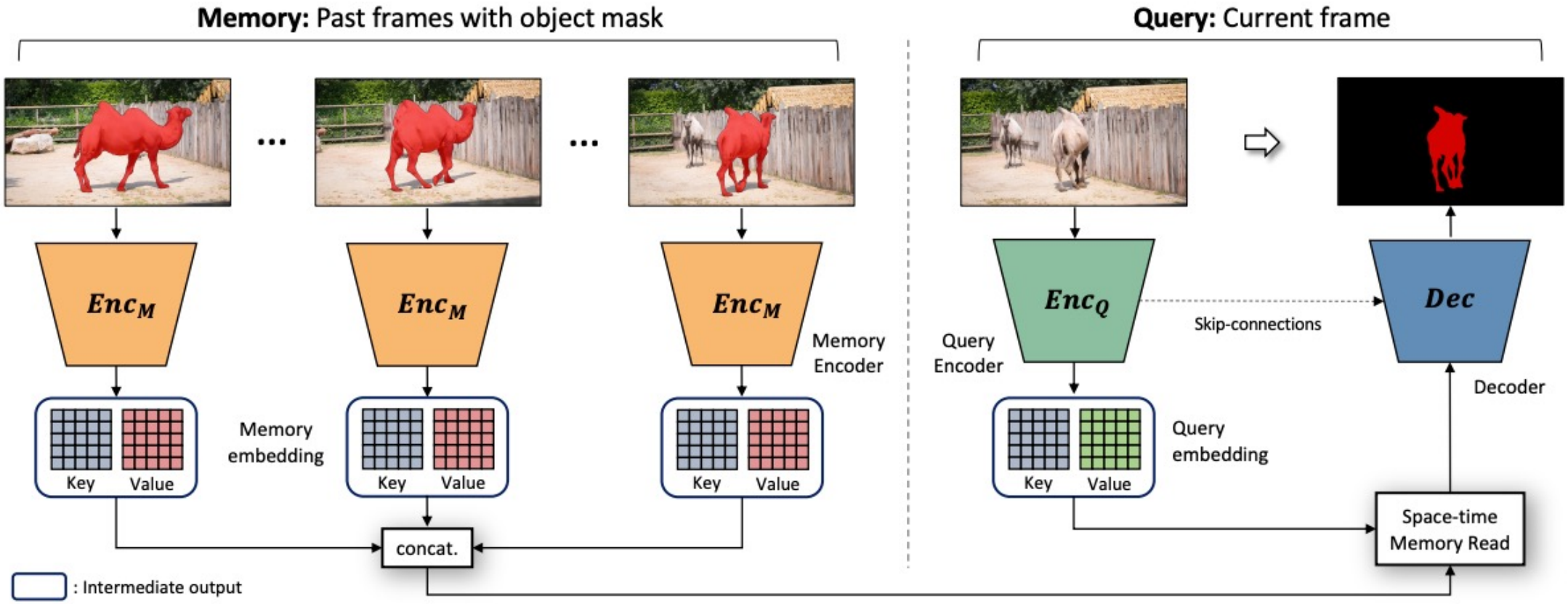
Look Before You Match: Instance Understanding Matters in Video Object Segmentation

Junke Wang[1,2], Dongdong Chen[3], Zuxuan Wu[1,2†], Chong Luo[4], Chuanxin Tang[4], Xiyang Dai[3], Yucheng Zhao[4], Yujia Xie[3], Lu Yuan[3], Yu-Gang Jiang[1,2†]

[1]Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2]Shanghai Collaborative Innovation Center on Intelligent Visual Computing
[3]Microsoft Cloud + AI, [4]Microsoft Research Asia.
([†] denotes corresponding authors)

