01 VITA: Video Instance Segmentation via Object Token Association

Miran Heo*
Yonsei University

Sukjun Hwang*
Yonsei University

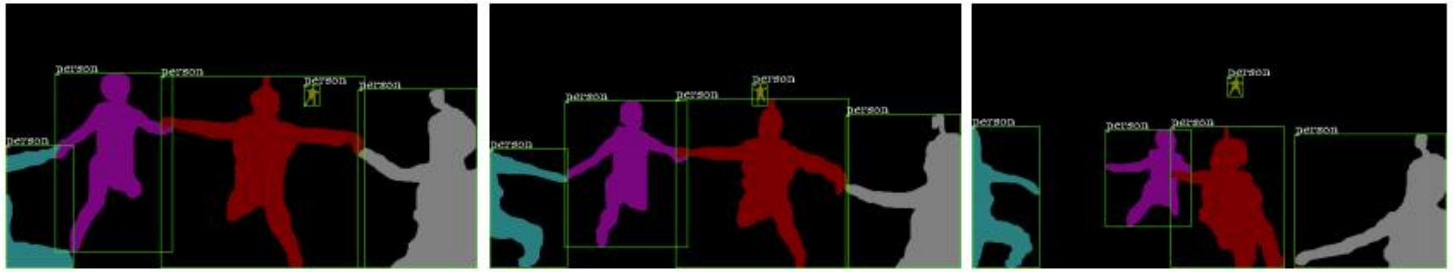
Seoung Wug Oh
Adobe Research

Joon-Young Lee
Adobe Research

Seon Joo Kim
Yonsei University

{miran, sj.hwang, seonjookim}@yonsei.ac.kr {seoh, jolee}@adobe.com

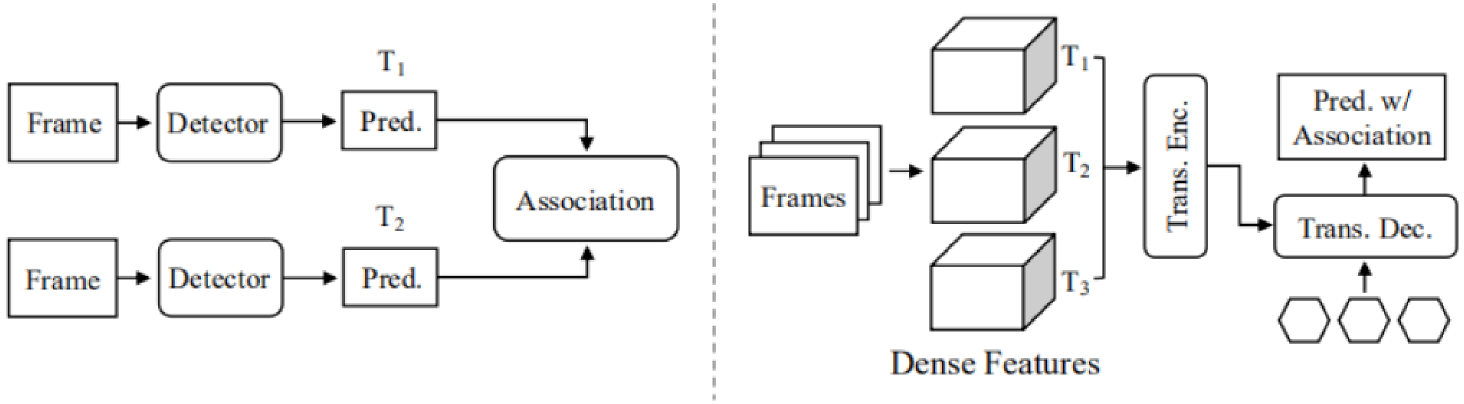
Video frames



Video instance annotations

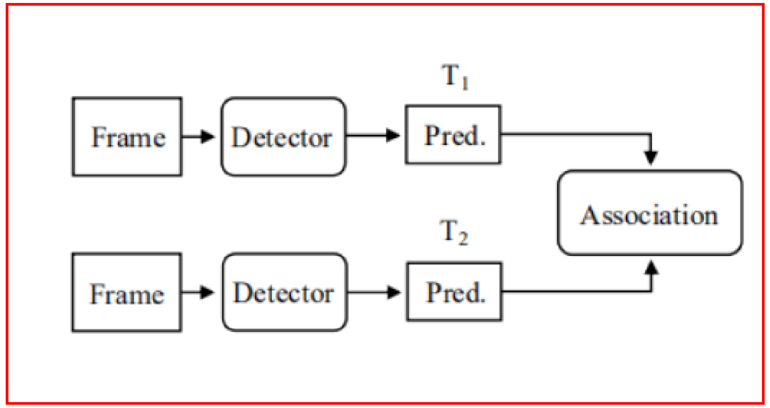
01 VITA: Video Instance Segmentation via Object Token Association

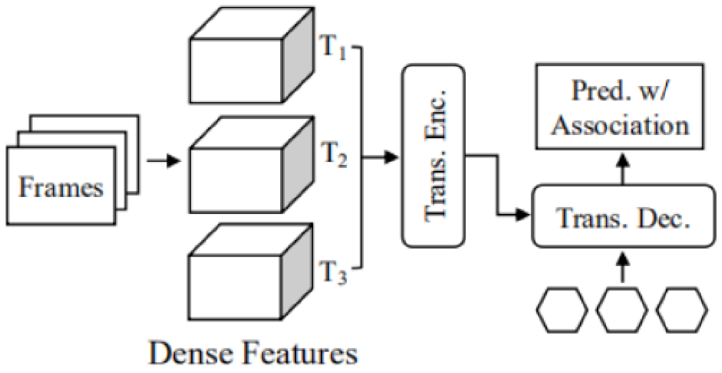
Online VIS approaches & Offline VIS approaches

