FineParser: A Fine-grained Spatio-temporal Action Parser for Human-centric Action Quality Assessment (Oral)

Jinglin Xu[1], Sibo Yin[2], Guohao Zhao[2], Zishuo Wang[2], Yuxin Peng[2†]
[1]School of Intelligence Science and Technology, University of Science and Technology Beijing
[2]Wangxuan Institute of Computer Technology, Peking University

PEKING UNIVERSITY
University of Science and Technology Beijing
Code Page · MIPL's Website · MIPL's GitHub · Personal HomePage
CVPR SEATTLE, WA JUNE 17-21, 2024

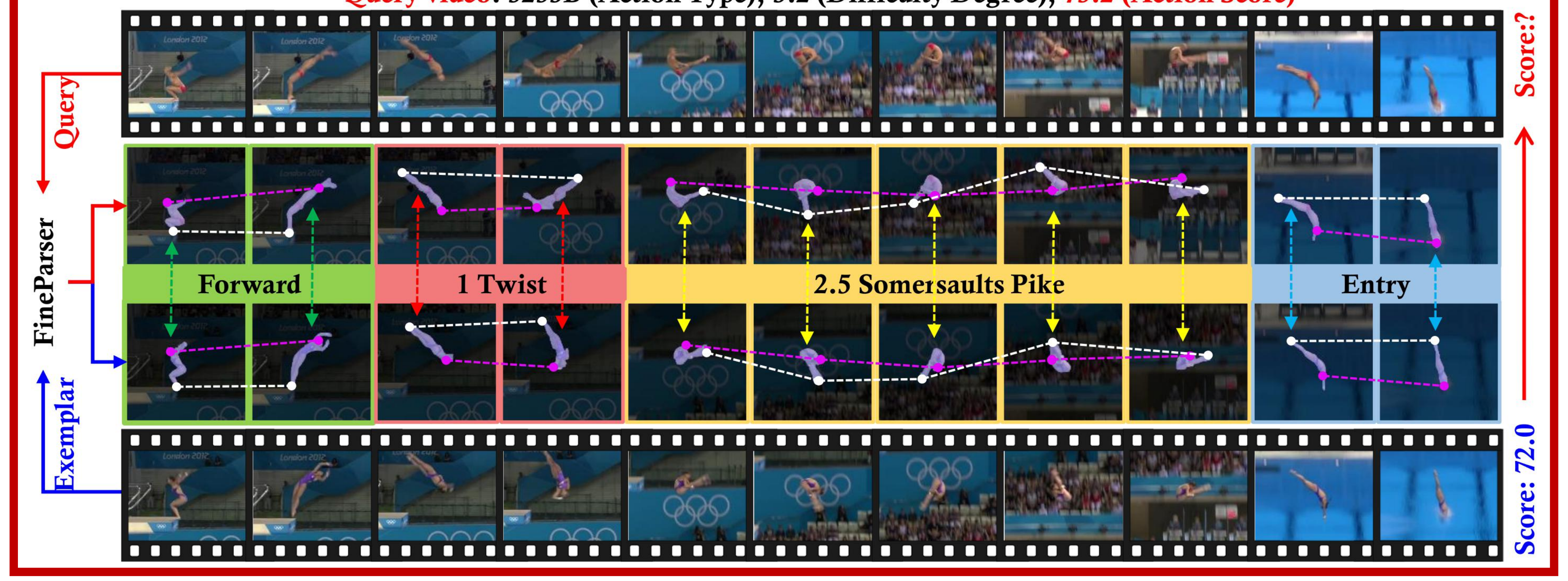
Motivation

- Action Quality Assessment (AQA) aims to evaluate the execution quality of a specific action by predicting a score after analyzing the performance of the action in a video. It is a crucial technique direction in video action understanding and holds extensive application prospects in fields such as healthcare and sports analysis.
- Due to the lack of fine-grained spatio-temporal action annotations, existing AQA methods mainly captured video-level representations and failed to parse actions across both spatial and temporal dimensions.

Contribution

- Human-centric foreground action mask annotations for FineDiving dataset, **FineDiving-HM**.
- A new fine-grained AQA method, FineParser, which **captures human-centric** foreground action representations and parses actions across **temporal and spatial** dimensions.

