Time Is MattEr:
Temporal Self-supervision for Video Transformers

Sukmin Yun [1] (presenter), **Jaehyung Kim** [1], **Dongyoon Han** [2], **Hwanjun Song** [2], **Jung-Woo Ha** [2], **Jinwoo Shin** [1]

[1]Korea Advanced Institute of Science and Technology (KAIST)

[2]NAVER AI Lab

KAIST

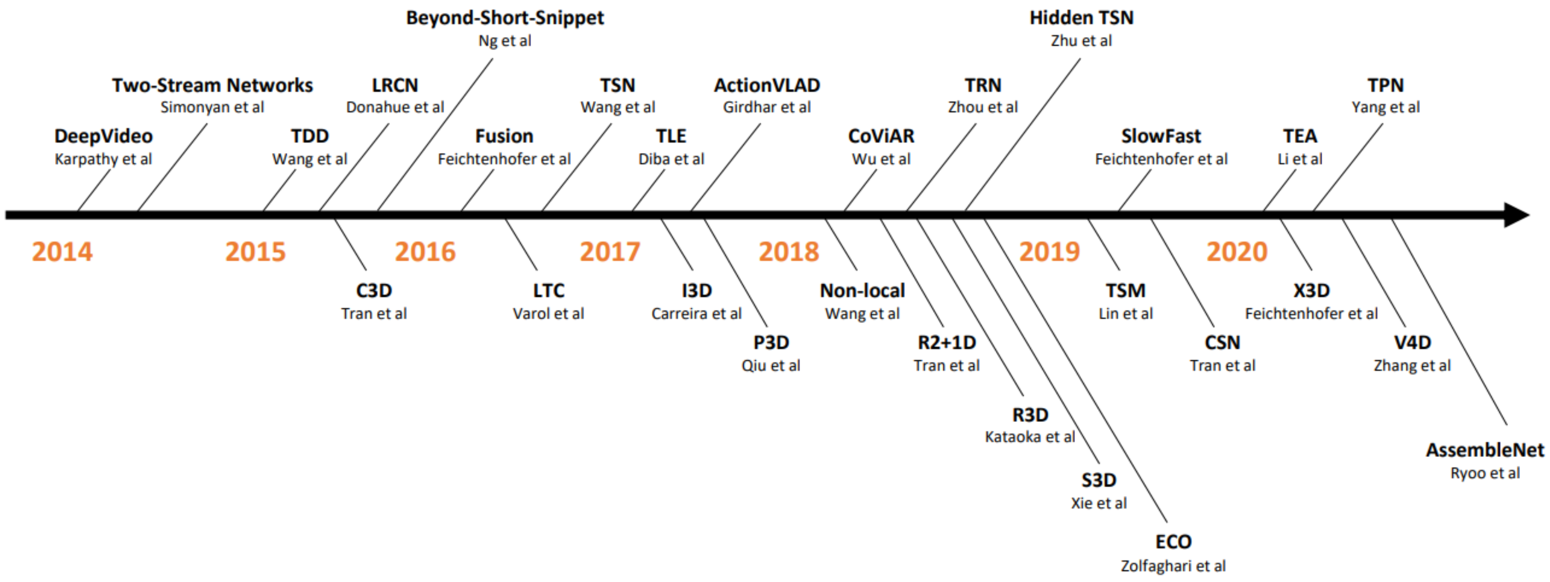
ICML
International Conference
On Machine Learning

NAVER AI LAB

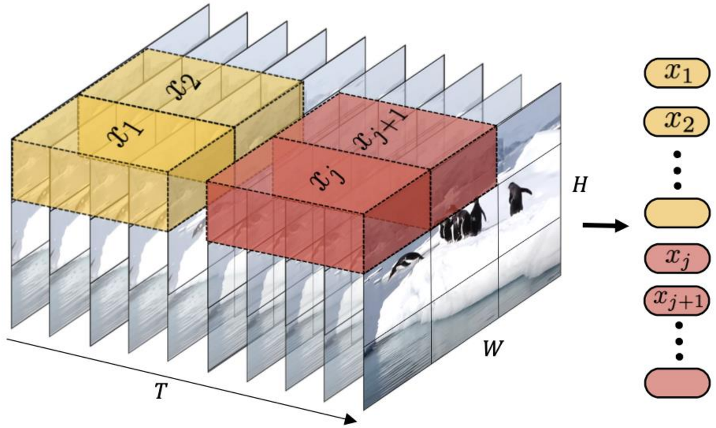
Transformers for Video Understanding

- Understanding **temporal dynamics** of video is an essential aspect of learning better video representations.
  - **Designing video-specific architectures** has been a common theme in learning better video representations
  - Recently, **transformer-based architectural designs** have been extensively explored for video tasks

Overview in architectural advances for video action recognition



Video Transformers



*source: [Zhu et al., 2020] A comprehensive study of deep video action recognition, arXiv
[Arnab & Dehghani et al., 2021] A Video Vision Transformer, ICCV 2021

Motivations

- However, it is still questionable whether **these architectural advances** are enough to fully capture the **temporal dynamics in a video**

- Video datasets often contain **action classes** can be recognized **without any temporal information!**

"A single frame is often informative enough to predict the label with good confidence" [Sevilla-Lara et al., 2021]

"Riding a bike" class in Kinetics [Key et al., 2017] dataset

[Key et al., 2017] The kinetics human action video dataset, 2017
[Sevilla-Lara et al., 2021] Only Time Can Tell: Discovering Temporal Data for Temporal Modeling, WACV 2021

Motivations

- However, it is still questionable whether **these architectural advances** are enough to fully capture the **temporal dynamics in a video**

- **Temporal classes** [Sevilla-Lara et al., 2021]**:** Temporal information is **essential** to discriminate the label

"moving something and something away from each other" class in Something-Something-v2 dataset [Goyal et al., 2017]

