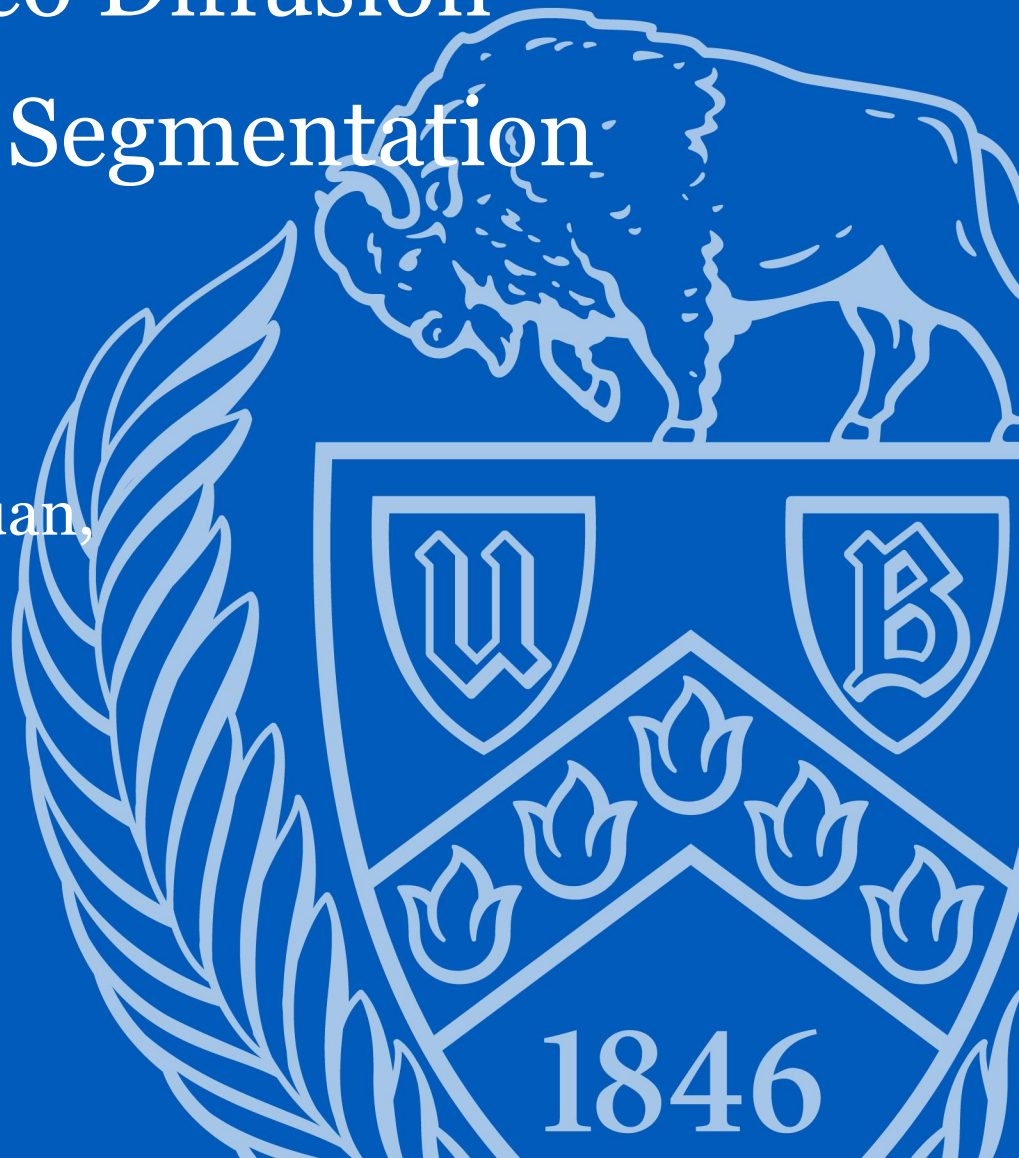
Exploring Pre-trained Text-to-Video Diffusion Models for Referring Video Object Segmentation

