2025.03.18

Presenter

Best Vision Paper

고재훈

IS IMAGENET WORTH 1 VIDEO?
LEARNING STRONG IMAGE ENCODERS
FROM 1 LONG UNLABELLED VIDEO

ICLR 2024, Honorable Mention

Table of Contents

I Introduction

II Dataset – Walking Tours (WTours)

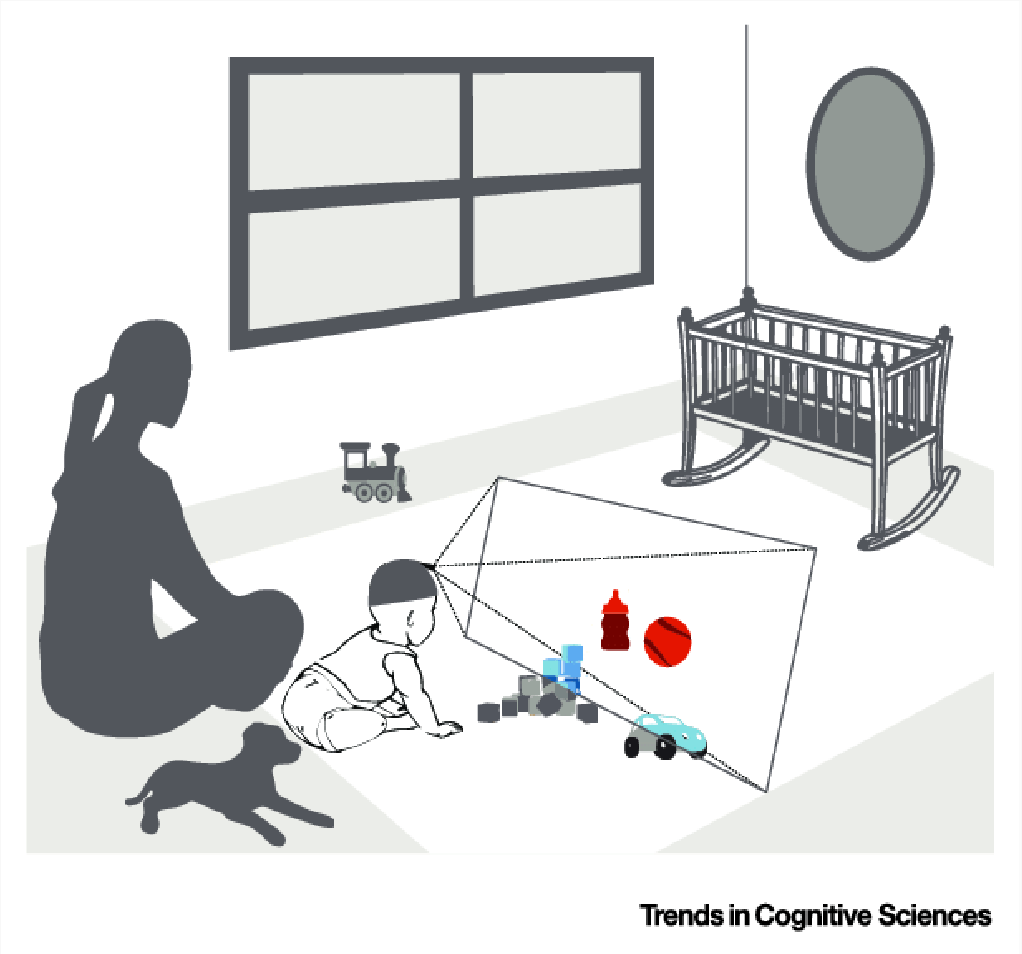
III Method – Discover and tRAck (DoRA)

IV Experiments

V Conclusion

Introduction

Motivation



Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). **"The Developing Infant Creates a Curriculum for Statistical Learning".** *Trends in Cognitive Sciences*

1. **Self-supervised learning (SSL)** enables large-scale pretraining without labeled data but still **relies heavily on static images.**
2. In contrast, **humans** develop visual understanding much faster by **continuously perceiving** their surroundings.
3. Existing video-based SSL struggles as it mainly uses object-centric internet videos.

Contribution

1. **WT(Walking Tours)** dataset is introduced as a first-person video dataset that closely mirrors human visual perception.

2. **DORA**, a novel SSL approach, is proposed to learn by tracking objects over time, inspired by human development.

3. It achieves **ImageNet-level performance** using only a single WT video, outperforming traditional image-based pretraining in segmentation and detection tasks.

II Walking Tours (WTours)