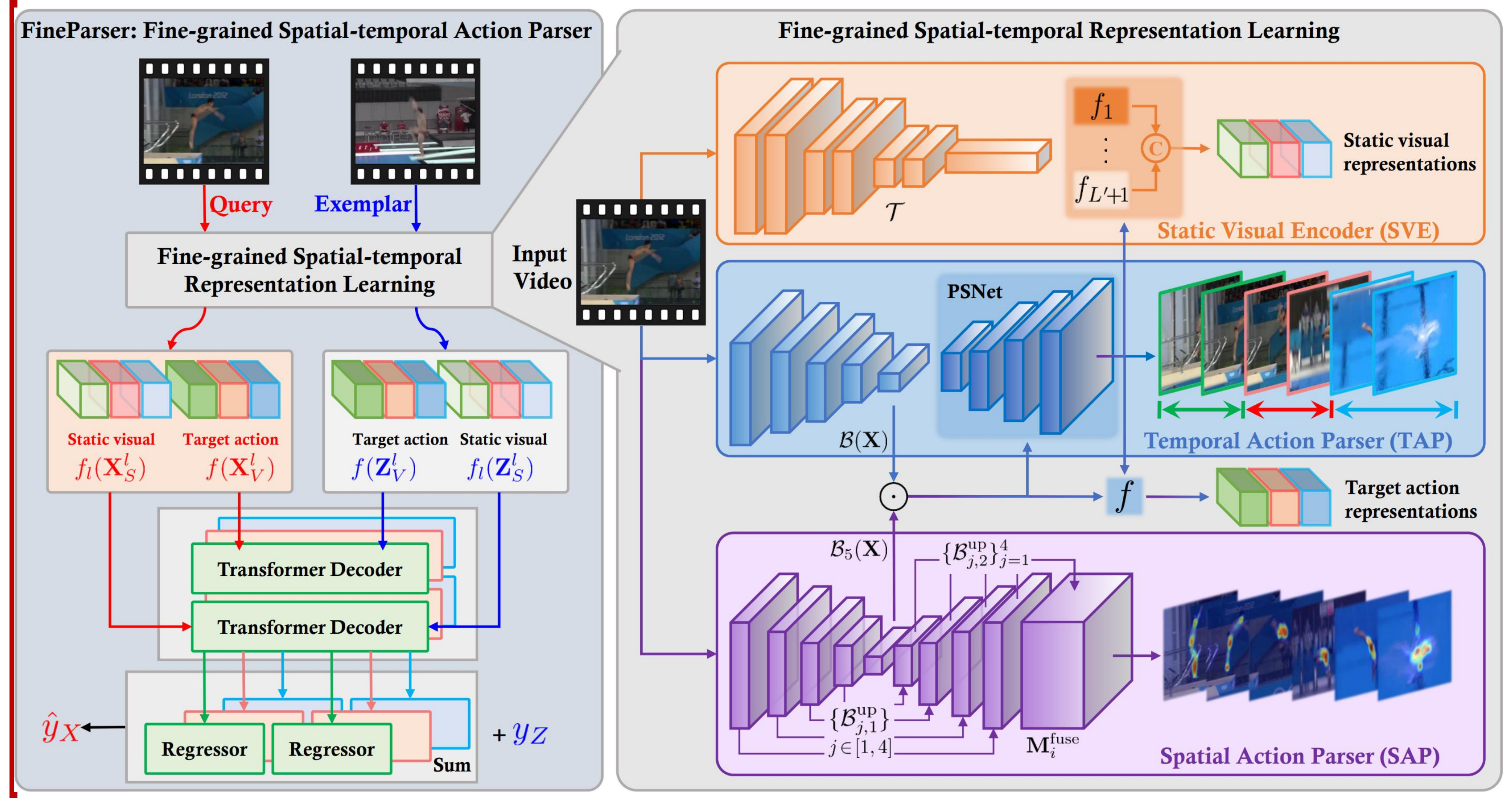
Query video: 5253B (Action Type), 3.2 (Difficulty Degree), 75.2 (Action Score)



Experiments

Methods	AQA Metrics		Methods	MTL-AQA	
	$\rho\uparrow$	$R\text{-}\ell_2\downarrow(\times 100)$		$\rho\uparrow$	$R\text{-}\ell_2\downarrow(\times 100)$
C3D-LSTM [26]	0.6969	1.0767	Pose+DCT [29]	0.2682	/
C3D-AVG [25]	0.8371	0.6251	C3D-SVR [26]	0.7716	/
MSCADC [25]	0.7688	0.9327	C3D-LSTM [26]	0.8489	/
I3D+MLP [31]	0.8776	0.4967	C3D-AVG-STL [25]	0.8960	/
USDL [31]	0.8830	0.4800	C3D-AVG-MTL [25]	0.9044	/
MUSDL [31]	0.9241	0.3474	USDL [31]	0.9231	0.4680
CoRe [37]	0.9308	0.3148	MUSDL [31]	0.9273	0.4510
TSA [36]	0.9324	0.3022	TSA-Net [34]	0.9422	/
FineParser	**0.9435**	**0.2602**	CoRe [37]	0.9512	0.2600
			FineParser	**0.9585**	**0.2411**