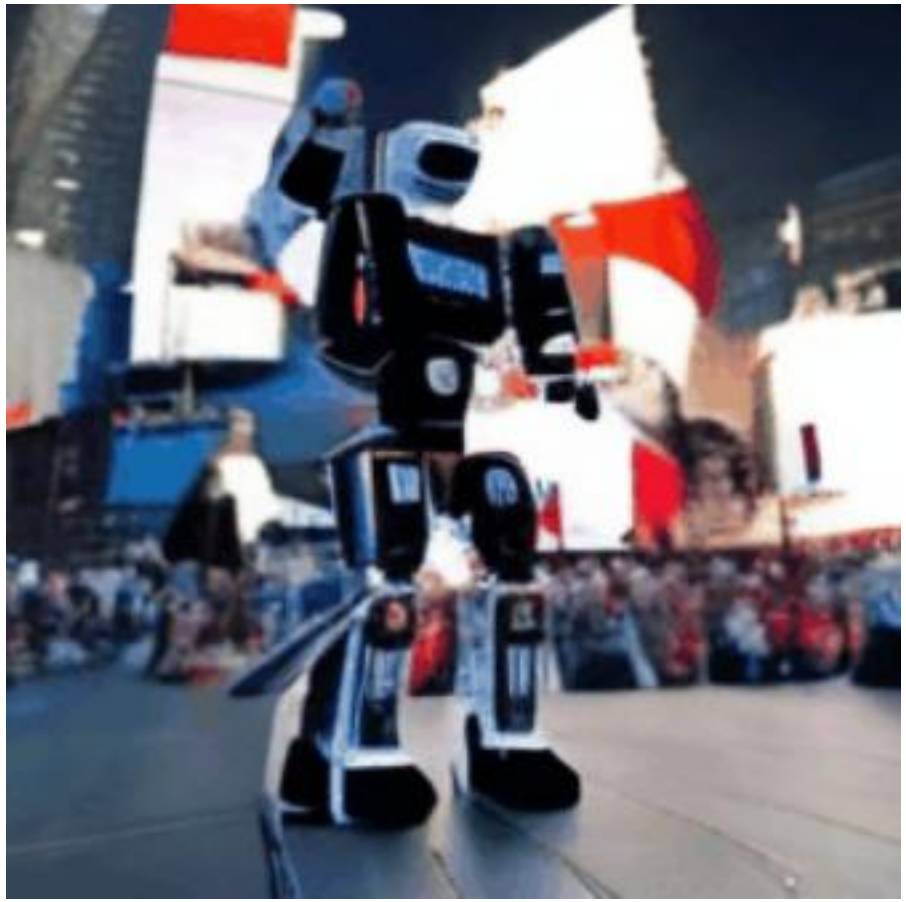
Zixin Zhu, Xuelu Feng, Dongdong Chen, Junsong Yuan,
Chunming Qiao, Gang Hua

Code: https://github.com/buxiangzhiren/VD-IT



University at Buffalo The State University of New York

Background: Text-to-Video (T2V) Diffusion Models[1]

