

Referring Video Object Segmentation via Language-aligned Track Selection

Seongchan Kim^{1*} Woojeong Jin^{2*} Sangbeom Lim^{1*} Heeji Yoon^{1*}
 Hyunwook Choi¹ Seungryong Kim^{2†}
¹Korea University ²KAIST

Abstract

Referring Video Object Segmentation (RVOS) seeks to segment objects throughout a video based on natural language expressions. While existing methods have made strides in vision-language alignment, they often overlook the importance of robust video object tracking, where inconsistent mask tracks can disrupt vision-language alignment, leading to suboptimal performance. In this work, we present Selection by Object Language Alignment (SOLA), a novel framework that reformulates RVOS into two sub-problems, track generation and track selection. In track generation, we leverage a vision foundation model, Segment Anything Model 2 (SAM2), which generates consistent mask tracks across frames, producing reliable candidates for both foreground and background objects. For track selection, we propose a light yet effective selection module that aligns visual and textual features while modeling object appearance and motion within video sequences. This design enables precise motion modeling and alignment of the vision language. Our approach achieves state-of-the-art performance on the challenging MeViS dataset and demonstrates superior results in zero-shot settings on the Ref-Youtube-VOS and Ref-DAVIS datasets. Furthermore, SOLA exhibits strong generalization and robustness in corrupted settings, such as those with added Gaussian noise or motion blur. Our project page is available at: <https://github.com/cvlab-kaist/SOLA>.

1. Introduction

Referring Video Object Segmentation (RVOS) [6, 8, 14, 25] has recently attracted significant research interest due to its potential applications in various fields, including interactive video editing and video surveillance. This task focuses on segmenting and tracking foreground objects throughout a video sequence based on a natural language expression.

RVOS, dealing with video data, requires a comprehensive

*These authors contributed equally.

†Corresponding author.

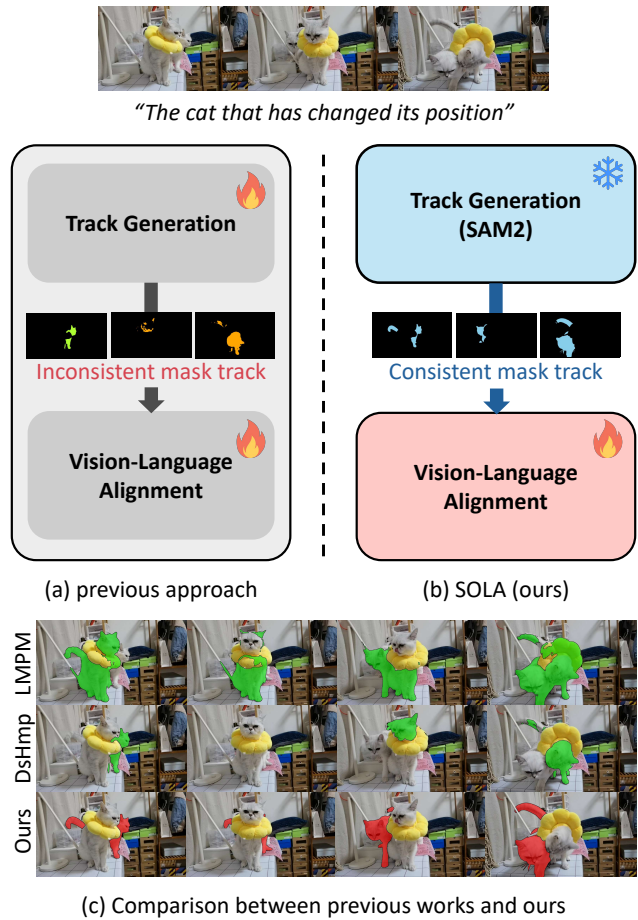


Figure 1. **Comparison between (a) previous approaches and (b) proposed SOLA framework.** Previous methods [6, 9] simultaneously train track generation and vision-language alignment, whereas our approach focuses solely on training the latter, given consistent mask track by SAM2. As a result, prior work often generates inconsistent mask tracks, which in turns shows limited performance as exemplified in (c), while our method produces more consistent outputs.

sive understanding of temporal dynamics across entire sequences to capture object dynamic motion. RVOS methods [3, 6, 9, 27] require not only mask generation, but also

precise matching of mask tracks to the given expression. These approaches generate object tokens frame-by-frame, for instance, using Mask2Former [4] and match these tokens to create candidate object tracks, aligned with sentence embeddings. However, these approaches require the model to simultaneously learn diverse capabilities, such as object tracking, motion modeling, and alignment between language and objects, all within an end-to-end framework. This often results in noisy tracks that fail to consistently track the same objects across frames, leading to inconsistent masks. Such inconsistencies hinder motion modeling, as the model struggles to capture continuous object motion throughout video tracks, despite advances in vision-language alignment.

We hypothesize that when high-quality mask tracks are ensured, motion modeling can become more reliable, allowing the model to focus exclusively on vision-language alignment within a simplified setting. In this paper, we reformulate the RVOS task into two sub-problems: *track generation* and *track selection*, while assuring high-quality mask tracks leveraging vision foundation model. Based on this reformulation, we propose a novel framework **Selection by Object Language Alignment (SOLA)** which consists of track generation followed by track selection. In the track generation, we leverage a vision foundation model, Segment Anything Model 2 (SAM2) [23], which is capable of generating consistent object mask tracks across frames, ensuring reliable tracking conditions. In the track selection, the focus shifts to motion modeling and vision-language alignment. Our novel track selection module directly utilizes the vision representation obtained during the track generation process and effectively bridges it with the language representation. Furthermore, this module effectively models motion and vision-language alignment, even with its light-weight structure, by leveraging high-level video object representations obtained from SAM2 [23]. The intuitive comparison between previous approaches and our proposed framework is shown in Figure 1.

In experiments, we evaluate our method on the standard RVOS benchmarks, such as MeViS [6], Ref-YouTube-VOS [16], and Ref-DAVIS [13] datasets. Our framework significantly outperforms prior state-of-the-art methods on MeViS [6]. The MeViS dataset especially challenges models to track temporal object sequences guided by complex language expressions, highlighting our method’s effectiveness in handling complex scenarios. Our framework further demonstrates strong zero-shot performance on Ref-YouTube-VOS [16] and Ref-DAVIS [13], highlighting the model’s robustness and generalization capabilities. Our method demonstrates both robust dense tracking and effective vision-language alignment, achieving outstanding performance both quantitatively and qualitatively. Additionally, we demonstrate the generalizability and robustness of

SOLA through the presented experiments.

Our main contributions are as follows:

- We propose a novel framework, SOLA, which reformulates the RVOS task into two streamlined sub-problems: track generation and track selection. By addressing the challenges in track generation using SAM2, our method shifts the emphasis to the track selection process, facilitating more effective and efficient performance optimization.
- We bridge the modality gap between vision and language representations from frozen models pre-trained in different modalities by introducing a lightweight language-aligned track selection module. This module effectively leverages the high-level video object representations of SAM2 [23] to achieve both motion modeling and vision-language alignment, ensuring efficient and accurate track selection.
- Our method achieves new state-of-the-art results on the MeViS dataset [6] and demonstrates superior performance on the Ref-YouTube-VOS [16] and Ref-DAVIS [13] datasets in zero-shot settings.

2. Related Work

Referring video object segmentation. Compared to image segmentation, RVOS poses a greater challenge, as it requires segmenting objects by capturing both action and appearance from video sequences based on a given expression. RVOS was first introduced by Gavriluk et al.[8] with the A2D-Sentences benchmark. Since then, RVOS has garnered significant attention, leading to the development of new benchmarks such as Ref-YouTube-VOS, Ref-DAVIS, and MeViS. Previous models, such as URVOS[25] and RefVOS [2], have advanced the field by incorporating cross-modal attention for per-frame segmentation, though they often overlook temporal information. Recently, query-based models utilizing Mask2Former [4], such as ReferFormer [27], MTTR [3], LMPM [6], and DsHmp [9], have achieved impressive performance. Although previous methods have shown impressive performance on RVOS, significant challenges remain. Generating mask tracks using per-frame methods with simple cost functions [6, 9, 11] often fails to link corresponding objects in subsequent frames.

