DeRVOS: Decoupling Consistent Trajectory Generation and Multimodal Understanding for Referring Video Object Segmentation

Supplementary Material

6. RIS Dataset and Evaluation Metrics

Dataset. RefCOCO [82], RefCOCO+ [82], and RefCOCOg [56] are three representative benchmarks for referring image segmentation, all derived from the MS COCO [47] dataset. Both RefCOCO and RefCOCO+ were collected through a two-player interactive game. RefCOCO contains approximately 142K referring expressions and RefCOCO+ includes about 141K, both describing around 50K objects across approximately 20K images. Unlike RefCOCO, RefCOCO+ prohibits the use of positional words in the expressions. RefCOCOg consists of about 85K more natural and complex expressions, referring to roughly 55K objects. These expressions feature longer and semantically richer descriptions, spread across approximately 26K images.

Evaluation Metrics. For RIS, we adopt the mean Intersection over Union (mIoU) as the evaluation metric. The mIoU measures the average overlap between the predicted mask $\hat{M}$ and the ground truth mask M, and is defined as:

$$\text{IoU} = \frac{|M \cap \hat{M}|}{|M \cup \hat{M}|}, \quad \text{mIoU} = \frac{1}{N}\sum_{i=1}^{N} \text{IoU}_i, \tag{7}$$

where N denotes the total number of samples. A higher mIoU indicates better segmentation quality and more accurate alignment between prediction and ground truth.

7. Additional Implementation Details

We freeze the DVIS++ [88] and adopt the AdamW [49] optimizer. For MeViS [22], the learning rate is set to 5×10^{-6} for the multimodal encoder and 2×10^{-5} for the remaining modules. The training process requires only 5 epochs. For Ref-YouTube-VOS [63], the learning rate is 2.5×10^{-6} for the multimodal encoder and 1×10^{-5} for the remaining modules. During the pre-training stage, the learning rate is set to 8×10^{-6} for the multimodal encoder and 4×10^{-5} for the other parts, with a decay factor of 0.1 applied after the 8th epoch, and a total of 12 epochs. For the RIS task and image-level pretraining before fine-tuning on Ref-YouTube-VOS, we use pretrained weights from the VIPSeg [54] dataset, freezing the DVIS++ model except for the Refiner module, which is fine-tuned to optimize mask quality. After pretraining, we follow the pretraining settings and fine-tune the model on Ref-YouTube-VOS. All experiments are conducted on four NVIDIA RTX 4090 GPUs. For ablation studies, the input resolution is set to 384×384, and the multimodal encoder input resolution is set to 224×224. We uniformly set the training clip length to $T = 8$. During MeViS training, we adopt a random frame sampling strategy, where 8 frames are randomly selected from each video. In addition, a filtering strategy is applied to avoid entirely empty samples, which could otherwise increase training difficulty.

8. Additional Experimental Results

Performance on Long-RVOS Dataset. As shown in Table 9, we compare the performance of different methods on the Long-RVOS test set for long-video referring segmentation. We follow the experimental setup of ReferMo [44] but do not use motions. The Long-RVOS dataset contains three scene types: Static, Dynamic, and Hybrid, which evaluate the model's segmentation capability under different motion patterns. Experimental results demonstrate that compared to ReferDINO [45], our method improves the overall metrics by 5.6, 0.2, and 3.5 in $\mathcal{J}\&\mathcal{F}$, tIoU, and vIoU, respectively. Compared to the specialized method ReferMo [44], our method achieves improvements of 1.1, 0.1, and 2.2 in overall $\mathcal{J}\&\mathcal{F}$, tIoU, and vIoU, respectively. Notably, our method shows particularly significant improvements in Dynamic scenes, with $\mathcal{J}\&\mathcal{F}$ gains of 5.8 and 1.9 compared to ReferDINO and ReferMo, respectively. These results further validate that our method possesses stronger generalization capability in complex scenarios.

