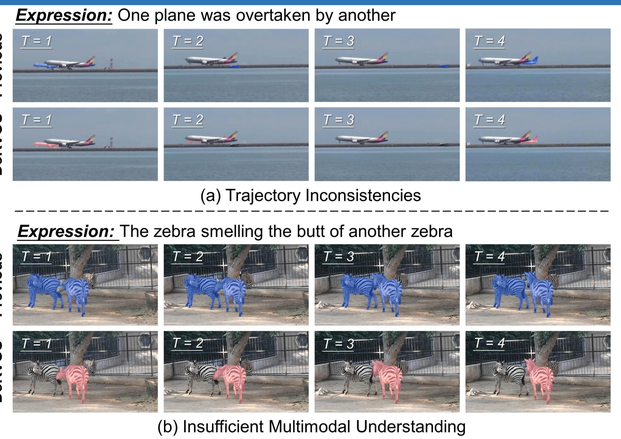
CVPR JUNE 3-7, 2026 DENVER COLORADO

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

Wenxuan Cheng*, Ming Dai*, Huimin Lu, Wankou Yang†

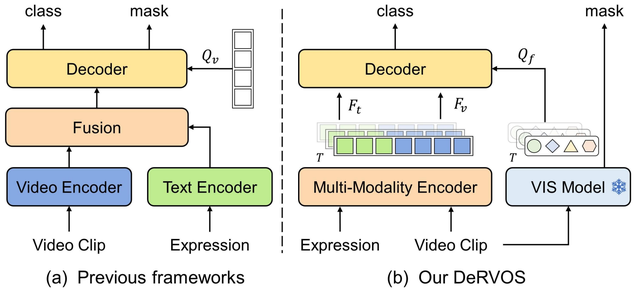
SOUTHEAST UNIVERSITY 东南大学 1902 南京

Motivation



(a) Trajectory Inconsistencies

(b) Insufficient Multimodal Understanding



(a) Previous frameworks (b) Our DeRVOS

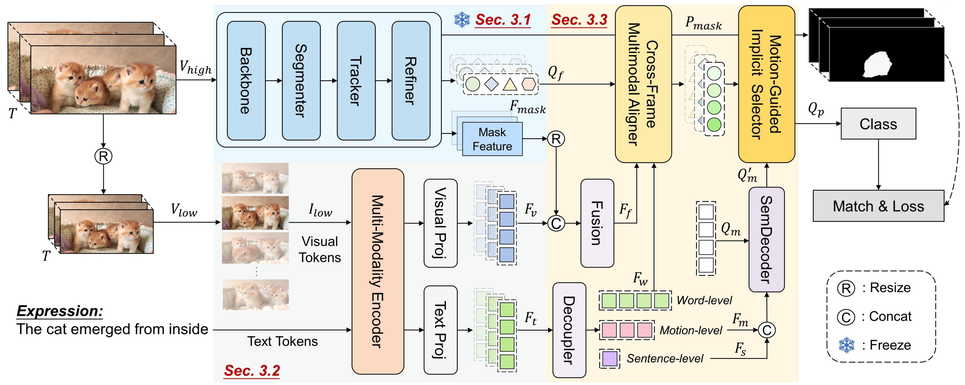
Visualization comparison

(a) In occlusion and fast-motion scenarios, previous methods suffer from trajectory inconsistencies, resulting in fragmented object tracking.
(b) Under expressions involving actions and relationships, previous methods exhibit insufficient multimodal understanding, leading to ambiguous target referencing.

Structural comparison

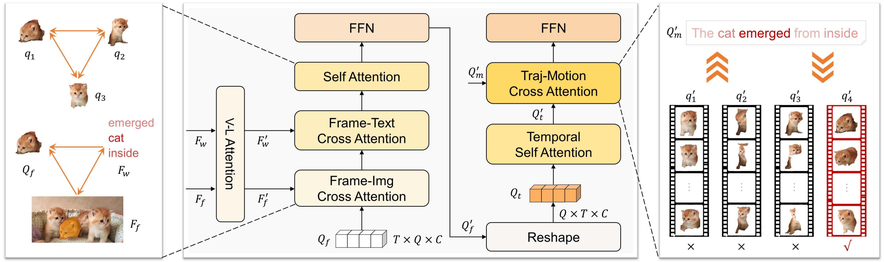
(a) Previous methods follow a logically multi-stage, query-based pipeline.
(b) DeRVOS integrates consistent trajectory generation and multimodal understanding at the upstream stages, streamlining the task to focus on modelling the relationship between referring expressions and instance trajectories.