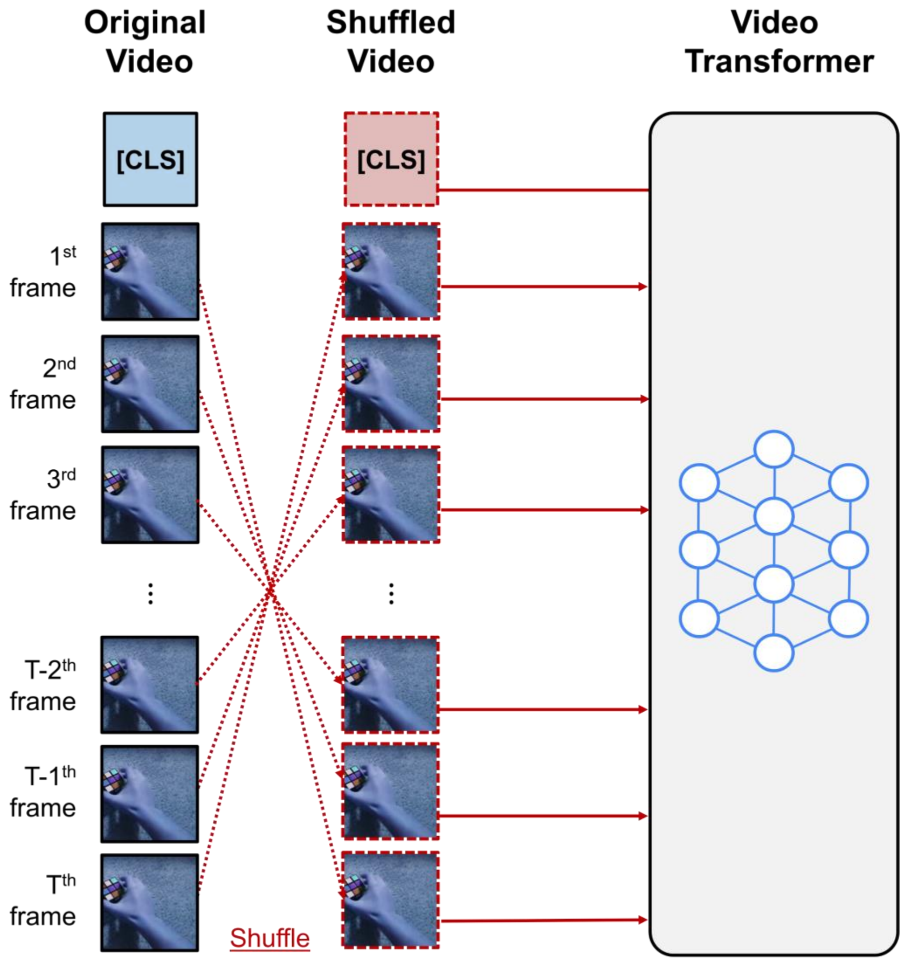
  - Discriminating moving objects **away** from or **closer** to each other classes requires temporal understanding

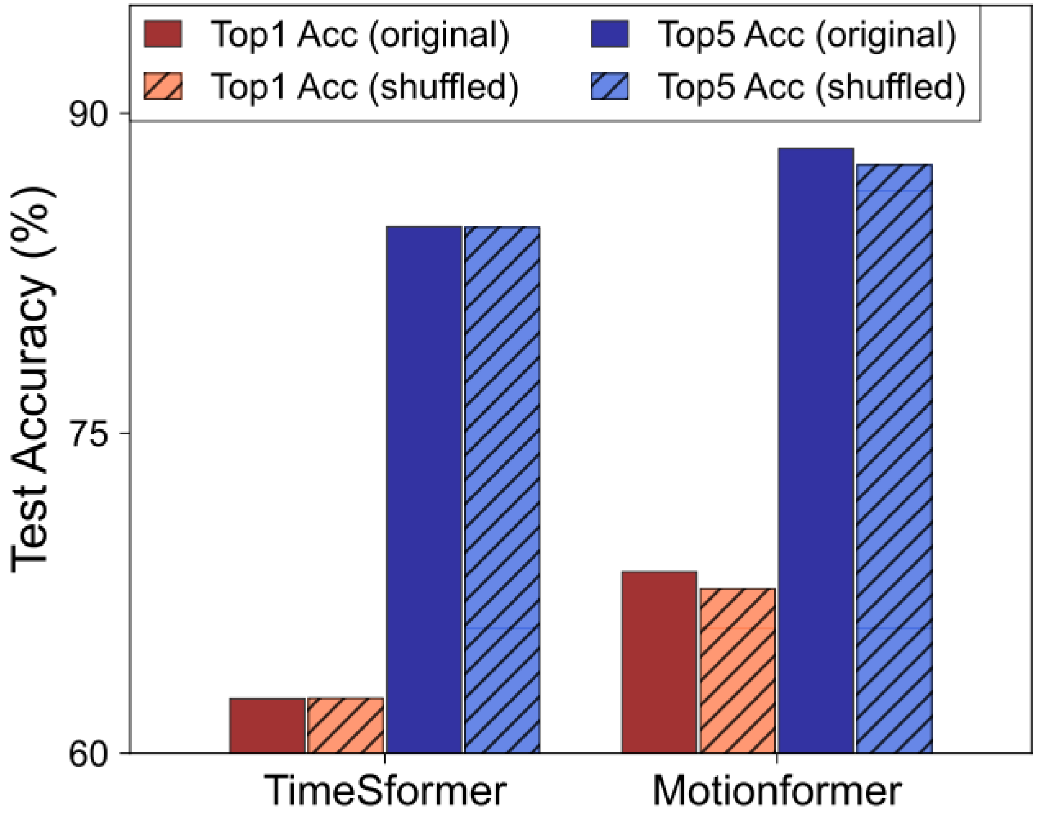
- **Static classes** [Sevilla-Lara et al., 2021]**:** Temporal information is **redundant** to discriminate the label

