Long-RVOS: A Comprehensive Benchmark for Long-term Referring Video Object Segmentation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Referring video object segmentation (RVOS) aims to identify, track and segment the objects in a video based on language descriptions, which has received great attention in recent years. However, existing datasets remain focus on short video clips within several seconds, with salient objects visible in most frames. To advance the task towards more practical scenarios, we introduce **Long-RVOS**, a large-scale benchmark for long-term referring video object segmentation. Long-RVOS contains 2,000+ videos of an average duration exceeding 60 seconds, covering a variety of objects that undergo occlusion, disappearance-reappearance and shot changing. The objects are manually annotated with three different types of descriptions to individually evaluate the understanding of static attributes, motion patterns and spatiotemporal relationships. Moreover, unlike previous benchmarks that rely solely on the per-frame spatial evaluation, we introduce two new metrics to assess the temporal and spatiotemporal consistency. We benchmark 6 state-of-the-art methods on Long-RVOS. The results show that current approaches struggle severely with the long-video challenges. To address this, we further propose ReferMo, a promising baseline method that integrates motion information to expand the temporal receptive field, and employs a local-to-global architecture to capture both short-term dynamics and long-term dependencies. Despite simplicity, ReferMo achieves significant improvements over current methods in long-term scenarios. We hope that Long-RVOS and our baseline can drive future RVOS research towards tackling more realistic and long-form videos. Our dataset and code will be released.

1 Introduction

Referring Video Object Segmentation (RVOS) [2, 7, 44] is an emerging task that aims to identify, track and segment the object in the video based on a natural language description. Unlike traditional semi-supervised VOS models that require first-frame masks as the object prompt, RVOS models rely solely on text descriptions to segment the target. Considering its potential applications like video editing, growing efforts have been devoted to this field [7, 15, 25, 30, 33]. Recently, the advent of multi-modal large language models [17, 27, 51] and segment anything models [21, 35] has further accelerated this progress [1, 47, 50, 54].

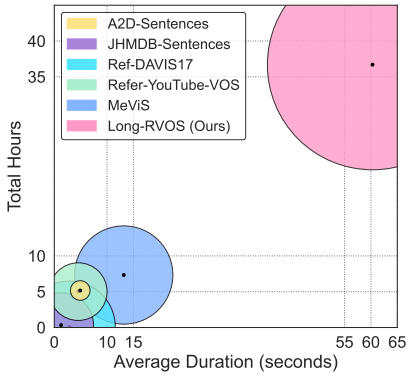


Figure 1: Duration comparison of current RVOS datasets. The circle size indicates the number of frames.

Despite these advances, current RVOS datasets [7, 11, 20, 36] remain limited to short video clips lasting only a few

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Figure 2: Examples from Long-RVOS dataset, with frame indices displayed in the upper left, and selected objects masked in orange ■. Long-RVOS contains extensive long-term videos, where the objects always undergo occlusion, disappearance-reappearance and shot changing. In addition, the objects are annotated with three different types descriptions: *static*, *dynamic* and *hybrid*.

seconds, with target objects clearly visible in most frames. For state-of-the-art (SOTA) methods, in order to capture the target object effectively, it is inevitable to integrate the text and spatiotemporal information throughout the video. However, when the video becomes longer, the number of distractors also increase accordingly, making it more challenging to perform sufficient spatiotemporal reasoning and capture the key information. Especially in RVOS, many text descriptions (e.g., "the cat jumps down") only refer to a brief fragment in the video. In addition, due to the GPU memory limitation, existing methods typically sample 4∼8 frames per video for training, but use all the frames during inference. As the video length increases, the gap between training and inference phases may become more pronounced. Despite these concerns, due to the lack of a long-term RVOS dataset, the exact challenges posed by longer videos remain unclear.

Another concern lies in the evaluation metrics. Existing RVOS benchmarks [7, 11, 20, 36] typically evaluate performance by simply averaging the frame-wise segmentation metrics (e.g., $\mathcal{J}\&\mathcal{F}$). However, in real-world videos, the target objects do not appear in every frame, due to occlusion and constrained camera views. Therefore, a robust RVOS model should exhibit a sound temporal consistency. This means it should not only accurately segment the target when it is present, but also be able to predict its absence by outputting an empty mask. However, this capability of temporal consistency can not be adequately reflected by current metrics.

To address these gaps, this work proposes **Long-RVOS**, a large-scale benchmark for long-term video object segmentation. Long-RVOS is the first minute-level dataset in RVOS field, designed to tackle various realistic long-video challenges such as frequent occlusion, disappearance-reappearance and shot changing, as shown in Figure 1 and Figure 2. Additionally, we introduce two new metrics for better evaluation of temporal consistency: tIoU, which measures the temporal overlap between predicted and ground-truth mask sequences; and vIoU, which further measures the spatiotemporal volume overlap between them. We benchmark 6 state-of-the-art (SOTA) methods on Long-RVOS. The results demonstrate that while notable progress has been achieved in existing short-term benchmarks, these SOTA models still significantly struggle in realistic long-term scenarios, in both frame-level segmentation and video-level temporal consistency.

To tackle the challenges posed by Long-RVOS, we present a baseline method **ReferMo**, which integrates additional motion frames to expand the temporal receptive field during training, and employs a local-to-global architecture to perceive both static attributes, short-term dynamics and long-term dependencies. Specifically, ReferMo decomposes each video into a sequence of clips, each consisting of a high-resolution keyframe and multiple low-resolution motion frames. Then, it perceives the static appearance and short-term motion within local video clip, and captures the global target in long-term context via inter-clip interactions. In this way, the temporal receptive field is

Table 1: Statistical overview of representative RVOS datasets. Long-RVOS features the longest video duration and the most diverse object classes. Besides, Long-RVOS offers explicit text description types for finer-grained evaluation.

