XII Jornada de Jóvenes Investigadores/as del I3A

Instituto Universitario de Investigación
en Ingeniería de Aragón
Universidad Zaragoza

Saliency Prediction in 360º Videos with Transformers

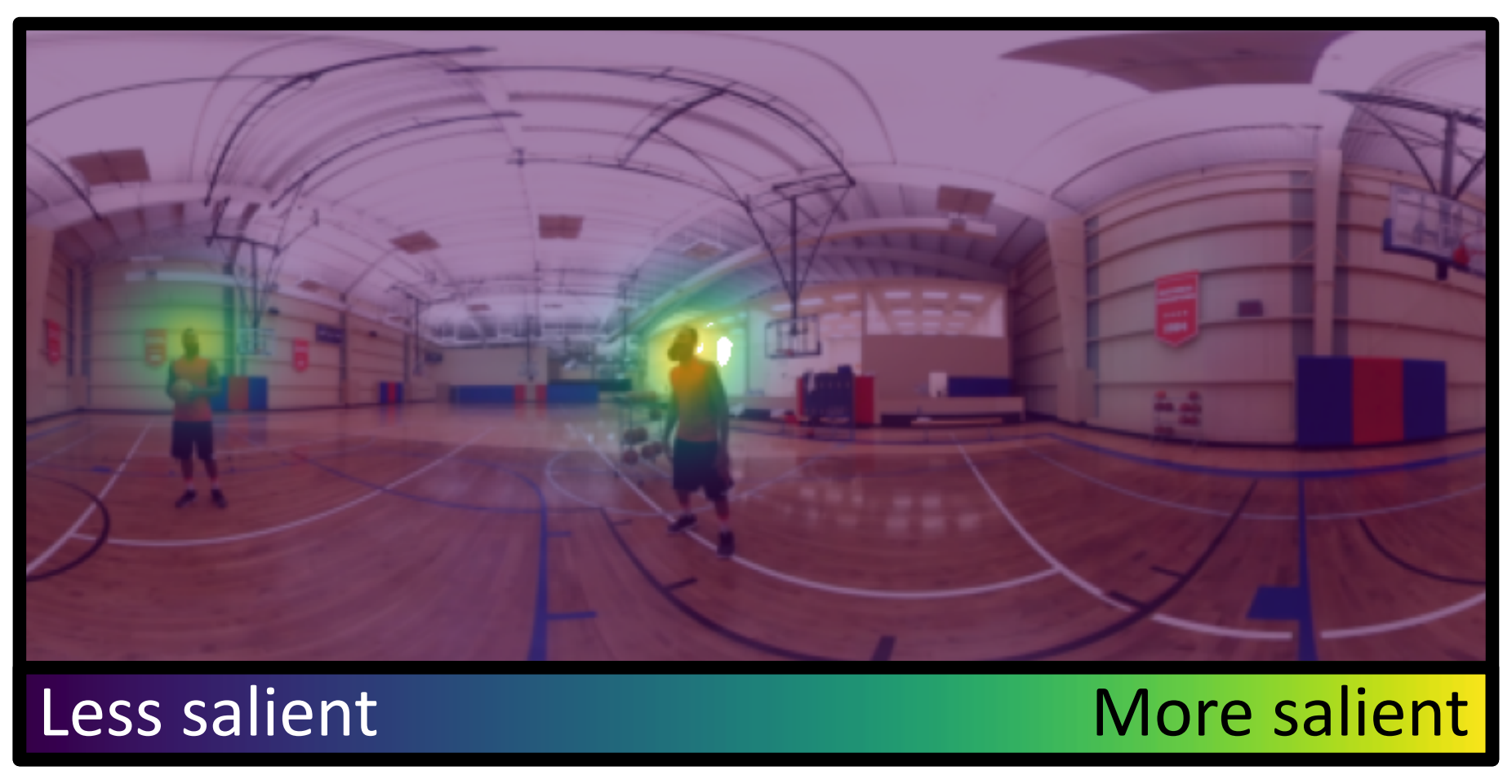
Mateo Vallejo, Diego Gutiérrez, Edurne Bernal

