Video Instance Segmentation using Inter-Frame Communication Transformers

Sukjun Hwang[1] Miran Heo[1] Seoung Wug Oh[2] Seon Joo Kim[1]

[1]Yonsei University [2]Adobe Research

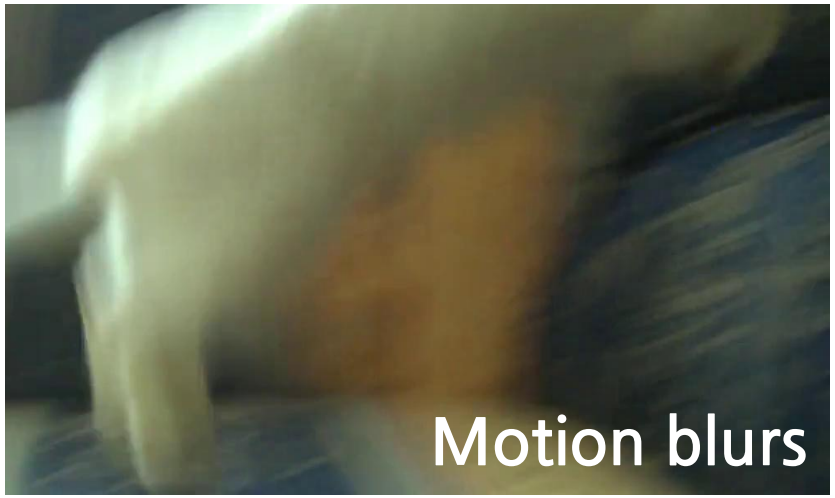
Video Instance Segmentation

Given a video, predict **spatio-temporal** masks of instances



masks predicted by our model

Typical failure cases in videos



Utilizing clip information

Cannot easily detect the lizard if receiving **per-frame** input

Utilizing clip information

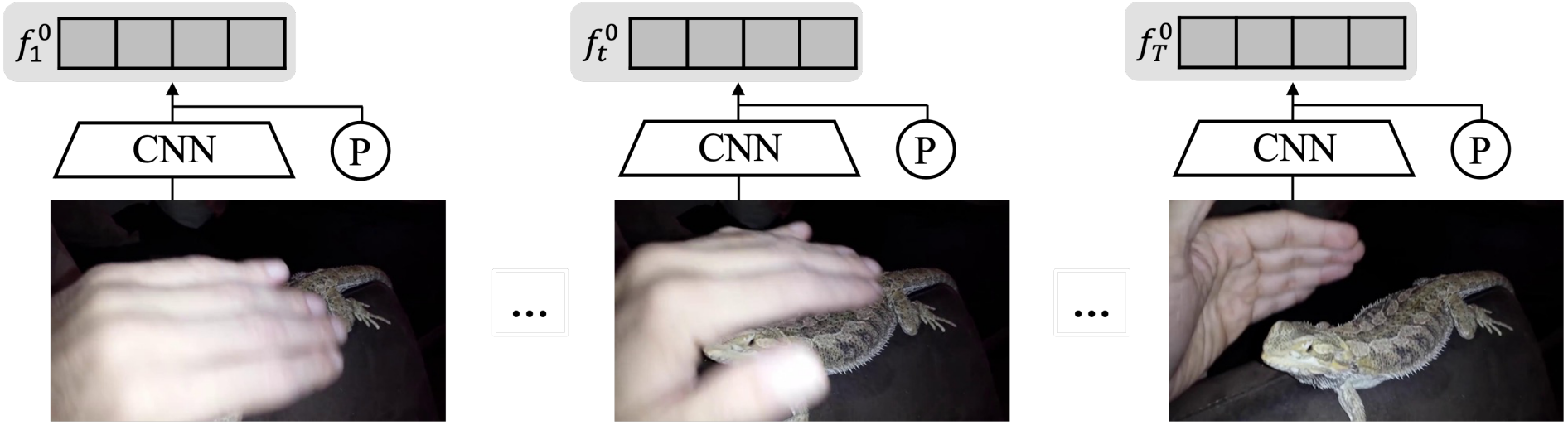
Cannot easily detect the lizard if receiving **per-frame** input