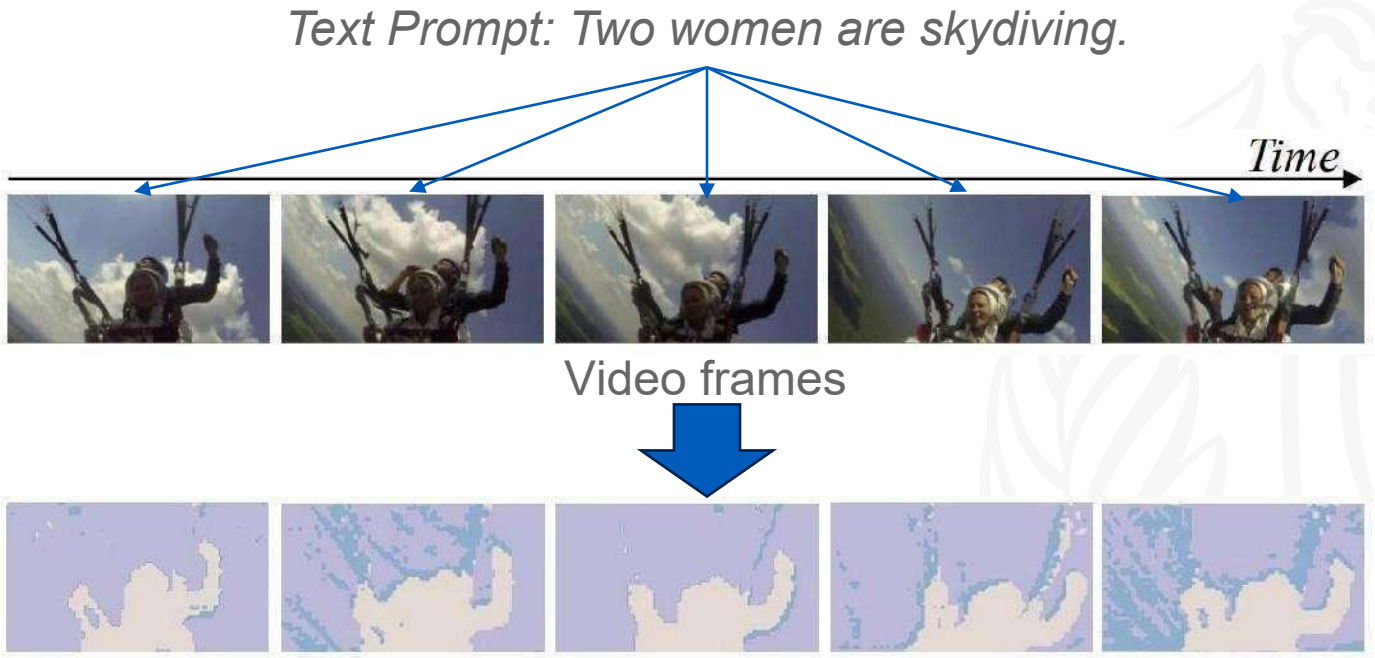


Robot dancing in times square.

Clown fish swimming through the coral reef.