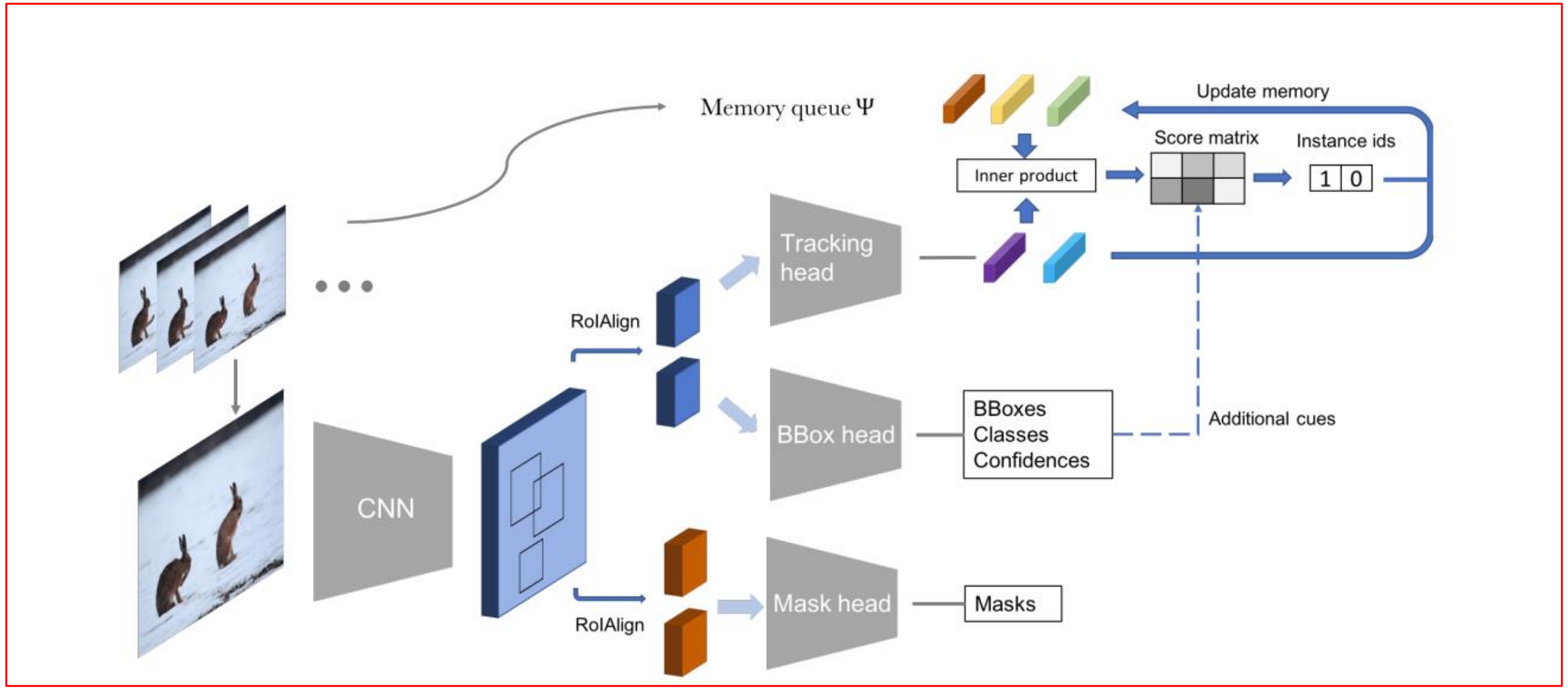


01 VITA: Video Instance Segmentation via Object Token Association

Online VIS approaches & Offline VIS approaches



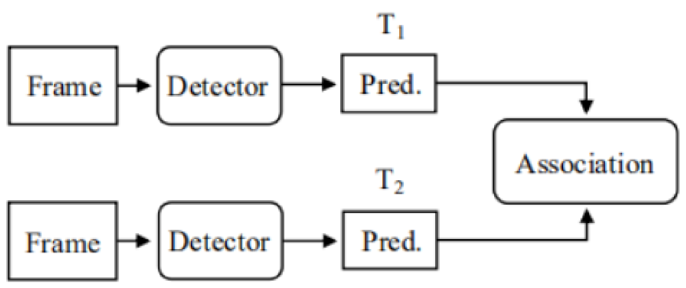


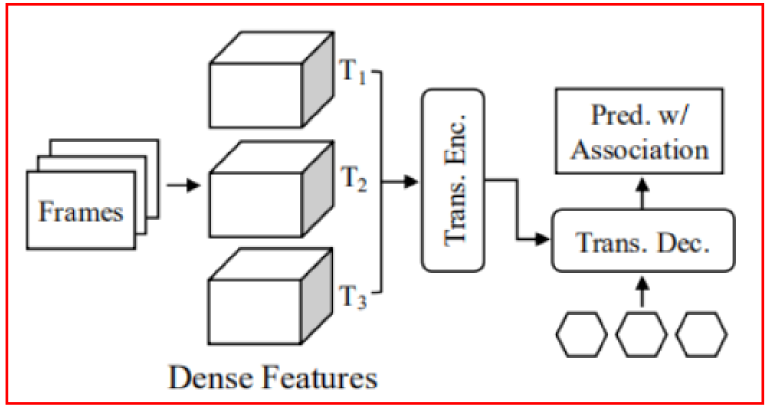


Video Instance Segmentation [ICCV2019]