Architecture



Overview of DeRVOS.
The RVOS task is decoupled into two branches: consistent trajectory generation (blue) and multimodal understanding (gray), responsible for cross-frame consistency modeling and cross-modal semantic comprehension, respectively. These are subsequently connected via the trajectory alignment and implicit selection module (yellow) to model the matching relationship between instance trajectories and referring expressions.

Architecture



(a) Cross-Frame Multimodal Aligner (b) Motion-Guided Implicit Selector

Structure of the cross-frame multimodal aligner and the motion-guided implicit selector.
The former aligns object representations with multimodal features through a tripartite interaction among the object, image, and text, while the latter implicitly selects target trajectories using motion-aware semantic cues.

Experiment

Method	Reference	MeViS (*val*) $\mathcal{J}\&\mathcal{F}$	MeViS (*val*) $\mathcal{J}$	MeViS (*val*) $\mathcal{F}$	MeViS (val^u) $\mathcal{J}\&\mathcal{F}$	MeViS (val^u) $\mathcal{J}$	MeViS (val^u) $\mathcal{F}$
Compared with expert methods							
MTTR [6]	[CVPR'22]	30.0	28.8	31.2	-	-	-
ReferFormer [70]	[CVPR'22]	31.0	29.8	32.2	-	-	-
LMPM [22]	[ICCV'23]	37.2	34.2	40.2	40.2	36.5	43.9
DsHmp [30]	[CVPR'24]	46.4	43.0	49.8	55.3	51.0	60.4
SSA [54]	[CVPR'25]	48.6	44.0	53.2	56.9	51.7	62.2
DMVS [25]	[CVPR'25]	48.6	44.2	52.9	58.3	52.6	63.9
ReferDINO [44]	[ICCV'25]	49.3	44.7	53.9	-	-	-
DeRVOS	-	**51.8**	**48.1**	**55.4**	**60.6**	**56.1**	**65.1**
Compared with LVLM methods							
VISA-7B [74]	[ECCV'24]	43.5	40.7	46.3	57.8	52.3	63.1
VideoLISA-3.8B [1]	[NeurIPS'24]	44.4	41.3	47.6	-	-	-
GLUS [45]	[CVPR'25]	51.3	48.5	54.2	-	-	-
DeRVOS †	-	**56.0**	**52.5**	**59.4**	**61.5**	**57.3**	**65.6**

Table 1. Comparison with state-of-the-art models on MeViS val and val^u datasets.

Method	Reference	Backbone	Ref-Youtube-VOS $\mathcal{J}\&\mathcal{F}$	Ref-Youtube-VOS $\mathcal{J}$	Ref-Youtube-VOS $\mathcal{F}$	Ref-DAVIS17 $\mathcal{J}\&\mathcal{F}$	Ref-DAVIS17 $\mathcal{J}$	Ref-DAVIS17 $\mathcal{F}$
ReferFormer [70]	[CVPR'22]	Video-Swin-B	62.9	61.3	64.6	61.1	58.1	64.1
OnlineRefer [68]	[ICCV'23]	Swin-L	63.5	61.6	65.5	64.8	61.6	67.7
HTML [29]	[ICCV'23]	Video-Swin-B	63.4	61.5	65.2	62.1	59.2	65.1
SgMg [50]	[ICCV'23]	Video-Swin-B	65.7	63.9	67.4	63.3	60.6	66.0
TempCD [63]	[ICCV'23]	Video-Swin-B	65.8	63.6	68.0	64.6	61.6	67.6
SOC [47]	[NeurIPS'23]	Video-Swin-B	66.0	64.1	67.9	64.2	61.0	67.4
DsHmp [30]	[CVPR'24]	Video-Swin-B	67.1	65.0	69.1	64.9	61.7	68.1
MUTR [75]	[AAAI'24]	Video-Swin-B	67.5	65.4	69.6	66.4	62.8	70.0
SSA [54]	[CVPR'25]	CLIP	64.3	62.2	66.4	67.3	64.0	70.7
ReferDINO [44]	[ICCV'25]	Swin-B	69.3	67.0	71.5	68.9	65.1	72.9
DeRVOS	-	BEiT3-B	**70.0**	**68.0**	**71.9**	**70.9**	**68.3**	**73.5**

Table 2. Comparison with state-of-the-art models on Ref-Youtube-VOS and Ref-DAVIS17 datasets.