Pre-training and Fine-tuning Effects on RIS. As shown in Table 10, we systematically evaluate the effect of different pre-training weights and fine-tuning strategies on referring image segmentation across the RefCOCO/+/g [52, 56, 82] datasets. The results show that even with frozen pre-trained weights, the model exhibits efficient video segmentation performance. However, further fine-tuning of the Refiner module significantly improves performance, yielding an improvement of approximately 4 percentage points regardless of whether the pre-training originates from OVIS [59] or VIPSeg [54]. This suggests that refining mask boundaries and enhancing local consistency are crucial for the RIS task. Notably, models initialized with VIPSeg weights achieve better results on the RefCOCO+/g datasets, highlighting stronger generalization and cross-modal semantic understanding. Finally, we select the initialization weights from the VIPSeg dataset and fine-tune the configuration of the Refiner module as the final setup for RIS and the pre-training process on the Ref-YouTube-VOS [63] dataset.

Model Parameters and Performance. As shown in Table 11, we compare the parameter scale and performance of different methods on the Ref-YouTube-VOS [76] dataset. Compared to specialized RVOS approaches such as ReferFormer [73] and SgMg [53], our method incorporates a slightly larger number of parameters. However, it achieves significantly better performance, demonstrating the strong potential of the decoupled structure in enhancing both trajectory generation and multimodal understanding capabilities. Meanwhile, compared with LVLM approaches such as VISA [77], VideoLISA [1], and RGA3-3B [67], our method uses substantially fewer parameters while still achieving competitive results. These results highlight that the proposed approach strikes a better balance between model complexity and task performance.

Effect of Training Clip Lengths. As illustrated in Fig. 6, we analyze the impact of different training clip lengths T on performance over the MeViS [22] val^{u} subset. Shorter clips limit the model's

Table 9. Comparison on Long-RVOS test set. The best performance is shown in bold.

Method	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
SOC [50]	39.3	71.8	33.9	38.8	73.2	34.2	37.7	71.9	32.5	38.6	72.3	33.5
MUTR [78]	42.8	72.6	38.7	41.2	73.5	37.7	42.4	72.3	38.1	42.2	72.8	38.2
ReferDINO [45]	50.9	73.6	46.0	45.4	73.8	41.5	48.7	73.1	44.0	48.4	73.5	43.9
ReferMo [44]	55.8	73.6	47.5	49.3	**74.2**	42.4	53.3	72.9	45.4	52.9	73.6	45.2
DeRVOS	**56.7**	**73.9**	**50.0**	**51.2**	74.0	**45.3**	**53.7**	**73.3**	**46.7**	**54.0**	**73.7**	**47.4**

Table 10. Impact of different pre-training weights and fine-tuning strategies on RIS performance with RefCOCO/+/g datasets.

Weight Source	Refiner Fine-tuning	Referring Image Segmentation RefCOCO	RefCOCO+	RefCOCOg
OVIS [59]	✗	76.7	71.7	70.9
OVIS	✓	80.6	74.5	75.9
VIPSeg [54]	✗	76.0	71.4	73.2
VIPSeg	✓	80.2	74.9	76.1

Table 11. Model parameter and performance analysis on Ref-YouTube-VOS dataset.

Structure	Reference	All Params	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
ReferFormer [73]	[CVPR'22]	0.3B	62.9	61.3	64.6
SgMg [53]	[ICCV'23]	0.24B	65.7	63.9	67.4
VISA [77]	[ECCV'24]	13B	61.5	59.8	63.2
VideoLISA [1]	[NeurIPS'24]	3.8B	61.7	60.2	63.3
RGA3-3B [67]	[ICCV'25]	7B	68.5	66.8	70.1
DeRVOS	-	0.56B	**70.0**	**68.0**	**71.9**

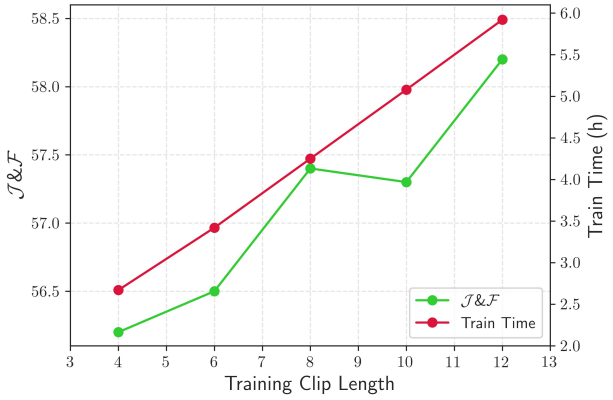


Figure 6. Effect of training clip length T on performance evaluated on the MeViS val^u dataset.

ability to capture long-term temporal dependencies, which may lead to degraded performance when handling videos with complex motion dynamics. In contrast, longer clips enable the model to learn richer temporal information and better understand continuous target variations over time, but they also increase computational cost and training difficulty, and may introduce redundant or noisy information that hinders convergence. Experimental results show that on the MeViS val^u subset, performance reaches a local peak at $T = 8$ and achieves the best performance at $T = 12$.