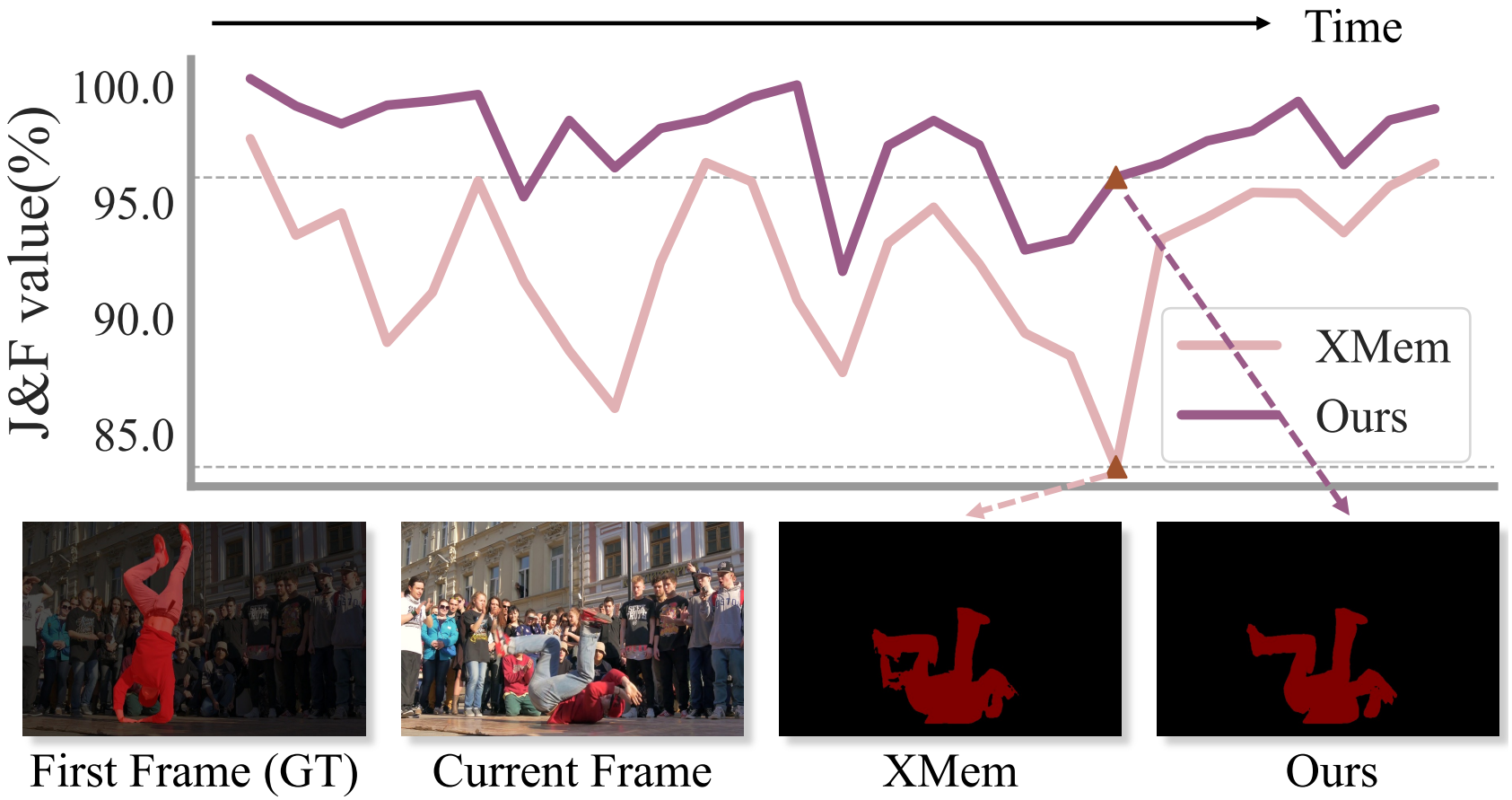
Tag: TUE-AM-135

Memory-based Video Object Segmentation



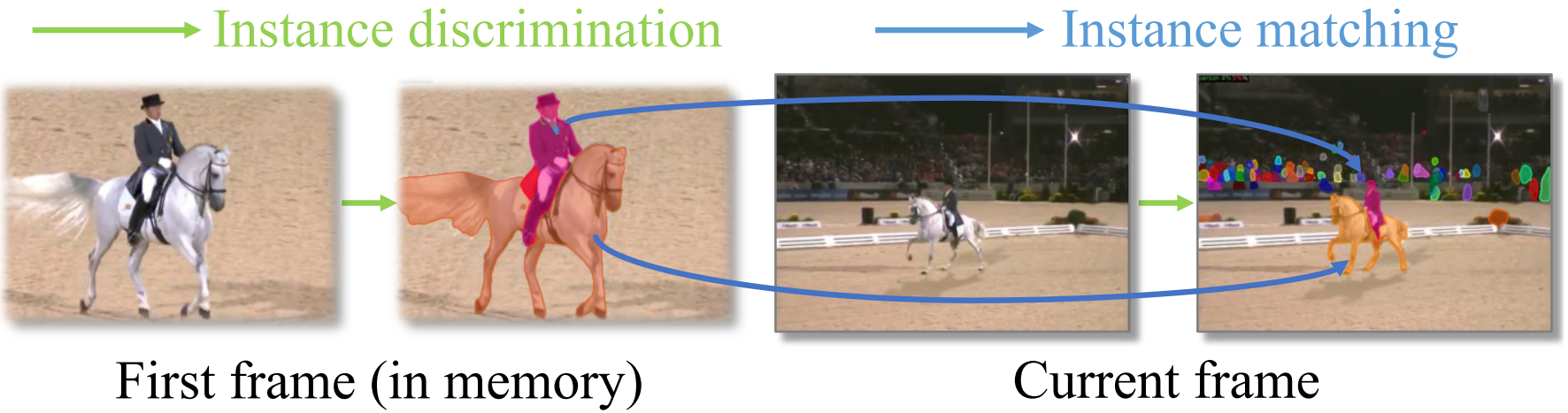
A feature memory is maintained to store the past frames