[Goyal et al., 2017] The "something something" video database for learning and evaluating visual common sense, ECCV 2017
[Sevilla-Lara et al., 2021] Only Time Can Tell: Discovering Temporal Data for Temporal Modeling, WACV 2021

Motivations: Observations

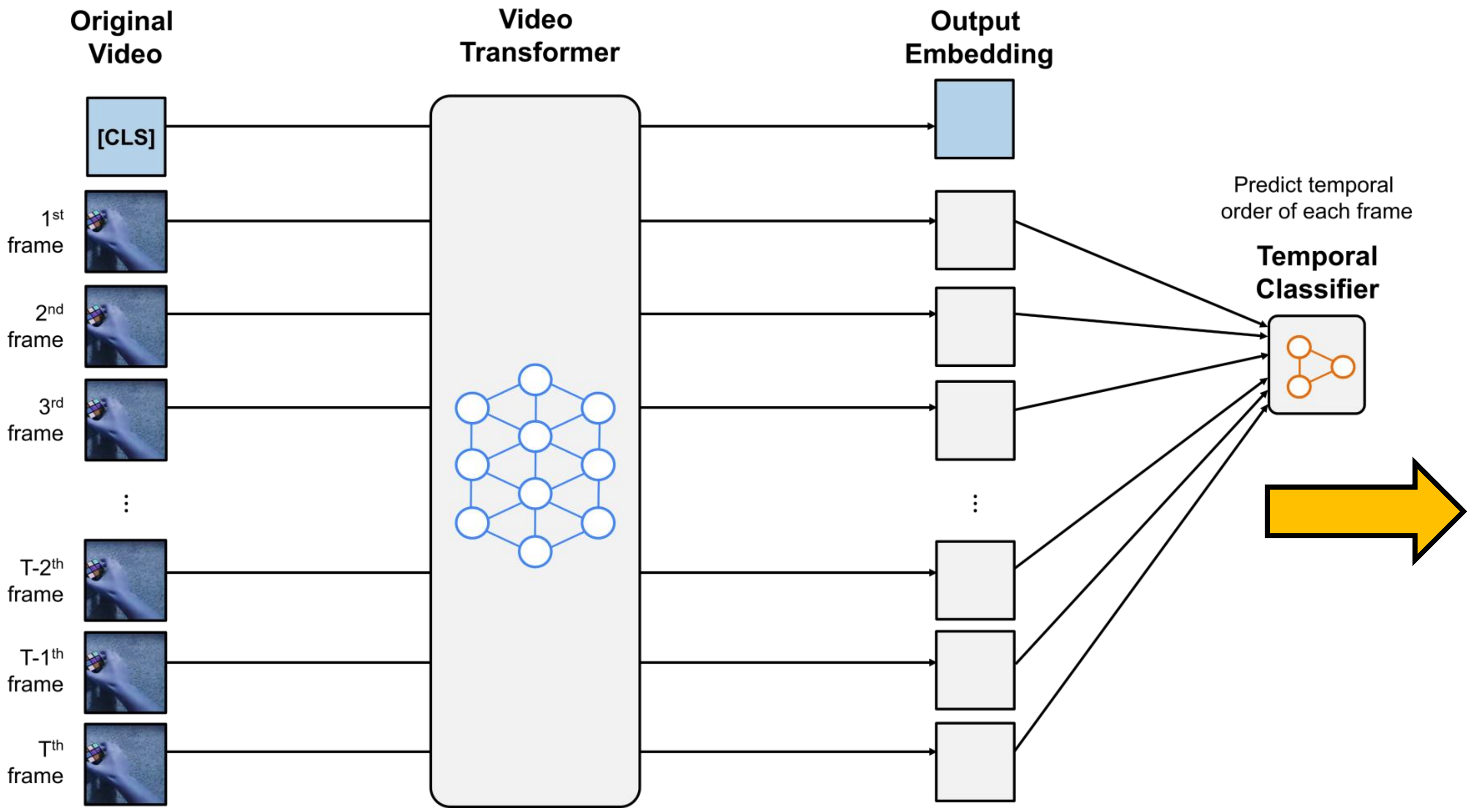
- **Video Transformers** are still **biased to learn spatial dynamics** rather than temporal ones
  - Video Transformers often predict a video action correctly even when input video frames are **randomly shuffled**