Incorporating **clip-wise information** is the key to the higher accuracy

Inter-Frame Communication Transformers

Each frame is encoded by passing a **backbone** network

+ **spatial positional embeddings**



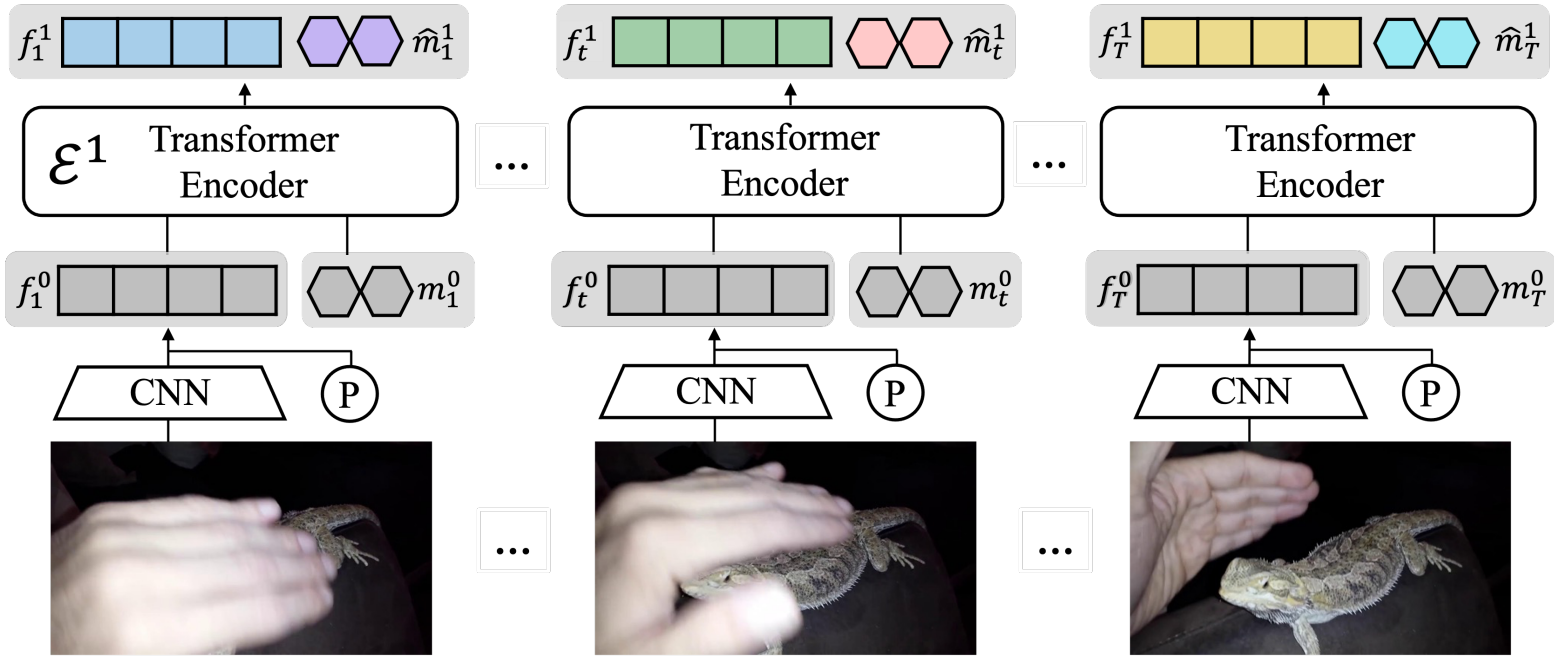
Inter-Frame Communication Transformers

Encode-Receive layer simultaneously encodes