Multi-modal foundation model. Vision foundation models such as CLIP [22], ALIGN [12], and Grounding DINO [17] provide strong backbones that show outstanding performance across a wide range of vision task. Trained on massive, often noisy image-text datasets, these models include specific encoders for each modality, generating embeddings for both images and text. With a contrastive objective, they are optimized to align embeddings of matching image-text pairs within random batches. This enables effective zero-shot applications, such as image-text retrieval

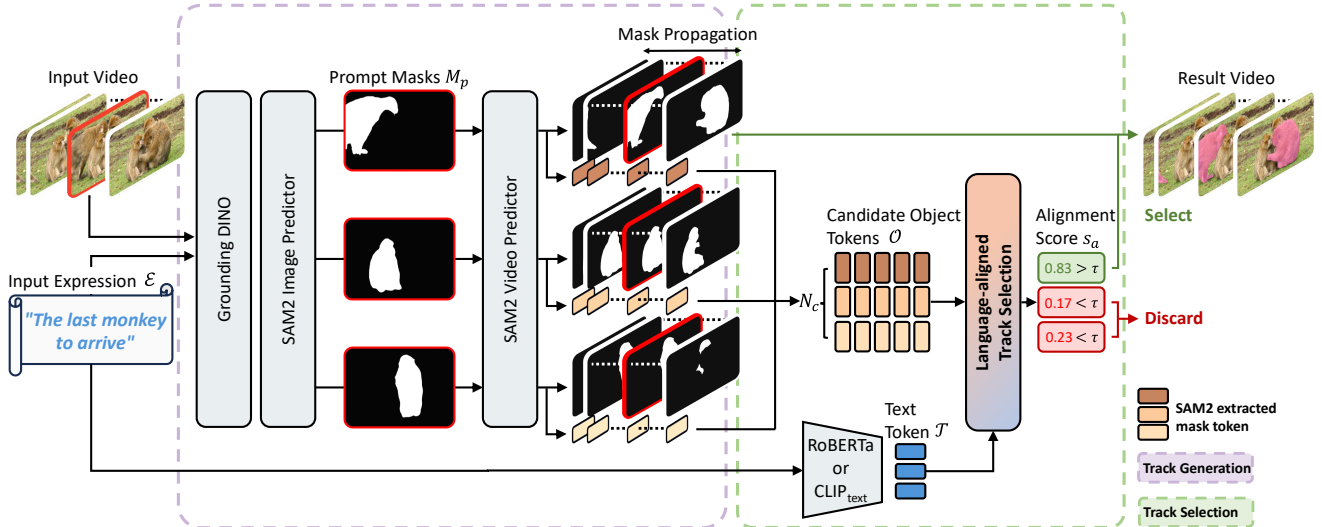


Figure 2. **Overall pipeline of the proposed SOLA framework.** Our core idea is to redefine referring video object segmentation as two sub-problems: track generation and track selection. We first generate candidate mask tracks with the Segment Anything Model 2 [23], ensuring consistent and clear mask tracks. Then, our light-weight language-aligned track selection module efficiently selects the referred mask tracks through motion modeling and object-language alignment. During the inference stage, we leverage the visual grounding model, Grounding DINO [17], for efficient candidate track generation.

and classification via text prompts, achieving robust performance across diverse domains.

Segment Anything Model (SAM). 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. Known for its strong zero-shot transfer capabilities, SAM has proven highly adaptable across a wide range of vision applications, including object segmentation, image editing, and reconstruction. A key feature of SAM is its flexibility in interpreting various input formats, such as points, bounding boxes, and text, allowing users to provide segmentation guidance through multiple modalities and making the model both highly usable and versatile. Building on SAM, SAM2 [23] 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.

In our approach, we directly incorporate these foundation models for consistent and accurate mask prediction, avoiding the need to train additional models with learnable parameters. This strategy allows our method to leverage enhanced generalization capabilities.

3. Method

For given expression \mathcal{E} and T frames of video clip $\mathcal{I} = \{I_t\}_{t=1}^T$, where $I_t \in \mathbb{R}^{C \times H \times W}$, with H , W , and C denoting height, width and channels of each frame, respectively, the objective of RVOS is to generate binary mask tracks

$\mathcal{M} = \{M_t\}_{t=1}^T$, where $M_t \in \{0, 1\}^{H \times W}$ corresponds to the referred object. RVOS poses a significant challenge, as the model must be capable of dense mask tracking and simultaneously ensure alignment between these predictions and natural language descriptions.

In this paper, we propose a novel framework consisting of track generation and track selection, by redefining RVOS as a separated sub-problems of mask tracking and language alignment. The overview of the proposed approach, named SOLA, is shown in Figure 2. Specifically, we first generate candidate mask tracks and their feature representations by leveraging the well known generalized mask tracker, SAM2 [23], as detailed in Section 3.2. Next, we select mask tracks based on their semantic correspondence with given expression. For track selection, we introduce light-weight track selection module, in Section 3.3. This module integrates language and object mask features to effectively determine their correlation.

3.1. Preliminary - SAM2

SAM2 [23] is a promptable video segmentation model that consists of an image encoder, a mask decoder, a prompt encoder, and a memory encoder. Below, we provide an overview of SAM2 to support the understanding of SOLA.

Prompt Encoder. SAM2 inherits the prompt encoder design from SAM [15], allowing it to handle video-based mask predictions. The prompt encoder supports three types of user inputs: points, bounding boxes, and masks. 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 to update the memory.

Memory Module. SAM2 incorporates a memory module to condition the features of the current frame on both previous frames and user-provided prompts. Each memory entry consists of two elements: (1) the spatial embedding combined with the predicted mask and (2) the object-level pointer (i.e. mask token). By cross-attending to this memory, the model ensures that current frame features capture both fine-grained correspondences and high-level semantic information.

3.2. Track generation

Object representation in SAM2. Revisiting the architecture of SAM2 [23], the object pointer stored in the spatio-temporal memory bank serves as an auxiliary high-level semantic representation of the objects to be segmented. Based on this, we hypothesize that the sequence of object pointers for each segment, gathered from the entire video, retains temporal semantic and motion information of the corresponding region. Therefore, as we generate N mask track candidates $\mathcal{M} \in \{0, 1\}^{N \times T \times H \times W}$ by SAM2, we simultaneously extract object pointers and concatenate them along the temporal dimension. We define these as object tokens, $\mathcal{O} \in \mathbb{R}^{N \times T \times d}$, of \mathcal{M} , where d is the feature dimension. These object tokens serve as high-level video object representation for determining alignment with language.

Grid point prompt. Considering that SAM2 [23] is a promptable segmentation model, a key consideration is how to effectively prompt SAM2 to ensure it generates all the potential mask tracks. Since some target objects in the dataset only appear momentarily in video, we adopt a strategy of selecting frames at predefined frame intervals across the entire video as a prompt frame I_p for mask generation. This approach allows us to propagate forwards and backwards to obtain the mask tracks, ensuring that no object is missed.

During the training phase, to enhance the robustness of the model, we extract grid point-originated tracks for diversity, regardless of foreground or background. We first provide N_g grid points $\mathcal{P} = \{P^i\}_{i=1}^{N_g}$ along with frame I_p as input to SAM2, and generate N_g of binary masks M_p as

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

where $\text{SAM2}_{\text{Image}}(\cdot)$ is SAM2 image predictor. The prompt masks, M_p are recursively propagated through the input

video \mathcal{I} into the SAM2 video predictor $\text{SAM2}_{\text{Video}}(\cdot)$, to obtain mask tracks \mathcal{M} and object tokens \mathcal{O} as

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

Grounded prompt. Generating tracks based on grid points produces an unnecessarily large number of mask tracks, resulting in increased latency. Thus at inference stage, we leverage visual grounding model, specifically Grounding DINO (GDINO) [17], enabling it to localize primarily around object candidates relevant to the given expression \mathcal{E} in I_p .