Structure	Reference	RefCOCO	RefCOCO+	RefCOCOg
Compared with Non-LVLM methods				
SimVG-Seg [14]	[NeurIPS'24]	77.8	72.2	72.2
SAMWISE [13]	[CVPR'25]	76.8	67.1	67.3
CoHD [48]	[ICCV'25]	78.1	72.0	70.8
DeRIS-B [16]	[ICCV'25]	**82.0**	**75.6**	**76.3**
Compared with LVLM methods				
GSVA-13B [72]	[CVPR'24]	79.2	70.3	75.7
VISA-7B [74]	[ECCV'24]	72.4	59.8	65.5
RGA3-3B [64]	[ICCV'25]	78.9	71.3	74.7
DeRVOS	-	80.2	74.9	76.1

Table 3. Comparison with state-of-the-art models on the RefCOCO/+/g val sets for RIS.

Structure	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Text-Direct Integrator	56.1	51.4	60.9
Video-Query Integrator	56.2 ↑0.1	51.6 ↑0.2	60.8 ↓0.1
CFMA	56.5 ↑0.4	52.3 ↑0.9	60.7 ↓0.2
TAIS	57.4 ↑1.3	52.9 ↑1.5	61.8 ↑0.9

Table 5. Impact of connection structures between generation and understanding branches, evaluated on the MeViS val^u set.

Dataset	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
YouTube-VIS 2019 [76]	55.4	51.5	59.3
YouTube-VIS 2021 [77]	54.1	50.0	58.2
VIPSeg [51]	53.8	50.3	57.2
OVIS [56]	57.4	52.9	61.8

Table 7. Ablation study on the effect of DVIS++ pre-training across VIS datasets, evaluated on the MeViS val^u set.

CTG	MU	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
M2F [11]		55.5	50.4	60.6
VITA [31]	BEiT3-B	56.0 ↑0.5	51.3 ↑0.9	60.7 ↑0.1
DVIS++ [85]		57.4 ↑1.9	52.9 ↑2.5	61.8 ↑1.2
	BERT-B [20]	55.9	50.9	61.0
DVIS++	ViLT-B [38]	56.6 ↑0.7	51.5 ↑0.6	61.7 ↑0.7
	BEiT3-B [65]	57.4 ↑1.5	52.9 ↑2.0	61.8 ↑0.8

Table 4. Analysis of Consistent Trajectory Generation (CTG) and Multimodal Understanding (MU) capabilities, evaluated on the MeViS val^u set.

Resolution	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	Train time (h)
320 × 320	56.9	52.2	61.6	3.8
384 × 384	57.4 ↑0.5	52.9 ↑0.7	61.8 ↑0.2	4.2 ↓0.4
448 × 448	59.3 ↑2.4	54.5 ↑2.3	64.0 ↑2.4	4.9 ↓1.1
640 × 640	60.9 ↑4.0	56.3 ↑4.1	65.4 ↑3.8	5.8 ↓2.0

Table 6. Ablation study on the input resolution of the consistent trajectory generation branch, conducted on the MeViS val^u set.

Method	Learnable params	GPU memory	Training time	MeViS $\mathcal{J}\&\mathcal{F}$	YT-VOS $\mathcal{J}\&\mathcal{F}$	DAVIS $\mathcal{J}\&\mathcal{F}$
DsHmp	346.2M	34G	20 hours	46.4	67.1	64.9
DeRVOS	179.4M	18G	9 hours	51.8	70.0	70.9

Table 8. Quantitative comparison of performance and efficiency between DeRVOS and DsHmp, evaluated on the MeViS, Ref-YouTube-VOS, and Ref-DAVIS17 datasets.

Conclusion

Conclusion. We propose DeRVOS, a novel RVOS framework that decouples consistent trajectory generation from multimodal understanding, combining temporally stable object tracking with strong visual-text comprehension. It leverages a pretrained trajectory generation model to obtain consistent object representations and fine-grained masks, and a pretrained multimodal encoder for image-level visual-text alignment. A trajectory alignment and implicit selection module further connects the two branches, enabling motion-aware semantic guidance. Experiments demonstrate its superiority over existing methods on RVOS and RIS benchmarks.