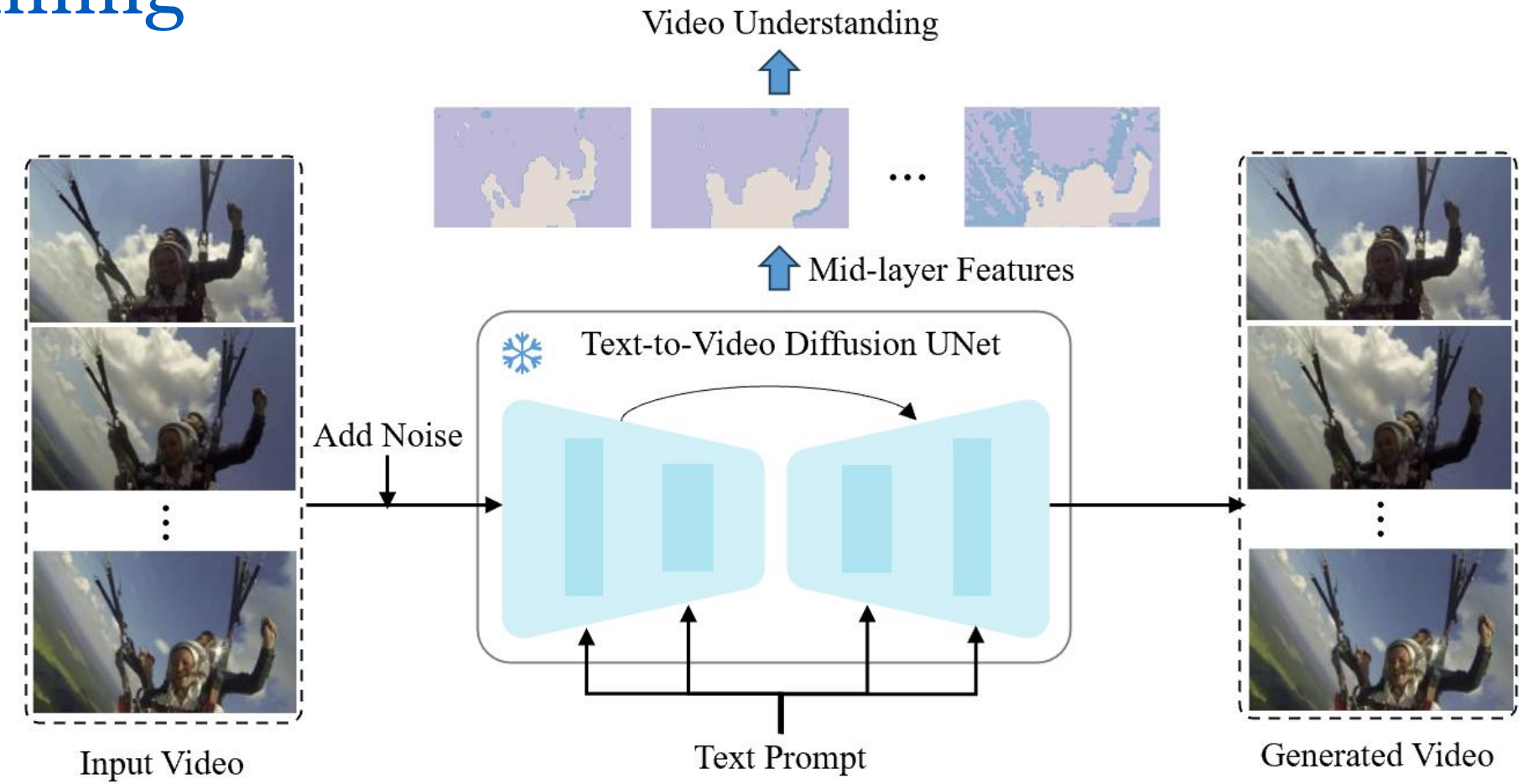
[1] https://modelscope.cn/models/iic/text-to-video-synthesis/summary