Motivation

- **Saliency** is the quality which makes some stimuli stand out and capture **human attention**.
- **Visual saliency** is used for studying **human visual behaviour** and **guiding user attention** in **Virtual Reality** applications.
- **Current methods** for saliency prediction in 360º videos have difficulties representing their **long-term temporal dependencies**.
- **Our method** is capable of representing the **spatio-temporal dependencies** of 360º videos by employing the **global attention** mechanism of the **Transformer** architecture.

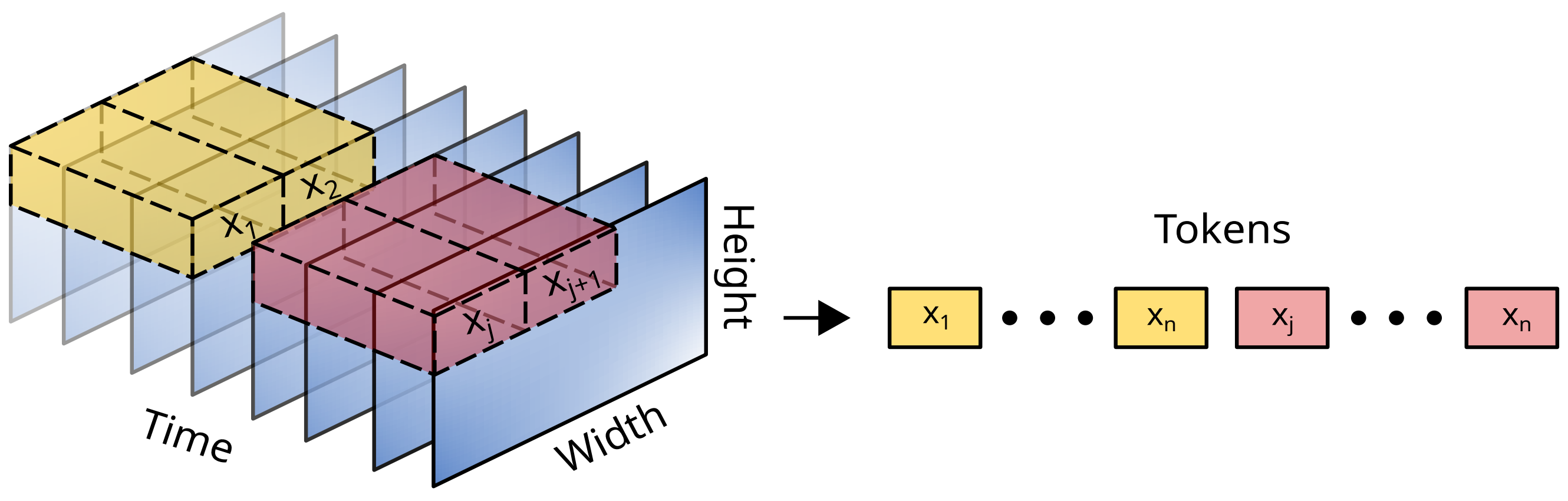
Example frame from 360º video [3]