match with the current frame.

Drawbacks of Memory-based VOS



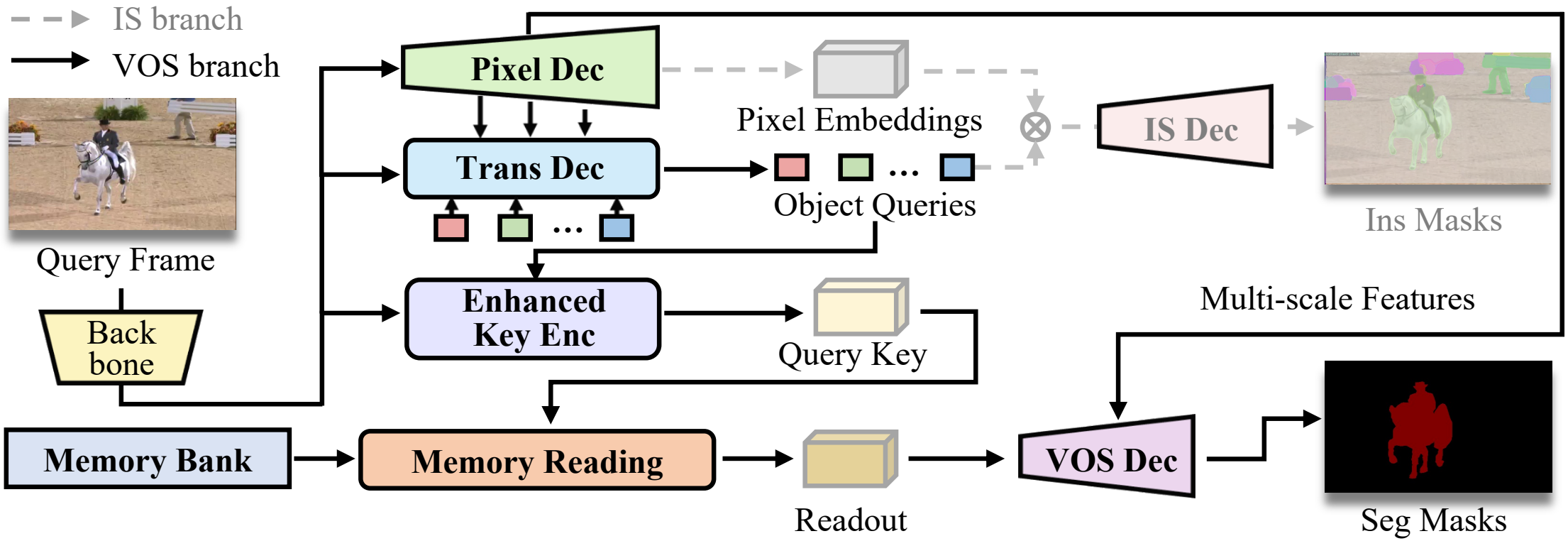
The dense matching is vulnerable to the appearance variations and object deformation.