frame-wise information

frame features f and memory tokens m

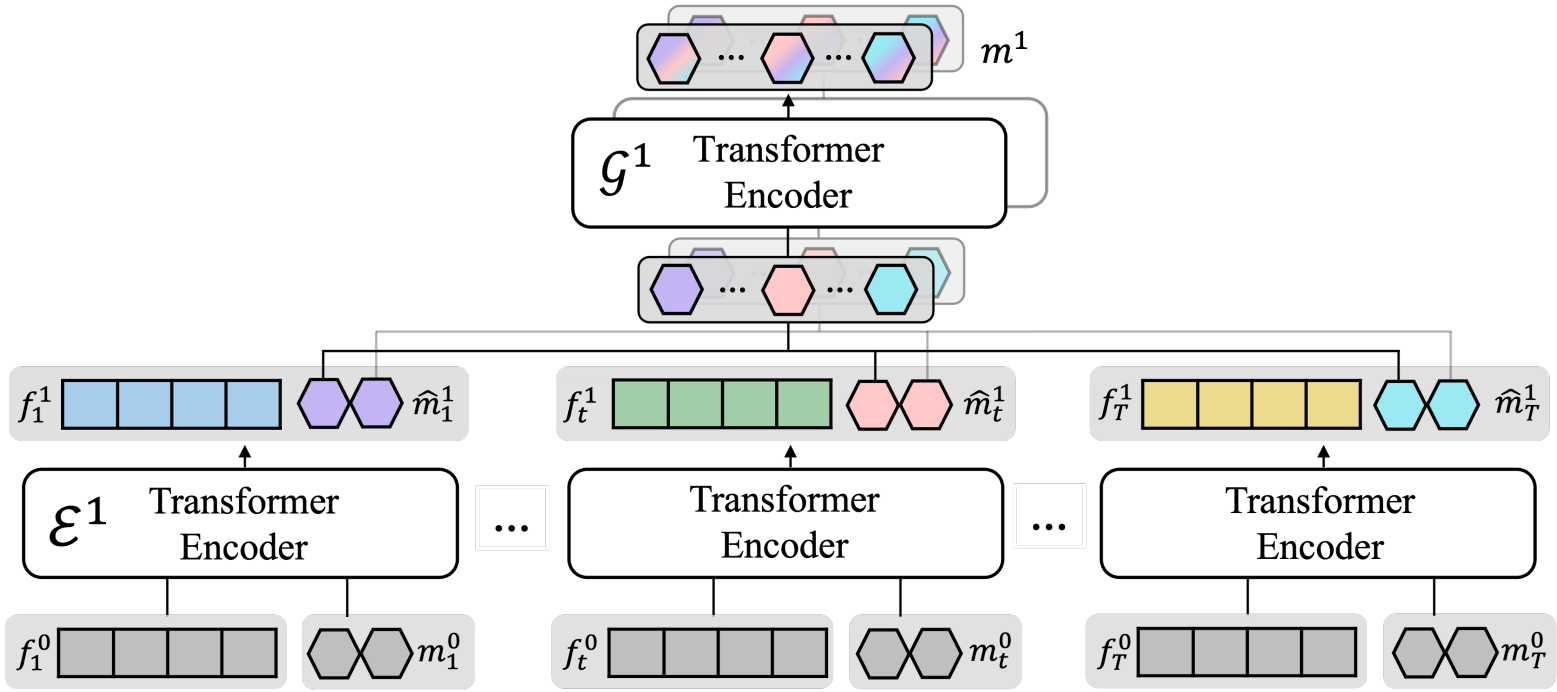
clip-level context



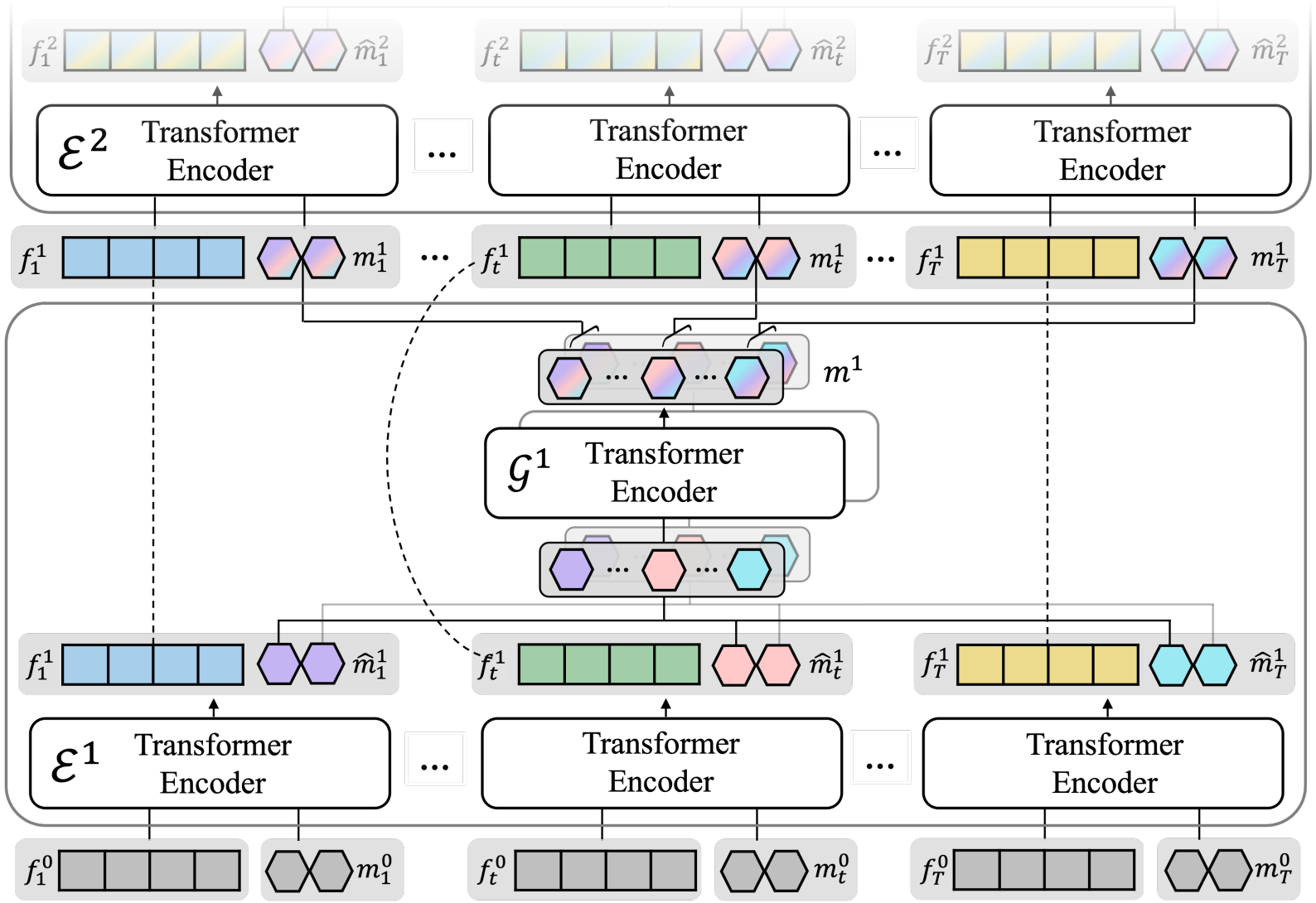
Inter-Frame Communication Transformers

Memory tokens are grouped by memory indices

Gather-Communicate layer aggregates **frame-wise information** and builds **communications** in between