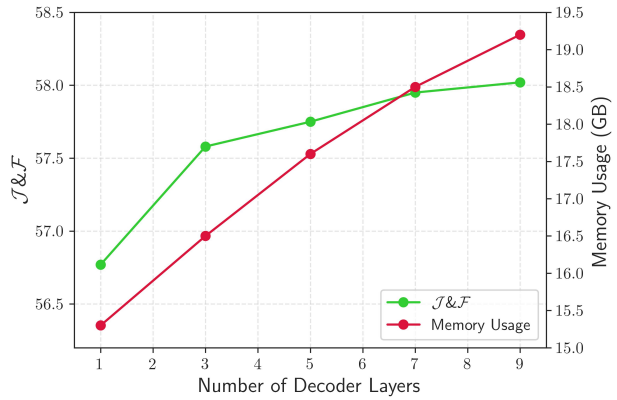


Figure 7. Impact of the number of decoder layers N_d on performance evaluation on the MeViS val^u dataset.

In terms of training time, the training cost increases linearly with the length of the training clips, resulting in higher computational overhead. Ultimately, to balance training cost and performance, we selected $T = 8$ as the final configuration.

Effect of the Number of Decoder Layers. In designed cross-frame multimodal aligner and motion-guided implicit selector, the number of decoder layers has a significant impact on modeling the relationship between trajectories and referring expressions. A shallow decoder may fail to capture the deep semantic relationships between complex trajectory features and referring expressions, thereby affecting the matching between trajectories and expressions. However, a deeper decoder increases training time and memory consumption significantly, while also introducing the risk of overfitting. To systematically assess the impact of decoder layers N_d on model performance, we conducted an ablation study on the MeViS dataset with a fixed batch size of $BS = 2$, as shown in Fig. 7. As the number of decoder layers increases, model performance gradually improves. However, beyond three layers, the performance gains plateau. To achieve an effective balance between performance and computational cost, we selected $N_d = 3$ as the optimal decoder configuration. This choice ensures strong model performance while controlling computational overhead.

Effect of Confidence Threshold. We further analyze the impact of the confidence threshold on RVOS performance and conduct systematic experiments on the MeViS [22] dataset, as shown in Fig. 8. The confidence threshold is applied during inference to filter trajectories, directly affecting both the quantity and quality of detected targets. A higher threshold helps remove low-

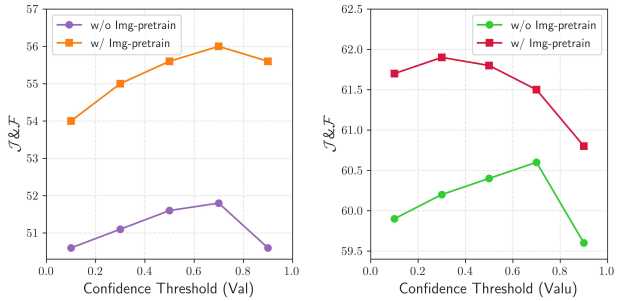


Figure 8. Effect of confidence threshold on the MeViS dataset before and after image-level pretraining.

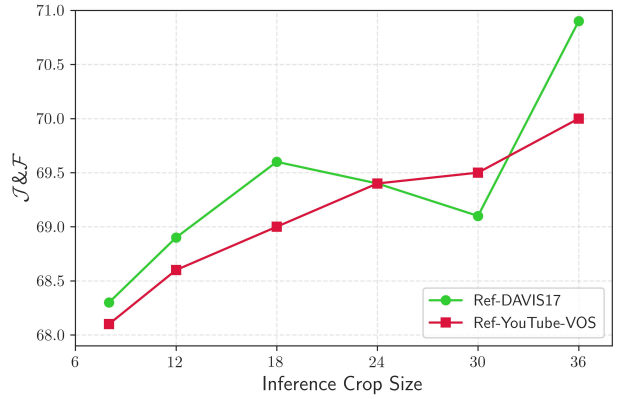


Figure 9. Effect of inference crop size on Ref-YouTube-VOS and Ref-DAVIS17 datasets. The inference crop size refers to the number of consecutive frames fed into the model at once.