The way Humans handle VOS



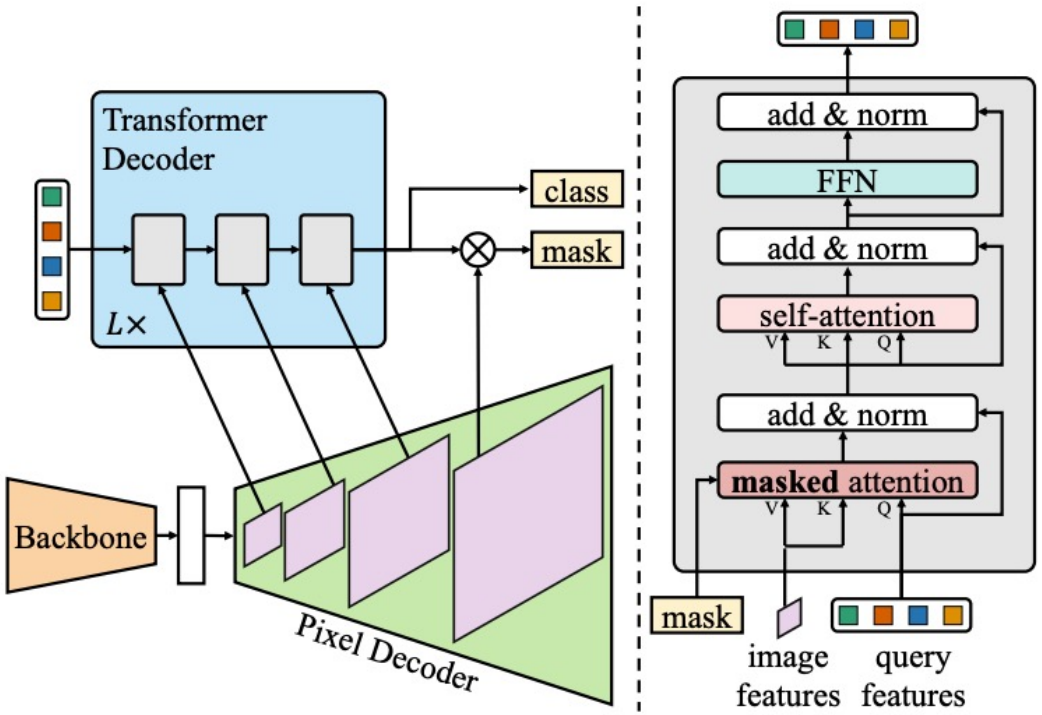
In the absence instance understanding, pure matching is difficult to generate accurate predictions for regions that are invisible in reference frame by pure matching.

Architecture of ISVOS



A two-branch network consisting of an **instance segmentation** and a **VOS branch** is presented.