Inter-Frame Communication Transformers



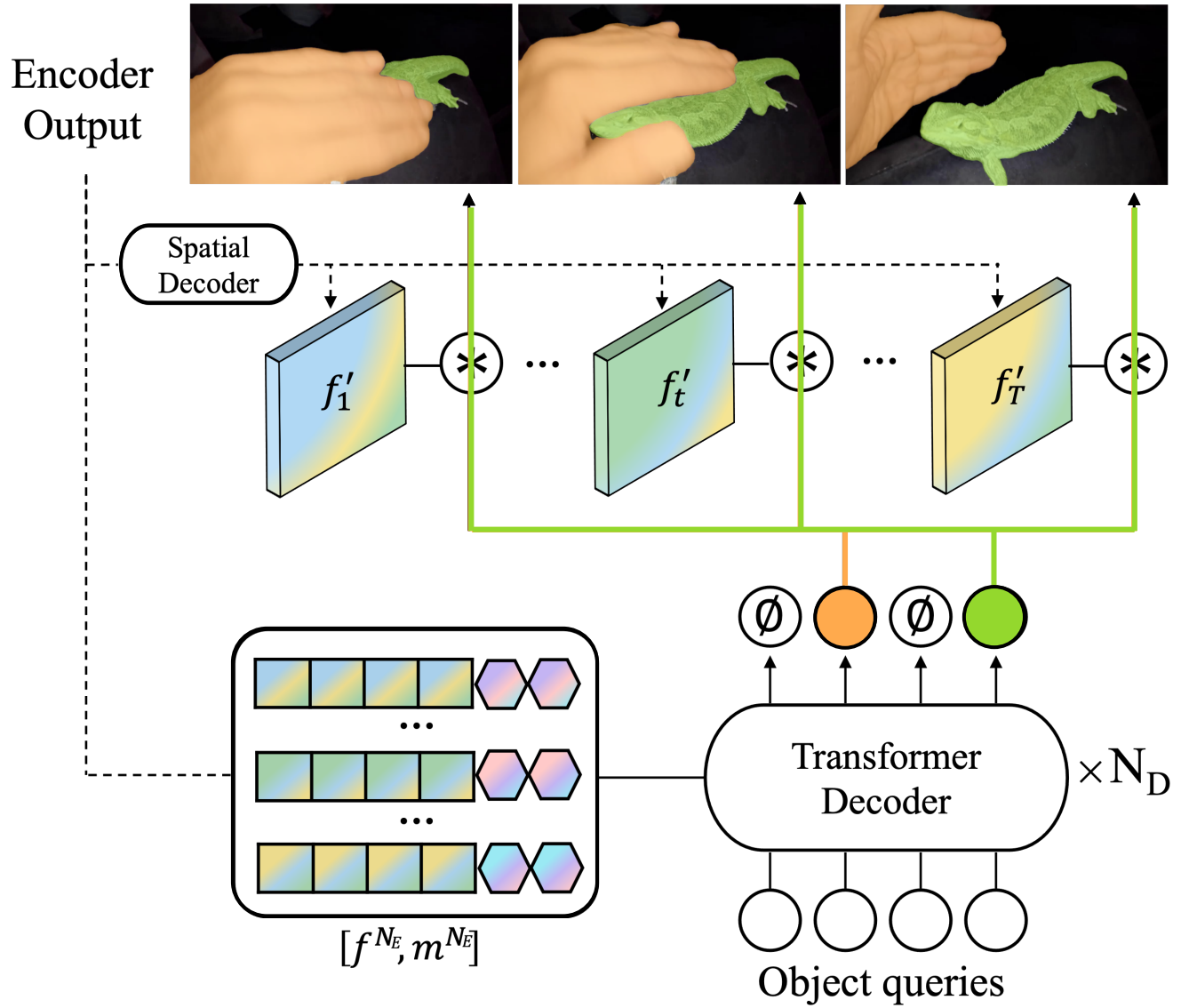
Clip-wise Segmentation

Encoder outputs are used by
- Spatial Decoder
- Transformer Decoder

