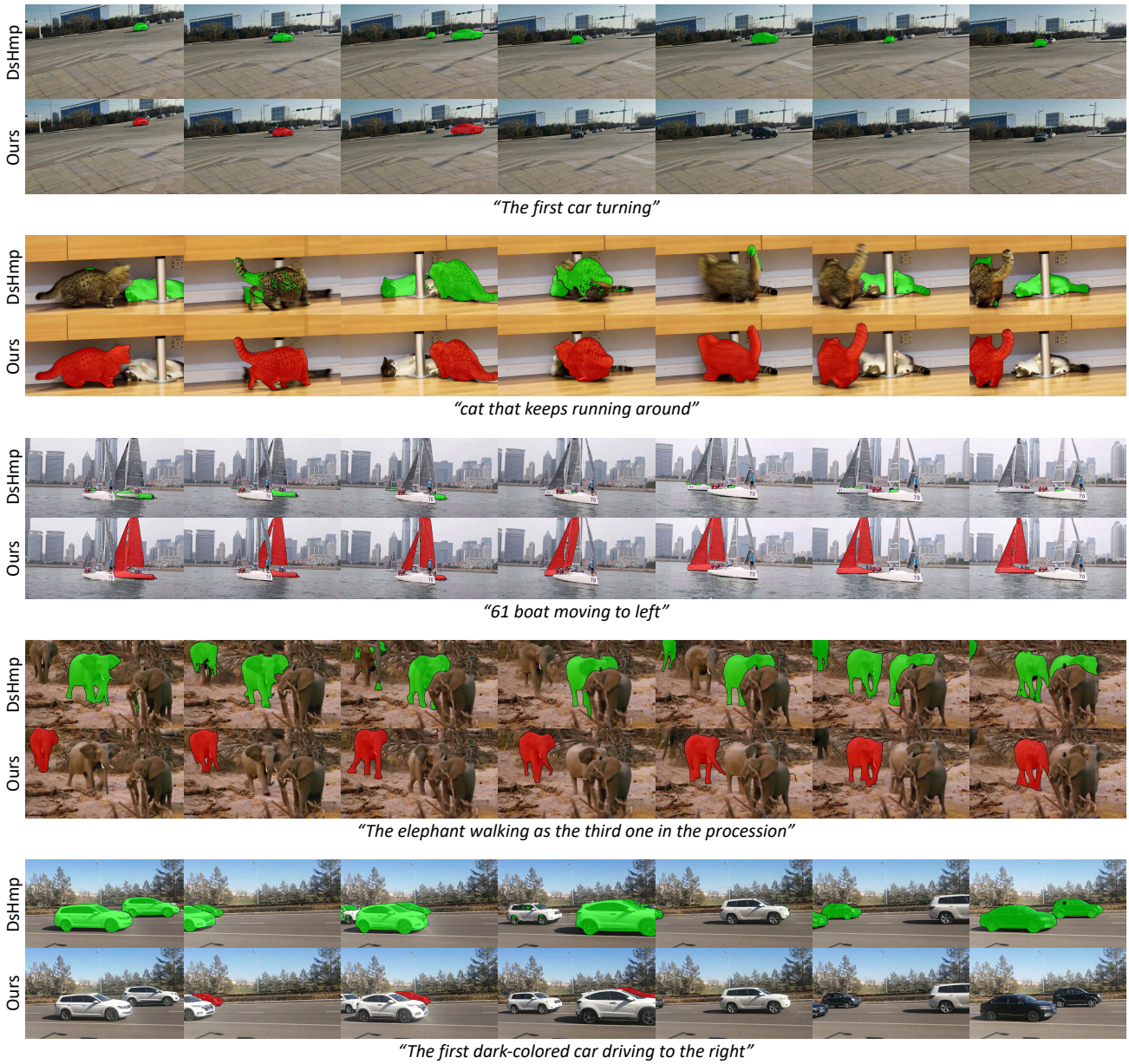

Figure 1. **Qualitative results on MeViS.** Our proposed method outperforms previous state-of-the-art approach (He & Ding, 2024) in terms of mask quality and tracking ability, while ensuring accurate segmentation of the corresponding object based on the given expression.

Figure 2. **Qualitative results on MeViS.** Our proposed method outperforms previous state-of-the-art approach (He & Ding, 2024) in terms of accurate selection of the corresponding object, while ensuring accurate segmentation of the corresponding object based on the given expression.