Dataset	Year	Videos	Mean duration	Total duration	Mean frames	Masks	Objects	Object classes	Text	Text type
A2D-Sentences [11]	2018	3,782	4.9s	5.2h	3.2	58k	4,825	6	6,656	✗
JHMDB-Sentences [11]	2018	928	1.3s	0.3h	34.3	32k	928	1	928	✗
Ref-DAVIS17 [20]	2018	90	2.9s	0.1h	69.0	14k	205	78	1,544	✗
Refer-YouTube-VOS [36]	2020	**3,978**	4.5s	5.0h	27.2	131k	7,451	94	15,009	✗
MeViS [7]	2023	2,006	13.2s	7.3h	79.0	443k	**8,171**	36	**28,570**	✗
Long-RVOS (ours)	2025	2,193	**60.3s**	**36.7h**	**361.7**	**2.1M**	6,703	**163**	24,689	✓

expanded from multiple frames to multiple clips, but the training cost does not increase significantly. Despite simplicity, ReferMo achieves significant improvements over existing RVOS approaches, serving a promising baseline for long-term referring video object segmentation.

Our contributions are summarized as follows: (i) We build Long-RVOS, the first large-scale benchmark for long-term RVOS. In Long-RVOS, we provide explicit description types and introduce new metrics to enable more comprehensive evaluation. (ii) We benchmark 6 state-of-the-art RVOS approaches on Long-RVOS, and propose a promising baseline ReferMo to address the challenges in long-video scenarios. These contributions establish a foundation for developing more robust RVOS models to handle the realistic long-term videos.

2 Related Works

RVOS Benchmarks. Given an object description, RVOS aims to identify, tracking and segment the referring object throughout the video. This task was initially introduced by Gavrilyuk et al. [11] and Khoreva et al. [20] in 2018, and has gradually become a popular topic in vision-language understanding. Gavrilyuk et al. [11] built A2D-Sentences and JHMDB-Sentences datasets, which focus on distinguishing different actors in a video through the descriptions about appearance and actions. Khoreva et al. [20] built Ref-DAVIS17 [20], which covers more diverse object types. Later, Ref-Youtube-VOS [36] was developed to further expand the benchmark scale in this field. Recently, MeViS [7] was proposed to highlight the importance of motion understanding in RVOS task. Despite the efforts, these benchmarks remain limited to short video clips lasting only a few seconds, with target objects clearly visible in most frames. Besides, they also lack sufficient evaluation mechanisms to consider the models' specific capabilities in various aspects.

RVOS Approaches. Recent RVOS approaches are mainly based on Transformer-based end-to-end architecture, represented by MTTR [2] and ReferFormer [44]. For an effective and consistent object identification across the frames, follow-up works [14, 15, 30, 39] focus on integrating more object-level temporal information. ReferDINO [25] further improves the object-level visual-language understanding by inheriting the object grounding capability of GroundingDINO [28]. Meanwhile, the recent emergence of segment anything models, i.e., SAM [21] and SAM2 [35], provides unique opportunity for downstream segmentation tasks. Some frontier studies [1, 5, 26, 47, 50] explore to incorporate SAM and SAM2 into RVOS approaches, achieving significant improvements on existing benchmarks. For example, VideoLISA [1] incorporates large language models with SAM for reasoning video segmentation. SAMWISE [5] integrate text prompts into SAM2 by inserting trainable adapters. While these models achieve great progress in current short-video benchmarks, their abilities and robustness in handling real-world long videos is still unclear.

Long-term Video Understanding. Real-world videos are always long, untrimmed, and involves multiple events. To promote research into long-term video understanding, many large-scale benchmarks [3, 10, 31, 43] have been constructed. However, these benchmarks are mainly constructed for video question answering and temporal action localization, containing only sparse annotations such as timestamps, action labels and captions. To support object-level long-term understanding, some datasets including VidOR [37] and LaSOT [9] also provide dense annotations of bounding boxes. However, long-video datasets with pixel-level dense annotations are still very scarce. Recently, LVOS [16] is built for long-term video object segmentation. However, it is limited in scale and lacks text annotation. In this work, we build Long-RVOS, the first large-scale benchmark for long-term video object segmentation, providing both pixel-wise annotations and diverse object descriptions.

3 Long-RVOS: A Comprehensive Benchmark for Long-term RVOS

3.1 Video Collection

Previous RVOS datasets [7, 11, 20, 36] were typically constructed by providing text annotations on their corresponding VOS datasets (e.g., DAVIS17 [34], YouTube-VOS-2019 [46] and MOSE [8]). However, the existing long-term VOS datasets like LVOS [16] are limited in scale (containing only 720 videos), and most videos feature only one object target. Therefore, in order to establish a large-scale and diverse RVOS benchmark, we bypass the existing VOS datasets and turn to integrate multi-source long video datasets. Specifically, we build Long-RVOS based on three long-video datasets: TAO [6], VidOR [37], and Ego-Exo4D [12]. Moreover, TAO is a federated dataset combining multiple sources like Charades [38], LaSOT [9], ArgoVerse [4], AVA [13], YFCC100M [41], BDD-100K [49], and HACS [53]. We select videos and objects based on the following criteria:

