

Appendix

A More Dataset Statistics

To further highlight the challenges posed by Long-RVOS, we present a statistical analysis of video attributes, with definitions provided in Table 4. As shown in Table 5, compared to the current largest dataset MeViS [7], Long-RVOS involves numerous long-video challenges, including frequent object occlusion (POC, FOC, and OV) and long-term object disappearance-reappearance (LRA). In addition, the videos in Long-RVOS exhibit significant object motion (CM and MB) and appearance changes (VC, ARC and SV), making the dataset more akin to real-world scenarios.

| Attribute | Full Name | Definition |
|------------|------------------------|---|
| POC | Partial Occlusion | The target object is partially occluded in the sequence. |
| FOC | Full Occlusion | The target object is fully occluded in the sequence. |
| OV | Out-of-view | The target leaves the video frame completely. |
| LRA | Long-term Reappearance | Target object reappears after disappearing for at least 100 frames. |
| VC | View Change | Viewpoint affects target appearance significantly. |
| ARC | Aspect Ratio Change | The ratio of bounding box aspect ratio is outside the range [0.5, 2]. |
| SV | Scale Variation | The ratio of any pair of bounding-box is outside of range [0.5, 2.0]. |
| CM | Camera Motion | Abrupt motion of the camera. |
| MB | Motion Blur | The boundary of target object is blurred because of camera or object fast motion. |

Table 4: Definitions of the video attributes.

| Dataset | POC | FOC | OV | LRA | VC | ARC | SV | CM | MB |
|-------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| MeViS [7] | 54.8 | 15.1 | 28.7 | 0.1 | 10.0 | 88.2 | 78.7 | 49.2 | 18.8 |
| Long-RVOS (Ours) | 60.5 | 36.2 | 61.0 | 11.5 | 25.9 | 96.2 | 93.6 | 60.7 | 28.7 |

Table 5: The percentage (%) of videos featuring specific attributes.

B More Implementation Details

Motion Extraction. Following Video-LaVIT [18], we rely on motion vectors instead of the expensive dense optical flow. Formally, given a video clip, we partition each frame into 16×16 non-overlapping macroblocks. Then, motion vectors \vec{m} of the t -th frame are estimated by matching the macroblocks between the adjacent frames I_t and I_{t-1} :

$$\vec{m}(p, q) = \arg \min_{i, j} \|I_t(p, q) - I_{t-1}(p - i, q - j)\|, \quad (6)$$

where $I(p, q)$ denotes the pixel values of the macroblock at location (p, q) , and (i, j) denotes the coordinate offset between the centers of the two macroblocks. Empirically, the extraction of motion vectors runs at 748 FPS on our device (Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz), enabling real-time processing of long videos.

Global Interaction. This module performs temporal attention over the inter-frame object features, enabling temporal reasoning and understanding. Since this is a common module in RVOS approaches [7, 25, 30], we follow the object-consistent temporal enhancer (OTE) of ReferDINO [25] rather than designing a new one from scratch. For clarity, we briefly revisit OTE here. Given T -frame object features $\{O_t\}_{t=1}^T$ where $O_t \in \mathbb{R}^{N_q \times d}$, OTE utilizes the Hungarian algorithm [22] to align the N_q objects between adjacent frames based on the pairwise cosine similarity. After that, it performs temporal self-attention over the aligned object features and cross-attention with the sentence features \tilde{E} . We refer the readers to the original paper [25] for additional details.