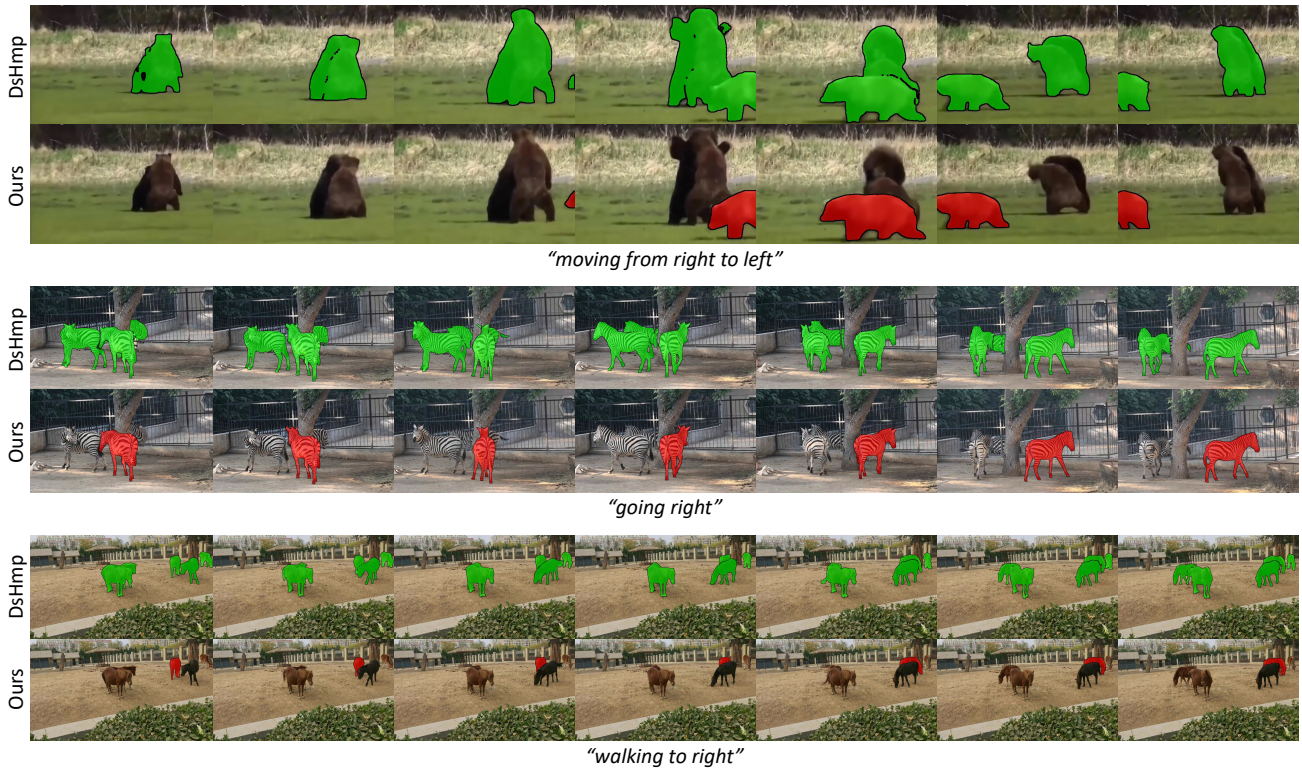

Figure 3. **Qualitative results on MeViS.** Our proposed method outperforms previous state-of-the-art approach (He & Ding, 2024) in terms of accurate selection of the corresponding object, while ensuring accurate segmentation of the corresponding object based on the given expression. Notably, despite the given expression focusing solely on motion information, our model effectively handles the task without relying on appearance cues.