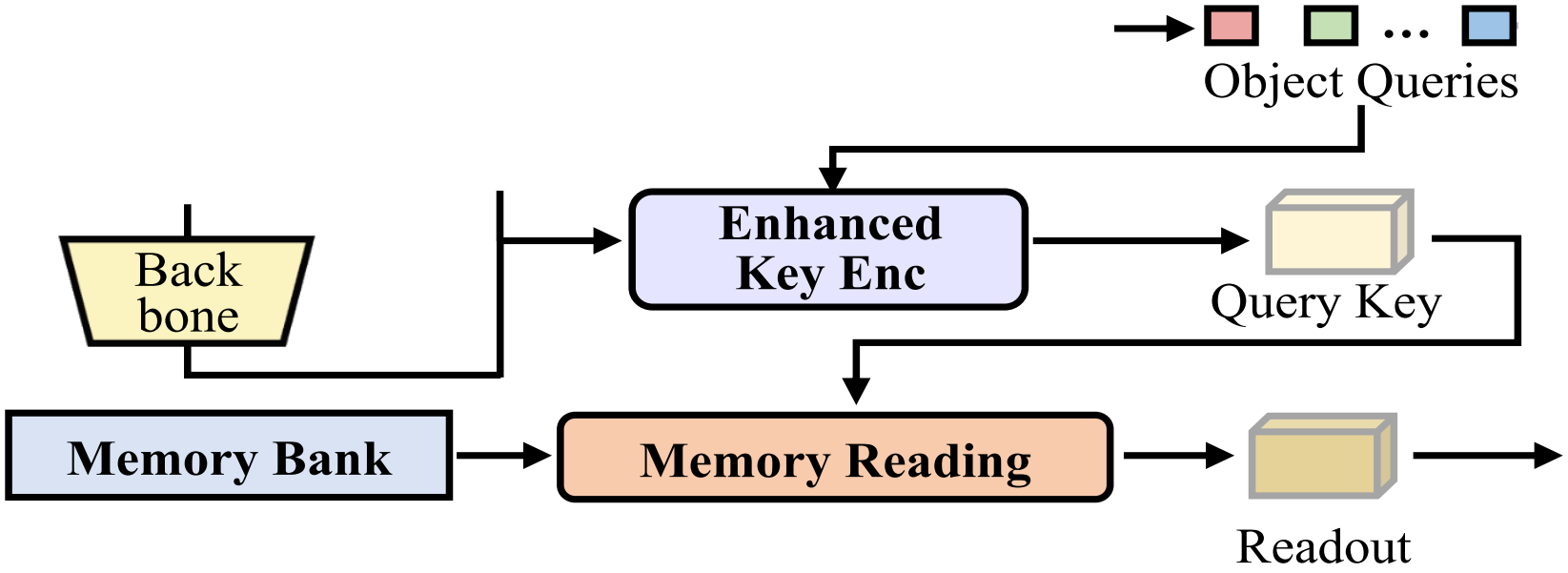
Instance Segmentation Branch



- **Pixel Decoder** takes backbone features as input and generates a high-resolution per-pixel embedding, as well as a feature pyramid.
- **Transformer Decoder** gathers the local information in the feature pyramid to a set of learnable object queries.

The IS branch is built upon a query-based instance segmentation model Mask2Former.

Video Object Segmentation Branch



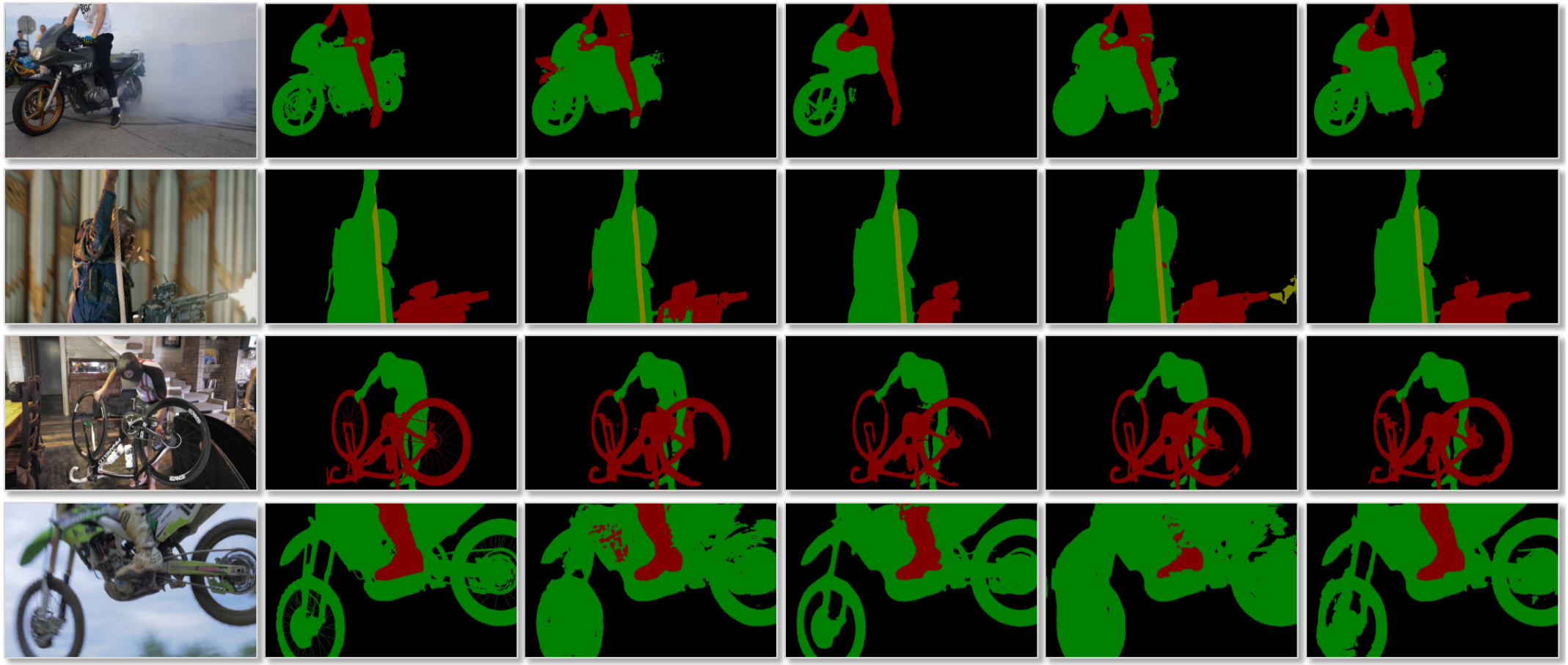
- **Enhanced Key Encoder** enhances the backbone feature with the updated object queries to generate the query key of current frame.
- **Memory Reading** first measures the similarity between the query key and the memory key, and then calculates the weighted summation between affinity matrix and memory value to obtain the readout features.