In our framework, GDINO takes I_p and an expression \mathcal{E} as input and predicts N_D bounding boxes B_p as follows:

$$B_p = \text{GDINO}(I_p; \mathcal{E}), \quad (3)$$

where $B_p \in \mathbb{R}^{N_D \times 4}$. B_p localize objects within the frame that are likely associated with \mathcal{E} . B_p are then used by $\text{SAM2}_{\text{Image}}(\cdot)$ to recursively generate corresponding segments M_p , which are propagated to $\text{SAM2}_{\text{Video}}$.

$$\begin{aligned} M_p &= \text{SAM2}_{\text{Image}}(I_p; B_p) \\ \mathcal{O}, \mathcal{M} &= \text{SAM2}_{\text{Video}}(\mathcal{I}; M_p). \end{aligned} \quad (4)$$

Since GDINO [17] tends to primarily localize foreground objects, relying solely on GDINO can result in insufficient contextual information. To address this, we additionally extract a subset of grid-originated object tokens by prompting on the first, middle, and last frames of the video.

3.3. Language-aligned track selection

Once we have successfully gathered N_c of candidate mask tracks and their consistent feature representations, object tokens, 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 visual and linguistic features and outputs scores reflecting mask track-expression correspondence. We define these scores as alignment score s_a , representing the probability of selection. Thus, the module takes object tokens and text token \mathcal{T} as input, and produces $s_a \in \mathbb{R}^{N_c}$ along with an aligned object token $\mathcal{O}_a \in \mathbb{R}^{N_c \times d}$. Here, \mathcal{O}_a , a temporally aggregated, language-aligned object token, serves as the video-level object representation.

$$\mathcal{O}_a, s_a = \text{TS}(\mathcal{O}; \mathcal{T}), \quad (5)$$

where $\text{TS}(\cdot)$ denotes track selection module and $\mathcal{T} \in \mathbb{R}^{N_c \times N_w \times d}$ represents the embedding of N_w tokenized words extracted by the text encoder.

As depicted in Figure 3, the track selection module is composed of initial short-term motion encoder followed by L object-language alignment layers and final language aligned object aggregation block.

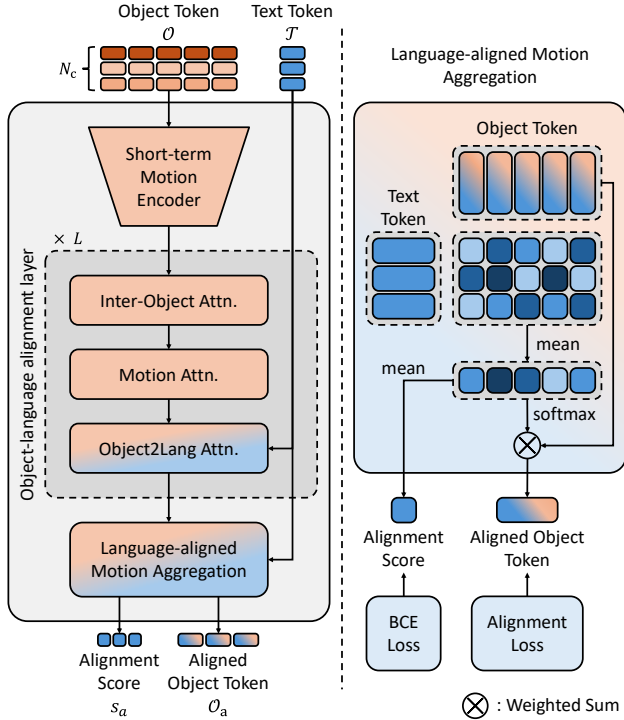


Figure 3. **Architecture of the language-aligned track selection module.** This selection module effectively encodes the dynamics of objects and aligns them with language.

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 to achieve effective alignment of two modalities. 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 $\mathcal{O}^i \in \mathbb{R}^{T \times d}$.

Object-language alignment layer. As the main component of track selection module, the object-language alignment layers sequentially perform 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 inter-object attention and motion attention, respectively. Both inter-object and motion attention are standard self-attention [26], but each operate along a different dimension for distinct pursuit. Inter-object attention is applied to all objects within the same frame $\mathcal{O}_t \in \mathbb{R}^{N_c \times d}$. As we aforementioned in Section 3.2, track selection module is trained using grid-originating tracks including both foreground and background tracks evenly. Thus inter-object attention captures both object relations and object-background interactions, leading to a comprehensive under-

standing of the global context. We discover its importance in Section 4.5. In contrast, motion attention aims to aggregate long-term motion information for each object across the video, operating on the temporal dimension of each object token \mathcal{O}^i .

With the motion information of object tokens enhanced through the preceding inter-object and motion attention, we inject text information using object-to-language cross-attention as previous methods [6, 9]. This cross attention layer efficiently fuses text features into object tokens \mathcal{O} , where the query, key are the \mathcal{O} and \mathcal{T} , thereby generating the language aware object token \mathcal{O}' . Finally, \mathcal{O}' is utilized as an input for language-aligned object aggregation block.

Language-aligned object aggregation. The language-aligned object aggregation block receives language-aligned \mathcal{O}' and produces s_a as well as an \mathcal{O}_a that serves as the object representative. We define $\mathcal{O}_a \in \mathbb{R}^{N_c \times d}$, which serves as the object representative, as weighted sum of each object token using the frame weighting matrix w_a , where w_a is specified by:

$$w_a = \text{softmax}(\text{Avg}(\mathcal{O}' \otimes \mathcal{T})), \quad (6)$$

where \otimes means matrix multiplication and $\text{Avg}(\cdot)$ represents the mean along the last dimension. Thus, we can obtain \mathcal{O}_a and s_a as follows:

$$\begin{aligned} \mathcal{O}_a &= w_a \otimes \mathcal{O}', \\ s_a &= \text{sigmoid}(\text{Avg}(\mathcal{O}' \otimes \mathcal{T})). \end{aligned} \quad (7)$$

During inference stage, 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 .

Training objective. The overall loss \mathcal{L} is a combination of Binary Cross-Entropy (BCE) loss \mathcal{L}_{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}_{BCE} is applied to enforce alignment between the object features and the language, as follows:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=0}^N \left(y^i \log(s_a^i) + (1 - y^i) \log(1 - s_a^i) \right), \quad (8)$$

where y^i represents the ground truth 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} . Specifically $y^i = 1$ if the Intersection over Union (IoU) between \mathcal{M}^i and the segmentation mask of the target object is greater than a predefined threshold; otherwise $y^i = 0$. This guides the model to accurately determine whether an object token derived from a ground truth segment aligns with the expression when fused with the language feature, thereby enforcing correct language-object alignment.

The alignment loss $\mathcal{L}_{\text{align}}$ is a modified form of contrastive loss, which encourages each aligned object tokens \mathcal{O}_a 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 token \mathcal{T} , while the negative anchor $\mathcal{A}_n \in \mathbb{R}^d$ consists of N_{neg} learnable embeddings. $\mathcal{L}_{\text{align}}$ is defined as follows:

$$\mathcal{L}_{\text{align}} = -\frac{1}{N} \sum_{i=0}^N (y^i \mathcal{L}_{\text{pos}}(\mathcal{O}_a^i) + (1 - y^i) \mathcal{L}_{\text{neg}}(\mathcal{O}_a^i)), \quad (9)$$

where

$$\begin{aligned} \mathcal{L}_{\text{pos}} &= d(\mathcal{O}_a^i, \mathcal{A}_p) - \sum_{j=1}^{N_{\text{neg}}} d(\mathcal{O}_a^i, \mathcal{A}_n^j), \\ \mathcal{L}_{\text{neg}} &= d(\mathcal{O}_a^i, \mathcal{A}_n^{k^*}) - d(\mathcal{O}_a^i, \mathcal{A}_p) - \sum_{j=1, j \neq k^*}^{N_{\text{neg}}} d(\mathcal{O}_a^i, \mathcal{A}_n^j). \end{aligned} \quad (10)$$

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 token to the aligned object token.

4. Experiments

4.1. Datasets and evaluation metrics

Dataset. We evaluated our method on three video datasets: MeViS [6], Ref-YouTubeVOS [16], 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-YouTubeVOS is the largest RVOS dataset, containing 3,978 videos with approximately 13,000 annotations. Ref-DAVIS builds upon DAVIS17 [21] by incorporating linguistic annotations for a variety of objects, featuring a total of 90 videos.