01 VITA: Video Instance Segmentation via Object Token Association

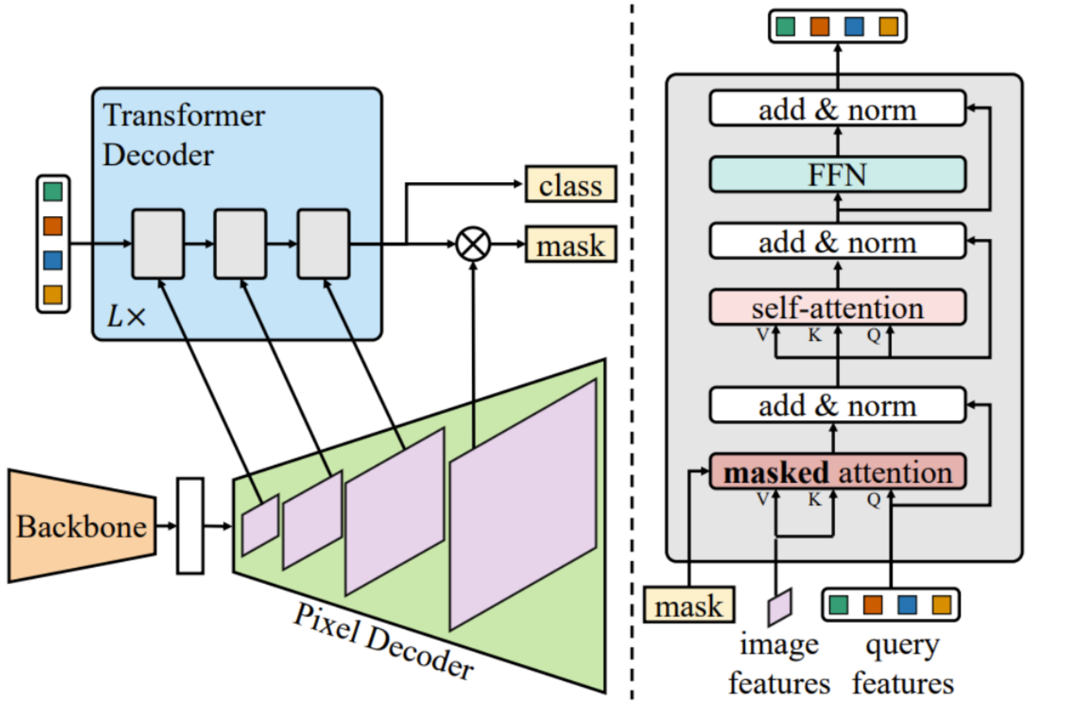
Online VIS approaches & Offline VIS approaches





Advantages：
1) they have a greater receptive field to the temporal axis
2) they can avoid error propagation derived from hand-crafted association algorithms.

Disadvantages：
such methods show difficulties in handling **long sequences** as the myriad of dense reference features hinders the Transformer layers from retrieving relevant information.



Masked-attention Mask Transformer for Universal Image Segmentation [CVPR2022]

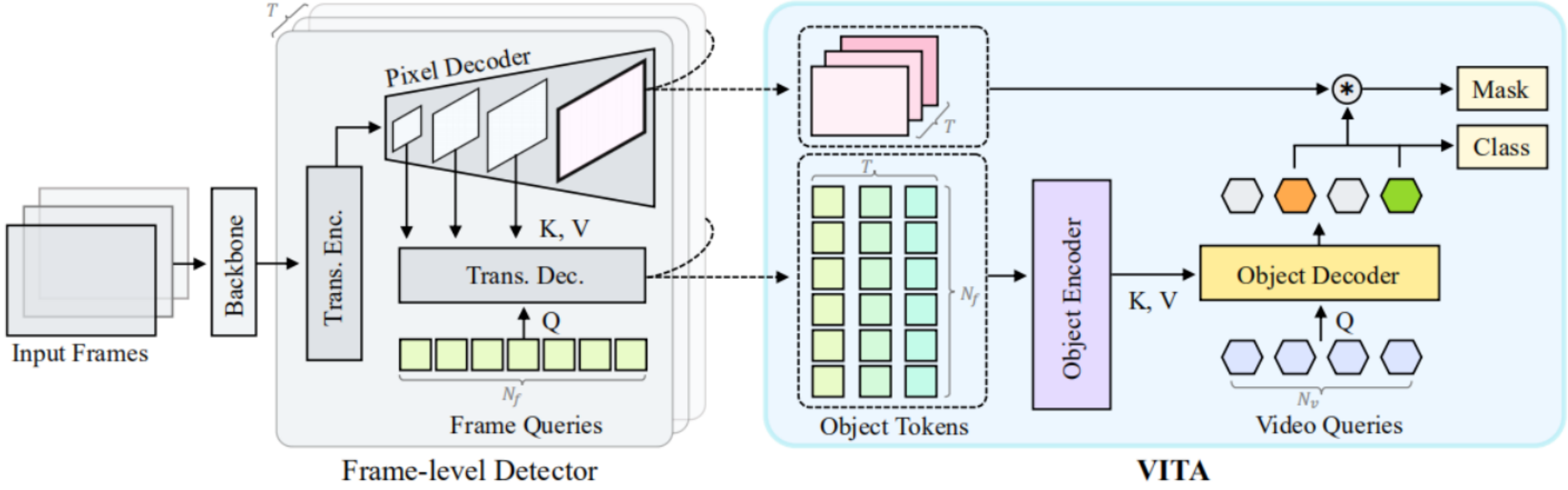
01 VITA: Video Instance Segmentation via Object Token Association

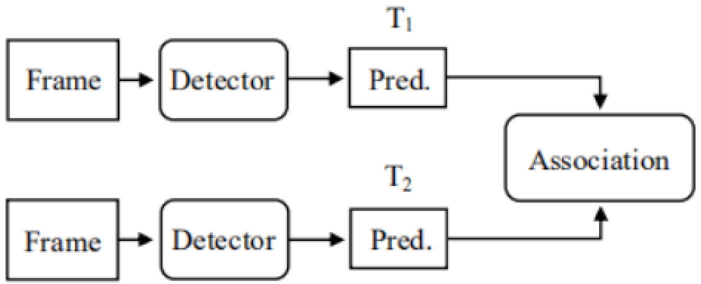
VITA approach

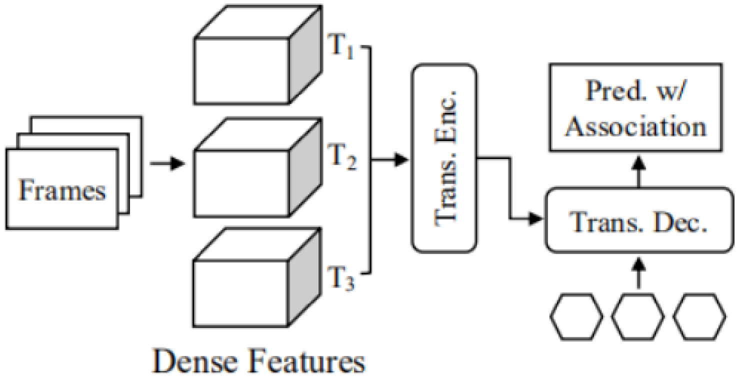
Hypotheses:

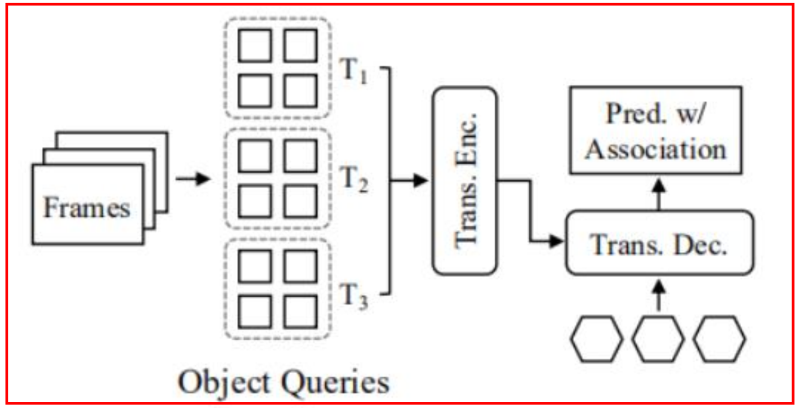
1) an image object detector can fully embody the context of an object into a feature vector (or a token);
2) a video can be represented by the relationship between the objects.

VITA aims to parse an input video from the collection of object tokens without the necessity of referencing dense spatio-temporal backbone