Training. Unlike previous RVOS methods, ReferMo relies only on the keyframe ground-truth annotations for efficient training. Formally, given a text description and a video of T_c clips, ReferMo outputs the prediction sequences $\{\mathbf{p}_i\}_{i=1}^{N_q}$ for the N_q object queries, where each sequence $\mathbf{p}_i = \{\hat{s}_i^t, \hat{b}_i^t, \hat{m}_i^t\}_{t=1}^{T_c}$ describes the binary classification probability, bounding box and mask of the i -th object query on t -th keyframe. Our training pipeline follows the practice in previous approaches [25, 30, 45]. Suppose $\mathbf{y} = \{s^t, b^t, m^t\}_{t=1}^{T_c}$ as the ground truth of keyframes, we select the prediction sequence with the lowest matching cost as the positive and assign the remaining sequences as negative. The matching cost is defined as follows:

$$\mathcal{L}_{\text{total}}(\mathbf{y}, \mathbf{p}_i) = \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}(\mathbf{y}, \mathbf{p}_i) + \lambda_{\text{box}} \mathcal{L}_{\text{box}}(\mathbf{y}, \mathbf{p}_i) + \lambda_{\text{mask}} \mathcal{L}_{\text{mask}}(\mathbf{y}, \mathbf{p}_i). \quad (7)$$

The matching cost is computed on individual frames and normalized by T_c . Here, \mathcal{L}_{cls} is the focal loss that supervises the binary classification prediction. \mathcal{L}_{box} sums up the L1 loss and GIoU loss. $\mathcal{L}_{\text{mask}}$ is the combination of DICE loss, binary mask focal loss and projection loss [42]. λ_{cls} , λ_{box} and λ_{mask} are scalar weights of individual losses. The model is optimized end-to-end by minimizing the total loss $\mathcal{L}_{\text{total}}$ for positive sequences and only the classification loss \mathcal{L}_{cls} for negative sequences.

Inference. ReferMo produces instance mask for the referring object on keyframes and then employs SAM2 [35] for subsequent frames. Specifically, for the prediction sequences $\{\mathbf{p}_i\}_{i=1}^{N_q}$, we select the best sequence with the highest average classification score as follows:

$$\sigma = \arg \max_{i \in [1, N_q]} \frac{1}{T_c} \sum_{t=1}^{T_c} \hat{s}_i^t \quad (8)$$

Then, the output mask sequence on keyframes is formed as $\{\mathbf{m}_\sigma^t\}_{t=1}^{T_c}$. For the t -th video clip, we use the keyframe prediction \mathbf{m}_σ^t as the mask prompt for SAM2, which tracks the target across the subsequent frames within the clip.

C More Benchmark Results

In Table 6, we present the benchmark results on Long-RVOS validation set. The results show that our baseline ReferMo achieves consistent improvements over previous RVOS methods, especially those built on SAM or SAM2.

| Method | Year | Static | | | Dynamic | | | Hybrid | | | Overall | | |
|--------------------|------|----------------------------|-------------|-------------|----------------------------|-------------|-------------|----------------------------|-------------|-------------|----------------------------|-------------|-------------|
| | | $\mathcal{J}\&\mathcal{F}$ | 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 | 34.5 | 70.0 | 30.5 | 32.6 | 70.9 | 29.2 | 33.2 | 70.2 | 29.5 | 33.4 | 70.4 | 29.7 |
| MUTR [49] | 2024 | 44.1 | 70.6 | 38.8 | 39.7 | 71.1 | 35.6 | 43.8 | 70.8 | 38.9 | 42.6 | 70.8 | 37.8 |
| ReferDINO [25] | 2025 | 53.5 | 72.8 | 46.3 | 45.8 | 71.6 | 39.5 | 52.1 | 72.2 | 45.5 | 50.5 | 72.2 | 43.8 |
| With SAM / SAM2 | | | | | | | | | | | | | |
| VideoLISA [1] | 2024 | 31.6 | 67.5 | 25.4 | 27.4 | 68.9 | 23.8 | 32.0 | 68.2 | 26.2 | 30.4 | 68.2 | 25.1 |
| GLUS [26] | 2025 | 37.8 | 69.1 | 36.1 | 37.5 | 70.2 | 35.9 | 37.1 | 69.4 | 35.1 | 37.5 | 69.5 | 35.7 |
| SAMWISE [5] | 2025 | 35.8 | 70.3 | 31.2 | 32.2 | 70.3 | 28.8 | 33.0 | 70.2 | 29.3 | 33.7 | 70.3 | 29.4 |
| ReferMo | 2025 | 55.8 | 72.6 | 47.3 | 48.7 | 72.2 | 41.5 | 53.9 | 71.9 | 46.5 | 52.9 | 72.3 | 45.2 |

Table 6: Benchmark results on Long-RVOS validation set.