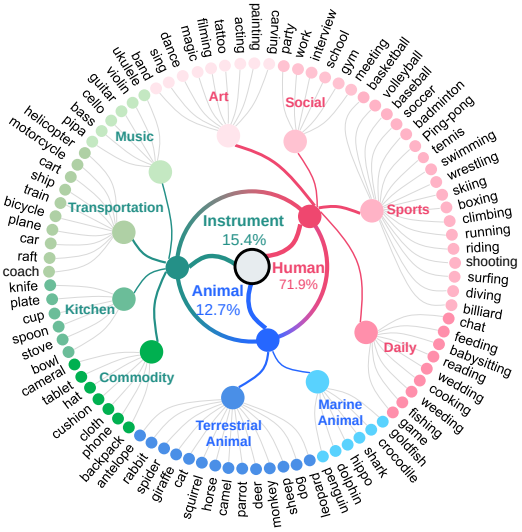


Figure 3: Overview of object categories and scenes in Long-RVOS.

- The video duration exceeds 20 seconds.
- Objects that belong to background, ambiguous or unknown categories are excluded.
- Each selected video must contain more than two valid objects, and at least one object is not continuously visible.

With these criteria, we have initially collected over 3K videos and 8K objects as candidates. After careful inspections on quality, we finally select 2,193 videos and 6,703 objects to build Long-RVOS.

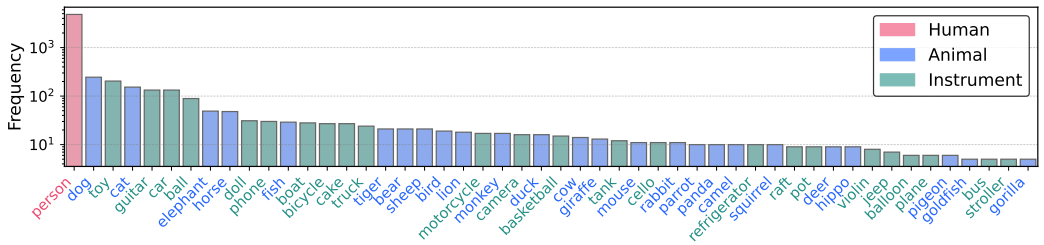
3.2 Dataset Annotation.

Text Annotation. We develop an online platform for annotating object descriptions. This platform randomly samples a video from our dataset and displays it, with all target objects highlighted by bounding boxes. To ensure the diversity of annotations, each video can be sampled repeatedly at most three times. The annotators consisting of 20 college students are asked to watch the videos and provide the following three types of descriptions for each object:

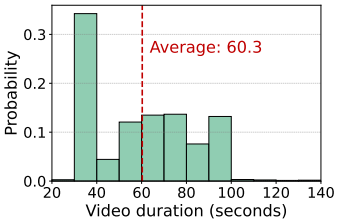
- **Static type** includes appearance (e.g., colors and shapes), relative position (e.g., "the left cat"), and environmental context (e.g., "on the grass").
- **Dynamic type** includes motions, changes over time (e.g., in position or state) and interactions with other entities (e.g., "the cat chasing a mouse").
- **Hybrid type** integrates both static and dynamic attributes to provide comprehensive object cues.

The key annotation principle is that every single description, regardless of type, must clearly distinguish the target object from others. For objects that cannot be distinguished by only static or dynamic attributes, the corresponding type of annotation can be skipped. After this annotation phase, we have collected over 30K text descriptions. These annotations and the corresponding videos are then sent to a validation team formed by three experts for quality verification. Any descriptions that violate our principle are directly removed. Besides, we do not use techniques like synonym replacement to artificially scale up the text annotations, keeping the dataset clear and authentic to support reliable RVOS training. Finally, we gather 24,689 high-quality descriptions for building Long-RVOS.

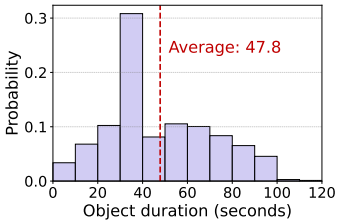
Mask Annotation. Our source datasets [6, 12, 37] have provided sparse bounding-box annotations. For each object, we segment the video into clips based on the annotated frames. Then, we utilize SAM2 [35], the state-of-the-art VOS model, to track the objects within each clip and produce high-quality masks, by regarding the annotated bounding box as the first-frame prompt. To ensure



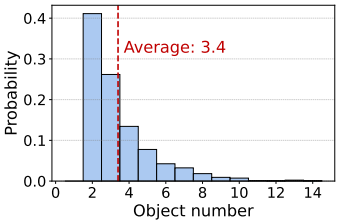
(a) Frequency distribution of the Top-50 object categories.



(b) Distribution of video duration.

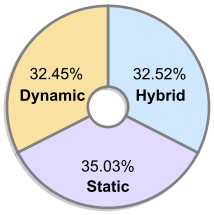


(c) Distribution of object duration.

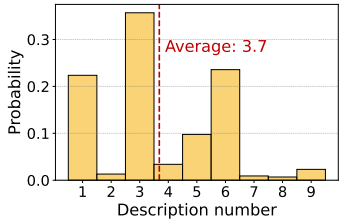


(d) Number of objects per video.

(e) Wordcloud of descriptions.



(f) Distribution of description types.



(g) Description number per object.

Figure 4: Representative statistics of Long-RVOS.

annotation quality, we conduct an iterative *check–correct* workflow. Specifically, the validation team checks every object's mask separately in the video, and marks the objects with inaccurate annotations. To facilitate the correction process, we develop an interactive annotation tool based on SAM2. This tool loads a marked object each time and visualizes its masks in the video. Nine annotators use our tool to refine the masks with point or box prompts, and remove masks from object-absent frames. The corrected results are then returned to the checking queue, and this *check–correct* loop repeats until all mask annotations are qualified.