Spatial Decoder is Instance-Agnostic

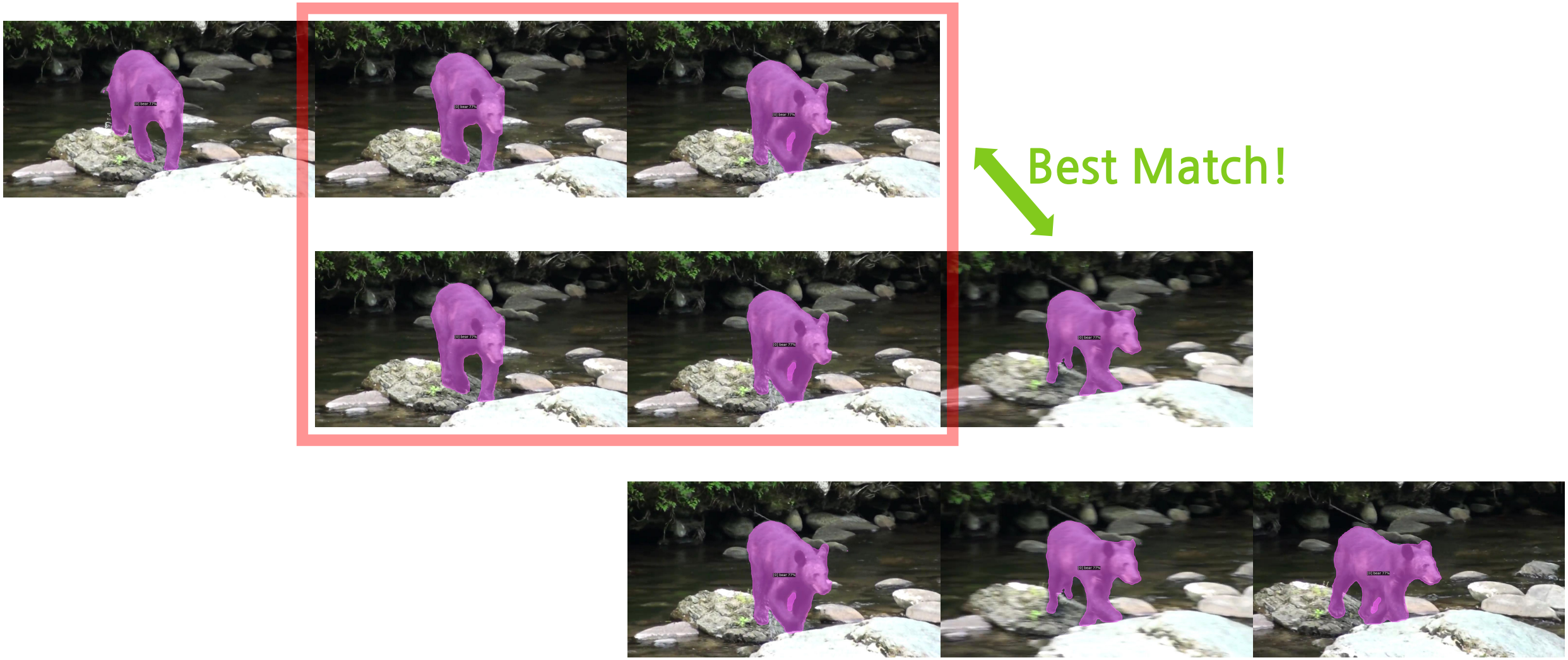
Each **dynamic convolutional weight** represents an instance within a **clip**