Evaluation metrics. Following prior research [6, 9, 19], 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

We structure our proposed pipeline into decoupled process: the track generation and the track selection. The training is conducted exclusively on the track selection.

Methods	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}
URVOS [25]	27.8	25.7	29.9
LBDT [7]	29.3	27.8	30.8
MTTR [3]	30.0	28.8	31.2
ReferFormer [27]	31.0	29.8	32.2
VLT+TC [5]	35.5	33.6	37.3
LMPM [6]	37.2	34.2	40.2
HTR [19]	42.7	39.9	45.5
DsHmp [9]	46.4	43.0	49.8
*VideoLISA [1]	44.4	41.3	47.6
*VideoGLaMM [20]	45.2	42.1	48.2
SOLA w/ RoBERTa	48.6	45.2	52.1
SOLA w/ CLIP	<u>47.3</u>	<u>43.6</u>	<u>50.9</u>

Table 1. **Quantitative comparison on MeViS [6].** The best results are in **bold** and the second best are underlined. * denotes llm-based methods.

Track generation. We employed GDINO [17] every fourth frame to generate prompt masks for potential objects. To ensure sufficient information for each potential object, we sequentially provided these prompt masks, starting with the largest mask, as inputs to SAM2 [23] to obtain mask tracks and object tokens. For each acquired mask track, we filtered out similar tracks based on their Intersection over Union (IoU) scores. Starting with the largest mask, we sequentially removed mask tracks with an IoU exceeding 0.7, following a process similar to Non-Maximum Suppression (NMS). We utilized the pre-trained Grounding DINO-T and the pre-trained SAM2 Hiera-L [24] models during track generation. To obtain diverse yet high-quality mask tracks, we filtered the mask tracks generated by SAM2. For the grid prompt, mask tracks were filtered based on stability score. Tracks with scores below the set thresholds were removed from the candidate tracks. We carefully defined thresholds for stability scores to ensure quality. Additionally, for GDINO, a box threshold of 0.2 was applied to filter confident bounding boxes.

Language-aligned track selection module. While the track selection, we train the track selection module exclusively on the MeViS dataset, as all experiments on the Ref-YouTube-VOS and Ref-DAVIS datasets are conducted in a zero-shot setting. The text encoder utilizes both RoBERTa [18] and CLIP [22], with training conducted over 13 epochs for the RoBERTa-based model configuration and 11 epochs for the CLIP-based configuration. The loss function components are weighted with $\lambda_1 = 1.0$ and $\lambda_2 = 0.3$, and the initial learning rate is set to $5e-6$, gradually decreasing throughout training. The training process requires approximately 8 hours on a single RTX 3090 GPU.

Methods	# of trainable parameters	Ref-Youtube-VOS			Ref-DAVIS		
		$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}
LMPM [6]	66.4M	31.5	30.0	32.9	39.9	36.7	43.2
ReferFormer [12]	70.3M	35.0	34.2	35.8	40.5	36.8	44.2
DsHmp [9]	92.4M	45.8	<u>43.7</u>	47.9	42.6	37.8	47.3
SOLA w/ RoBERTa	<u>32.9M</u>	<u>46.6</u>	42.7	<u>50.6</u>	<u>45.4</u>	<u>43.0</u>	47.7
SOLA w/ CLIP	19.7M	51.3	47.8	54.8	45.5	43.3	<u>47.6</u>

Table 2. **Zero-shot quantitative comparison on Ref-YouTubeVOS [16] and Ref-DAVIS [13].** The best results are in **bold**, and the second best are underlined. The models are trained on the training set of MeViS [6] and evaluated on Ref-YouTubeVOS and Ref-DAVIS.

Methods	Algorithm	$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}
LMPM [6]	Motion blur	33.3	31.2	35.4
ReferFormer [12]		26.3	25.4	27.1
DsHmp [9]		38.0	35.0	41.1
SOLA w/ RoBERTa		39.8	36.6	43.0
SOLA w/ CLIP		<u>38.6</u>	<u>35.4</u>	<u>41.9</u>
LMPM [6]	Gaussian noise	36.0	33.4	38.6
ReferFormer [12]		26.9	24.0	29.9
DsHmp [9]		43.4	39.5	47.2
SOLA w/ RoBERTa		44.4	40.5	48.3
SOLA w/ CLIP		<u>43.7</u>	<u>39.9</u>	<u>47.6</u>

Table 3. **Quantitative result on a corrupted version of MeViS [6].** The best results are in **bold** and the second best are underlined. 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.

Composing corresponding mask tracks As described in Sec 3.3, when the alignment score of an object token predicted by the proposed track selection module exceeds the threshold τ , the corresponding mask track is selected. Tracks with alignment scores above the threshold are merged at the frame level to produce a single integrated mask track. In our experiments, we set τ to 0.5 for the RoBERTa implementation and 0.3 for the CLIP implementation, respectively.

Labeling. To annotate ground truth labels from grid prompts, we label each candidate object mask track based on its Intersection over Union (IoU) with ground-truth object mask tracks. If the IoU between a candidate mask track and any ground-truth track in the set exceeds 0.7, we label the candidate as a positive sample for the given expression; otherwise, it is labeled as a negative sample, i.e. background token.

4.3. Quantitative results

Main results. Table 1 presents the quantitative results of our method in a fully-supervised setting. We evaluated comprehensively on the MeViS dataset, widely regarded as the most challenging dataset in the RVOS field. Our method achieves the state-of-the-art performance, underscoring its effectiveness.

Zero-shot evaluation. Since our method leverages the object tokens obtained from SAM2, we also conducted a zero-shot experiment to demonstrate the generalization capability of our approach. We trained our model on the MeViS dataset and tested it on the Ref-YouTube-VOS and Ref-DAVIS datasets. As shown in Table 2, SOLA achieved superior performance, surpassing the previous state-of-the-art method. Moreover, with respect to the number of learnable parameters used in training, our method demonstrates both efficiency and effectiveness.

Robustness on corrupted data. 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 substantiates its robustness and highlights its suitability for practical applications. Table 3 presents the quantitative results, showing that our proposed method outperforms previous approaches [6, 9, 12] even under image corruption scenarios. This robustness is attributed to reformulating the RVOS into two sub-problems, one of which addresses the track generation problem by effectively leveraging the generalized mask tracker, SAM2.

4.4. Qualitative results

In Figure 4, our proposed SOLA method demonstrates exceptional temporal consistency and precise vision-language alignment. The model accurately captures both appearance cues—such as “The cat” and “The cow” attributes—and