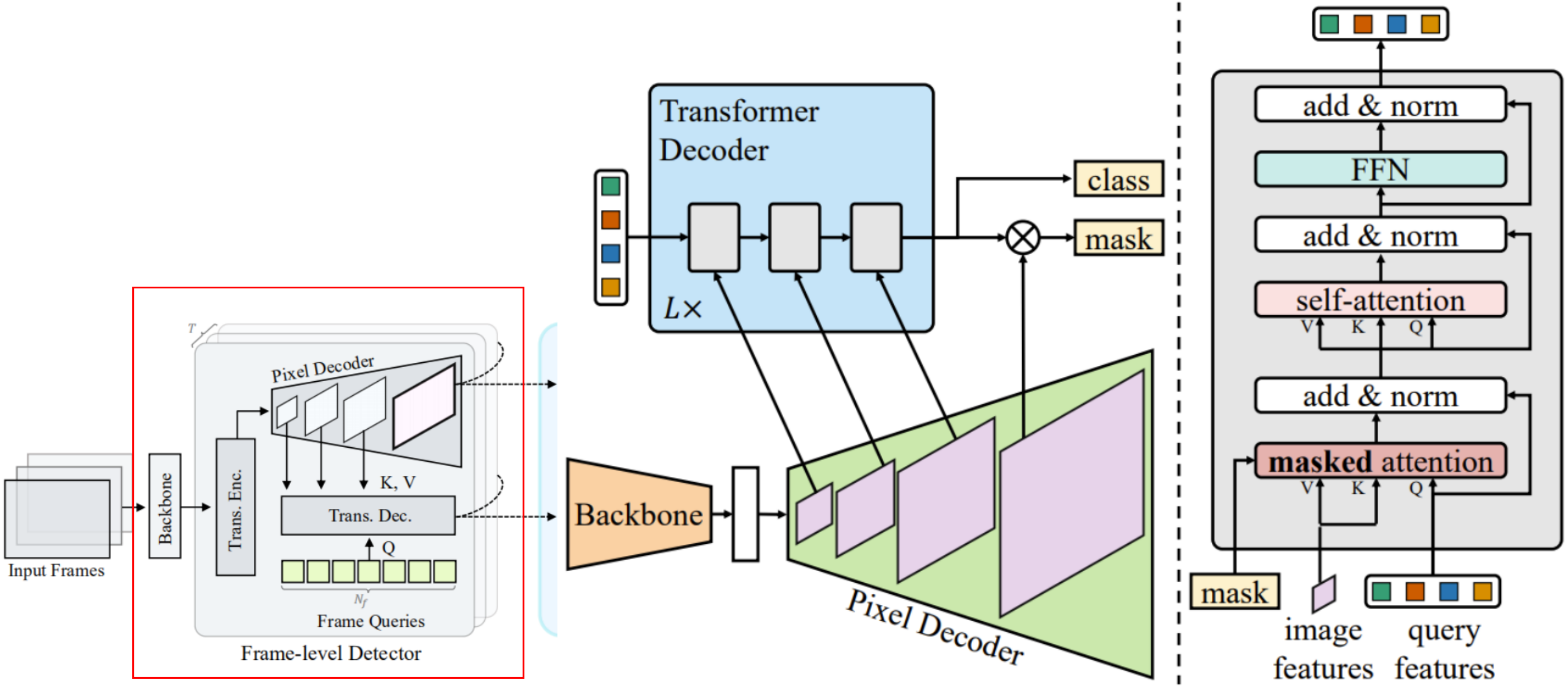






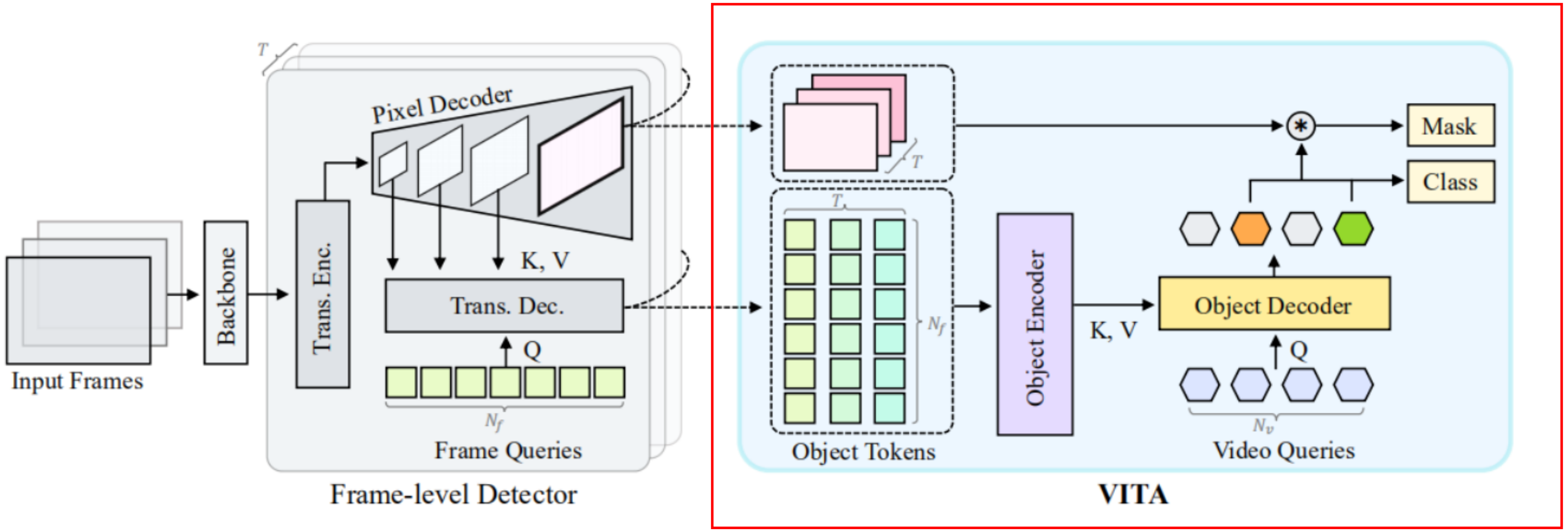
01 VITA: Video Instance Segmentation via Object Token Association

- **Frame-level Detector**
  - frame-independent manner; no inter-computation between frames is involved



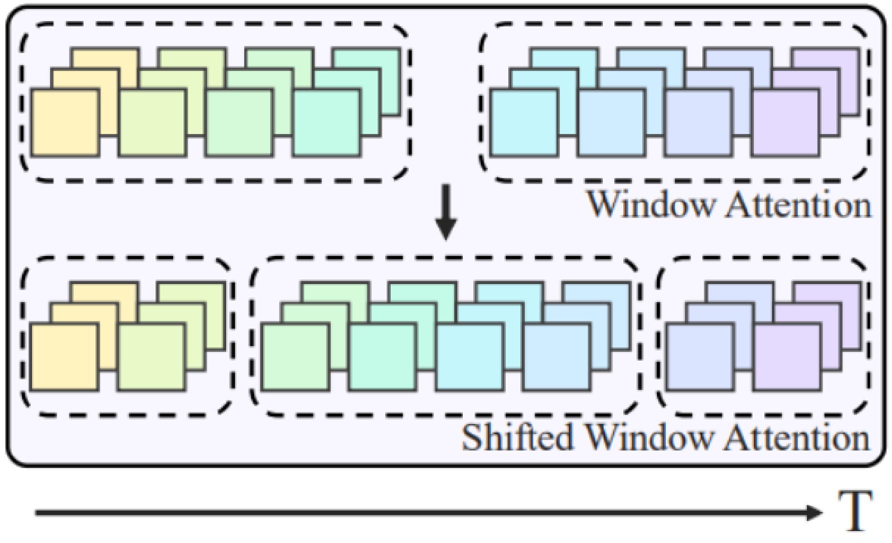
01 VITA: Video Instance Segmentation via Object Token Association

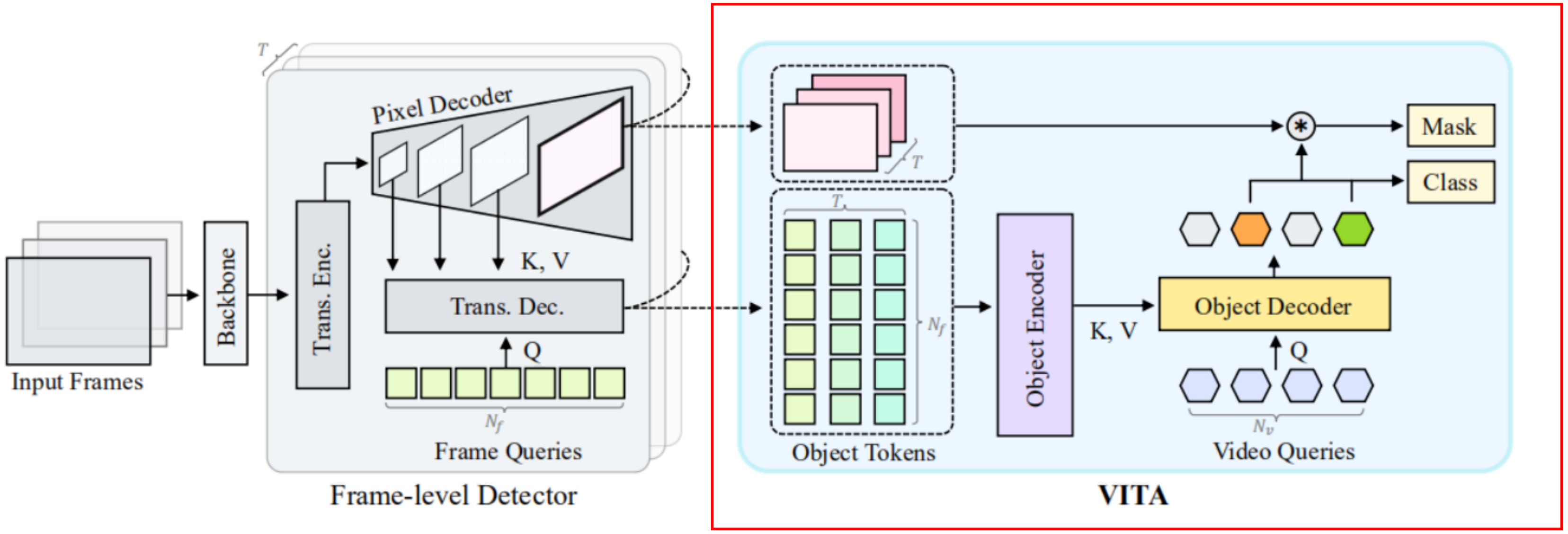
- **Object Encoder**
  - Build temporal communication by employing self-attention along the temporal axis
- **Object Decoder and Output heads**
  - **Q:** N_v learnable queries
  - **K, V:** object tokens