Segmentation and **Tracking** at once within a **clip**



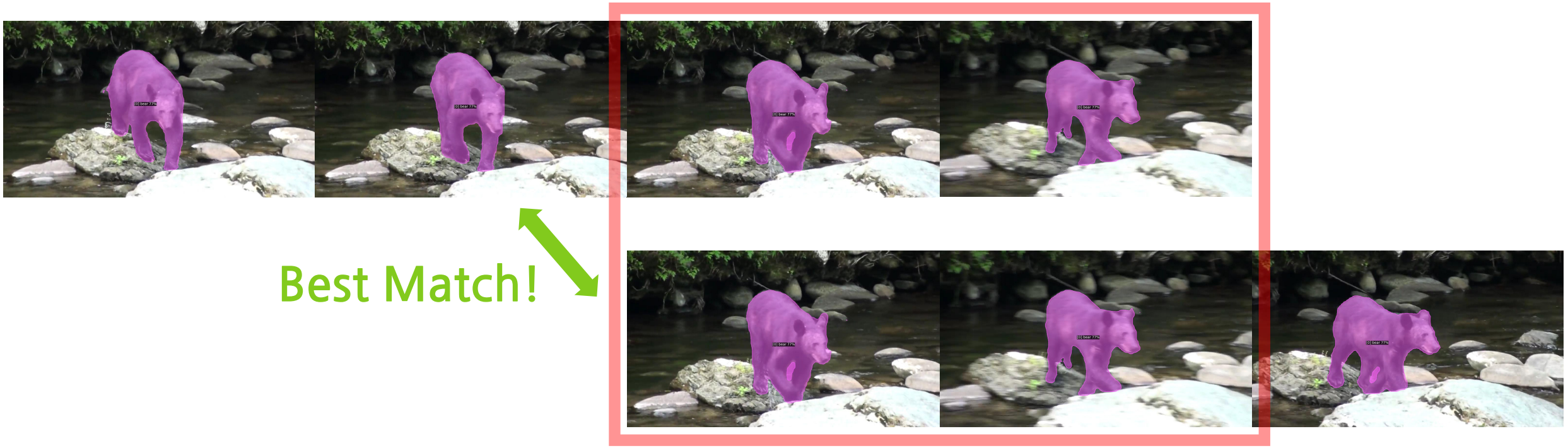
Clip-level Instance Tracking

Use space-time soft IoU of intersecting frames



Clip-level Instance Tracking

Use space-time soft IoU of intersecting frames



Experiments - Encoder Comparison

All components are the same except for the **encoder**

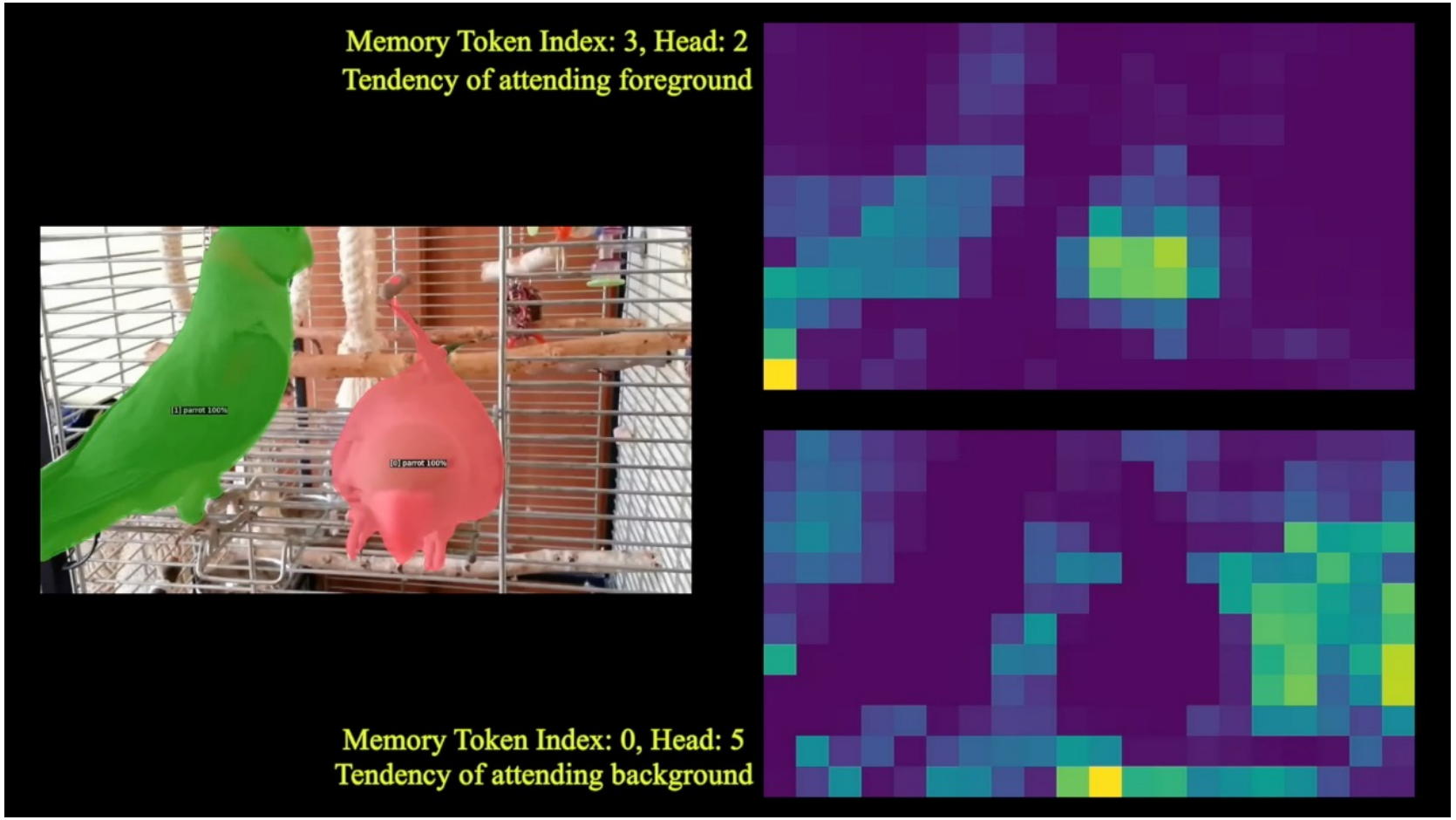
Alleviates computational overheads of self-attention using the **memory tokens**

Communication Type	Complexity per Layer	FLOPs (G)[1] 360 × 640 T=5	360 × 640 T=36	720 × 1280 T=5	720 × 1280 T=36
No Comm	$\mathcal{O}(C^2THW + CT(HW)^2)$	5.17	37.23	24.62	177.29
Full THW	$\mathcal{O}(C^2THW + C(THW)^2)$	6.94	148.70	50.63	1815.38
Decompose T-HW	$\mathcal{O}(C^2THW + CT(HW)^2 + CT^2HW)$	8.33	60.24	36.73	265.50