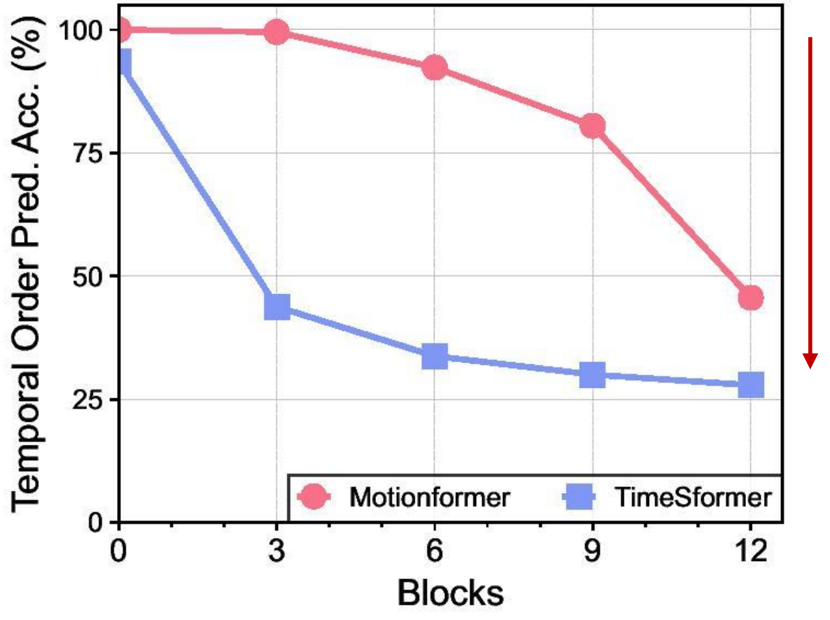




Motivations: Observations

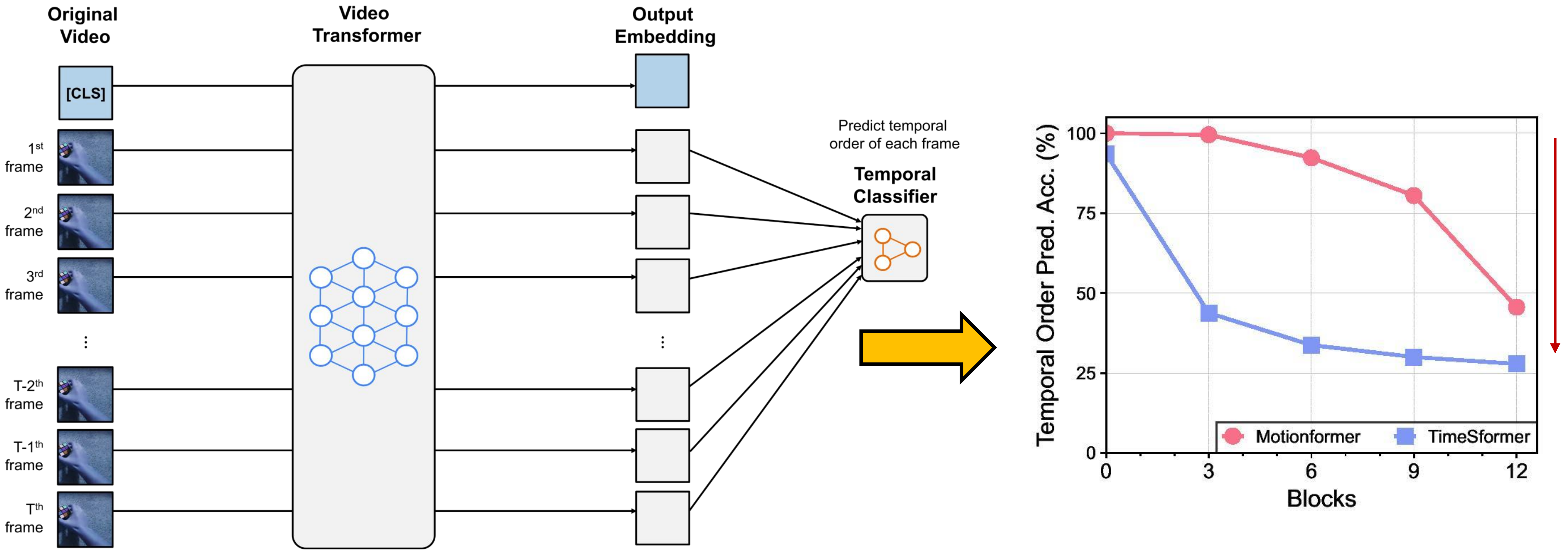
- **Video Transformers** are still **biased to learn spatial dynamics** rather than temporal ones
  - Video Transformers also **fail to capture the temporal order** of video frames as their layers go deeper





Motivations: Observations

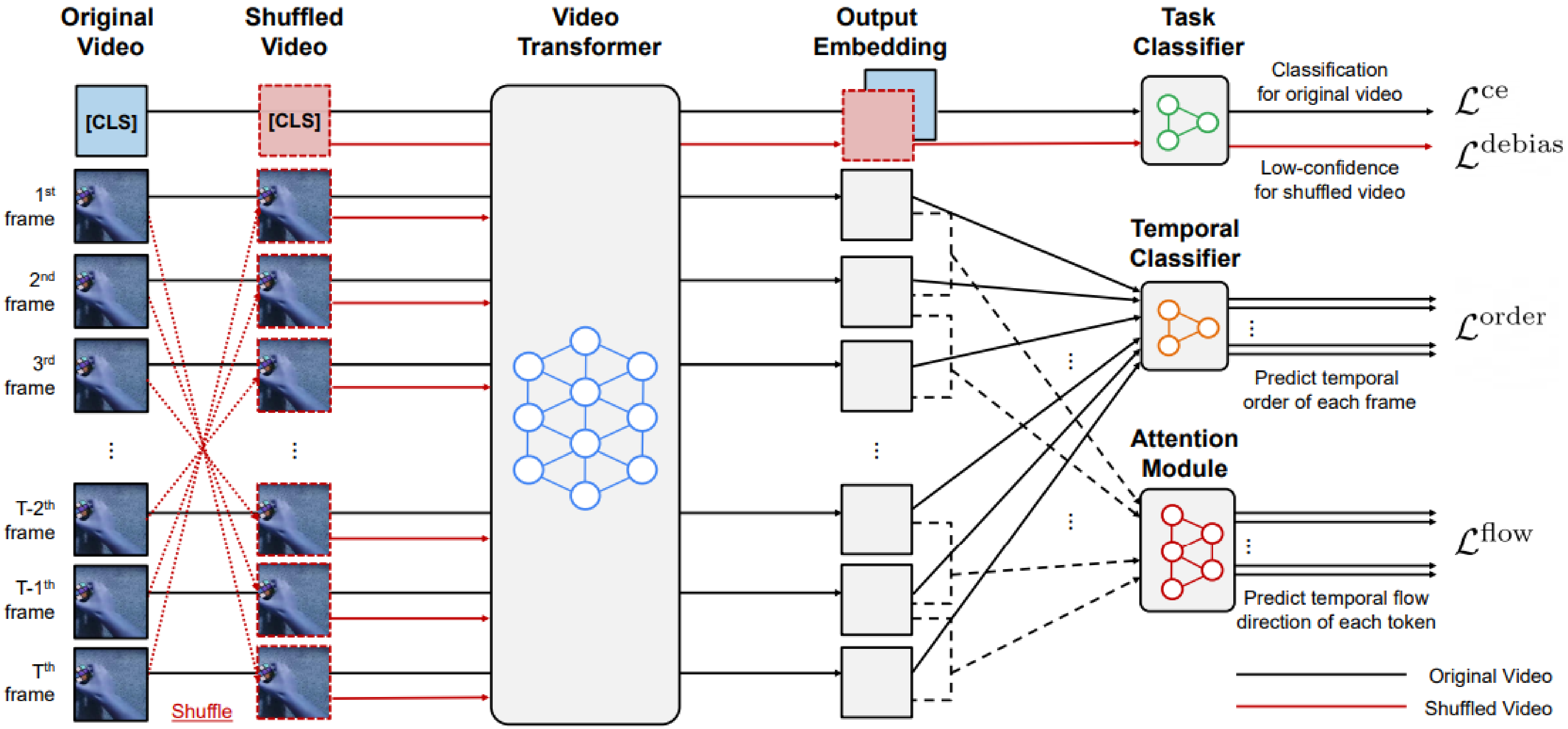
- **Video Transformers** are still **biased to learn spatial dynamics** rather than temporal ones
  - Video Transformers also **fail to capture the temporal order** of video frames as their layers go deeper