01 VITA: Video Instance Segmentation via Object Token Association

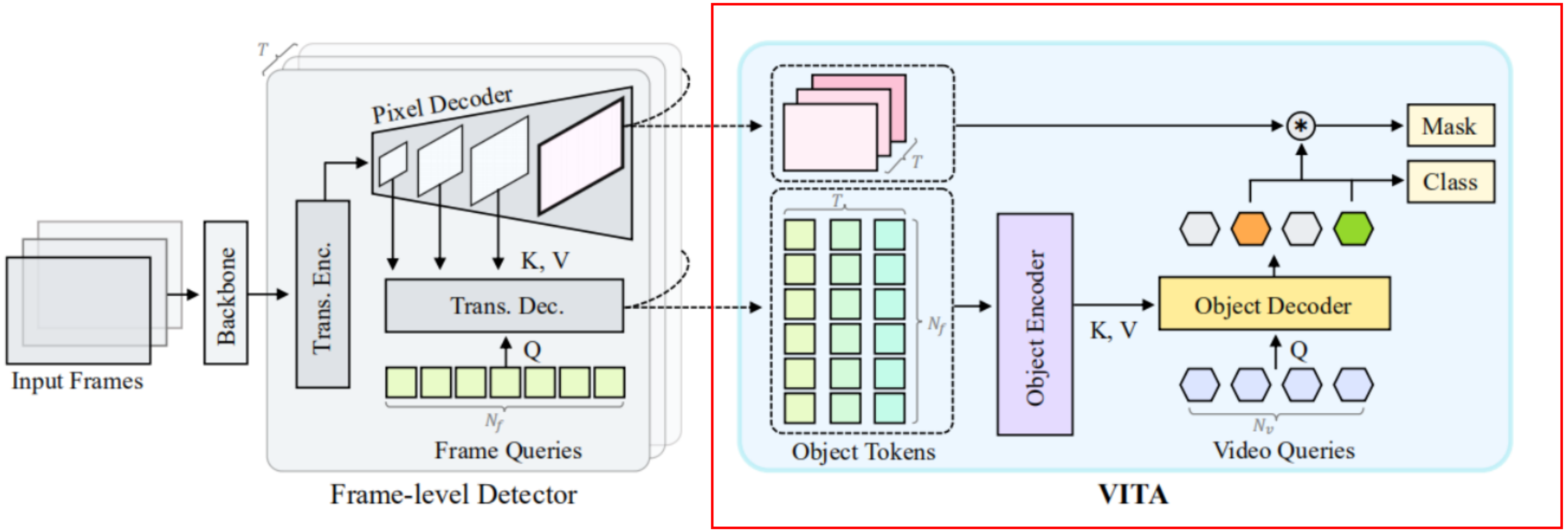
- **Object Encoder**
  - Build temporal communication by employing self-attention along the temporal axis
- **Object Decoder and Output heads**
  - **Q:** N_v learnable queries
  - **K, V:** object tokens

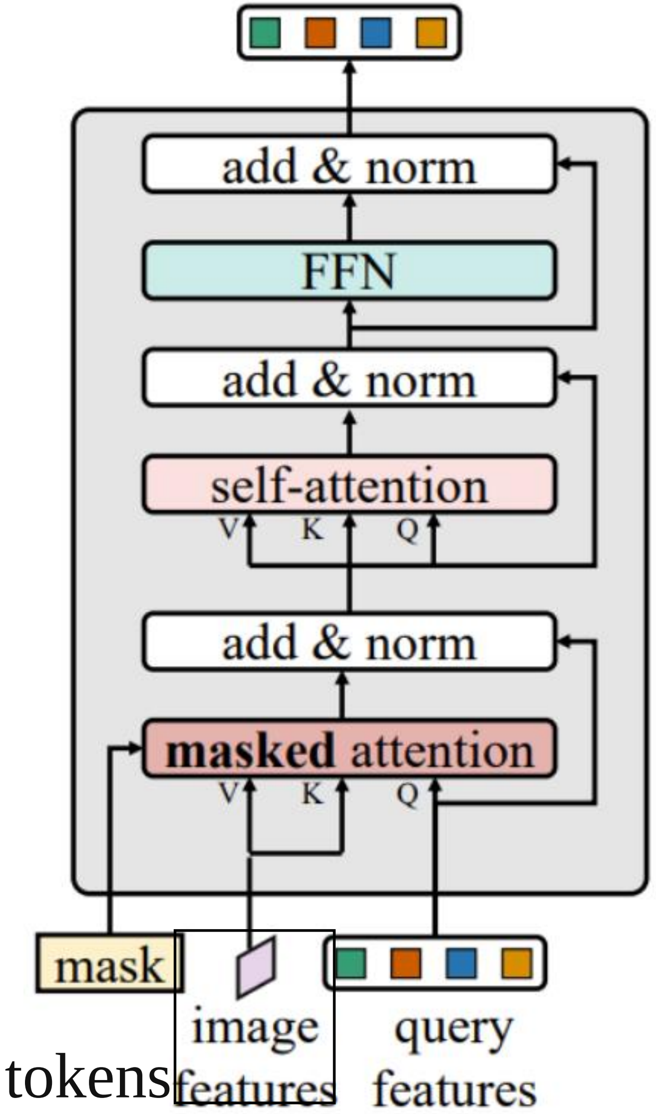




01 VITA: Video Instance Segmentation via Object Token Association

- **Object Encoder**
  - Build temporal communication by employing self-attention along the temporal axis
- **Object Decoder and Output heads**
  - **Q:** N_v learnable queries
  - **K, V:** object tokens