Method


FineParser: Fine-grained Spatial-temporal Action Parser
Query
Exemplar
Fine-grained Spatial-temporal Representation Learning
Input Video
Static visual $f_l(\mathbf{X}_S^l)$
Target action $f(\mathbf{X}_V^l)$
Target action $f(\mathbf{Z}_V^l)$
Static visual $f_l(\mathbf{Z}_S^l)$
Transformer Decoder
Transformer Decoder
$\hat{y}_X$
Regressor
Regressor
Sum
$+y_Z$
Fine-grained Spatial-temporal Representation Learning
$\mathcal{T}$
f_1
$f_{L'+1}$
Static visual representations
Static Visual Encoder (SVE)
PSNet
$\mathcal{B}(\mathbf{X})$
Temporal Action Parser (TAP)
f
Target action representations
$\mathcal{B}_5(\mathbf{X})$
$\{\mathcal{B}_{j,2}^{up}\}_{j=1}^4$
$\{\mathcal{B}_{j,1}^{up}\}$
$j\in[1,4]$
$\mathbf{M}_i^{fuse}$
Spatial Action Parser (SAP)

- **Spatial Action Parser (SAP):** SAP employs a 3D CNN-based backbone[1] to capture multi-scale representations from input video and generates foreground action masks.

$$\mathbf{M}_{j,i}^{\mathrm{up}_1}=\mathcal{B}_{j,1}^{\mathrm{up}}(\mathcal{B}_j(\mathbf{X}_i)),\ \mathbf{M}_{j,i}^{\mathrm{up}_2}=\mathcal{B}_{j,2}^{\mathrm{up}}(\mathcal{B}_j(\mathbf{X}_i)),$$
$$\mathbf{M}_i^{\mathrm{fuse}}=\mathrm{Conv3d}(\mathrm{Concat}(\{\mathbf{M}_{j,i}^{\mathrm{up}_1}\}_{j=1}^4)),$$

where M_i^{fuse} is the target actions mask of X_i.

- **Temporal Action Parser (TAP):** TAP utilizes procedure segmentation network[2] to predict the switches of sub-actions.

$$\hat{t}_k=\underset{\frac{T}{L'}(k-1)<t\leq\frac{T}{L'}k}{\arg\max}\ \mathcal{S}(\mathbf{X}_V)[t,k],$$

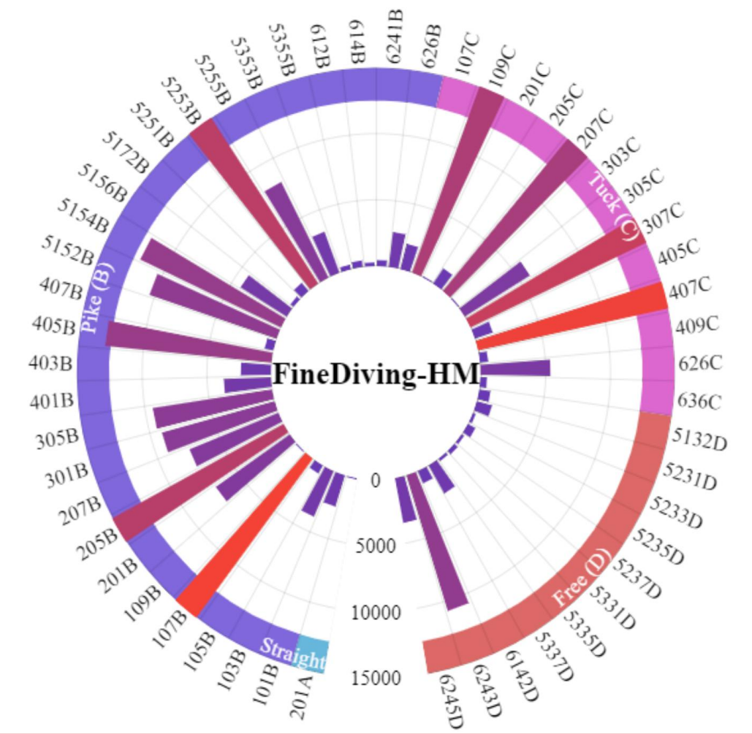
where $\hat{t}_k$ is the predicted k-th step transition.