quality trajectories but may also discard some correct ones. On the MeViS val subset, regardless of whether image-level pretraining is applied, the performance exhibits a "rise–then–fall" trend as the threshold increases. This reflects the inherent precision–recall trade-off in trajectory selection: at lower thresholds, the model retains more candidate trajectories to ensure recall; as the threshold increases, erroneous trajectories are gradually eliminated, leading to performance improvement; however, when the threshold becomes too high, some correct trajectories are also filtered out, causing performance degradation. On this subset, the optimal performance is achieved when the threshold is around 0.7. On MeViS val^{u} subset, the model without image-level pretraining shows a similar trend, while the pretrained model achieves its peak performance at a lower threshold of 0.3. This shift indicates that image-level pretraining enhances the discriminative ability of the model features, enabling more confident separation between correct and incorrect trajectories even under lower thresholds. Furthermore, on both subsets, the pretrained model consistently outperforms its non-pretrained counterpart, demonstrating that image-level pretraining not only improves generalization but also strengthens the model's robustness in trajectory quality assessment. Finally, based on the performance across both subsets, we selected a threshold of 0.7 as the final configuration.

Quantitative Analysis of Inference Crop Size. During video inference, each video is divided into multiple inference blocks, which are processed independently by the model, and the results are then concatenated to form the final prediction. Here, the inference crop size refers to the number of consecutive frames fed into the model at once. The crop size directly affects the model's ability to capture temporal information, thereby significantly influencing overall performance. As illustrated in Fig. 9, we evaluate the performance under different crop sizes on the Ref-YouTube-VOS [76] and Ref-DAVIS17 [37] datasets. Overall, the results indicate that larger crop sizes progressively improve performance, reaching a peak of 69.1% on both datasets when the crop size is 36. This demonstrates that larger inference blocks provide richer cross-frame information, facilitating the model to capture continuous temporal variations of targets and enhancing both temporal modeling and segmentation accuracy.

Qualitative Analysis of Inference Crop Size. The inference crop size has a significant impact on temporal consistency. We compare models with crop sizes of $Cs = 18$ and $Cs = 30$ on the MeViS [22] validation set, and visualize the segmentation results at intervals of five frames, as illustrated in Fig. 10. During inference, we divide the entire video into several independent inference crops and then concatenate the results to form the final prediction. Since there are no explicit connections or heuristic association mechanisms between these segments, temporal information cannot be shared across different blocks. Therefore, the choice of crop size directly affects the model's ability to capture motion semantics and maintain trajectory consistency over time. In scenarios with rapid motion, the referring expressions often contain dynamic action descriptions, which may appear only within a single inference block. If the crop size is too small, temporal cues cannot be propagated between adjacent segments, leading to discontinuities in action understanding and segmentation failures in subsequent frames. As shown in Fig. 10(a), the expression refers to all airplanes, which is a static category. The model can rely mainly on visual appearance to complete segmentation, achieving stable results under different crop sizes. However, for the motion-oriented expression "turning around and walking backward" in Fig. 10(b), when $Cs = 18$, the action only appears in the first inference block. The following segments lack contextual continuity, resulting in segmentation failure for $T > 18$. In contrast, setting $Cs = 30$ enables the model to capture continuous semantic cues over a longer temporal range, thereby maintaining consistent trajectories. Fig. 10(c) presents another phenomenon: the expression "first car turning" describes a stage-specific action. When $Cs = 18$, the model focuses only on the car currently turning within the active segment. In subsequent segments, since the car has left the scene, the model mistakenly identifies another turning car as the target, causing semantic drift. Increasing the inference crop size to $Cs = 30$ allows the model to associate longer temporal contexts, but also introduces irrelevant instances, leading to the inclusion of incorrect targets at $T = 10$ that persist thereafter. These observations indicate that the inference crop size not only influences the completeness of temporal context but also affects the stability of semantic propagation. Smaller crop sizes tend to cause temporal discontinuities, while larger crop sizes may introduce accumulated semantic noise. Establishing an effective mechanism to connect different inference blocks, as well as achieving a