How to improve the quality of learned video representations via better temporal modeling? 🤷‍♂️

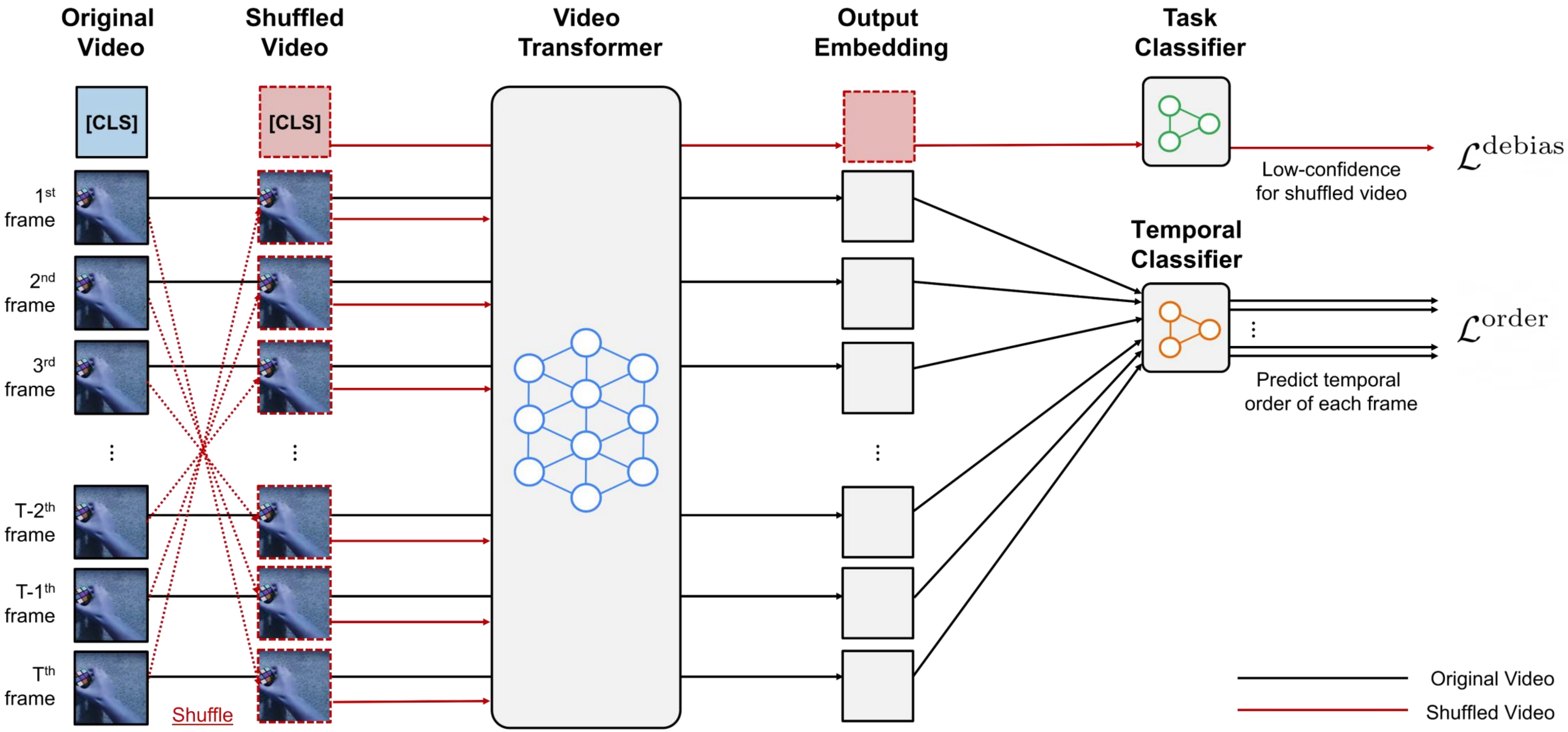
TIME: Overview

- **Goal:** How to lean better temporal dynamics?
  1. Debiasing the spurious correlation learned from spatial dynamics
  2. Enhancing the correlation toward temporal dynamics