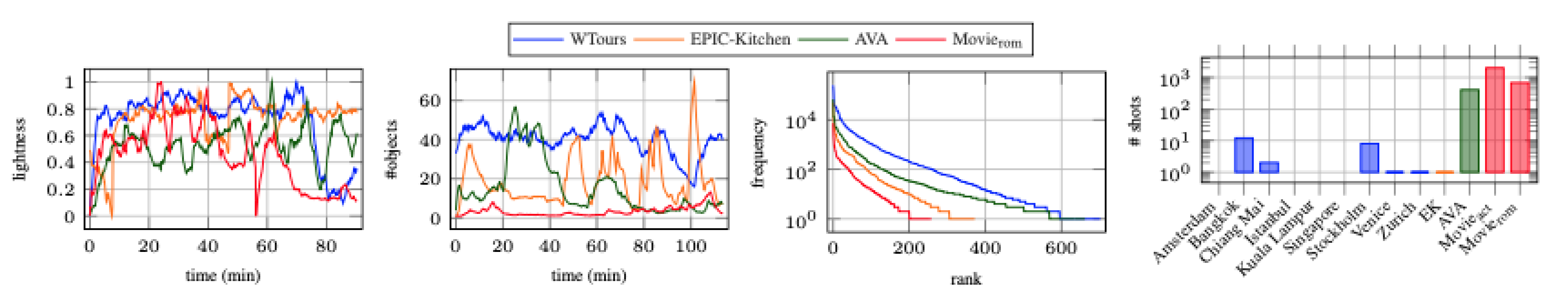
Comparison with other video datasets

Dataset	Domain	Ego	Pre	Bal	Annot	Avg. Dur (sec)	Dur (hr)	#Videos	Frame resolution
Diverse Pretraining									
Kinetics-400 (Kay et al., 2017)	Actions	✗	✓	✓	Class	10.2	851	400	340 × 255
WebVid-2M (Bain et al., 2021)	Open	✗	✓	✗	Weak	18	13k	–	320 × 240
HowTo100M (Miech et al., 2019)	Instructions	✗	✓	✗	Weak	4	135k	–	–
Egocentric									
Epic-Kitchens (Damen et al., 2022)	Cooking	✓	✗	✗	Loc.	510	100	37	1920 × 1080
Ego-4D (Grauman et al., 2022)	Daily	✓	✗	✗	Loc.	1446	120	931	1920 × 1080
Meccano (Ragusa et al., 2023)	Industry	✓	✗	✗	Loc.	1247	849	20	1920 × 1080
Assembly-101 (Sener et al., 2022)	Assembly	✓	✗	✗	Loc.	426	167	362	1920 × 1080
ImageNet-aligned									
R2V2 (Gordon et al., 2020)	ImageNet	✗	✓	✓	Class	–	–	–	467 × 280
VideoNet (Parthasarathy et al., 2022)	ImageNet	✗	✓	✓	Class	10	3055	–	
Walking Tours (ours)	Urban	✓	✓	✗	None	5880	23	10	3840 × 2160

- Compared to existing video datasets, this dataset contains **longer, higher-resolution, and more continuous** videos.
- Unlike curated datasets with controlled class balance, it is **open-ended, scalable, and does not require manual labeling.**
- **A rich variety of objects per frame** makes it well-suited for representation learning.

II Walking Tours (WTours)