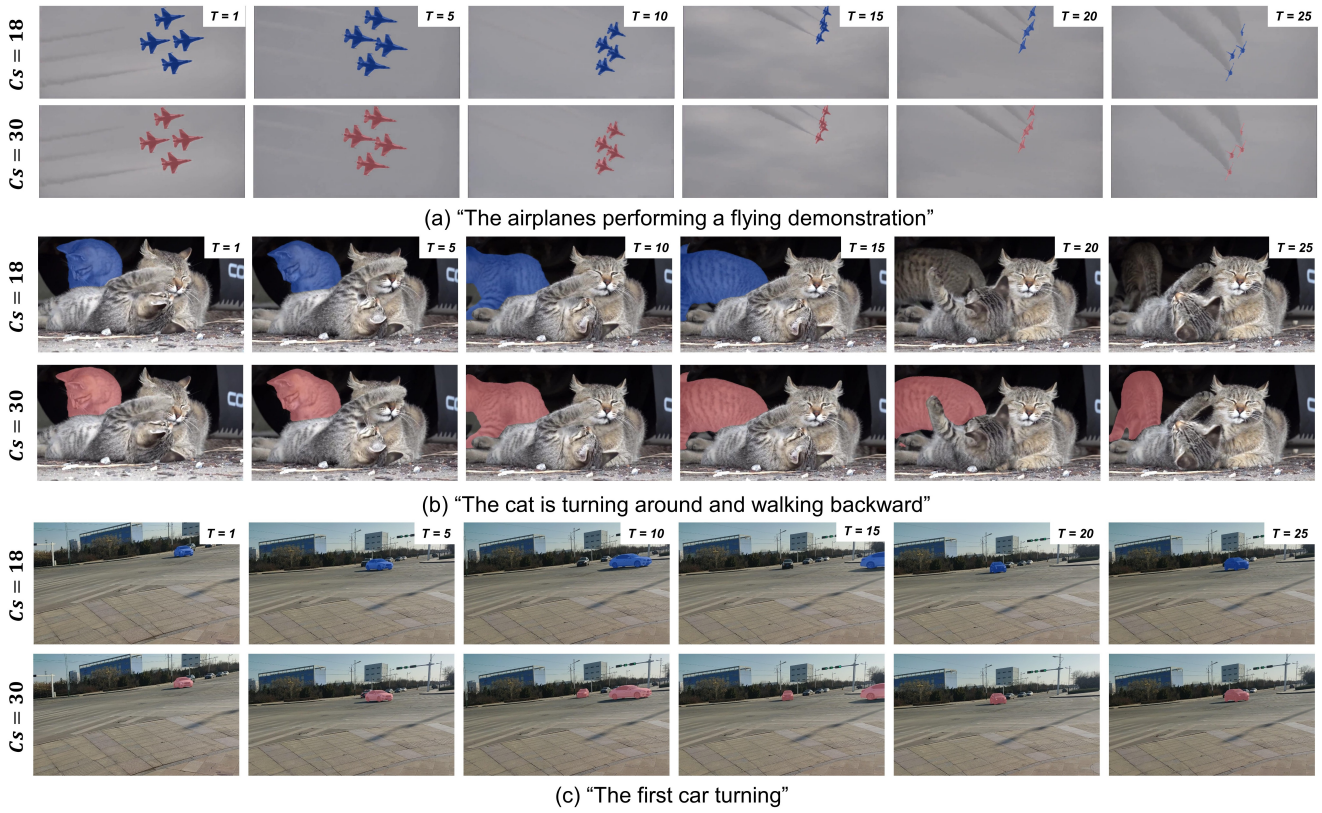
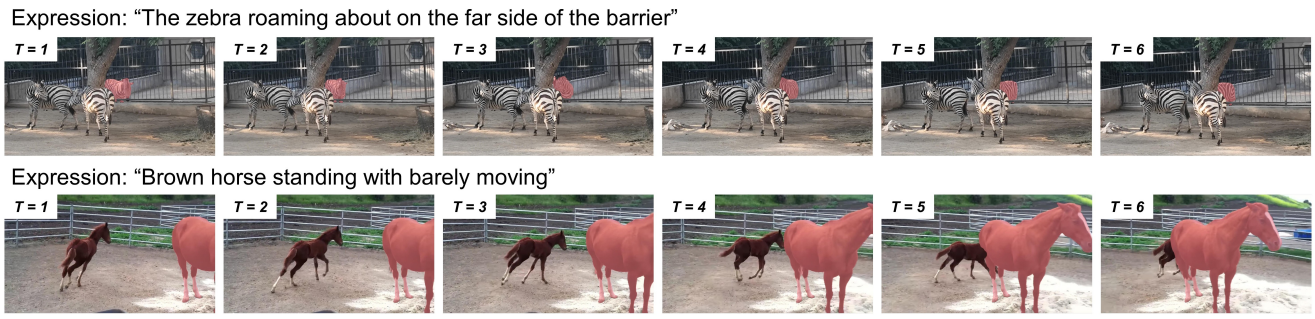