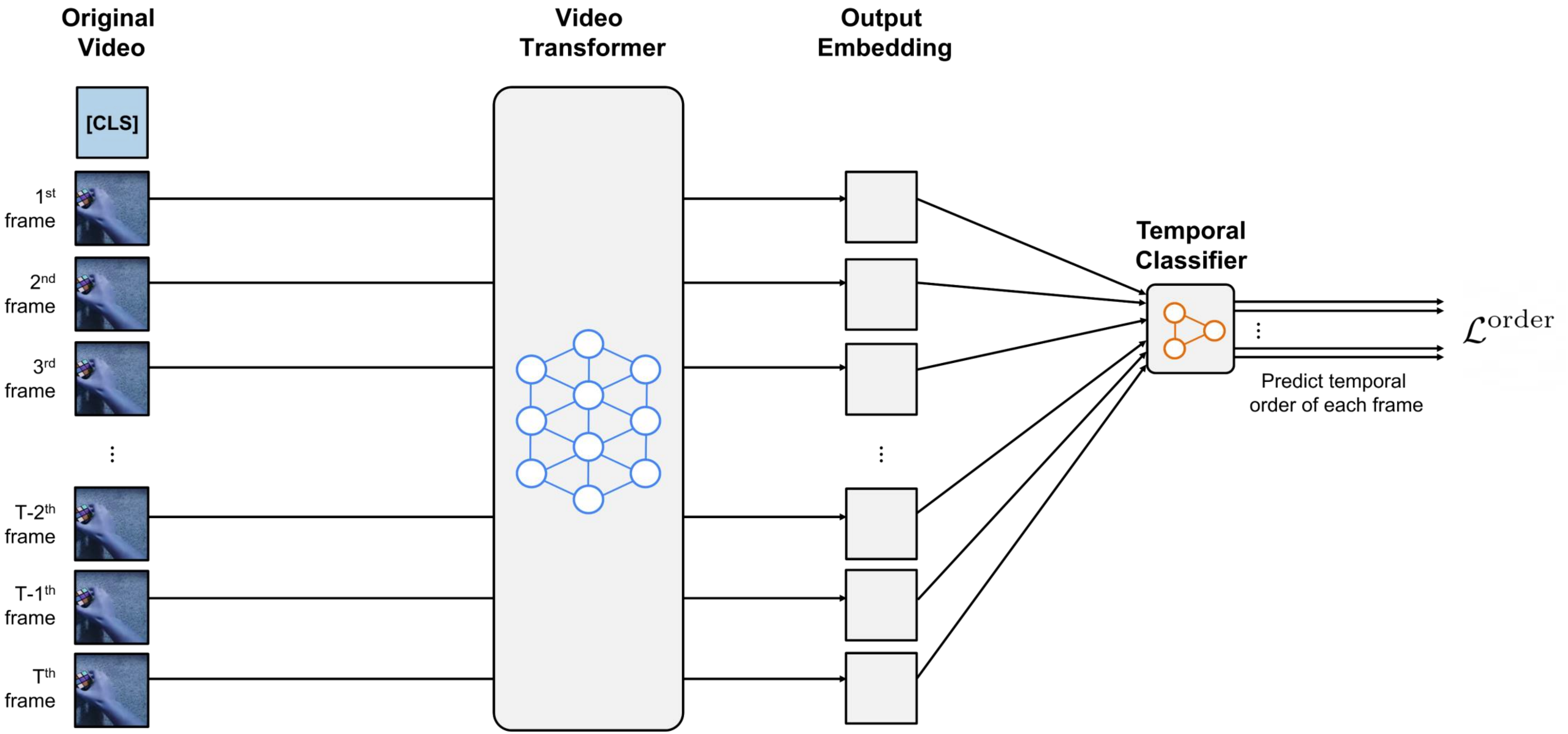
TIME: Frame-level Self-supervision

- **Goal:** How to lean better temporal dynamics?
  1. Debiasing the spurious correlation learned from spatial dynamics
  2. Enhancing the correlation toward temporal dynamics