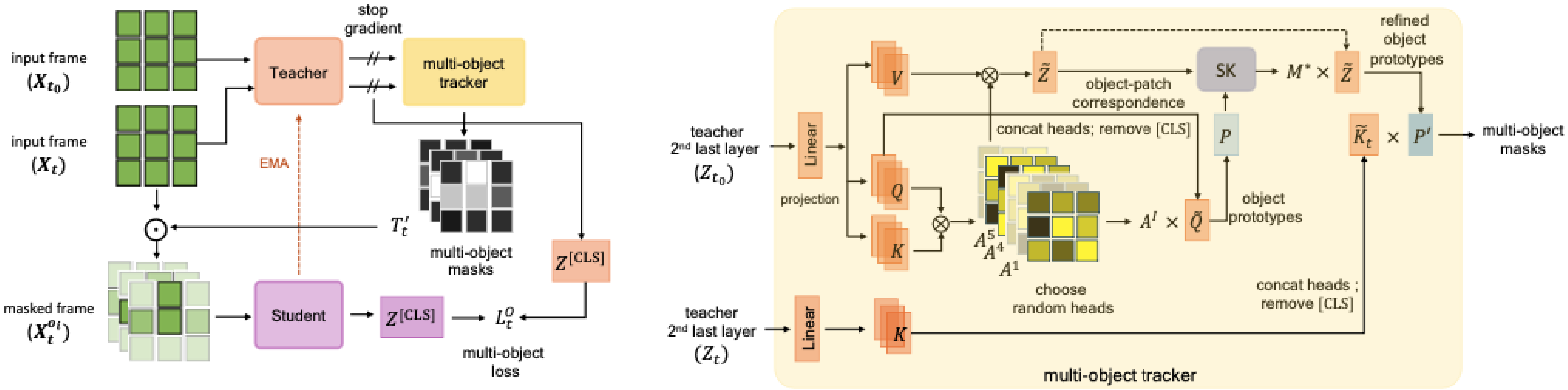
Dataset Analysis



(a) Lightness (b) #Objects per frame (c) Total #classes (d) Shots

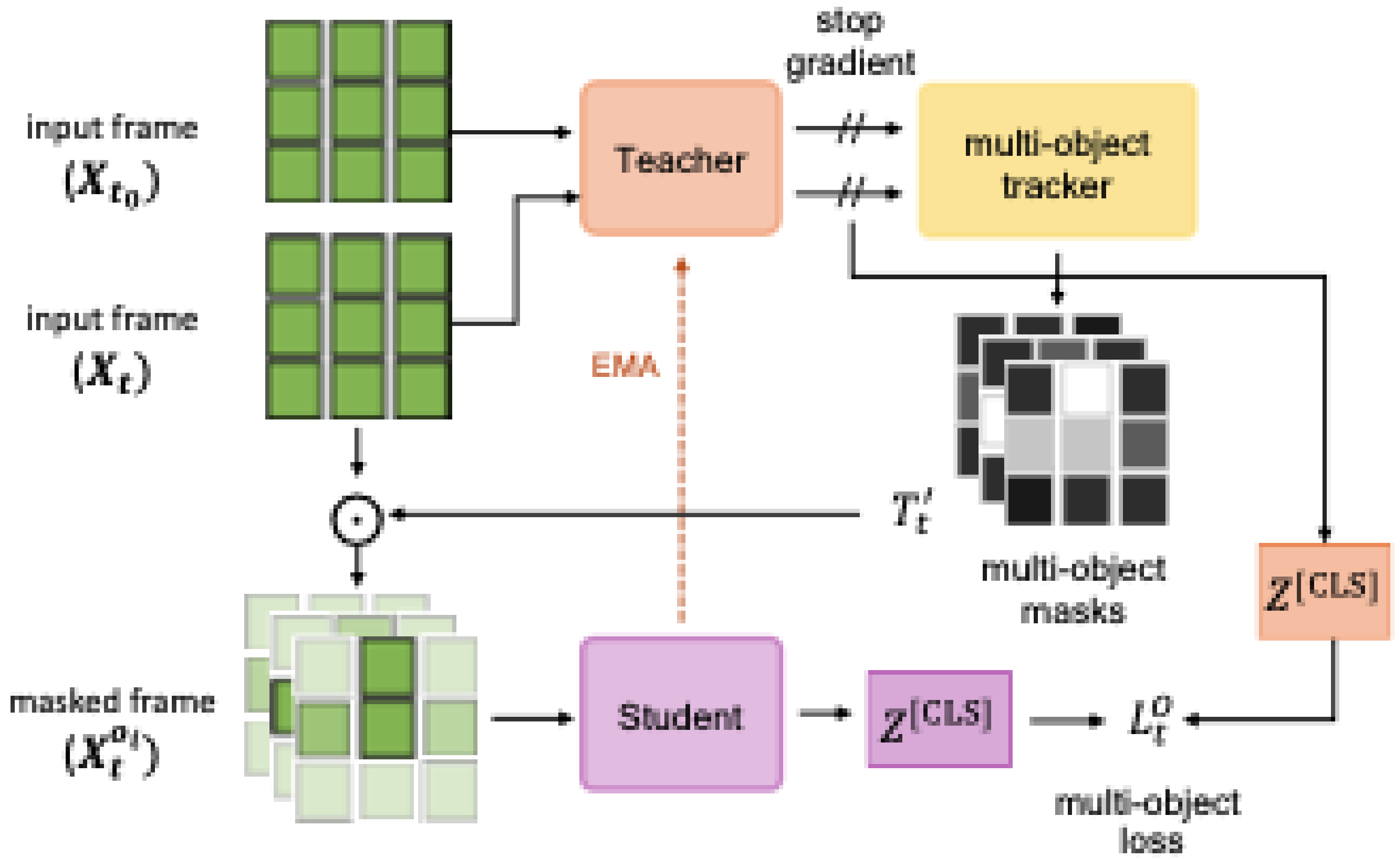
- **Gradual brightness transitions** occur, unlike other datasets that exhibit random fluctuations.
- A **higher number of unique object classes per frame** contributes to greater semantic richness.
- **Minimal shot transitions** provide a more continuous learning experience compared to movie datasets with frequent scene cuts.

Overview



input frame
(X_{t_0})
input frame
(X_t)
masked frame
($X_t^{o_i}$)
Teacher
stop
gradient
multi-object
tracker
EMA
T'_t
multi-object
masks
$Z^{[CLS]}$
Student
$Z^{[CLS]}$
L_t^O
multi-object
loss
teacher
2nd last layer
(Z_{t_0})
Linear
projection
V
Q
K
$\tilde{Z}$
object-patch
correspondence
SK
$M^* \times \tilde{Z}$
refined
object
prototypes
concat heads; remove [CLS]
P
A^5
A^4
A^1
$A^I \times \tilde{Q}$
object
prototypes
choose
random heads
$\tilde{K}_t \times P'$
multi-object
masks
concat heads ;
remove [CLS]
teacher
2nd last layer
(Z_t)
Linear
K
multi-object tracker