3.3 Dataset Statistics

A detailed comparison with five existing RVOS datasets is shown in Table 1. Notably, Long-RVOS offers significantly longer video duration than existing datasets. In addition, it contains the largest number of object classes and mask annotations. The large scale of Long-RVOS supports comprehensive training and evaluation of RVOS models.

Diverse Objects and Scenes. Long-RVOS is constructed by integrating multiple sources of video datasets, achieving a wide variety of objects and scenes, as illustrated in Figure 3. These sources include indoor videos from Charades [38], outdoor videos from LaSOT [9], movie scenes from AVA [13], egocentric videos from Ego-Exo4D [12], and more diverse videos from other datasets [37, 41, 53]. In total, Long-RVOS contains 163 object categories, significantly surpassing the existing RVOS datasets. As shown in Figure 4 (a), while Long-RVOS primarily focuses on human instances (71.9%), it also covers a diverse range of animals (12.7%) and instruments (15.4%). In Figure 4 (b)-(d), we present further statistics on the videos and objects in Long-RVOS. Notably, the object number of each video spans from 2 to 14, preventing over-reliance on the most salient object and highlighting text-guided segmentation. With such extensive diversity, Long-RVOS can serve a comprehensive benchmark for RVOS research, facilitating the development of more real-world applications.

Diverse Descriptions. In real-world applications, user queries are always unpredictable. They might refer to salient attributes or instantaneous actions. To enable more comprehensive evaluation of model capabilities, Long-RVOS introduces three distinct types of text descriptions — *static*, *dynamic*, and *hybrid*. By explicitly categorizing these types, Long-RVOS prevents evaluation bias toward specific attribute cues (e.g., color or position), ensuring a fair and robust assessment. We present the detailed statistics of text descriptions in Figure 4 (c)-(g). Critically, Long-RVOS maintains a balanced distribution of text types, and the description number for each object can vary from 1 to 9. These properties encourage comprehensive learning of diverse object attributes. With its explicit type annotations and diverse object descriptions, Long-RVOS provides a comprehensive benchmark for training and evaluating RVOS models in more realistic scenarios.

3.4 Evaluation Metrics

Previous RVOS benchmarks tend to evaluate model performance with the frame-wise spatial metrics, such as $\mathcal{J}\&\mathcal{F}$. Here, $\mathcal{J}$ denotes the Intersection-over-Union (IoU) between the predicted and ground-truth masks, $\mathcal{F}$ measures the contour accuracy, and $\mathcal{J}\&\mathcal{F}$ is their average over all the frames. However, these metrics focus solely on the per-frame segmentation quality, neglecting the temporal consistency. A robust RVOS model should accurately segment the target when it is present and correctly output an empty mask when it is absent. Inspired by the field of spatiotemporal video grounding [40, 52], we additionally introduce two new metrics, tIoU and vIoU, in Long-RVOS to individually evaluate the temporal and spatiotemporal performance.

Formally, let $\hat{M}_t, M_t \in \{0,1\}^{H\times W}$ denote the predicted and ground-truth masks at t-th frame, respectively, where $t \in [1, T]$. The frame-index sets of non-empty masks are defined as $\hat{\mathcal{T}} = \{t \mid \|\hat{M}_t\|_0 > 0\}$ (for predictions) and $\mathcal{T} = \{t \mid \|M_t\|_0 > 0\}$ (for the ground-truth), where the ℓ_0-norm $\|\cdot\|_0$ denotes the count of non-zero elements. Then, tIoU is obtained by computing their IoU:

$$\mathrm{tIoU} = \frac{T_i}{T_u}, \quad \text{where } T_i = \hat{\mathcal{T}} \cap \mathcal{T} \text{ and } T_u = \hat{\mathcal{T}} \cup \mathcal{T}, \tag{1}$$

and vIoU computes the volume IoU between predicted and ground-truth mask sequences:

$$\mathrm{vIoU} = \frac{1}{T_u} \sum_{t\in T_i} \mathcal{J}_t, \quad \text{where } \mathcal{J}_t = \frac{\hat{\mathcal{M}}_t \cap \mathcal{M}_t}{\hat{\mathcal{M}}_t \cup \mathcal{M}_t}. \tag{2}$$

By combining the spatial metric $\mathcal{J}\&\mathcal{F}$, temporal metric tIoU and spatiotemporal metric vIoU, Long-RVOS establishes a rigorous evaluation protocol for RVOS research.

4 ReferMo: A Baseline Approach

As illustrated in Figure 5, ReferMo decomposes the video into a sequence of clips, each consisting of a high-resolution keyframe and subsequent low-resolution motion frames. Then, it perceives the static appearance and short-term motion within local video clip, and captures the object target in long-term context by integrating the cross-clip information. Critically, ReferMo only predicts target masks over the keyframes, and the masks on the remain frames are generated by a pretrained object tracker (e.g., SAM2 [35]). In this way, ReferMo achieves a trade-off between training costs and long-term understanding without processing a large number of high-resolution frames.