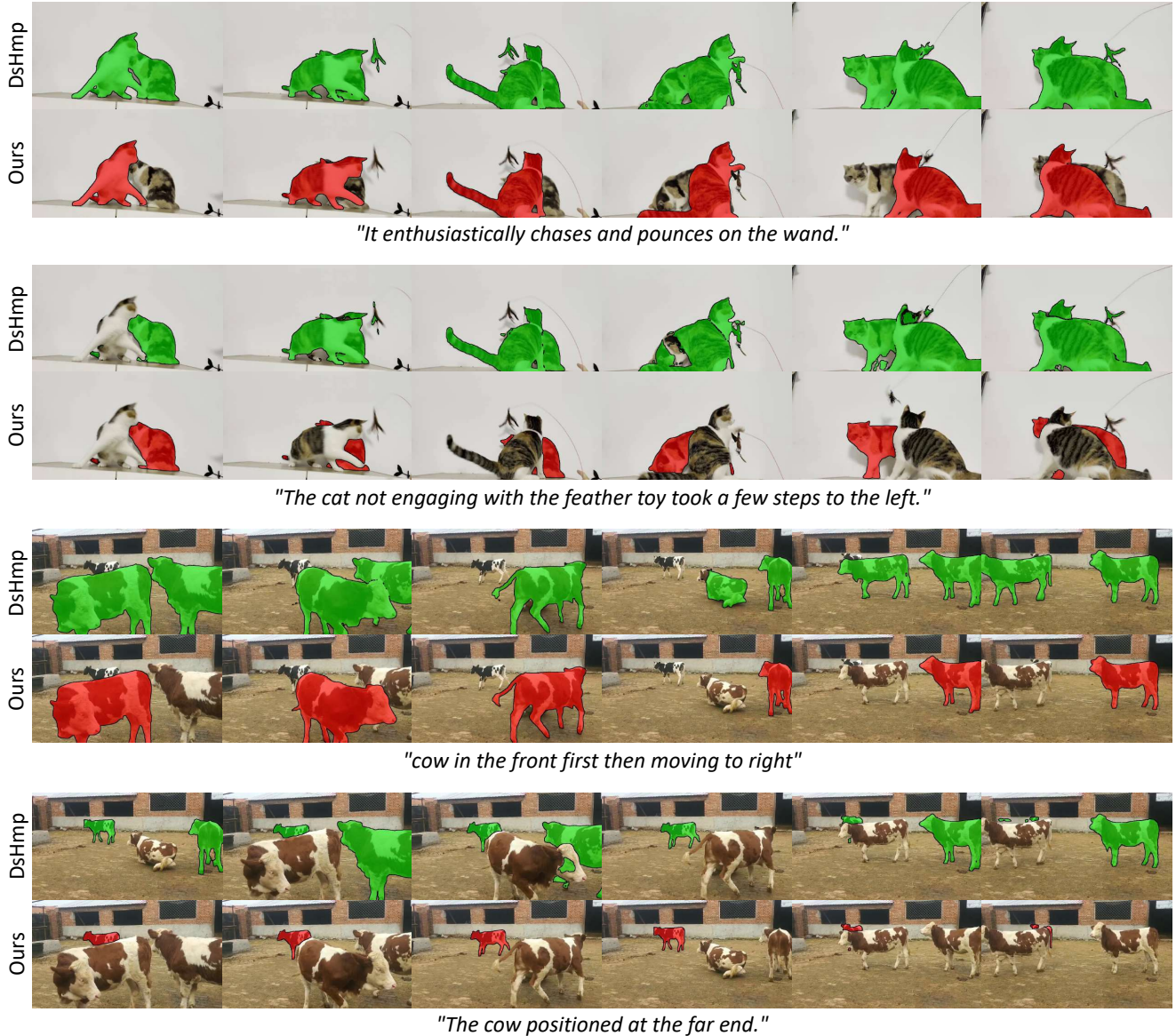


Figure 4. Qualitative results of our model on MeViS.

complex motion cues, including “moving to right”. This capability extends even to challenging scenarios in zero-shot settings, where our model generalizes effectively across diverse video domains. This robustness is achieved by leveraging the knowledge from vision foundational model, SAM2, which provides rich object tokens that reinforce our model’s understanding of both static and motion-based language cues. Compared to previous approaches, our method consistently maintains mask track coherence and adapts flexibly to varied visual contexts, even in cases where expressions rely solely on motion (e.g., “chases”, “pounces”). These results highlight SOLA’s unique ability to preserve object continuity and match language cues

with visual motion, a challenge that many existing RVOS models struggle to address. Overall, our model’s qualitative performance underscores the advantages of a lightweight yet effective vision-language alignment module in achieving high-fidelity segmentation across a range of challenging scenarios. We show our additional qualitative results in Appendix A.

4.5. Ablation study

We conducted several ablation studies on the MeViS dataset using RoBERTa as the text encoder to evaluate the effectiveness of our proposed selection method and the impact of background token presence. Additionally, we per-

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 the proposed selection method.

formed further ablation studies on associated loss functions and components within the language-aligned track selection module.

Effect of the proposed selection method. The quantitative results in Table 4 demonstrate that our track-selection-based method effectively addresses the challenges of RVOS. **w/o selection module** indicates an approach solely based on Grounding DINO [17] without specific track selection, while **w/ selection module** represents our framework, SOLA, which selects the referred track based on motion-encoded, language-aligned object tokens.

Existence of background object tokens. The quantitative results presented in Table 5a underscore the critical role of incorporating background object tokens during both training and inference. Given that the inter-object attention mechanism within the track selection module is designed to capture relationships among diverse object tracks, the inclusion of background object tokens in both training and inference significantly enhances performance. This comprehensive interaction between foreground and background objects proves indispensable, enabling a more holistic video-level understanding of language.

Ablation on losses. In Table 5b, we evaluate the model’s performance under different loss configurations. When using only BCE loss (**w/o \mathcal{L}_{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}_{align}**). This result indicates that alignment loss enhances the model’s discriminative ability, improving its alignment with the given expression.

Ablation on different types of attention. Table 5c presents the model’s performance with different types of attention layers. When only motion attention is used, the model aggregates long-term temporal information across frames, enhancing motion modeling but lacking insight into relationships among objects within each frame. Conversely, using only inter-object attention encodes spatial relationships among all objects in each frame, including both foreground and background, providing a comprehensive spatial context but lacking temporal continuity and object motion information. The configuration that combines both motion and inter-object attention represents our full method, enriching the model’s understanding of both spatial and temporal aspects and resulting in a more robust representation of object tokens.

bg. tokens (train)	bg. tokens (inference)	Metrics		
		$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}
\times	\times	45.7	42.4	48.9
\checkmark	\times	47.5	43.9	51.1
\checkmark	\checkmark	48.6	45.2	52.1

(a) Effects of including background object tokens.

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

(b) Different loss functions.

inter-object attn.	motion attn.	Metrics		
		$\mathcal{J}\&\mathcal{F}$	\mathcal{J}	\mathcal{F}
\times	\checkmark	44.3	41.6	47.0
\checkmark	\times	44.9	42.2	47.0
\checkmark	\checkmark	48.6	45.2	52.1

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

# 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

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

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

Ablation on the number of object-language alignment layers Table 5d shows the results of using different numbers of attention block layers. With a single layer, the model has limited attention capacity, potentially missing finer spatial-temporal relationships. Our two-layer setup achieves a balanced representation, effectively capturing object interactions and motion patterns. However, using three layers can result in overfitting, leading to slight performance degradation.

5. Conclusion and Discussion

We propose SOLA, a novel framework that simplifies the RVOS task into a selection problem. It leverages a foundation model to generate consistent candidate mask tracks and selects the most relevant one based on a given expression. Through our lightweight vision-language module, which captures appearance and motion, SOLA selects the mask track that best corresponds to the expression. Combined with state-of-the-art video segmentation models, our approach achieves leading performance on RVOS benchmarks, demonstrating exceptional generalization ability.

References

- [1] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *arXiv preprint arXiv:2409.19603*, 2024. 6
- [2] Miriam Bellver, Carles Ventura, Carina Silberer, Ioannis Kazakos, Jordi Torres, and Xavier Giro-i Nieto. A closer look at referring expressions for video object segmentation. *Multimedia Tools and Applications*, 82(3):4419–4438, 2023. 2
- [3] 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. 1, 2, 6
- [4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2
- [5] 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
- [6] 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, 5, 6, 7, 12, 13, 14, 15, 17, 18
- [7] 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
- [8] Kirill Gavriluk, 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
- [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. 1, 2, 5, 6, 7, 12, 13, 14, 15, 16
- [10] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 7
- [11] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. *Advances in Neural Information Processing Systems*, 35:31265–31277, 2022. 2
- [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. 2, 7
- [13] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *ACCV*, 2018. 2, 6, 7
- [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. 3
- [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, 6, 7
- [17] 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, 3, 4, 6, 9
- [18] 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. 6
- [19] 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. 6
- [20] Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. Videoglamm: A large multimodal model for pixel-level visual grounding in videos. *arXiv preprint arXiv:2411.04923*, 2024. 6
- [21] 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
- [22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 6
- [23] 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. 2, 3, 4, 6, 12
- [24] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu

- Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International Conference on Machine Learning*, pages 29441–29454. PMLR, 2023. [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](#), [2](#), [6](#), [12](#), [16](#)
- [26] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. [5](#)
- [27] 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. [1](#), [2](#), [6](#)
- [28] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. [19](#)

Referring Video Object Segmentation via Language-aligned Track Selection

Supplementary Material

A. Additional qualitative results

We provide the qualitative results for our model, emphasizing its effectiveness in addressing the challenges of referring video object segmentation (RVOS).