Preliminaries

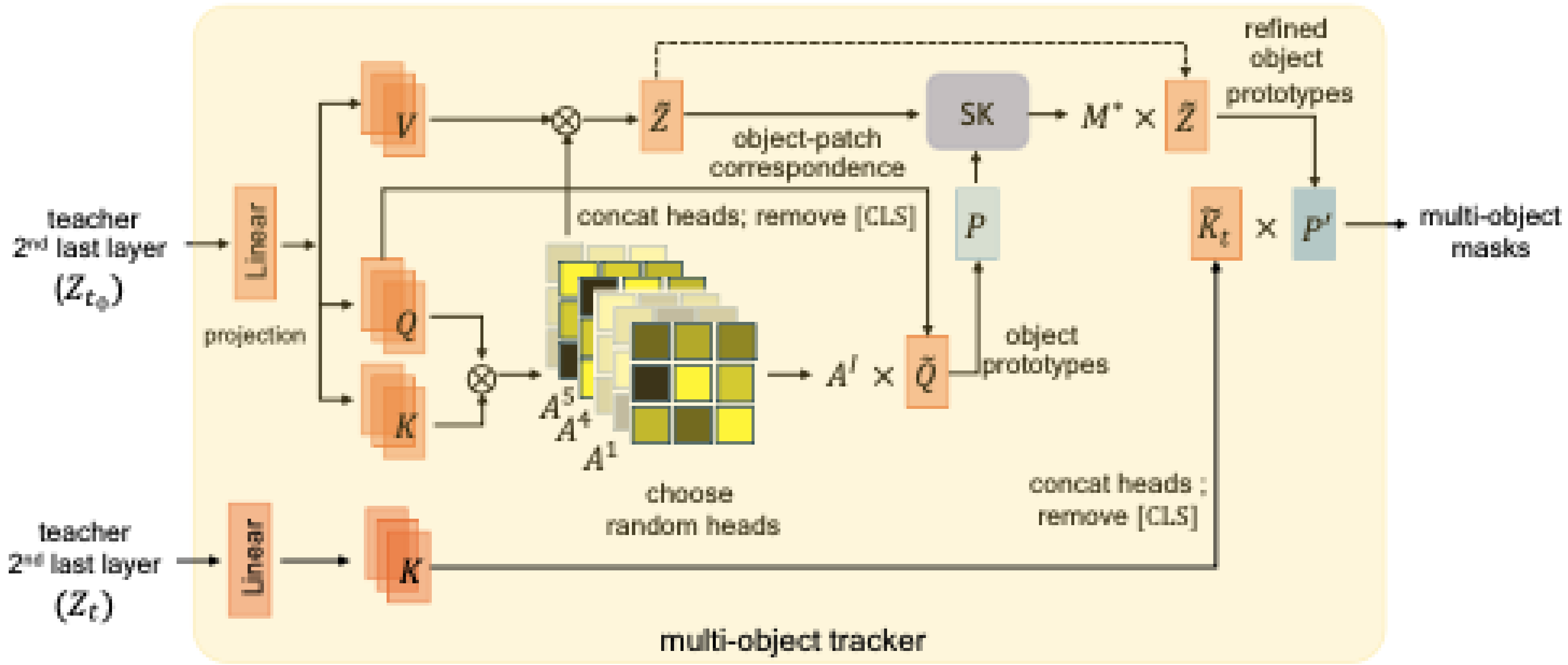


Video Clip $\mathbf{X}_t \in \mathbb{R}^{h \times w \times c}$ for $t \in \{1, \ldots, T\}$

# of patches $n = hw/p^2$

EMA $\theta' \leftarrow \alpha\theta' + (1-\alpha)\theta$

Preliminaries



Output Embeddings $Z_t = g_\theta(\mathbf{X}_t) \in \mathbb{R}^{(n+1)\times d} = [Z^{[\mathrm{CLS}]}; \tilde{Z}]$

[CLS] Embeddings $Z^{[\mathrm{CLS}]} \in \mathbb{R}^{1\times d}$

Patch Embeddings $\tilde{Z} \in \mathbb{R}^{n\times d}$

Discovering objects with multi-head attention

Query & Key Embeddings for MHA

$$Q, K \in \mathbb{R}^{(n+1)\times d}$$

$$Q^i, K^i \in \mathbb{R}^{(n+1)\times d/h} \text{ for } i = 1, \ldots, h$$

Self-Attention Matrix

$$A^i := \operatorname{softmax}\left(Q^i (K^i)^\top / \sqrt{d}\right) \in \mathbb{R}^{(n+1)\times(n+1)}$$

[CLS] – Attention Vector

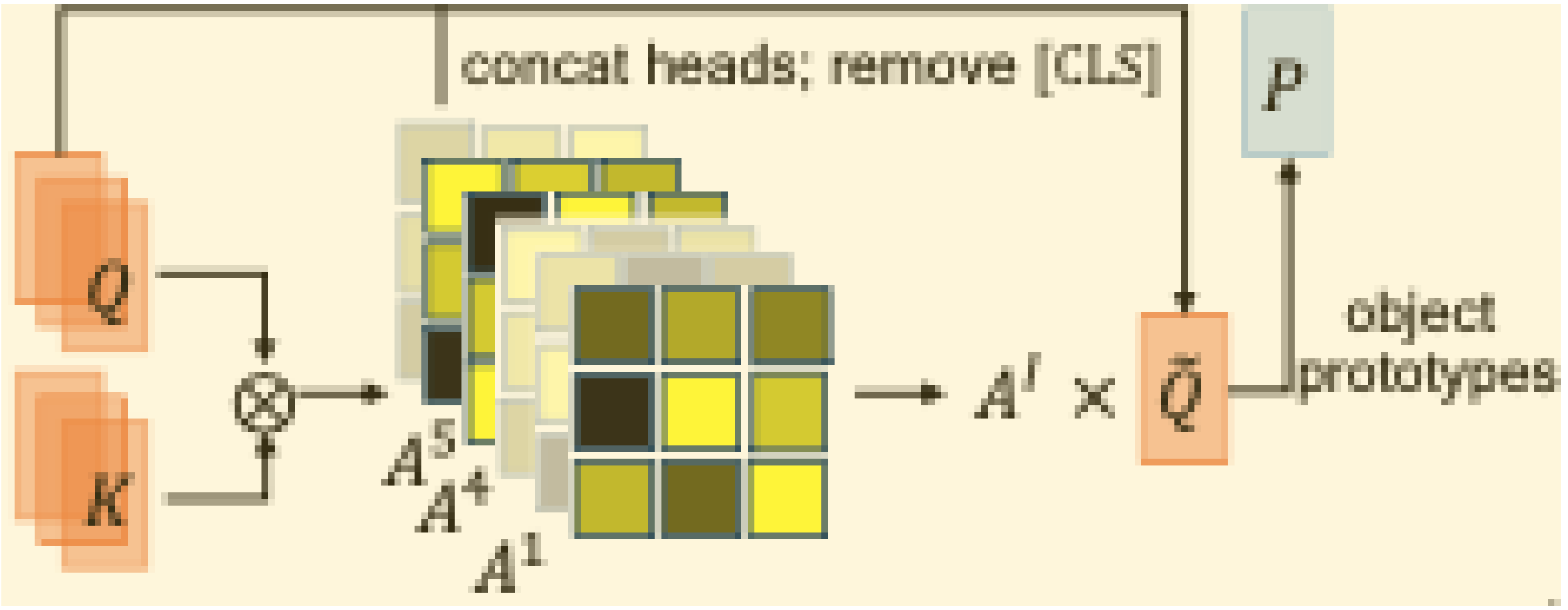
$$A^{[\text{CLS}]} := [a_{1,2}, \ldots, a_{1,n}] \in \mathbb{R}^{1\times n}$$