Motivation: Temporal Consistency Prior via Diffusion Pretraining



T2V diffusion pretraining ➔ Consistent text prompts guide all video frames ➔ Temporal consistency ➔ Improved downstream video understanding

Motivation: Temporal Consistency Prior via Diffusion Pretraining



T2V diffusion pretraining ➔ Consistent text prompts guide all video frames ➔ Temporal consistency ➔ Improved downstream video understanding

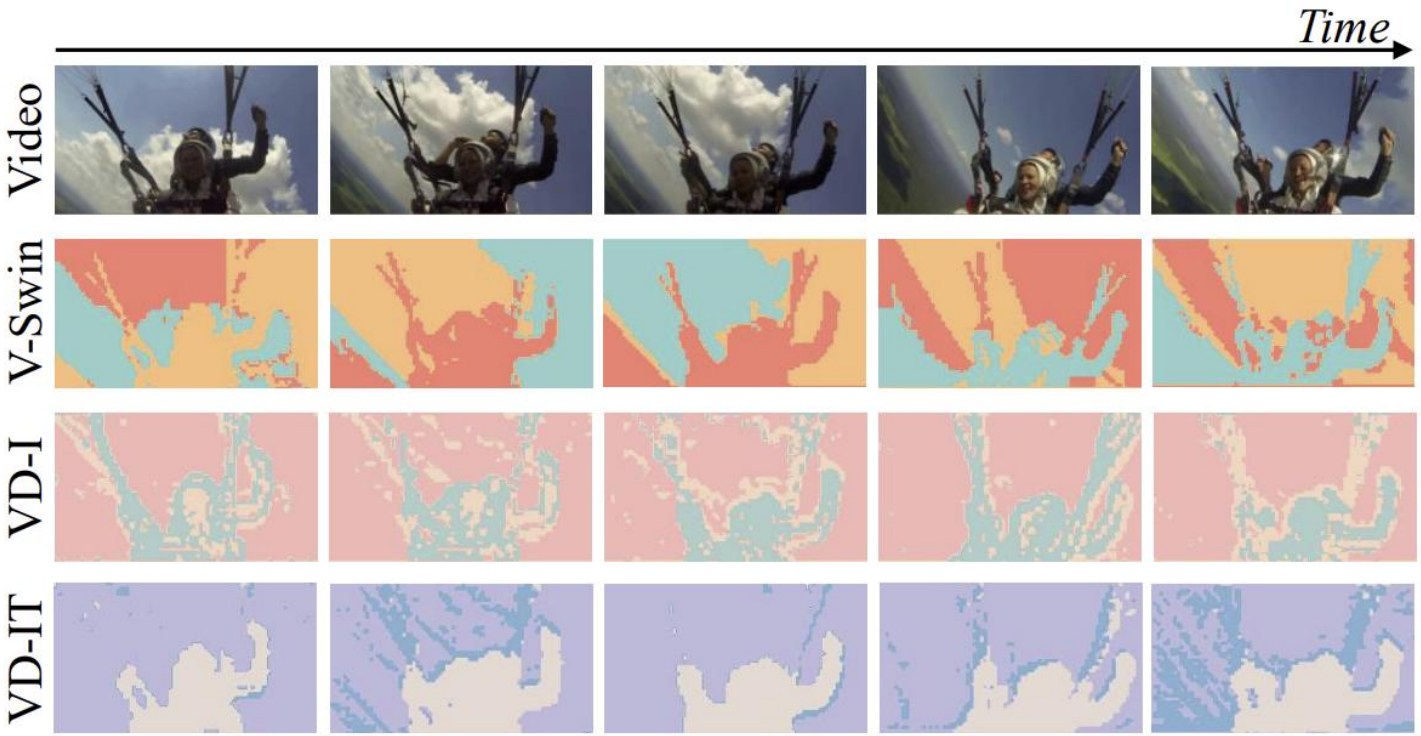
Background: Referring Video Object Segmentation