A.1. Qualitative results on MeViS

Figure A.1 presents the qualitative results on MeViS [6], comparing the performance of DsHmp [9] with our proposed method, SOLA. Our approach consistently demonstrates superior capability in accurately segmenting 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.*”). Even in such scenarios, our proposed selection module performs motion modeling based on consistent mask tracks, enabling the model to effectively encode motion information for language alignment. Consequently, it can establish correspondence with the expression using motion cues from the language alone, independently of appearance-based features.

A.2. 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 highlight our model’s remarkable capability to generalize across diverse videos and expressions, despite not having seen the dataset during training. Such zero-shot generalization underscores the strength of our approach in leveraging pre-trained knowledge of SAM2 [23] and aligning it efficiently with natural language expressions.

A.3. Qualitative results on MeViS with image corruption.

Figures A.5 and A.6 visualize the results presented in Table 3 of the main paper. These results demonstrate that SOLA consistently retains its ability to generate high-quality final mask outputs regardless of the level of distortion in the input data.

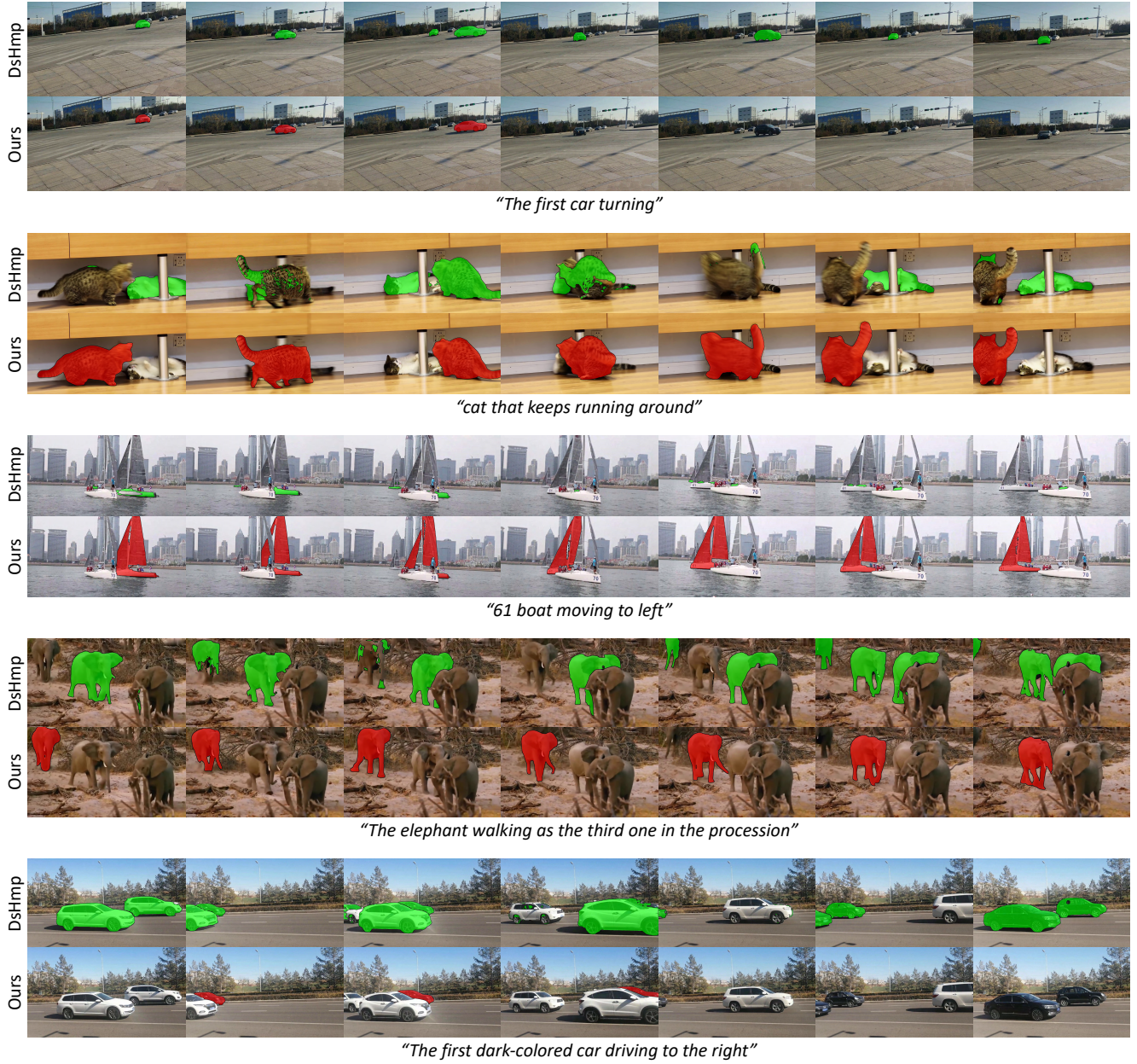
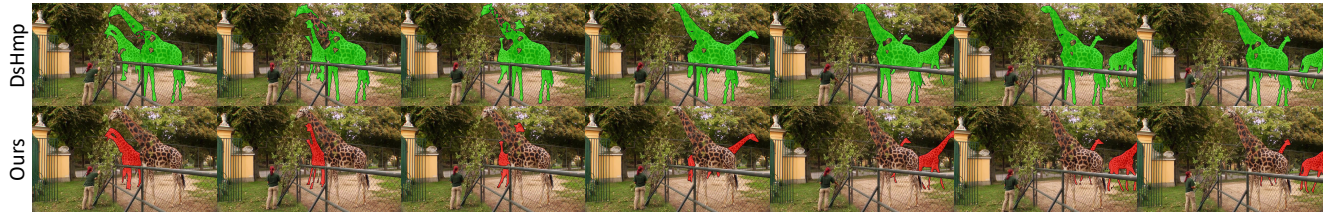
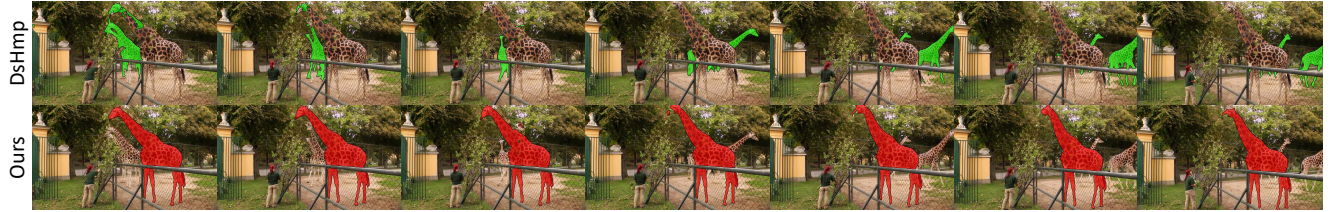


Figure A.1. **Qualitative results on MeViS [6].** Our proposed method outperforms previous state-of-the-art approaches [9] in terms of mask quality and tracking ability, while ensuring accurate segmentation of the corresponding object based on the given expression.



"The two giraffes turning around and leaving."



"Giraffe standing in place and grazing"



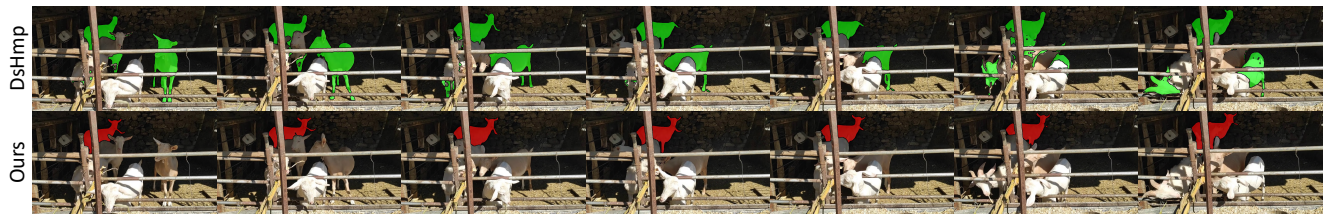
"baby tiger without moving position"



"The small tiger progressing to the area behind the big tiger"



"goat moving from rightmost to the middle"



"The distant sheep, grazing at the corner of the wall"

Figure A.2. **Qualitative results on MeViS [6].** Our proposed method outperforms previous state-of-the-art approaches [9] in terms of mask quality and tracking ability, while ensuring accurate segmentation of the corresponding object based on the given expression.