Figure 10. Visualization comparison of inference crop sizes $Cs = 18$ and $Cs = 30$ on the MeViS val subset. Visualizations are shown at an interval of 5 frames. Cs denotes the crop size.

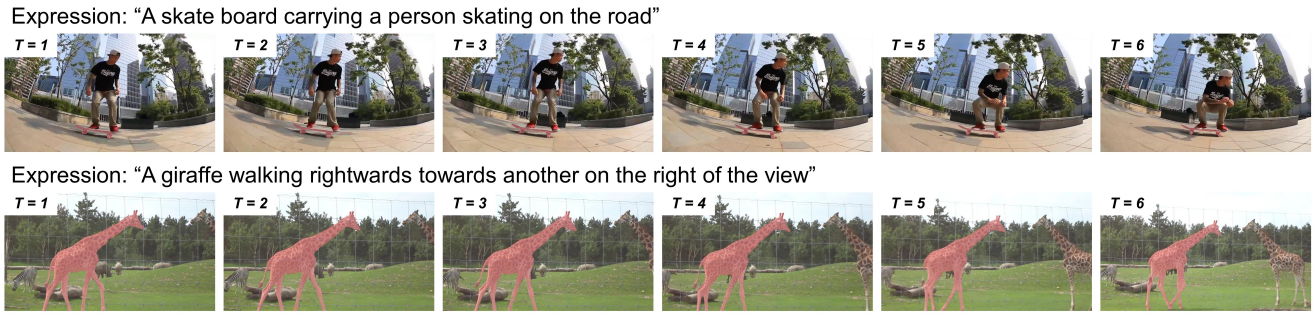
balance between retaining temporal information and suppressing semantic interference, will be the focus of our future work.

9. Additional Visualizations

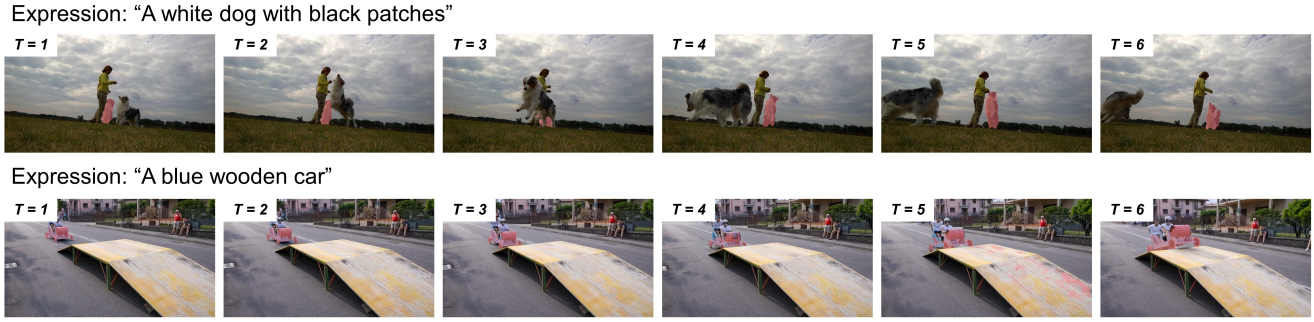
As illustrated in Fig. 11, the qualitative results of DeRVOS are presented on the MeViS [22], Ref-YouTube-VOS [76], and Ref-DAVIS17 [37] datasets. The results demonstrate that our approach robustly handles complex motions and diverse referring expressions across different scenarios, highlighting its effectiveness and generalization ability across datasets and tasks.



(a) MeViS



(b) Ref-YouTube-VOS



(c) Ref- DAVIS17

Figure 11. Qualitative results on the MeViS, Ref-YouTube-VOS, and Ref-DAVIS17 datasets.