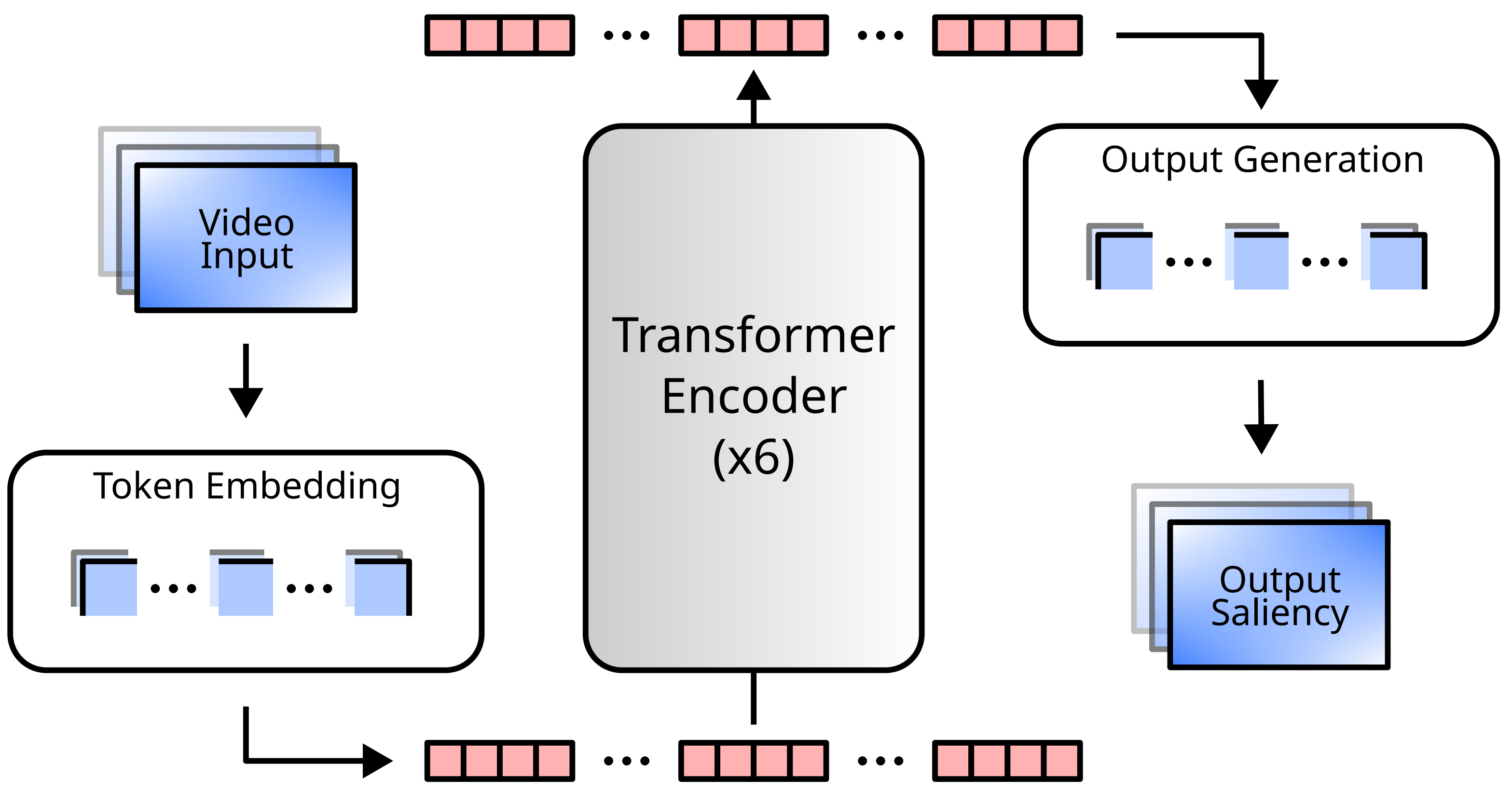
Visual attention (saliency)