4.1 Video Decomposition

Typically, a long-term video is composed of multiple shots, and the video frames within each shot often show significant temporal redundancy. This redundancy can be efficiently described by motion information to reduce the frame-by-frame computations. Inspired by Video-LaVIT [18], we employ the MPEG-4 [23] compression technique to extract keyframe and motion information from the videos. More sophisticated (but expensive) keyframe selection strategies [42, 45] can also be explored, but they are not the primary focus of this work. In MPEG-4, a video is decomposed into multiple clips, where each clip consists of a keyframe $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ and the motion vectors $\mathcal{M} \in \mathbb{R}^{T\times \frac{H}{16}\times \frac{W}{16}\times 2}$ of its subsequent T frames. Unlike the dense optical flow, these motion vectors can be directly extracted during the compressed video decoding process, making them well-suited for processing large-scale, long-term videos. The details of motion extraction process are provided in the supplementary.

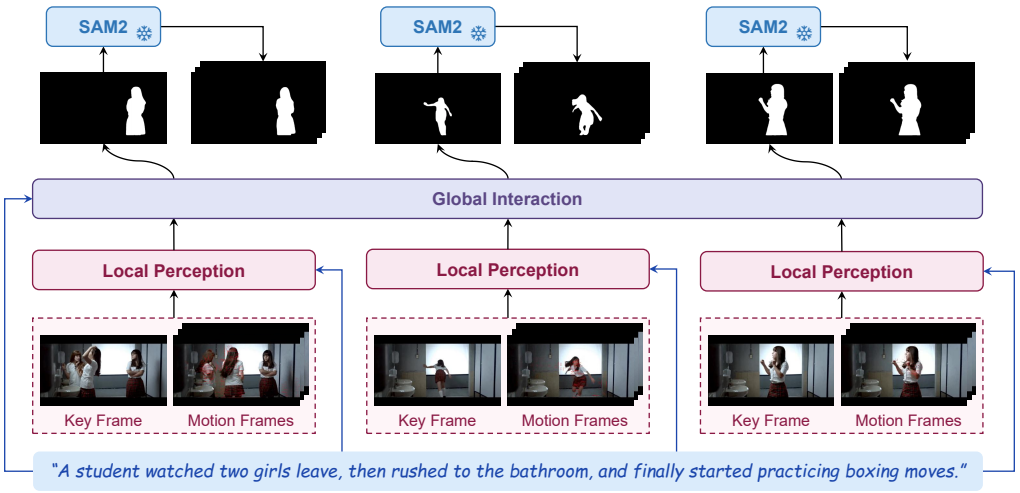


Figure 5: Overview of ReferMo. A video is decomposed into clips (keyframe + motion frames). ReferMo perceives the static attributes and short-term motions within each clip, then aggregates inter-clip information capture the global target. Notably, ReferMo is supervised by only keyframe masks, and SAM2 is only used at inference for target tracking in subsequent frames.

4.2 From Local Perception to Global Interaction

Different from the previous RVOS methods [25, 30, 48] that perform vision-language fusion on each single frame, we introduce motion representations to enable clip-level vision-language fusion. For each video clip, as shown in Figure 6, the local perceiver encodes the text, keyframe and motion information through three separate encoders, and then employs a multi-modal fuser to progressively aggregate these information for clip-level object extraction. By collecting the objects across different video clips, we perform global temporal interaction to enable consistent object prediction and long-term temporal understanding.

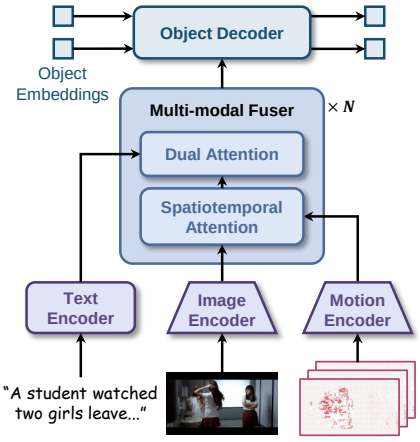


Figure 6: Overview of local perceiver.

Motion Encoder. The motion vectors are first embeded into a d-dimensional space via a linear projector. Then, the motion encoder performs self-attention separately along the spatial and temporal dimensions to extract the spatiotemporal motion features $M \in \mathbb{R}^{T \times \frac{H}{16} \times \frac{W}{16} \times d}$. Notably, we implement the spatial attention as deformable attention due to the large number of spatial tokens.

Image-Motion Fusion. Modern image encoders (e.g., Swin Transformer [29]) typically output multi-scale feature maps $I_i \in \mathbb{R}^{H_i \times W_i \times d}$, $i \in [1, 4]$. To match these spatial resolutions, we adopt a series of spatial convolutions with specific strides over the motion features M to produce multi-scale motion features $M_i \in \mathbb{R}^{T \times H_i \times W_i \times d}$. At each scale i, we treat the keyframe feature I_i as *query* and perform cross-attention along the temporal dimension to aggregate M_i into $\widetilde{M}_i \in \mathbb{R}^{H_i \times W_i \times d}$. To avoid undesired motion noise, we fuse the keyframe and motion features via the spatial-aware and channel-aware gating mechanisms:

$$M_i^* = (\underbrace{\sigma(I_i \cdot W_{down}^I)}_{\text{Spatial Gate}} \odot (\widetilde{M}_i \cdot W_{down}^M)) \cdot W_{up}, \tag{3}$$

$$F_i = I_i + \underbrace{\gamma_i}_{\text{Channel Gate}} \odot \max(M_i^*, 0)^2, \tag{4}$$

where $W_{down}^I, W_{down}^M \in \mathbb{R}^{d \times r}$ indicate the low-rank projectors that compress the features to a lower dimension r, and $W_{up} \in \mathbb{R}^{r \times d}$ is a projector to resort the dimension. σ denotes Sigmoid function and $\odot$ denotes Hadamard product. $\gamma \in \mathbb{R}^d$ is a learnable vector to modulate the channel-wise weights.