Referring Text: (1) A black shooting gun (2) A person shooting with a rifle

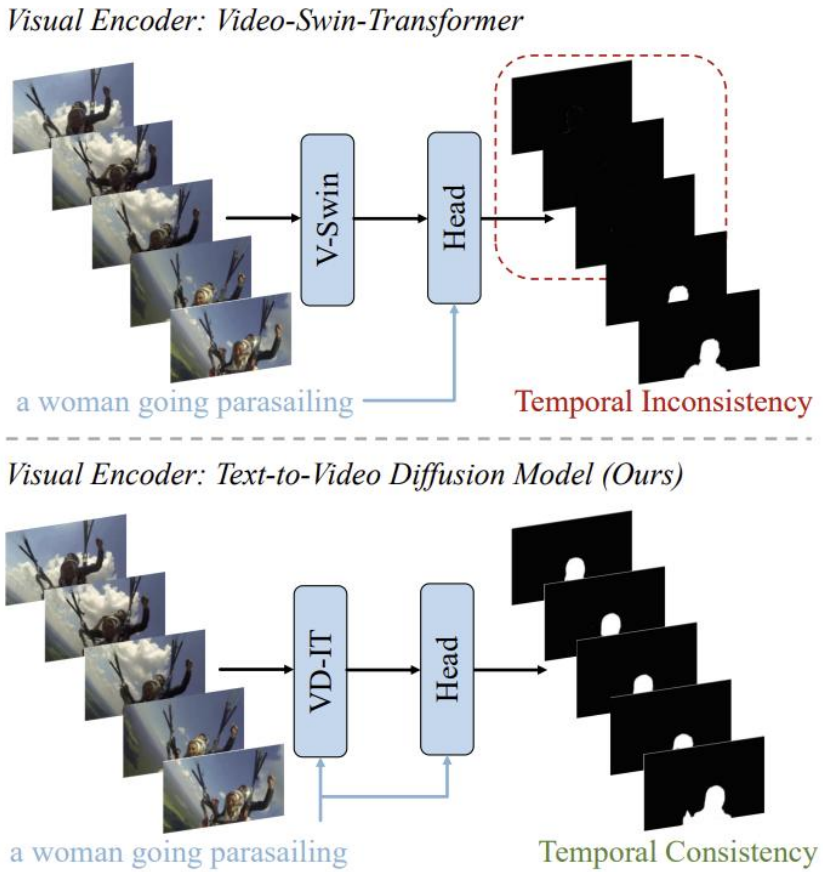
Input Video

Output Video

Introduction



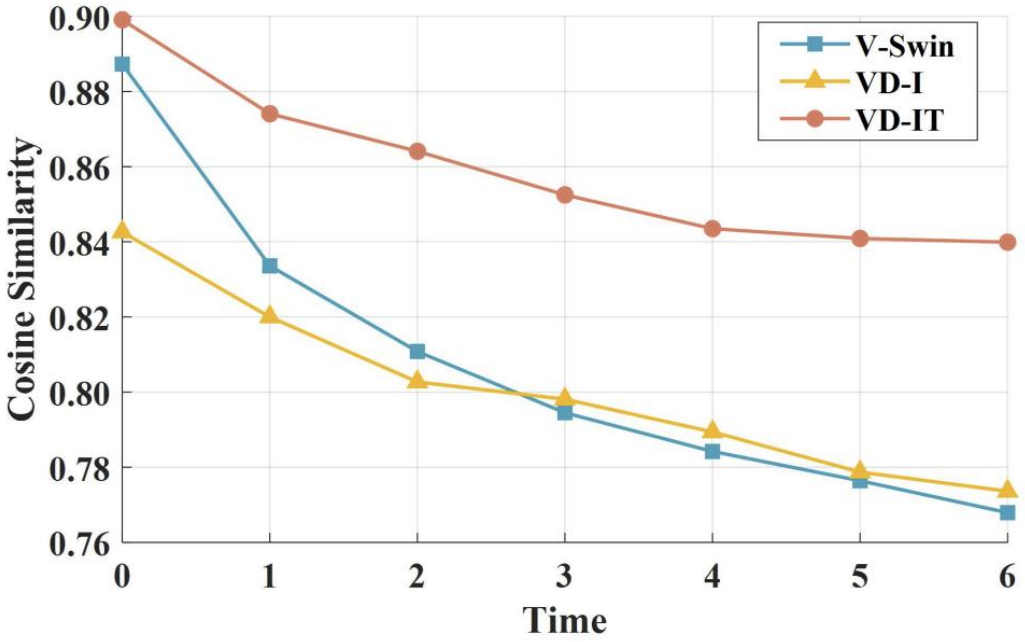
Analysis of learned features of existing methods that use discriminative backbone (Video Swin Transformer[2]) and our methods (VD-I and VD-IT) that use fixed pretrained generative T2V model.



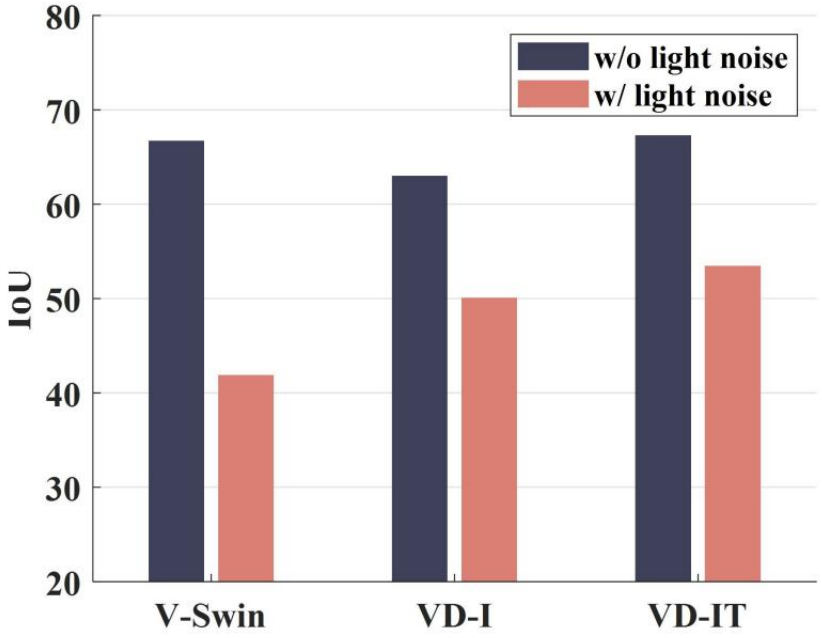
Temporal inconsistency in visual features will subsequently cause temporally inconsistent masks.

[2] Liu Z, Ning J, Cao Y, et al. Video swin transformer[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022