- **Static Visual Encoder(SVE):** SVE captures static features and more contextual information from single frames.
- **Fine-grained Contrastive Regression (FineReg):** FineReg evaluates contrastive scores from pairwise steps and static representations, which can be expressed as:

$$\hat{y}_X=\sum_{l=1}^{L'+1}\lambda_l(\mathcal{R}_V(\mathbf{D}_l^V)+\mathcal{R}_S(\mathbf{D}_l^S))+y_Z,$$

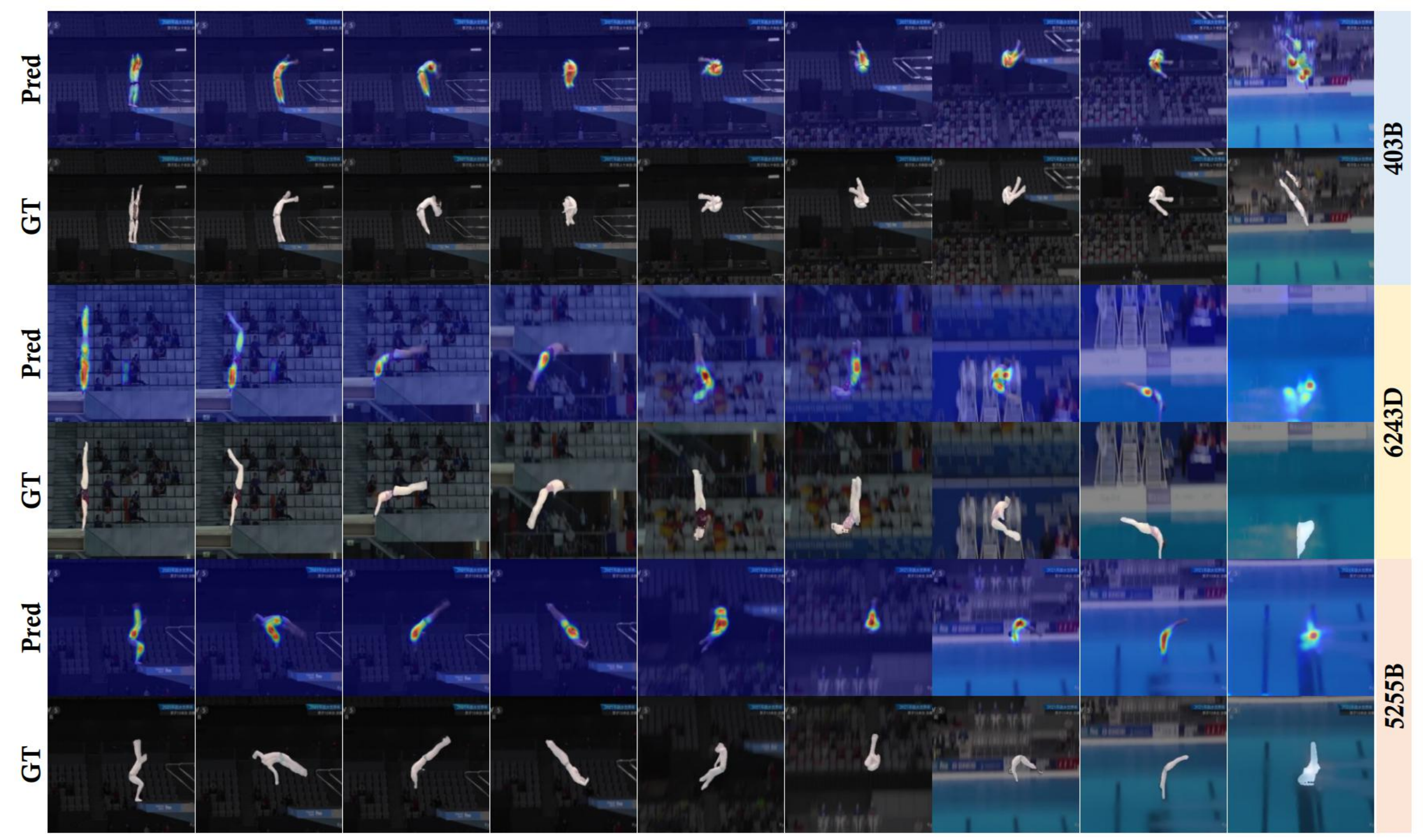
where y_Z is the score of examplar video and $\hat{y}$ is the predicted score.

FineDiving-HM



- Visualization and distribution FineDiving-HM dataset.
- We provided FineDiving-HM with **312,256 mask** frames across **3000** videos, in which each mask labels the target action region to distinguish the human-centric foreground and background.

Visualization



- Visualization of mask predictions produced by SAP.
- The predicted masks can focus on the target action regions in each frame, minimizing the impact of invalid backgrounds on action quality assessment.

[1] Jinglin Xu, et al., Finediving: A fine-grained dataset for procedure-aware action quality assessment. CVPR 2022
[2] Xumin Yu, et al., Group-aware contrastive regression for action quality assessment. ICCV 2021.