Vision-Language Fusion. We use the dual cross-attention modules [24, 28] for deep vision-language fusion. Formally, given the clip-level vision features $F \in \mathbb{R}^{N\times d}$ and the language features $E \in \mathbb{R}^{L\times d}$, where N and L individually denote their token number, we derive the cross-modal enhanced vision features $\widetilde{F}$ and language features $\widetilde{E}$ as follows:

$$\mathcal{A} = \frac{F \cdot E^{\top}}{\sqrt{d}}, \qquad \widetilde{F} = \mathrm{Softmax}(\mathcal{A}) \cdot E, \qquad \widetilde{E} = \mathrm{Softmax}(\mathcal{A}^{\top}) \cdot F. \tag{5}$$

For simplicity, the linear projections for multi-head attentions are omitted. The output features $\widetilde{F}$ and $\widetilde{E}$ are then fed into the object decoder to extract object features.

Global Interaction. To enable consistent object prediction and long-term temporal understanding, we collect the object features across video clips to perform global temporal interactions. Following ReferDINO [25], we use the Hungarian algorithm [22] to align the objects clip-by-clip. Then, we perform temporal self-attention over the aligned object features to achieve global modeling. For better modality alignment, we also infuse the language information $\widetilde{E}$ into the object features through a cross-attention layer. Finally, the interacted object features are output to the segmentation head for generating instance masks. Note that these masks are only predicted for the key frame within each clip, serving as object anchors for SAM2's mask propagation in subsequent frames.

5 Experiments

5.1 Experiment Setup

Dataset Split. Long-RVOS is a large-scale dataset containing 2,193 videos and 24,689 sentences, which are split into three subsets: a training set of 1,855 videos and 20,722 sentences, a validation set of 113 videos and 1,379 sentences, and a test set of 225 videos and 2,588 sentences.

Evaluation Metrics. We use three kinds of evaluation metrics: the spatial metric $\mathcal{J}\&\mathcal{F}$, the temporal metric tIoU and the spatiotemporal metric vIoU. Long-RVOS provides three types of descriptions: *static*, *temporal* and *hybrid*. We report performance for each type separately and overall. Additionally, we report the FPS for each competitor because efficiency is a major concern for long-video processing.

Implementation Details. We follow the default hyper-parameter settings of ReferDINO [25] and use Swin-Tiny as the backbone. For SAM2 [35], we use the $\mathrm{sam2.1\_hiera\_large}$ version. In MPEG-4 [23], each video clip typically consists of a keyframe and the motion vectors for up to 11 subsequent frames. During training, we randomly sample 6 clips and use 3-frame motion vectors. The input frames are resized to have the longest side of 640 pixels and the shortest side of 360 pixels during training and evaluation. Following the settings on MeViS [7], we do not use referring image segmentation datasets (e.g., RefCOCO/+/g [19, 32]) for pretraining. We train ReferMo on Long-RVOS dataset for 6 epochs, which take 24 hours on 8 Nvidia A6000 GPUs.

5.2 Benchmark Results

Overall Comparison. We compare ReferMo with six recent RVOS methods on Long-RVOS. All models in comparison are trained on Long-RVOS under consistent experimental settings for fairness. As demonstrated in Table 2, realistic long-video scenarios remain a significant challenge for current RVOS models. While the SAM2-based methods [5, 26] achieve SOTA performance on existing short-term benchmarks [7, 20, 36], they significantly struggle in Long-RVOS. This suggests that their improvements may primarily stem from SAM2's superior tracking and segmentation capabilities, rather than better language-guided object understanding. As the videos grow longer and more complex, it becomes more challenging to perform video-language reasoning and distinguish the objects, which leads to their performance degradation. In contrast, our baseline ReferMo integrates the static attributes, short-term dynamics and long-term dependencies to perform object-level visual-language reasoning, achieving significant improvements over existing methods. These findings highlight the need for both frame-level segmentation precision and video-level visual-language understanding to address the long-video challenges in Long-RVOS.

Fine-grained Evaluation. Long-RVOS provides three types of text descriptions to enable rigorous evaluation. For most models, the performance for static and hybrid types is comparable and largely better than that for dynamic type. This implies a strong bias in current RVOS models toward static

Method	Year	Static			Dynamic			Hybrid			Overall			FPS
		$\mathcal{J}\&\mathcal{F}$	tIoU	vIoU	$\mathcal{J}\&\mathcal{F}$	tIoU	vIoU	$\mathcal{J}\&\mathcal{F}$	tIoU	vIoU	$\mathcal{J}\&\mathcal{F}$	tIoU	vIoU	
Without SAM / SAM2														
SOC [30]	2023	34.8	67.7	28.4	34.9	68.7	28.8	35.1	68.0	28.5	34.9	68.1	28.6	53.8
MUTR [48]	2024	43.0	70.1	36.7	40.2	70.8	34.8	43.2	70.3	37.2	42.2	70.4	36.2	20.4
ReferDINO [25]	2025	50.7	**71.9**	42.8	45.9	**71.9**	38.9	49.2	**71.5**	41.7	48.7	**71.7**	41.2	46.4
With SAM / SAM2														
VideoLISA [1]	2024	34.3	69.6	28.9	31.0	69.7	26.9	33.9	69.4	28.6	33.1	69.6	28.2	6.6
GLUS [26]	2025	36.4	68.2	34.3	37.6	68.9	35.8	35.9	68.0	33.9	36.6	68.4	34.6	3.6
SAMWISE [5]	2025	36.6	68.4	29.2	34.3	68.6	28.1	33.8	69.4	28.4	35.6	68.4	28.6	7.0
ReferMo	2025	**53.5**	71.4	**44.0**	**48.1**	71.2	**40.1**	**52.2**	71.2	**43.6**	**51.3**	71.2	**42.6**	52.5

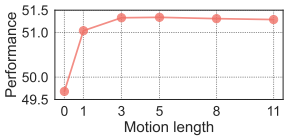
Table 2: Comparison of state-of-the-art RVOS models on Long-RVOS test set. FPS is estimated at 360P on Nvidia A6000 GPUs, excluding the video loading time.