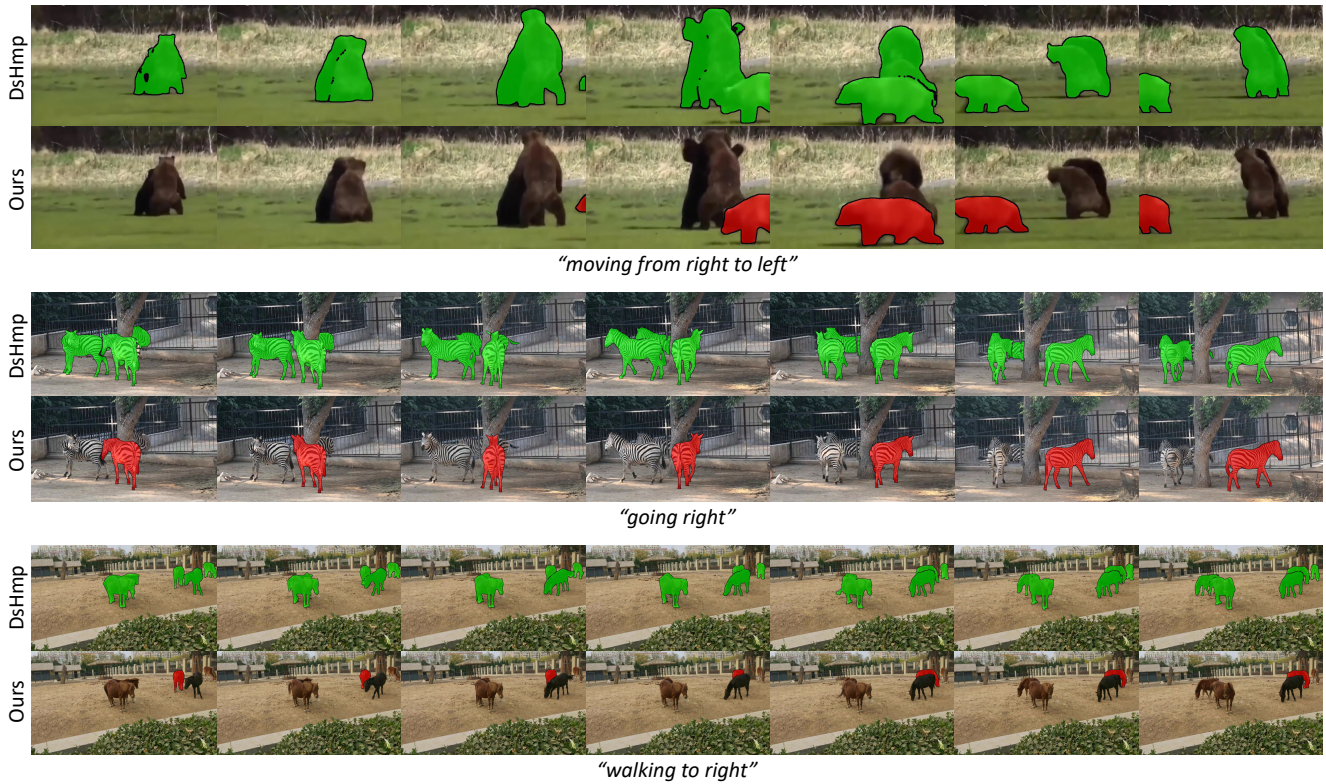


Figure A.3. **Qualitative results on MeViS [6].** Our proposed method outperforms previous state-of-the-art approaches [9] in terms of mask quality and tracking ability, while ensuring accurate segmentation of the corresponding object based on the given expression. Notably, despite the given expression focusing solely on motion information, our model effectively handles the task without relying on appearance cues.

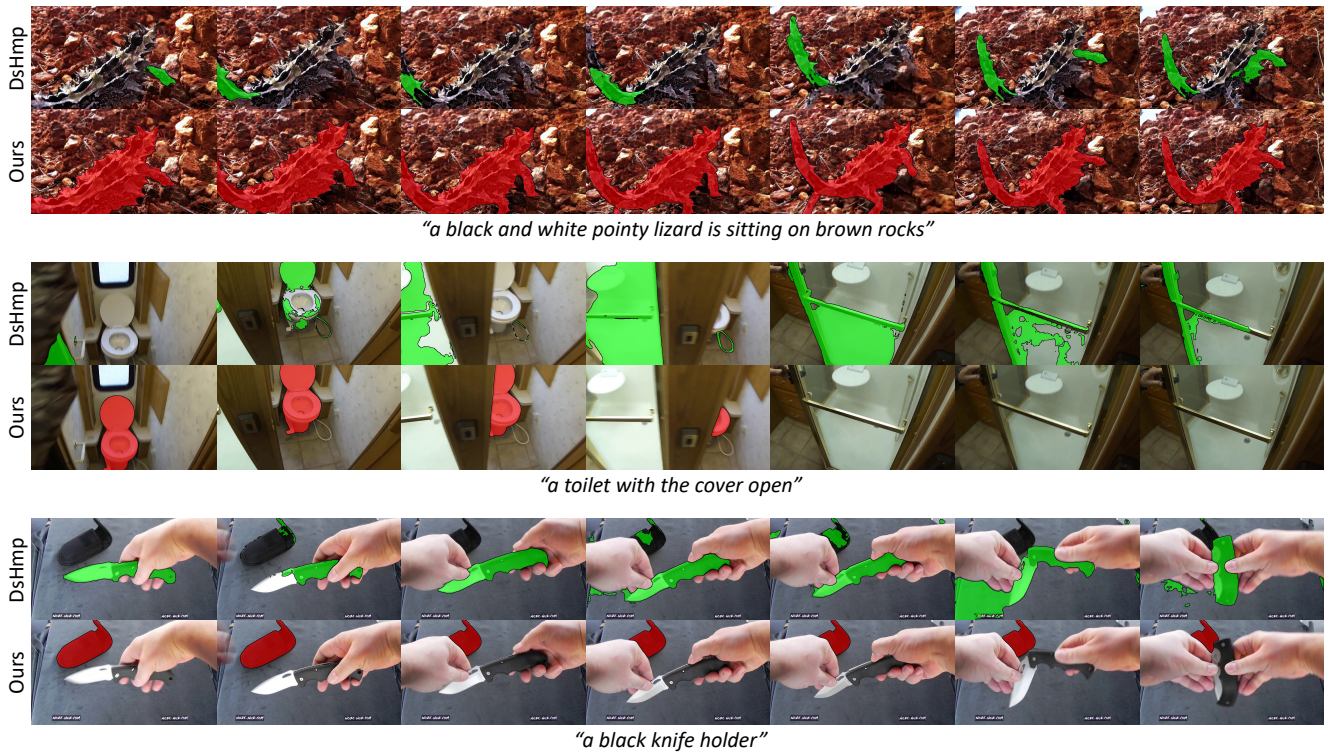
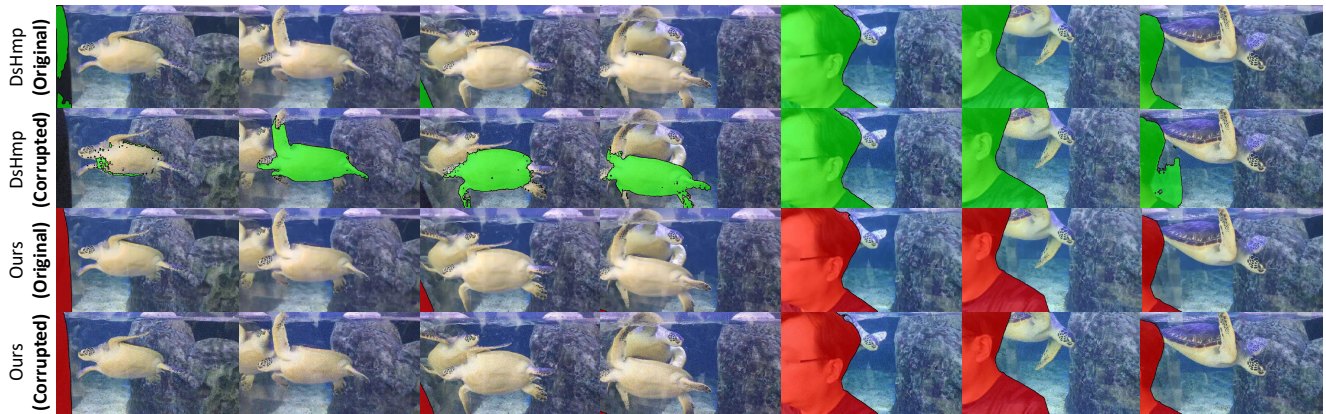
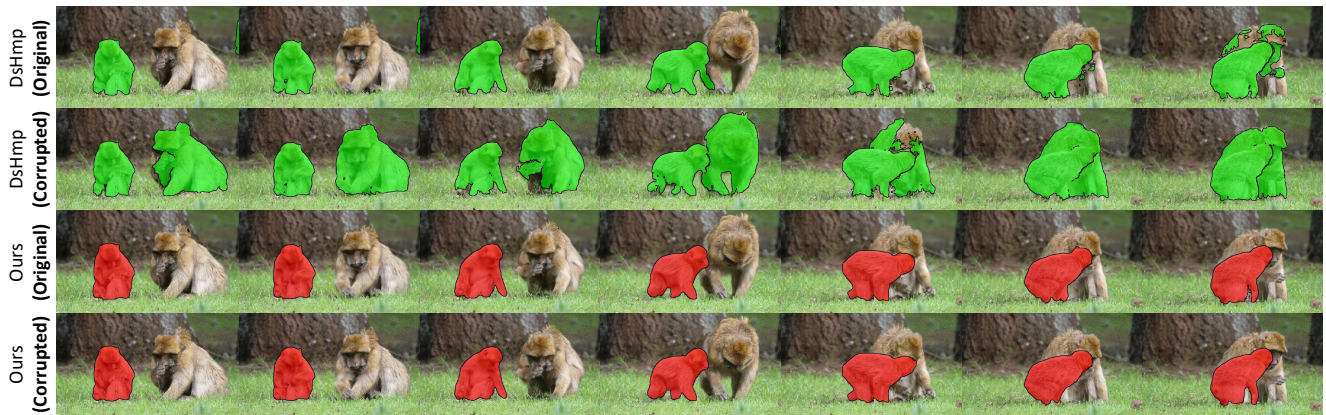