Random Subset of Attention Vectors

$$A^{\mathcal{I}} := [(A^{i_1})^{[\text{CLS}]}; \ldots; (A^{i_k})^{[\text{CLS}]}] \in \mathbb{R}^{k\times n}$$

Object Prototypes

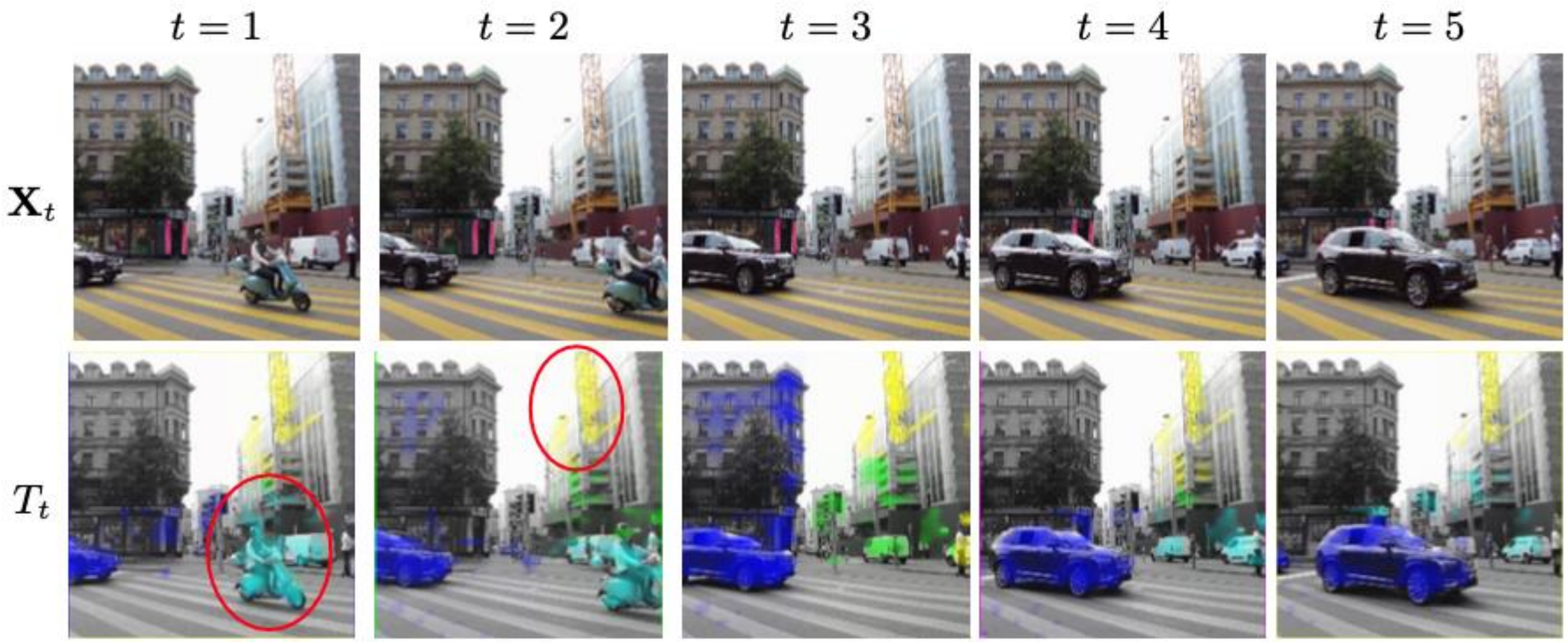
$$P := A^{\mathcal{I}} \tilde{Q} \in \mathbb{R}^{k\times d}$$



$$T_t := \operatorname{softmax}\left(P \tilde{K}_t^\top / \sqrt{d}\right) \in \mathbb{R}^{k\times n}$$

k – attention maps….?

Discovering objects with multi-head attention



Overlapping object regions detected!