Introduction



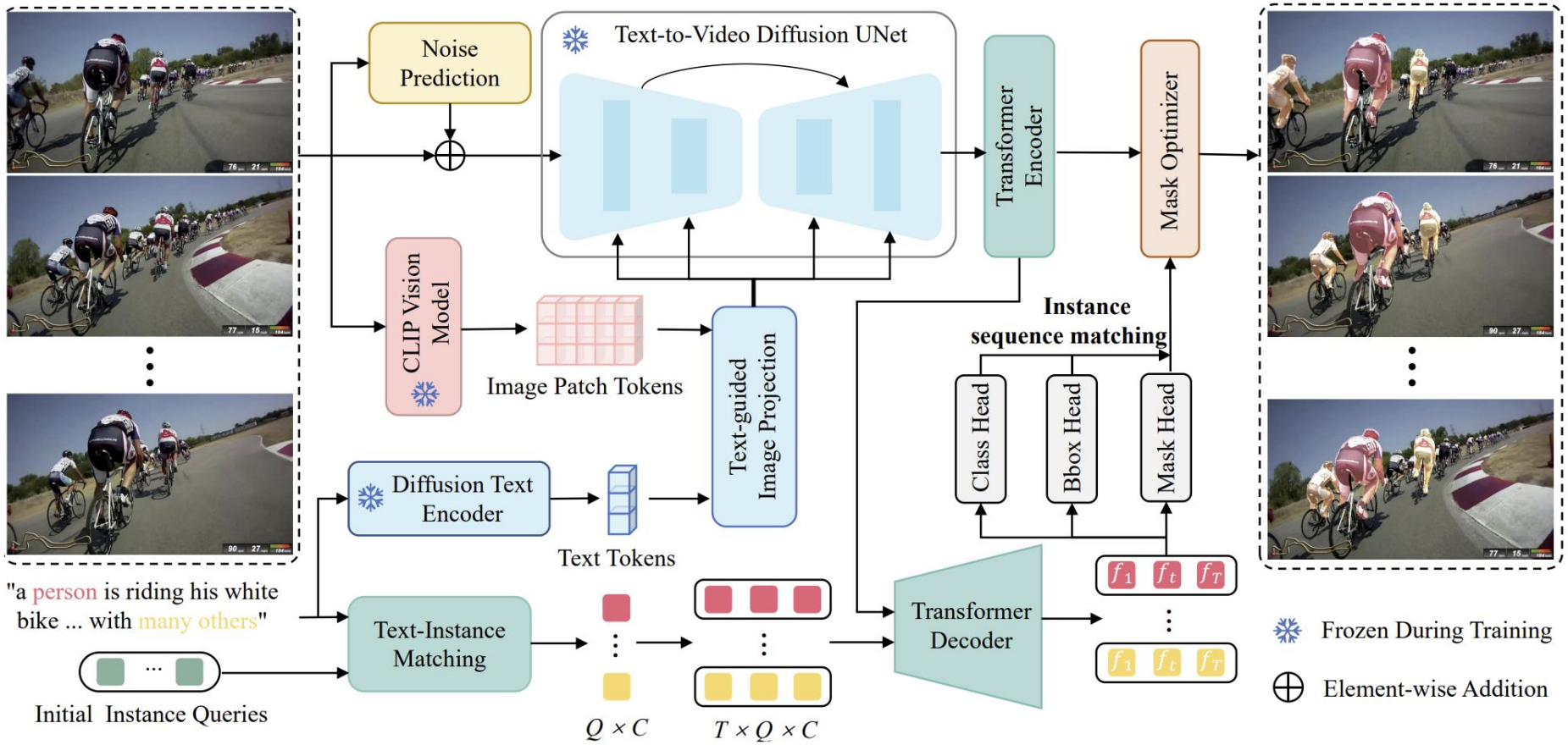
Temporal Semantic Consistency. Averaged over 1,000 samples from RefYoutube-VOS[3], the cosine similarity between the Region of Interest (RoI) features of the initial frame and the following seven frames is reported.



Robustness against light noise. We modify the brightness of various frames randomly and compare the IoU of segmentation results under changing lighting conditions. The results are reported on Ref-Youtube-VOS.

[3] Seo, S., Lee, J.Y., Han, B.: Urvos: Unified referring video object segmentation network with a large-scale benchmark. In: ECCV. pp. 208–223. (2020)

Method



This framework comprises two core components: visual feature extraction and the mask segmentation head.