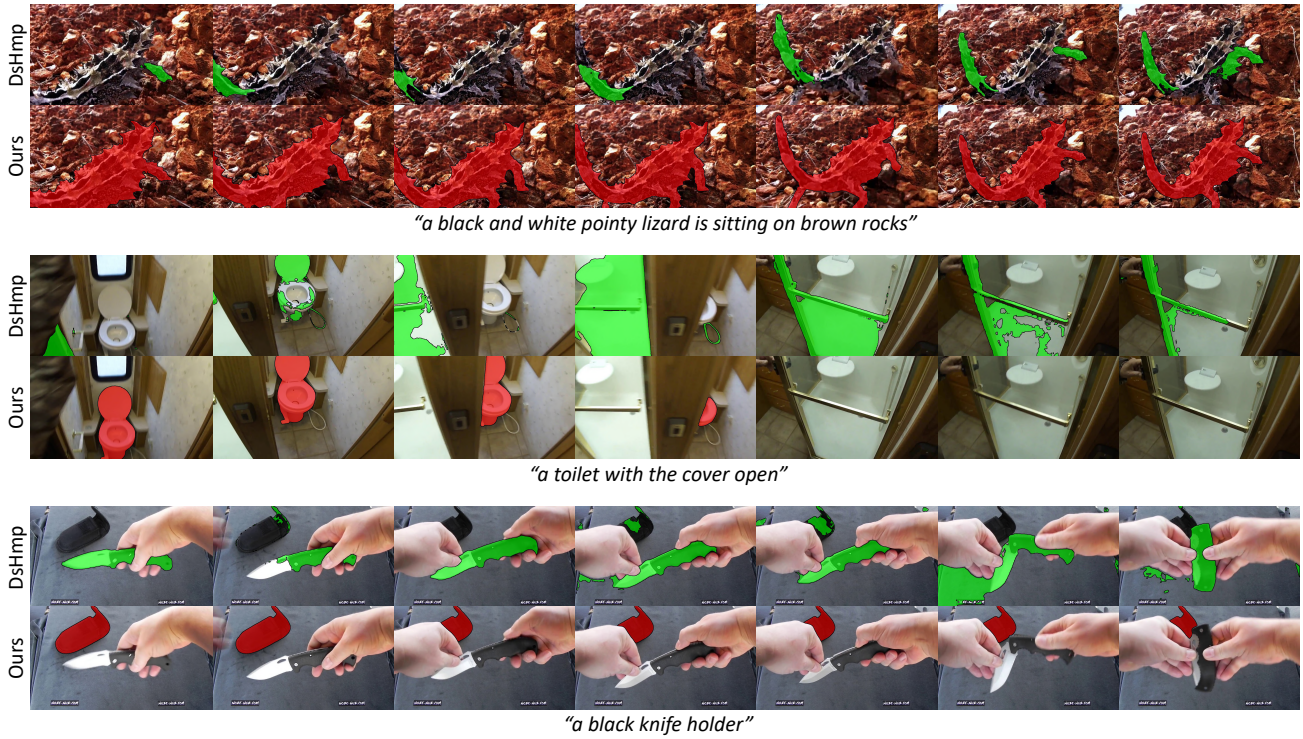

Figure 4. **Qualitative results on Ref-YouTube-VOS.** Our proposed method outperforms previous state-of-the-art approach (He & Ding, 2024) in terms of accurate selection of the corresponding object, while ensuring accurate segmentation of the corresponding object based on the given expression.