Establishing object-patch correspondences

Object Prototypes

$$P := A^{\mathcal{I}}\tilde{Q} \in \mathbb{R}^{k\times d}$$

Output of Teacher Network

$$Z = g_{\theta'}(\mathbf{X}_{t_0}) \in \mathbb{R}^{(n+1)\times d}$$

Transport Matrix

$$M \in \mathbb{R}^{k\times n}$$

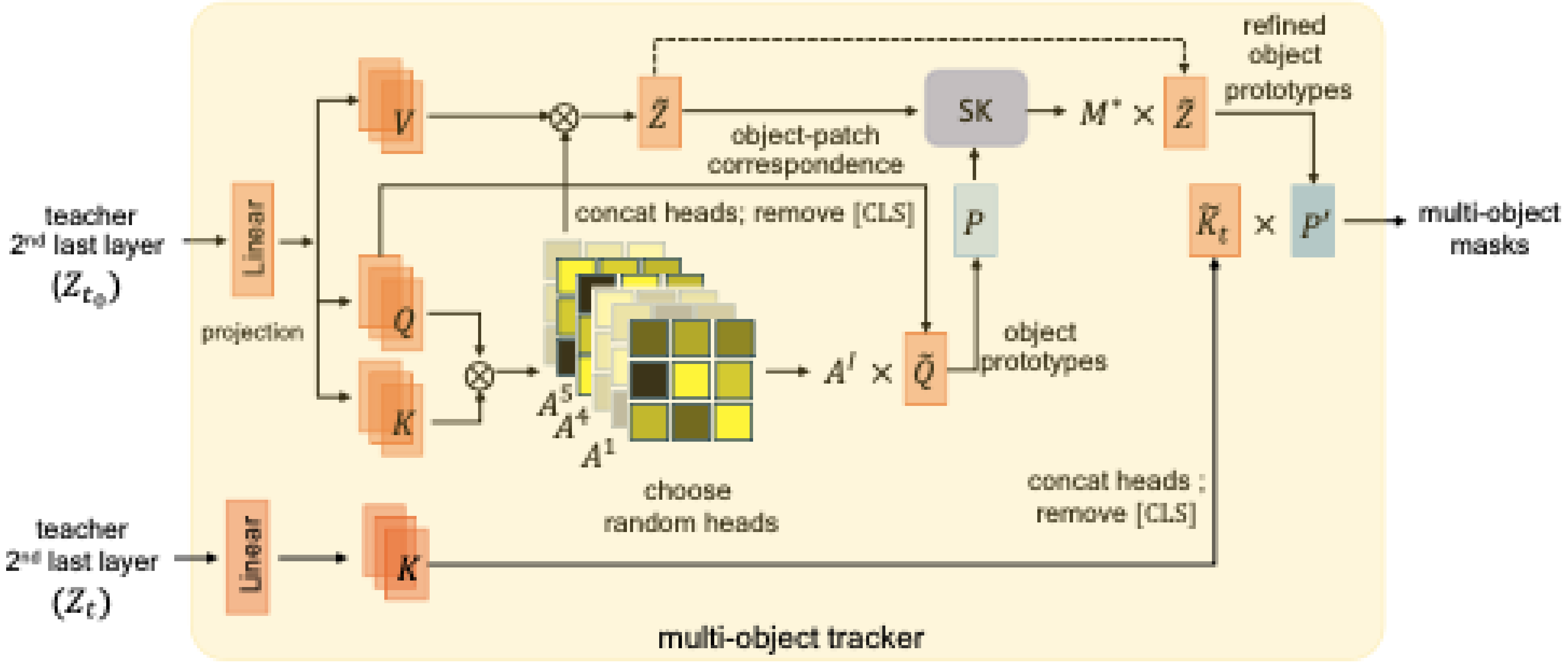
$$e^{-C/\epsilon}$$

Cost Matrix

$$C := -P\tilde{Z}^{\top} \in \mathbb{R}^{k\times n}$$

Optimal Transport Plan

$$M^* = \mathrm{SK}\left(\exp\left(P\tilde{Z}^{\top}/\epsilon\right)\right) \in \mathbb{R}^{k\times n}$$



Establishing object-patch correspondences

Optimal Transport Plan

$$M^* = \mathrm{SK}\left(\exp\left(P\tilde{Z}^\top/\epsilon\right)\right) \in \mathbb{R}^{k\times n}$$

Refined Object Prototypes

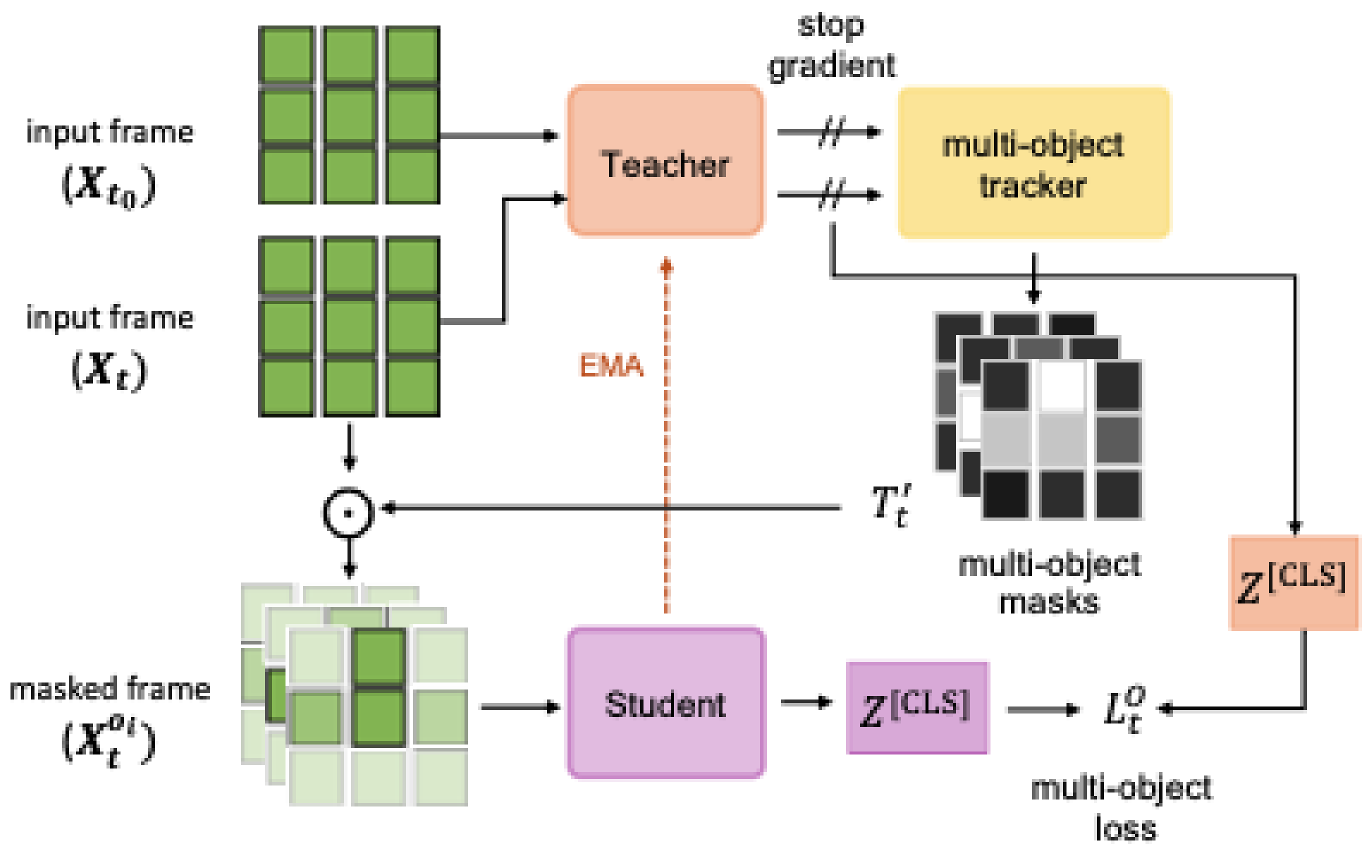
$$P' = M^*\tilde{Z} \in \mathbb{R}^{k\times d}$$