Experiments

Method	Backbone	Ref-YouTube-VOS				Ref-DAVIS17		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	FPS	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
CMSA [57]	ResNet-50	36.4	34.8	38.1	-	40.2	36.9	43.5
URVOS [38]	ResNet-50	47.2	45.3	49.2	-	51.5	47.3	56.0
CMPC-V [25]	I3D	47.5	45.6	49.3	-	-	-	-
PMINet [9]	ResNeSt-101	53.0	51.5	54.5	-	-	-	-
YOFO [17]	ResNet-50	48.6	47.5	49.7	10	53.3	48.8	57.8
LBDT [8]	ResNet-50	49.4	48.2	50.6	-	54.3	-	-
MLRL [50]	ResNet-50	49.7	48.4	51.0	-	52.8	50.0	55.4
MTTR [3]	Video-Swin-T	55.3	54.0	56.6	-	-	-	-
MANet [5]	Video-Swin-T	55.6	54.8	56.5	-	-	-	-
ReferFormer [52]	Video-Swin-T	56.0	54.8	57.3	50	-	-	-
SgMg [29]	Video-Swin-T	58.9	57.7	60.0	65	56.7	53.3	60.0
SgMg* [29]	Video-Swin-B	61.6	59.7	63.5	40	-	-	-
VD-IT (Ours)	Video Diffusion	**64.8**	**63.1**	**66.6**	21	**63.0**	**59.9**	**66.1**
Pre-training with RefCOCO/+/g								
ReferFormer [52]	Video-Swin-T	59.4	58.0	60.9	50	59.6	56.5	62.7
SgMg [29]	Video-Swin-T	62.0	60.4	63.5	65	61.9	59.0	64.8
ReferFormer [52]	Video-Swin-B	62.9	61.3	64.6	33	61.1	58.1	64.1
OnlineRefer [51]	Video-Swin-B	62.9	61.0	64.7	-	62.4	59.1	65.6
SgMg [29]	Video-Swin-B	65.7	63.9	67.4	40	63.3	60.6	66.0
VD-IT (Ours)	Video Diffusion	**66.5**	**64.4**	**68.5**	21	**69.4**	**66.2**	**72.6**

Comparison with the state-of-the-art methods on Ref-Youtube-VOS and Ref-DAVIS17[4]. * denotes that we run the official codes to get the results.
(These two datasets emphasize daily life videos.)

Method	Backbone	A2D-Sentences			JHMDB-Sentences		
		mAP	Overall	Mean	mAP	Overall	Mean
Hu et al. [14]	VGG-16	13.2	47.4	35.0	17.8	54.6	52.8
Gavrilyuk et al. [12]	I3D	19.8	53.6	42.1	23.3	54.1	54.2
ACAN [41]	I3D	27.4	60.1	49.0	28.9	57.6	58.4
CMPC-V [25]	I3D	40.4	65.3	57.3	34.2	61.6	61.7
ClawCraneNet [22]	ResNet-50/101	-	63.1	59.9	-	64.4	65.6
MTTR [3]	Video-Swin-T	46.1	72.0	64.0	39.2	70.1	69.8
ReferFormer [48]	Video-Swin-T	52.8	77.6	69.6	42.2	71.9	71.0
SgMg [29]	Video-Swin-T	56.1	78.0	70.4	44.4	72.8	71.7
ReferFormer [48]	Video-Swin-B	55.0	78.6	70.3	43.7	73.0	71.8
SgMg [29]	Video-Swin-B	58.5	79.9	72.0	45.0	73.7	72.5
VD-IT (Ours)	Video Diffusion	**61.4**	**81.5**	**73.2**	**46.5**	**74.4**	**73.4**

Quantitative comparison to state-of-the-art R-VOS methods on A2DSentences[5] and JHMDB-Sentences[5]. (These two datasets emphasize action videos)

[4] Khoreva, A., Rohrbach, A., Schiele, B.: Video object segmentation with language referring expressions. In: ACCV. pp. 123–141. (2019)
[5] Gavrilyuk, K., Ghodrati, A., Li, Z., Snoek, C.G.: Actor and action video segmentation from a sentence. In: CVPR. pp. 5958–5966 (2018)

Experiments

Referring Text: (1) A black shooting gun (2) A person shooting with a rifle

SgMg[6]

Ours

[6] Miao, B., Bennamoun, M., Gao, Y., Mian, A.: Spectrum-guided multi-granularity referring video object segmentation. In: ICCV. pp. 920–930 (2023)

Experiments

Referring Text: (1) Black and white bike (2) A man holding a bike

SgMg

Ours

Conclusion

In this paper, we present a pioneering exploration into leveraging video priors, specifically temporal consistency, acquired by pre-trained text-to-video diffusion models for video understanding tasks.

- Pioneering Use of Video Priors: We are the first to explore the use of temporal consistency priors acquired by pre-trained text-to-video diffusion models for video understanding tasks.
- Enhanced Temporal Consistency: Our research shows that these pre-trained diffusion models exhibit significantly better temporal consistency compared to conventional discriminatively fine-tuned video encoders.
- Innovative R-VOS Framework: We propose a new R-VOS framework, VD-IT, which incorporates several innovative designs that improve the quality of extracted features and boost overall performance.