Our Approach

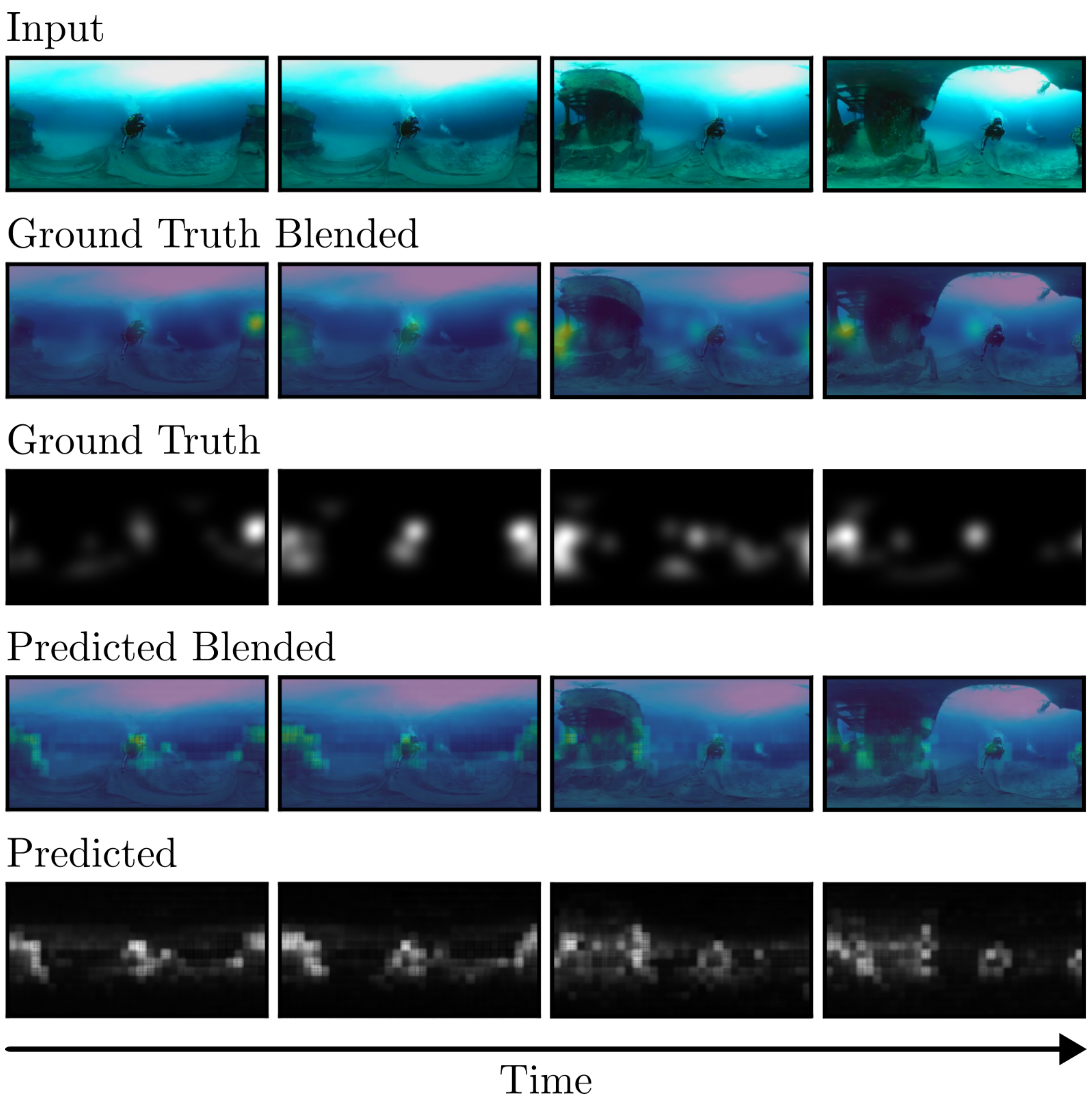
Based on the **Transformer** architecture [1], we employ the **global attention** mechanism in order to represent the **spatio-temporal dependencies** of visual attention in videos. We train our model with the **VR EyeTracking dataset** [3], by using **Mean-Square-Error (MSE)** as the loss function. To use the Transformer encoder, we first convert the input videos into a sequence of **spatio-temporal tokens** [1].





Results

Our model generates output saliency maps that **resemble human visual attention** in 360º videos.