01 VITA: Video Instance Segmentation via Object Token Association

- **Similarity loss**

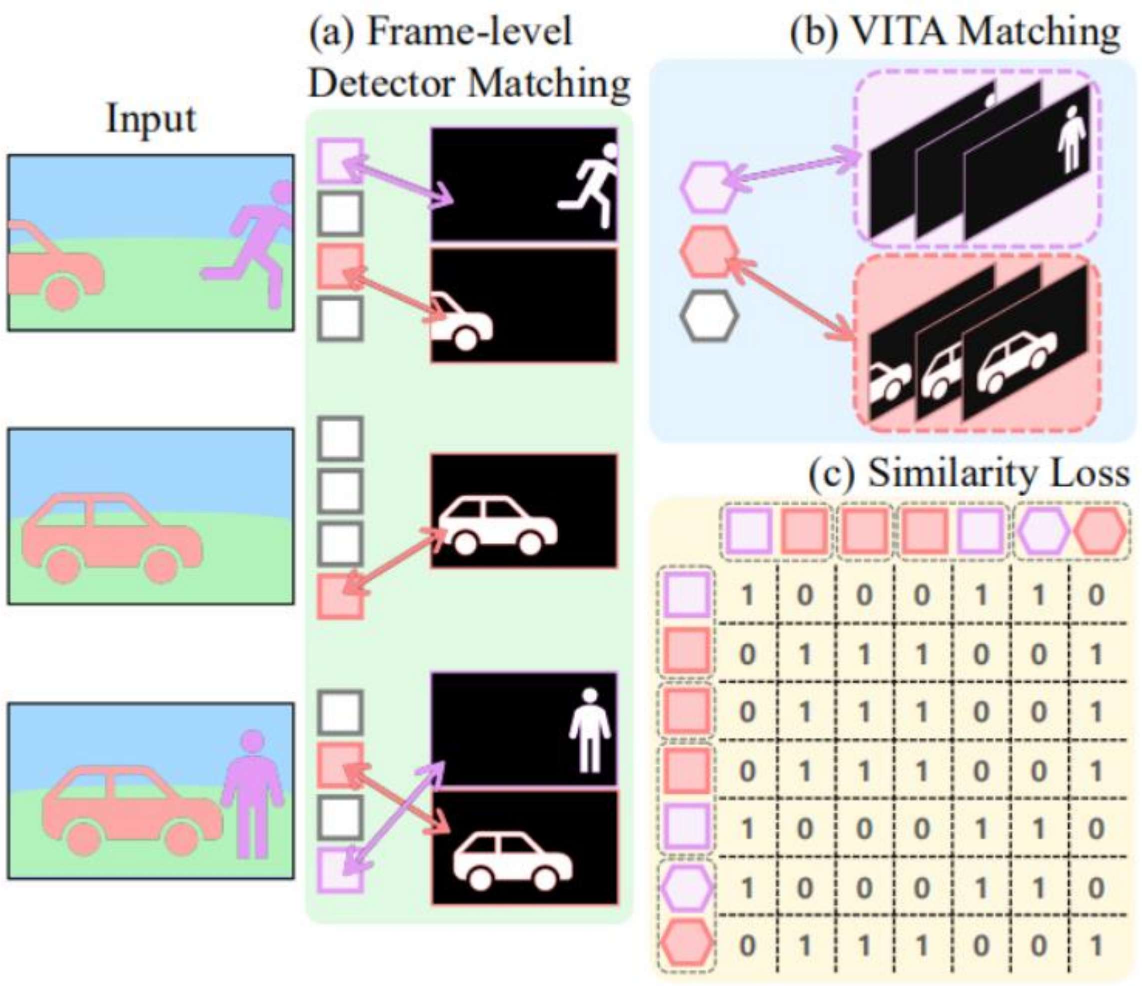


Figure 4: Similarity loss. ⬡ and □ indicate video query and frame query, respectively. Same color represents same GT instance ID.

- Embed the collection through a linear layer.
- Measure the similarity of all possible pairs using a simple matrix multiplication.
- Binary cross entropy