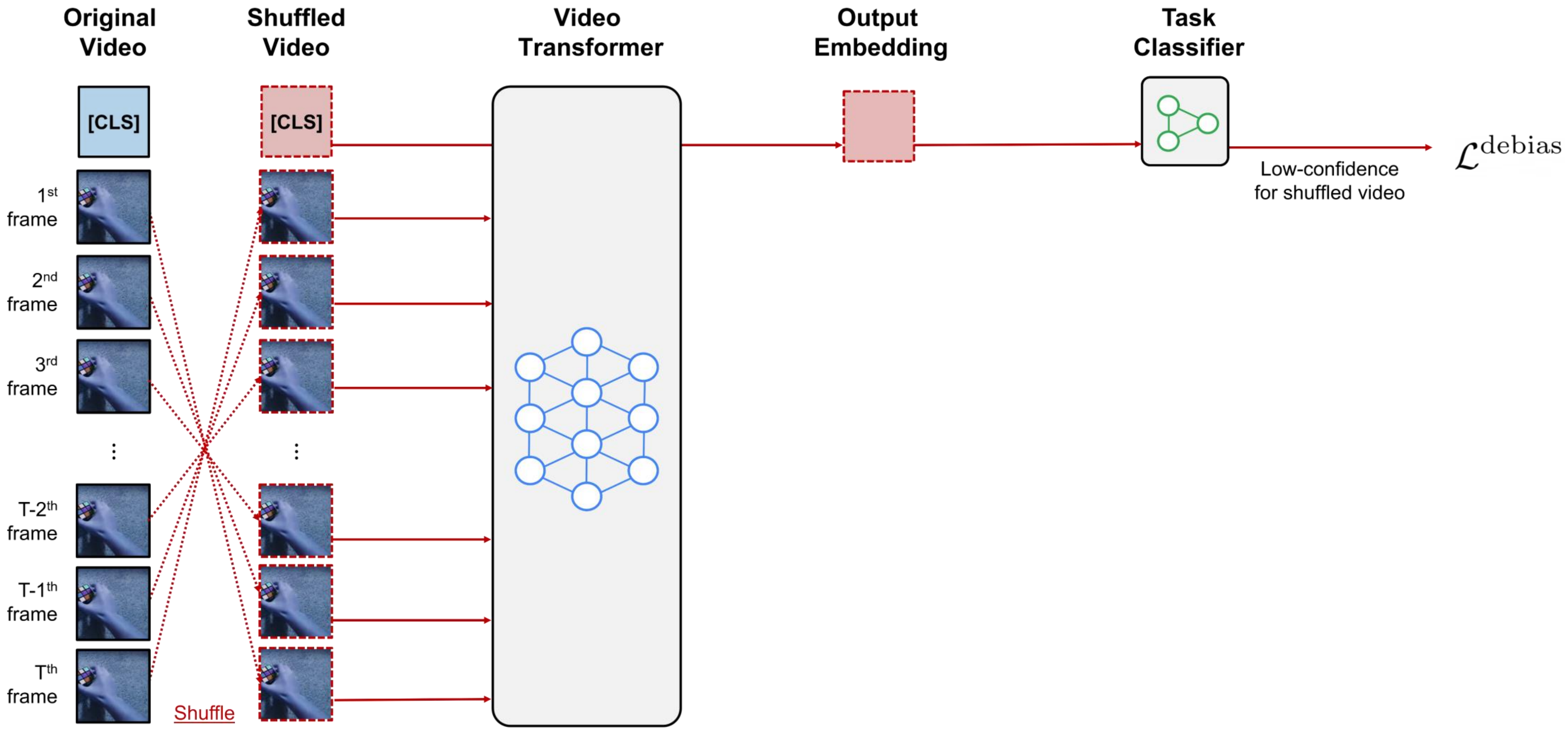
TIME: Frame-level Self-supervision

- **Goal:** How to lean better temporal dynamics?
  1. Debiasing the spurious correlation learned from spatial dynamics
  2. Enhancing the correlation toward temporal dynamics



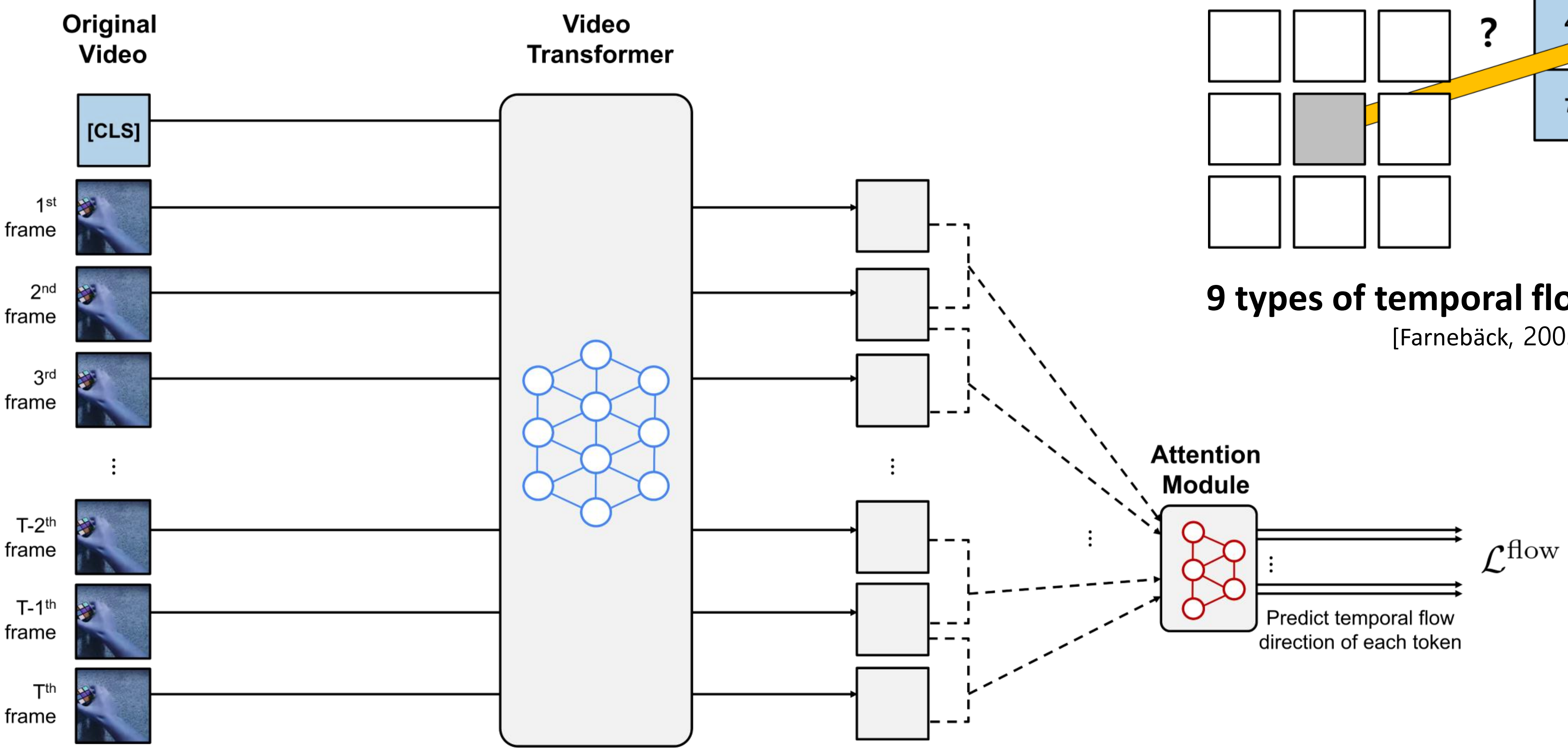
TIME: Frame-level Self-supervision

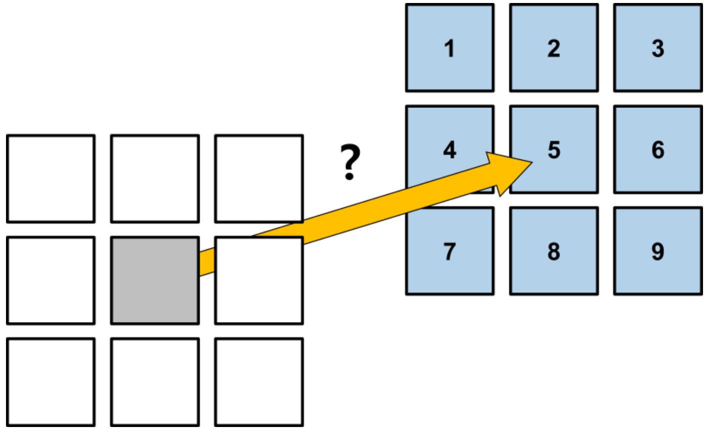
- **Goal:** How to lean better temporal dynamics?
  1. Debiasing the spurious correlation learned from spatial dynamics
  2. Enhancing the correlation toward temporal dynamics