Dataset		Point	Box	Mask
MeViS [7]	Valid_u	77.3	80.0	80.6
Long-RVOS	Valid	53.4	54.5	53.5
	Test	52.8	53.9	53.3

(a) Oracle analysis with SAM2.

Model	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
ReferDINO	49.1	47.6	50.6
ReferMo	49.6	48.0	51.2
-w/o motion	47.5	46.0	48.9

(b) Results on the keyframes.



(c) Different motion lengths.

Table 3: Oracle analysis and ablation studies.

attributes. Across different models, while the $\mathcal{J}\&\mathcal{F}$ scores show significant variance, their tIoU are relatively consistent. This reveals that existing RVOS models have little performance gap in temporal consistency, highlighting the need for effective tracking mechanisms to handle frequent target disappearance in long-term videos. ReferMo significantly outperforms other models across various types and metrics, except for tIoU, where it is slightly inferior to ReferDINO. We speculate that this is because ReferMo only performs language-guided reasoning on keyframes, resulting in suboptimal object identification on motion frames.

Oracle Analysis. We provide SAM2 with first-frame ground-truth object prompts and evaluate its tracking performance across different datasets. As shown in Table 3 (a), the oracle results for Long-RVOS (52.8~54.5 $\mathcal{J}\&\mathcal{F}$) are significantly lower than those for MeViS (77.3~80.6 $\mathcal{J}\&\mathcal{F}$). The notable performance gap of nearly 25% demonstrates the long-term challenges in Long-RVOS.

5.3 Ablation Studies

Results on Keyframes. In Table 3 (b), we compare the performance of ReferMo and ReferDINO [25] on the keyframes. We focus on the spatial metrics since the length of the keyframe sequence is short. Note that ReferDINO is trained on all frames, while our ReferMo is only trained on keyframes. However, ReferMo still outperforms ReferDINO by 0.5% in $\mathcal{J}\&\mathcal{F}$, owing to the integration of motion information. When ablating it, we see a significant 2.1% performance drop in $\mathcal{J}\&\mathcal{F}$. These results encourage further exploration of sparse-frame supervision for RVOS task.

Effect of Motion Information. We investigate the impact of varying the number of motion frames in ReferMo. As shown in Table 3 (c), the performance without motions is only 49.7 $\mathcal{J}\&\mathcal{F}$. However, even using just one motion frame yields +1.6% $\mathcal{J}\&\mathcal{F}$ improvements. Increasing the motion length to 3 frames improves $\mathcal{J}\&\mathcal{F}$ to 51.3, but further increasing only leads to marginal gains.

6 Conclusion

In this work, we introduce Long-RVOS, a large-scale benchmark for long-term referring video object segmentation, comprising over 2,000 videos averaging 60+ seconds to address the limitations of existing short-term datasets. To enable comprehensive and rigorous evaluation, we provide three types of descriptions and two novel metrics, tIoU and vIoU. Results on Long-RVOS indicate that current RVOS methods struggle severely in long-video scenarios. Furthermore, we propose ReferMo, a simple motion-enhanced baseline that significantly outperforms existing SOTA methods on long-term videos. We believe that Long-RVOS and ReferMo will provide a foundation for future research to develop robust RVOS models tackling real-world long-form videos.

References

[1] Z. Bai, T. He, H. Mei, P. Wang, Z. Gao, J. Chen, Z. Zhang, and M. Z. Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, 37:6833–6859, 2024. 1, 3, 9

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

[3] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 3

[4] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8748–8757, 2019. 4

[5] C. Cuttano, G. Trivigno, G. Rosi, C. Masone, and G. Averta. Samwise: Infusing wisdom in sam2 for text-driven video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3, 8, 9

[6] A. Dave, T. Khurana, P. Tokmakov, C. Schmid, and D. Ramanan. Tao: A large-scale benchmark for tracking any object. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 436–454. Springer, 2020. 4

[7] H. Ding, C. Liu, S. He, X. Jiang, and C. C. Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2694–2703, 2023. 1, 2, 3, 4, 8, 9

[8] H. Ding, C. Liu, S. He, X. Jiang, P. H. Torr, and S. Bai. Mose: A new dataset for video object segmentation in complex scenes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20224–20234, 2023. 4

[9] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 3, 4, 5

[10] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 3

[11] K. Gavrilyuk, A. Ghodrati, Z. Li, and C. G. Snoek. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5958–5966, 2018. 1, 2, 3, 4

[12] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 4, 5

[13] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 4, 5

[14] M. Han, Y. Wang, Z. Li, L. Yao, X. Chang, and Y. Qiao. Html: Hybrid temporal-scale multimodal learning framework for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13414–13423, 2023. 3

[15] S. He and H. 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, 3

[16] L. Hong, W. Chen, Z. Liu, W. Zhang, P. Guo, Z. Chen, and W. Zhang. Lvos: A benchmark for long-term video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13480–13492, 2023. 3, 4