IFC ($M = 8$)	$\mathcal{O}(C^2THW + CT(HW)^2)$	5.52	39.73	25.05	180.39

lower complexity
higher accuracy

	T=5 AP	T=5 AP_{75}	T=5 FPS	T=10 AP	T=10 AP_{75}	T=10 FPS	T=15 AP	T=15 AP_{75}	T=15 FPS	T=20 AP	T=20 AP_{75}	T=20 FPS
No Comm	37.4	39.9	38.1	38.8	41.6	40.8	39.3	41.7	46.7	39.6	41.9	52.9
Full THW	37.2	40.0	37.6	38.8	41.2	35.5	39.8	42.6	32.9	39.7	42.8	34.8
Decomp T-HW	37.2	39.8	35.7	38.3	40.9	37.9	38.5	41.5	42.6	39.0	41.9	49.4
IFC	39.0	42.7	36.3	39.6	43.0	38.9	39.8	43.0	43.7	40.4	43.4	50.2

Experiments - Memory Tokens



(b) Image instance segmentation on COCO `val` set

	AP^{coco}	AP^{coco}_{50}
w/o mem	35.0	56.6
w/ mem	35.1	56.5

(c) Number of memory tokens (AP)

	T=5	T=10	T=15	T=20
M=1	37.6	39.2	39.4	39.4
M=2	37.9	39.2	39.6	39.8
M=4	38.0	39.5	39.7	39.9
M=8	**39.0**	**39.6**	**39.8**	**40.4**
M=16	38.1	39.1	39.7	39.9