TIME: Token-level Self-supervision

- **Goal:** How to lean better temporal dynamics?
  1. Debiasing the spurious correlation learned from spatial dynamics
  2. Enhancing the correlation toward temporal dynamics





9 types of temporal flow directions

[Farnebäck, 2003]

*source: [Farnebäck, 2003] Two-Frame Motion Estimation Based on Polynomial Expansion, SCIA 2003

Experimental Results

- **Model architecture:** TimeSformer [Bertasius et al., 2021], Motionformer [Patrick et al., 2021] and X-ViT [Bulat et al., 2021]
  - All models are fine-tuned on the SSv2 dataset [Goyal et al., 2017] from the ImageNet-1k pretrained weights

- Our method consistently improves **all the backbone architectures** with a large margin
  - Our method could overcome failure modes in the Video Transformers

Model	Top-1	Top-5
TimeSformer (Bertasius et al., 2021)	62.1	86.4
TimeSformer + TIME	**63.7**	**87.8**
Motionformer (Patrick et al., 2021)	63.8	88.5
Motionformer + TIME	**64.7**	**89.3**
X-ViT (Bulat et al., 2021)	60.1	85.2
X-ViT + TIME	**63.5**	**88.1**

*source: [Bertasius et al., 2021] Is space-time attention all you need for video understanding, ICML 2021
[Patrick et al., 2021] Keeping Your Eye on the Ball: Trajectory Attention in Video Transformers, NeurIPS 2021
[Bulat et al., 2021] Space-time Mixing Attention for Video Transformer, NeurIPS 2021

Ablation Study: Temporal vs. Static

- **Model architecture:** TimeSformer [Bertasius et al., 2021], Motionformer [Patrick et al., 2021] and X-ViT [Bulat et al., 2021]
  - All models are fine-tuned on the SSv2 dataset [Goyal et al., 2017] from the ImageNet-1k pretrained weights

- The performances of **Shuffled** on the Static subset are often close to the **Original** ones (*i.e.*, poor **Gap**)
  - Static classes would allow video models to predict class labels without understanding temporal information

	SSv2 dataset			Temporal subset			Static subset		
Method	Original ↑	Shuffled ↓	Gap ↑	Original ↑	Shuffled ↓	Gap ↑	Original ↑	Shuffled ↓	Gap ↑
TimeSformer	62.1	41.3	20.8	84.9	57.0	27.9	84.1	84.1	0.0
TimeSformer + TIME	**63.7**	**25.3**	**38.4**	**90.2**	**22.1**	**68.1**	**86.9**	**69.3**	**17.6**

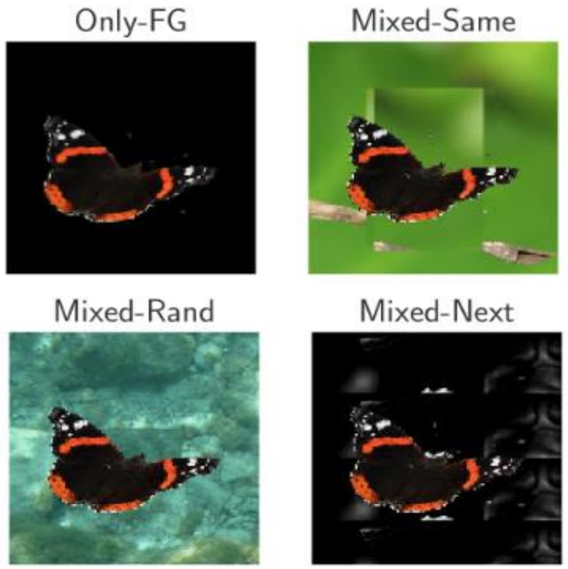
*source: [Bertasius et al., 2021] Is space-time attention all you need for video understanding, ICML 2021
[Patrick et al., 2021] Keeping Your Eye on the Ball: Trajectory Attention in Video Transformers, NeurIPS 2021
[Bulat et al., 2021] Space-time Mixing Attention for Video Transformer, NeurIPS 2021

Ablation Study: Image Domain

- Our approach can be extended to the **image domain** for alleviating background bias by replacing
  - Learning temporal order of frames with spatial order of patches
  - Debiasing spatial dynamics with image backgrounds

- Our method enhances the model **generalization** and **robustness** to background shifts
  - Backgrounds Challenge [Xiao et al., 2021] on ImageNet-9 dataset

Dataset	Baseline	Baseline + $\mathcal{L}_I^{\text{order}}$	Baseline + $\mathcal{L}_I^{\text{debias}}$	Baseline + $\mathcal{L}_I^{\text{TIME}}$
Original ↑	77.3	82.0 (+4.7)	79.0 (+1.7)	**83.3 (+6.0)**
Only-FG ↑	50.3	54.2 (+3.9)	52.7 (+2.4)	**58.9 (+8.6)**
Mixed-Same ↑	68.6	72.5 (+3.9)	69.7 (+1.1)	**74.0 (+5.4)**
Mixed-Rand ↑	43.7	48.4 (+4.7)	45.1 (+1.4)	**51.0 (+7.3)**
Mixed-Next ↑	39.9	43.6 (+3.7)	40.6 (+0.7)	**46.4 (+6.5)**
BG-Gap ↓	24.8	24.1 (−0.7)	24.6 (−0.2)	**23.0 (−1.8)**



Examples of background shifts
[Xiao et al., 2021]

*source: [Xiao et al., 2021] Noise or signal: The role of image backgrounds in object recognition, ICLR 2021

Summary

- Our work highlights the importance of **debiasing the spurious correlation** of visual transformer models with respect to the temporal or spatial dynamics

- We believe our work could inspire researchers to rethink the under-explored, yet important problem and provide a new research direction for improving video understanding

Thank you for your attention ☺