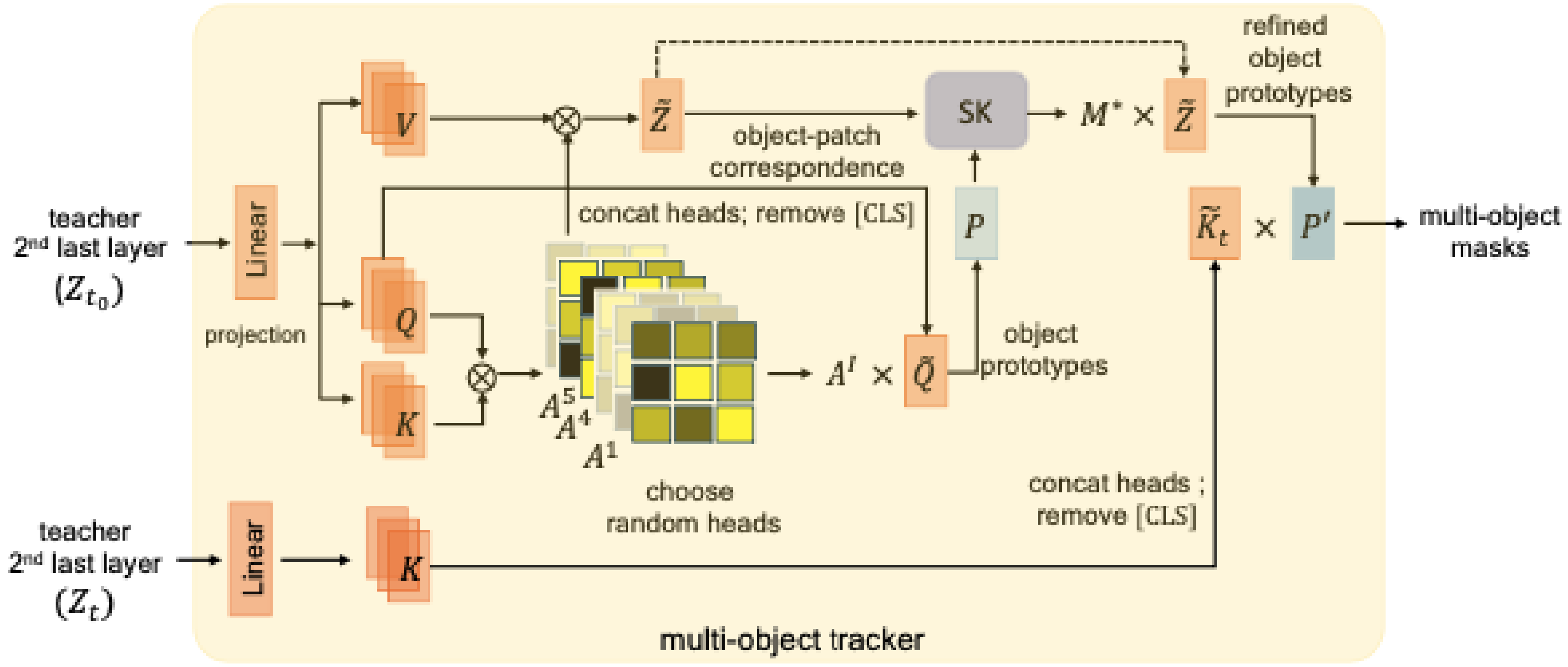
Refined Attention Map

$$T'_t := \mathrm{softmax}\left(P'\tilde{K}_t^\top/\sqrt{d}\right) \in \mathbb{R}^{k\times n}$$

III Discover and tRAck (DoRA)

Multi-object masking





Masked Video Clip

$$\mathbf{X}^{o_i} := \mathbf{X} \odot \mathbf{T}^i$$

Multi-Object Loss

$$L_t^{\mathrm{O}} := \sum_{u,v \in V} \mathbb{1}_{u \neq v} \sum_{i=1}^{k} f_{\theta'}(\mathbf{X}_t^u)^{[\mathrm{CLS}]} \log\left(f_\theta(\mathbf{X}_t^{v,o_i})^{[\mathrm{CLS}]}\right).$$

IV **Experiments**

Ablations

Method	Pretrain	#Frames (M)	LP	CorLoc
DINO	Movie_{rom}	0.19	34.9	51.5
DoRA	Movie_{rom}	0.19	**35.3**	**51.6**
DINO	K-400*	0.2	40.7	52.4
DoRA	K-400*	0.2	**43.0**	**55.2**
DINO	EK*	0.2	38.6	53.5
DoRA	EK*	0.2	**41.8**	**56.0**
DINO	$\text{WT}_{\text{Venice}}$	0.2	33.8	51.2
DoRA	$\text{WT}_{\text{Venice}}$	0.2	**44.5**	**56.2**