D More Ablation Studies

Effectiveness of Gating Image-Motion Fusion. ReferMo employs the spatial-aware gating (SG) and channel-aware gating (CG) mechanisms in image-motion fusion to avoid undesired motion noise. As shown in Table 7, directly adding keyframe and motion features results in significant performance degradation, as motion features can introduce object-irrelevant noise into the object features. By applying these two gating strategies, motion noise is effectively alleviated and the performance is significantly improved by 7.6 $\mathcal{J}\&\mathcal{F}$.

Effectiveness of Global Interaction. To validate the component choice of for global interaction, we compare OTE [25] with LMPM [7], which directly performs temporal self-attention over the object features. As shown in Table 8, removing the global interaction decreases the performance on keyframes by 2.4 $\mathcal{J}\&\mathcal{F}$, demonstrating the effectiveness of global interaction. In addition, OTE outperforms LMPM by 0.7 $\mathcal{J}\&\mathcal{F}$. Hence we select OTE as our global interaction module.

| SG | CG | $\mathcal{J}\&\mathcal{F}$ | \mathcal{J} | \mathcal{F} |
|----|----|----------------------------|---------------|---------------|
| | | 42.0 | 42.1 | 41.9 |
| ✓ | | 44.1 | 43.9 | 44.2 |
| | ✓ | 48.3 | 46.8 | 49.7 |
| ✓ | ✓ | 49.6 | 48.0 | 51.2 |

Table 7: Keyframe performance with different image-motion fusion strategy.

| Method | $\mathcal{J}\&\mathcal{F}$ | \mathcal{J} | \mathcal{F} |
|-------------|----------------------------|---------------|---------------|
| <i>None</i> | 47.2 | 45.6 | 48.7 |
| LMPM [7] | 48.9 | 47.3 | 50.5 |
| OTE [25] | 49.6 | 48.0 | 51.2 |

Table 8: Keyframe performance with different global interaction methods. *None* indicates that the global interaction is removed.



Figure 7: Qualitative comparison of our ReferMo with the SOTA model ReferDINO [25]. ReferMo performs better in long-term object consistency and segmentation quality.

E Visualization

In Figure 7, we provide the qualitative comparisons with the SOTA model ReferDINO [25] on Long-RVOS. These examples involve multiple long-term challenges, such as object occlusion, disappearance-reappearance and view changes. The results clearly show the effectiveness of our baseline ReferMo in long-term object consistency and segmentation quality.

F Limitations and Future Work

We currently focus on description-guided RVOS. It is interesting to broaden the benchmark scope to support more tasks, such as semi-supervised VOS [8, 34, 47], interactive VOS [21, 35], audio-guided VOS [49], video question answering with object segmentation [1, 48]. Besides, while our benchmark covers a variety of objects, it currently does not include background stuff classes (e.g., sky and river), which could be incorporated in future work for covering more diverse scenarios. For training efficiency, our simple baseline only employs SAM2 at inference. Future work may explore parameter-efficient fine-tuning techniques to enable end-to-end optimization with SAM2.

G Border Impacts

In this work, we propose a large scale benchmark to advance the RVOS task toward long-term, real-world scenarios. This benchmark potentially leads to the development of stronger and more practical RVOS models. These models hold significant potential for many real-world applications, such as video editing, human-robot interaction and automated video analysis. However, like many powerful AI technologies, RVOS models carry potential risks, including unauthorized surveillance or privacy-infringing tracking. Despite these concerns, we firmly believe that the task itself is neutral, and its positive implications outweigh potential risks when guided by ethical considerations and responsible deployment.