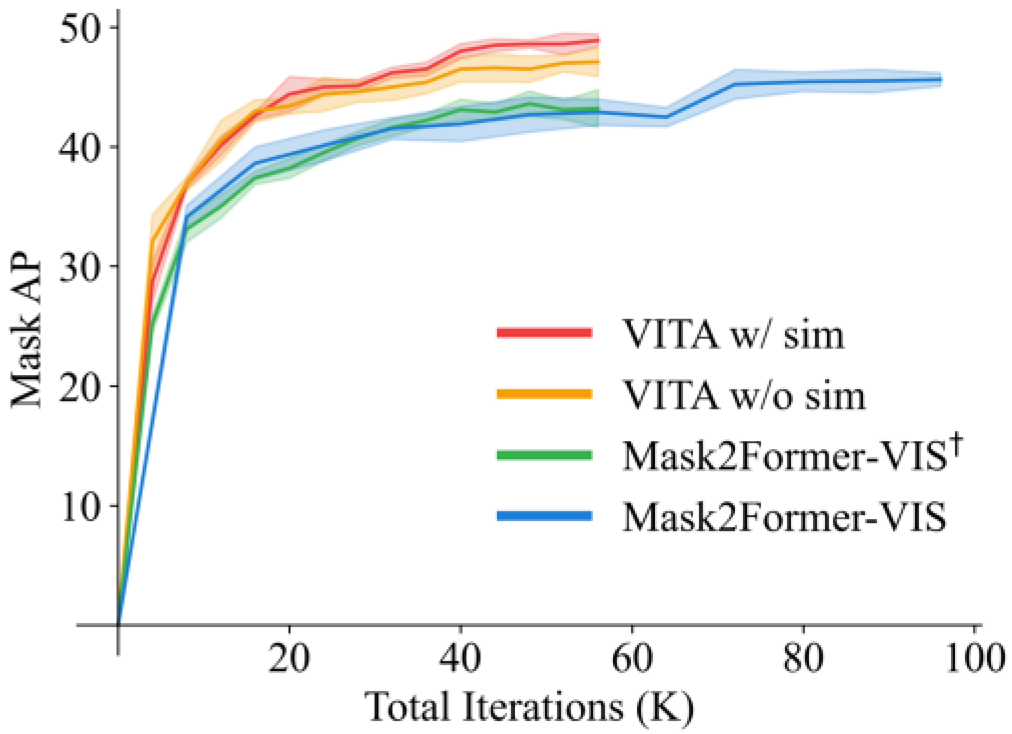


Table 1: Comparisons on YouTube-VIS 2019.

	Method	Backbone [13]	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
(Near) Online	MaskTrack R-CNN [31]	ResNet-50	30.3	51.1	32.6	31.0	35.5
	MaskTrack R-CNN [31]	ResNet-101	31.8	53.0	33.6	33.2	37.6
	CrossVIS [32]	ResNet-50	36.3	56.8	38.9	35.6	40.7
	CrossVIS [32]	ResNet-101	36.6	57.3	39.7	36.0	42.0
	PCAN [16]	ResNet-50	36.1	54.9	39.4	36.3	41.6
	PCAN [16]	ResNet-101	37.6	57.2	41.3	37.2	43.9
	EfficientVIS [28]	ResNet-50	37.9	59.7	43.0	40.3	46.6
	EfficientVIS [28]	ResNet-101	39.8	61.8	44.7	42.1	49.8
	VISOLO [11]	ResNet-50	38.6	56.3	43.7	35.7	42.5
Offline	VisTR [27]	ResNet-50	35.6	56.8	37.0	35.2	40.2
	VisTR [27]	ResNet-101	38.6	61.3	42.3	37.6	44.2
	IFC [14]	ResNet-50	41.2	65.1	44.6	42.3	49.6
	IFC [14]	ResNet-101	42.6	66.6	46.3	43.5	51.4
	TeViT [33]	MsgShifT	46.6	71.3	51.6	44.9	54.3
	SeqFormer [29]	ResNet-50	47.4	69.8	51.8	45.5	54.8
	SeqFormer [29]	ResNet-101	49.0	71.1	55.7	46.8	56.9
	SeqFormer [29]	Swin-L	59.3	82.1	66.4	51.7	64.4
	Mask2Former-VIS [6]	ResNet-50	46.4	68.0	50.0	-	-
	Mask2Former-VIS [6]	ResNet-101	49.2	72.8	54.2	-	-
	Mask2Former-VIS [6]	Swin-L	60.4	84.4	67.0	-	-
	VITA (Ours)	ResNet-50	49.8	72.6	54.5	49.4	61.0
	VITA (Ours)	ResNet-101	51.9	75.4	57.0	49.6	59.1
	VITA (Ours)	Swin-L	63.0	86.9	67.9	56.3	68.1

the tendency of offline methods with higher accuracy

Table 2: Comparisons with ResNet-50 backbone on YouTube-VIS 2021 and OVIS. † indicates using MsgShifT backbone.

Method	YouTube-VIS 2021					OVIS				
	AP	AP_{50}	AP_{75}	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
MaskTrack R-CNN [31]	28.6	48.9	29.6	26.5	33.8	10.8	25.3	8.5	7.9	14.9
CMaskTrack R-CNN [22]	-	-	-	-	-	15.4	33.9	13.1	9.3	20.0
STMask [18]	31.1	50.4	33.5	26.9	35.6	15.4	33.8	12.5	8.9	21.3
CrossVIS [32]	34.2	54.4	37.9	30.4	38.2	14.9	32.7	12.1	10.3	19.8
IFC [14]	35.2	55.9	37.7	32.6	42.9	-	-	-	-	-
VISOLO [11]	36.9	54.7	40.2	30.6	40.9	15.3	31.0	13.8	11.1	21.7
TeViT† [33]	37.9	61.2	42.1	35.1	44.6	17.4	34.9	15.0	11.2	21.8
SeqFormer [29]	40.5	62.4	43.7	36.1	48.1	-	-	-	-	-
Mask2Former-VIS [6]	40.6	60.9	41.8	-	-	-	-	-	-	-
VITA (Ours)	**45.7**	**67.4**	**49.5**	**40.9**	**53.6**	**19.6**	**41.2**	**17.4**	**11.7**	**26.0**

YouYube VIS2021: We hypothesize that the object-oriented design of VITA is more effective than typical dense Transformer decoders in addressing such challenging scenes.

OVIS: VITA is the first complete-offline approach to evaluate on OVIS valid set. (maximum 292 frames)

Table 3: Impact of local windows of varying sizes in Object Encoder.

W	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
3	49.4	72.2	54.4	48.6	60.9
6	49.8	72.6	54.5	49.4	61.0
12	50.0	73.0	54.7	49.0	60.8
All	50.1	72.4	54.7	49.0	60.6

Table 4: Maximum number of frames that can be processed at once using a single Titan XP.

Method		Max Frames 360×640	Max Frames 720×1280
VisTR [27]		46	12
IFC [14]		123	38
Mask2Former-VIS [6]		81	20
VITA (Ours)	$W = 3$	2677	
	$W = 6$	1392	
	$W = 12$	741	