[17] P. Jin, R. Takanobu, W. Zhang, X. Cao, and L. Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 1

[18] Y. Jin, Z. Sun, K. Xu, L. Chen, H. Jiang, Q. Huang, C. Song, Y. Liu, D. Zhang, Y. Song, et al. Video-lavit: Unified video-language pre-training with decoupled visual-motional tokenization. *arXiv preprint arXiv:2402.03161*, 2024. 6

[19] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 8

[20] A. Khoreva, A. Rohrbach, and B. Schiele. Video object segmentation with language referring expressions. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14*, pages 123–141. Springer, 2019. 1, 2, 3, 4, 8

[21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 3

[22] H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 8

[23] D. Le Gall. Mpeg: A video compression standard for multimedia applications. *Communications of the ACM*, 34(4):46–58, 1991. 6, 8

[24] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 8

[25] T. Liang, K.-Y. Lin, C. Tan, J. Zhang, W.-S. Zheng, and J.-F. Hu. Referdino: Referring video object segmentation with visual grounding foundations. *arXiv preprint arXiv:2501.14607*, 2025. 1, 3, 7, 8, 9

[26] L. Lin, X. Yu, Z. Pang, and Y.-X. Wang. Glus: Global-local reasoning unified into a single large language model for video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3, 8, 9

[27] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[28] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024. 3, 8

[29] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 7

[30] Z. Luo, Y. Xiao, Y. Liu, S. Li, Y. Wang, Y. Tang, X. Li, and Y. Yang. Soc: semantic-assisted object cluster for referring video object segmentation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 26425–26437, 2023. 1, 3, 7, 9

[31] K. Mangalam, R. Akshulakov, and J. Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 3

[32] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 8

[33] B. Miao, M. Bennamoun, Y. Gao, and A. Mian. Spectrum-guided multi-granularity referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 920–930, 2023. 1

[34] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 4

[35] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1, 3, 4, 6, 8

[36] S. Seo, J.-Y. Lee, and B. Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 208–223. Springer, 2020. 1, 2, 3, 4, 8

[37] X. Shang, D. Di, J. Xiao, Y. Cao, X. Yang, and T.-S. Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 279–287. ACM, 2019. 3, 4, 5

[38] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 510–526. Springer, 2016. 4, 5

[39] J. Tang, G. Zheng, and S. Yang. Temporal collection and distribution for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15466–15476, 2023. 3

[40] Z. Tang, Y. Liao, S. Liu, G. Li, X. Jin, H. Jiang, Q. Yu, and D. Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249, 2021. 6

[41] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 4, 5

[42] X. Wang, Y. Zhang, O. Zohar, and S. Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024. 6

[43] C.-Y. Wu and P. Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894, 2021. 3

[44] J. Wu, Y. Jiang, P. Sun, Z. Yuan, and P. Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 1, 3

[45] Z. Wu, C. Xiong, C.-Y. Ma, R. Socher, and L. S. Davis. Adaframe: Adaptive frame selection for fast video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1278–1287, 2019. 6

[46] N. Xu, L. Yang, Y. Fan, D. Yue, Y. Liang, J. Yang, and T. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 4

[47] C. Yan, H. Wang, S. Yan, X. Jiang, Y. Hu, G. Kang, W. Xie, and E. Gavves. Visa: Reasoning video object segmentation via large language models. *arXiv preprint arXiv:2407.11325*, 2024. 1, 3

[48] S. Yan, R. Zhang, Z. Guo, W. Chen, W. Zhang, H. Li, Y. Qiao, H. Dong, Z. He, and P. Gao. Referred by multi-modality: A unified temporal transformer for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6449–6457, 2024. 7, 9

[49] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2636–2645, 2020. 4

[50] H. Yuan, X. Li, T. Zhang, Z. Huang, S. Xu, S. Ji, Y. Tong, L. Qi, J. Feng, and M.-H. Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv*, 2025. 1, 3

[51] H. Zhang, X. Li, and L. Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, 2023. 1

[52] Z. Zhang, Z. Zhao, Y. Zhao, Q. Wang, H. Liu, and L. Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10668–10677, 2020. 6

[53] H. Zhao, A. Torralba, L. Torresani, and Z. Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8668–8678, 2019. 4, 5

[54] R. Zheng, L. Qi, X. Chen, Y. Wang, K. Wang, Y. Qiao, and H. Zhao. Villa: Video reasoning segmentation with large language model. *arXiv preprint arXiv:2407.14500*, 2024. 1

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In this paper, we introduce Long-RVOS, a large-scale benchmark for long-term referring video object segmentation with comprehensive evaluation. We further propose a simple yet effective baseline ReferMo to address the long-term challenges.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our ReferMo in Section 5.2. A separate "Limitations" section is provided in the Supplementary.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: This paper provides comprehensive descriptions of the dataset construction in Section 3 and the proposed baseline Section 4. The implementation details are present in Section 5.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will release the new dataset Long-RVOS and the source code of our baseline ReferMo.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 5.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The dataset statistics are provided in Section 3.3 and other experimental statistics are presented in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section 5.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes] ,

   Justification: We have made sure.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This paper establishes a foundation for long-term video object segmentation, which potentially enhances the development of realistic video applications, such as video editing and human-computer interaction. We have briefly discussed the positive impacts in our abstract and conclusion. A separate "Broader Impacts" section is provided in the Supplementary.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks. Our new benchmark is built upon existing, publicly available datasets.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use publicly available resources to build our dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We cite all the benchmarks and code repositories used.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.