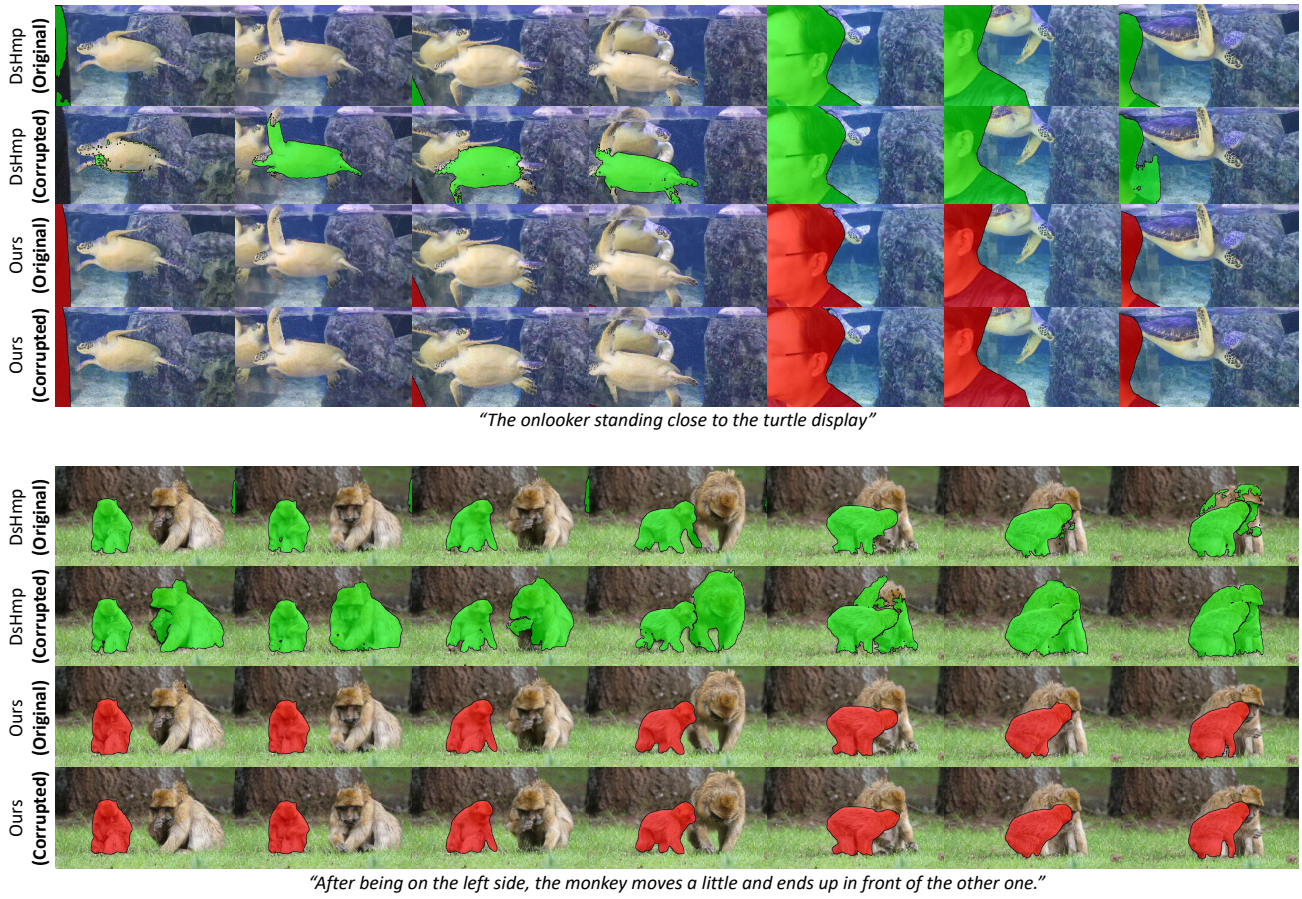

Figure 5. **Qualitative results on corrupted version of MeViS.** Despite the *gaussian noise* distortion, our method generates high-quality outputs, demonstrating its robustness and effectiveness in handling perturbed data. Compared to previous work, our results maintain their performance even under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

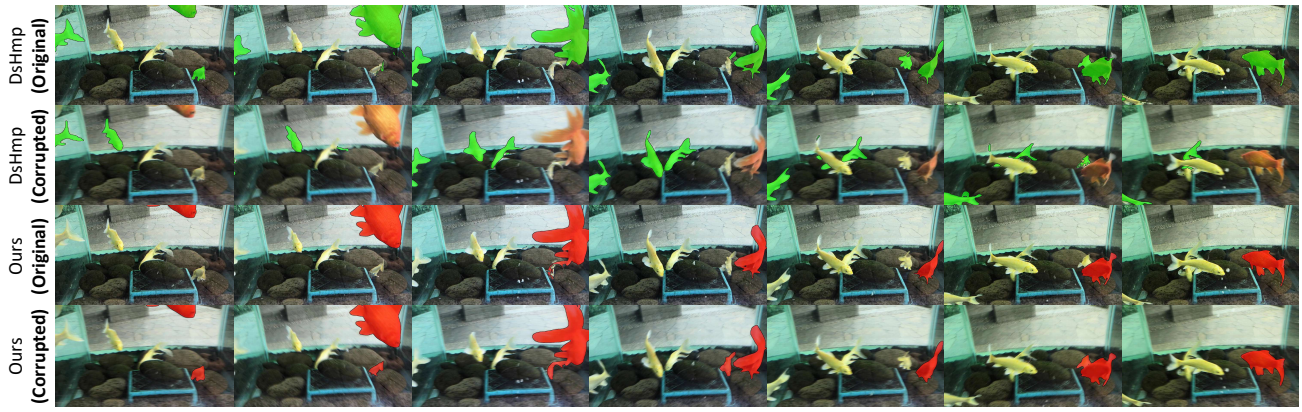

Figure 6. **Qualitative results on corrupted version of MeViS.** Despite the *motion blur* distortion, our method generates high-quality outputs, demonstrating its robustness and effectiveness in handling perturbed data. Compared to previous work, our results maintain their performance even under the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