Figure A.4. **Qualitative results on Ref-Youtube-VOS [25].** Our proposed method outperforms previous state-of-the-art approaches [9] in terms of mask quality and tracking ability, while ensuring accurate segmentation of the corresponding object based on the given expression.

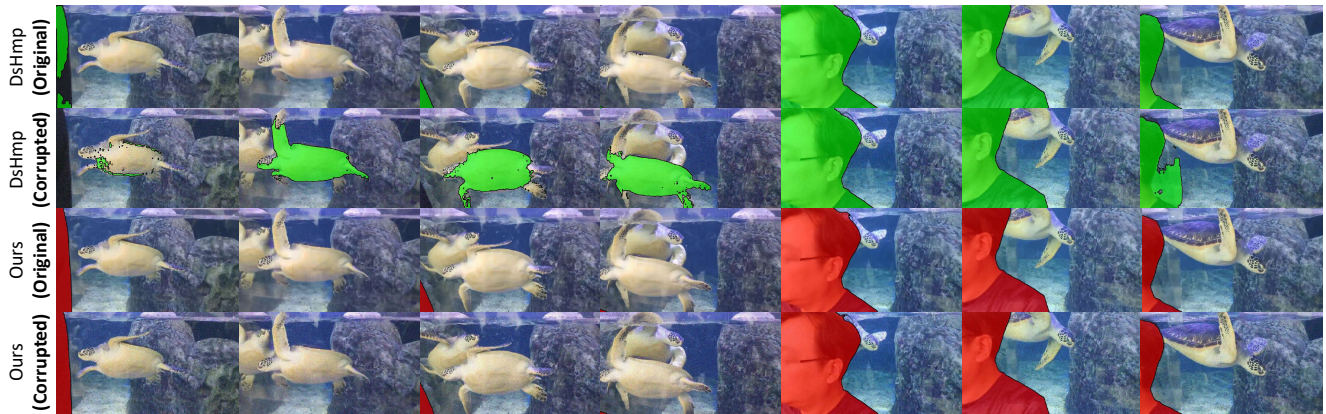


"The onlooker standing close to the turtle display"

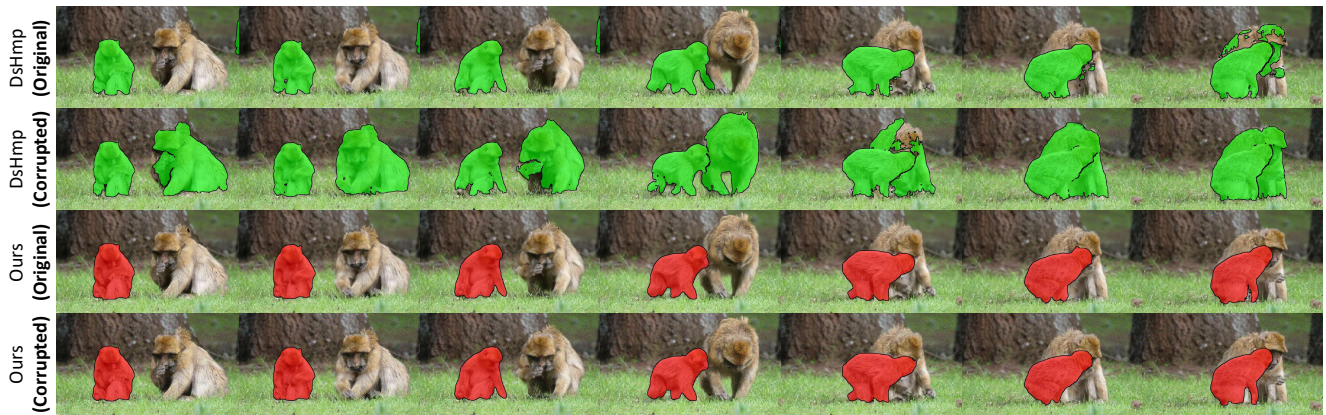


"After being on the left side, the monkey moves a little and ends up in front of the other one."

Figure A.5. **Qualitative results on corrupted version of MeViS [6].** 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.



"The onlooker standing close to the turtle display"



"After being on the left side, the monkey moves a little and ends up in front of the other one."

Figure A.6. **Qualitative results on corrupted version of MeViS [6].** 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.

B. Limitations and future works

B.1. Limitations

Moreover, 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. We aim to explore tuning the text encoder to capture features that are particularly beneficial for the RVOS task.

As discussed in the main paper, we further address additional limitations of SOLA. Since our framework leverages two different modal encoders from independent domains, aligning both modalities with scarce data may limit its performance. Specifically, MeViS only has 2,006 videos which is extremely smaller than other multi-modal benchmarks, specially video captioning datasets [28]. Additionally, while current state-of-the-art mask trackers demonstrate impressive performance, they still struggle to consistently track objects in scenarios with heavy occlusion and motion blur. Since the performance of SOLA relies heavily on the quality of the mask tracker, these challenges can lead to a degradation in SOLA’s overall performance. We anticipate that SOLA could achieve even better results as more advanced mask trackers are developed.

B.2. Future work

In this section, we discuss future work aimed at addressing the limitations of SOLA while ensuring its efficiency. Since our method generates candidate mask tracks by simultaneously utilizing grid and grounded prompts, running SOLA in real time is challenging due to enormous input prompts. However, tasks requiring real-time inference, such as autonomous driving, demand methods that are both efficient and fast. Additionally, it is crucial for SOLA to operate in an online manner, segmenting tracks without processing the entire video at once. We aim to explore future directions that overcome these limitations, enabling SOLA to perform real-time inference effectively.