Experimental Results

Method	w/ BL30K	DAVIS16 validation			DAVIS17 validation			YT2018 validation				
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{G}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$
STM [46]	✗	89.3	88.7	89.9	81.8	79.2	84.3	79.4	79.7	84.2	72.8	80.9
HMMN [54]	✗	90.8	89.6	92.0	84.7	81.9	87.5	82.6	82.1	87.0	76.8	84.6
RPCM [65]	✗	90.6	87.1	91.1	83.7	81.3	86.0	84.0	83.1	87.7	78.5	86.7
STCN [15]	✗	91.6	90.8	92.5	85.4	82.2	88.6	83.0	81.9	86.5	77.9	85.7
AOT [70]	✗	91.1	90.1	92.1	84.9	82.3	87.5	85.5	84.5	89.5	79.6	88.2
RDE [30]	✗	91.1	89.7	92.5	84.2	80.8	87.5	-	-	-	-	-
XMem [13]	✗	91.5	90.4	92.7	86.2	82.9	89.5	85.7	84.6	89.3	80.2	88.7
DeAOT [72]	✗	92.3	90.5	94.0	85.2	82.2	88.2	86.0	84.9	89.9	80.4	88.7
Ours	✗	**92.6**	**91.5**	**93.7**	**87.1**	**83.7**	**90.5**	**86.3**	**85.5**	**90.2**	**80.5**	**88.8**
MiVOS [14]	✓	91.0	89.6	92.4	84.5	81.7	87.4	82.6	81.1	85.6	77.7	86.2
STCN [15]	✓	91.7	90.4	93.0	85.3	82.0	88.6	84.3	83.2	87.9	79.0	87.3
RDE [30]	✓	91.6	90.0	93.2	86.1	82.1	90.0	-	-	-	-	-
XMem [13]	✓	92.0	90.7	93.2	87.7	84.0	91.4	86.1	85.1	89.8	80.3	**89.2**
Ours	✓	**92.8**	**91.8**	**93.8**	**88.2**	**84.5**	**91.9**	**86.7**	**86.1**	**90.8**	**81.0**	**89.0**

ISVOS achieves top-ranked performance on both single- and multi-object VOS benchmarks.

Visualizations



Take-home Messages

- This paper incorporates **instance understanding** for improved VOS through a two-branch network: an **IS branch** derives instance-aware representations and an **VOS branch** maintains a memory bank for spatial-temporal matching.
- We enhance the query key with the well-learned **object queries** from IS branch to inject the instance-specific information, with which the **instance-augmented matching** with memory bank is performed.
- In the future, ISVOS can be equipped with **efficient memory storage** to develop both accurate and efficient VOS models.

Thanks