(a) *Video datasets*

Method	k	LP	CorLoc
DINO	✗	33.8	51.2
DoRA	1	39.9	53.9
DoRA	2	43.1	55.7
DoRA	3	**44.5**	**56.2**
DoRA	4	39.2	53.8
DoRA	5	36.7	50.3
DoRA	6	35.8	48.8
DoRA	16	28.3	48.5
DoRA	32	27.1	46.8

(b) *#Objects k on $\text{WT}_{\text{Venice}}$*

Method	SK	Mask	LP	CorLoc
DINO	✗	✗	33.8	51.2
DoRA	✗	Random	33.0	49.8
DoRA	✗	Object	42.5	55.3
DoRA	✓	Random	29.9	46.7
DoRA	✓	Object	**44.5**	**56.2**

(c) *SK and masking on $\text{WT}_{\text{Venice}}$*

(a) WTours pretraining outperforms Kinetics-400 and movie-based datasets on downstream tasks.

(b) Tracking 3 objects (k=3) achieves the best performance by balancing detail and matchability.

(c) Multi-object masks with SK algorithm yield better results than random masking.

IV Experiments

Dense Scene Understanding

METHOD	EPOCHS	PRETRAIN	(a) SEMANTIC SEG.				(b) OBJECT DET.		(c) INSTANCE SEG.	
			mIoU	GAIN	Acc_m	GAIN	mAP	GAIN	mIoU	GAIN
ViT-S/16	100	none	25.1		33.3		28.6		24.3	
iBOT (Zhou et al., 2022a)	100	WT_{Venice}	33.9		43.3		37.6		33.0	
AttMask (Kakogeorgiou et al., 2022)	100	WT_{Venice}	33.6		42.7		36.5		32.5	
VITO (Parthasarathy et al., 2022)	300	VideoNet	39.4		–		44.0		–	
DINO (Caron et al., 2021)	100	IN-1k	33.9		44.3		39.9		35.1	
DoRA (ours)	100	WT_{all}	36.9		48.0		40.7		36.3	
DINO (Caron et al., 2021)	100	WT_{Venice}	32.4		43.7		37.1		32.1	
DoRA (ours)	100	WT_{Venice}	**35.4**	+3.0	**45.5**	+1.8	**39.5**	+2.4	**34.7**	+2.6

Consistent performance gains over DINO in semantic segmentation and object detection, even with fewer training frames.

Video Understanding

Method	Epochs	Pretrain	(a) Video object segmentation						(b) Object tracking					
			$(\mathcal{J}\&\mathcal{F})_m$	Gain	$\mathcal{J}_m$	Gain	$\mathcal{F}_m$	Gain	mAO	Gain	$SR_{0.5}$	Gain	$SR_{0.75}$	Gain
ViT (Dosovitskiy et al., 2020)	100	None	26.9		25.4		28.3		23.1		19.0		3.4	
iBOT (Zhou et al., 2022a)	100	WT_{Venice}	57.4		56.7		58.0		41.5		47.5		16.6	
DINO (Caron et al., 2021)	100	IN-1k	59.4		57.4		61.4		46.4		54.3		24.1	
DoRA (ours)	100	WT_{all}	57.6		55.1		60.2		45.9		53.4		23.7	
DINO (Caron et al., 2021)	100	WT_{Venice}	54.6		53.0		56.2		37.4		41.4		13.4	
DoRA (ours)	100	WT_{Venice}	**58.4**	+3.8	**56.4**	+3.4	**60.4**	+4.2	**41.4**	+4.0	47.2	+5.8	18.2	+4.8

DORA outperforms DINO in video object segmentation and multi-object tracking under challenging conditions.

Image classification and unsupervised object discovery

Method	Epochs	Pretrain	#Frames (M)	(a) Classification LP	Gain	k-NN	Gain	(b) Object discovery Jacc.	Gain	CorLoc	Gain
SimCLR (Chen et al., 2020)	100	WT_{Venice}	0.2	26.3		25.9		40.4		50.2	
SwAV (Caron et al., 2020)	100	WT_{Venice}	0.2	28.0		26.4		40.6		51.4	
iBOT (Zhou et al., 2022a)	100	WT_{Venice}	0.2	36.8		32.8		43.0		53.1	
AttMask (Kakogeorgiou et al., 2022)	100	WT_{Venice}	0.2	35.8		31.9		43.5		54.5	
VicReg (Bardes et al., 2021)	100	WT_{Venice}	0.2	36.5		30.1		42.7		52.1	
DINO (Caron et al., 2021)	100	WT_{Venice}	0.2	33.8		29.9		43.8		51.2	
DoRA (ours)	100	WT_{Venice}	0.2	**45.4**	+11.6	**33.8**	+3.9	**44.0**	+0.2	**56.2**	+5.0
DINO (Caron et al., 2021)	100	WT_{all}	1.5	36.6		31.1		42.9		55.8	
DoRA (ours)	100	WT_{all}	1.5	**45.3**	+8.7	**35.7**	+4.6	**44.3**	+1.4	**57.1**	+1.3

DORA surpasses DINO in image classification and object discovery, showing efficient learning even with limited video data.

Why Best Paper?

- Achieves ImageNet-level performance using a single video, highlighting the potential of data-efficient AI learning.
- Unlike curated datasets with controlled class balance, it is open-ended, scalable, and does not require manual labeling.

Limitation

- Optimized for specific types of videos, requiring further validation across diverse datasets.