(d) Index-wise memory decomposition

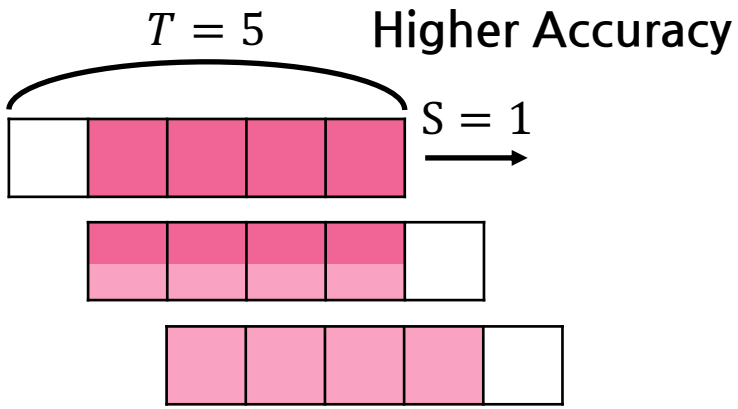
	T=5	T=10	T=15	T=20
Unified	38.1	38.9	39.7	39.9
Decomp	**39.0**	**39.6**	**39.8**	**40.4**

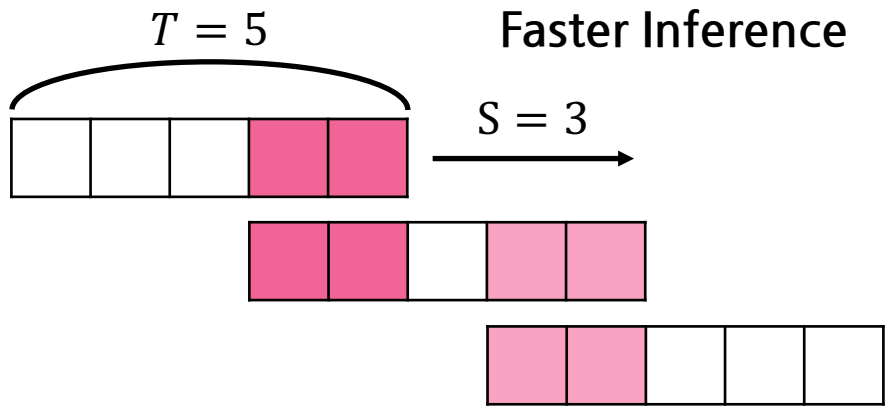
Experiments - Comparison w/ Previous Works

	Method (Settings)		Backbone [31]	FPS[2]	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
per-frame	MaskTrack R-CNN [1]		ResNet-50	26.1	30.3	51.1	32.6	31.0	35.5
	MaskTrack R-CNN [1]		ResNet-101	-	31.8	53.0	33.6	33.2	37.6
	SipMask [2]		ResNet-50	35.5	33.7	54.1	35.8	35.4	40.1
	SG-Net [4]		ResNet-50	-	34.8	56.1	36.8	35.8	40.8
	SG-Net [4]		ResNet-101	-	36.3	57.1	39.6	35.9	43.0
	CrossVIS [3]		ResNet-50	-	36.3	56.8	38.9	35.6	40.7
	CrossVIS [3]		ResNet-101	-	36.6	57.3	39.7	36.0	42.0
per-clip	STEm-Seg [32]		ResNet-101	3.0	34.6	55.8	37.9	34.4	41.6
	VisTR [11]	(T=36)	ResNet-50	51.1	35.6	56.8	37.0	35.2	40.2
	VisTR [11]	(T=36)	ResNet-101	43.5	38.6	61.3	42.3	37.6	44.2
	MaskProp [10]	(T=13)	ResNet-50	-	40.0	-	42.9	-	-
	MaskProp [10]	(T=13)	ResNet-101	-	42.5	-	45.6	-	-
	Ours$_{\text{near-online}}$	(T=5)	ResNet-50	46.5	39.0	60.4	42.7	41.7	51.6
	Ours$_{\text{offline}}$	(T=36)	ResNet-50	107.1	41.2	65.1	44.6	42.3	49.6
	Ours$_{\text{offline}}$	(T=36)	ResNet-101	89.4	42.6	66.6	46.3	43.5	51.4

(d) Effect of strides

		AP	AP_{75}	FPS
T = 5	S = 3	38.7	42.1	72.7
T = 10	S = 5	39.5	42.8	83.0
T = 15	S = 8	39.7	43.0	92.5
T = 20	S = 10	40.4	43.3	95.7





Video Instance Segmentation using Inter-Frame Communication Transformers

Sukjun Hwang[1] Miran Heo[1] Seoung Wug Oh[2] Seon Joo Kim[1]

[1]Yonsei University [2]Adobe Research