Input

Ground Truth Blended

Ground Truth

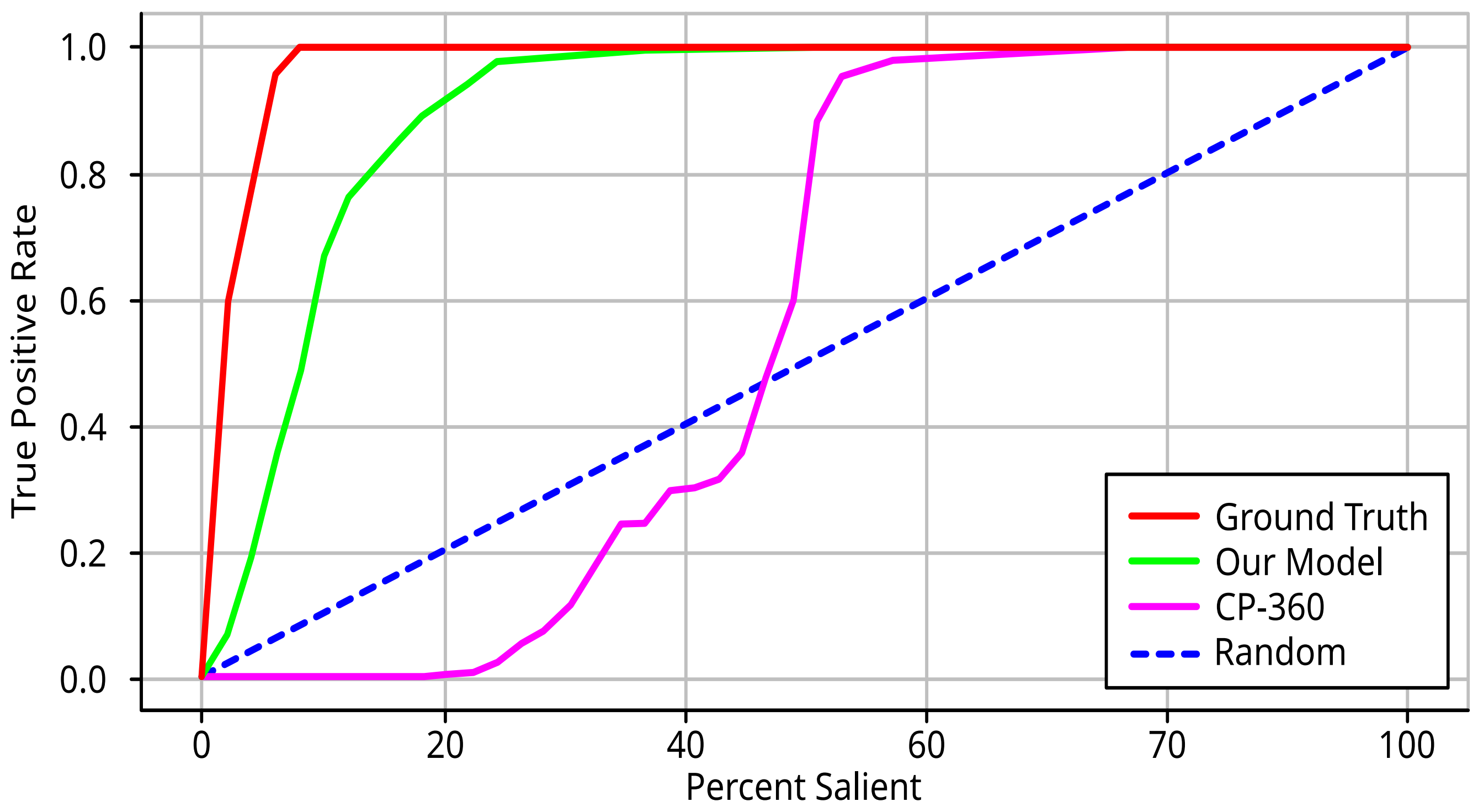
Predicted Blended

Predicted

Evaluation

We compare our model with a current **state-of-the-art** model [4] by measuring a series of metrics for the generated saliency maps and the ground truth saliency maps of the **Sports360 dataset** [2], outperforming it for **all metrics measured**.

	SIM↑	CC↑	KLD↓	NSS↑
OurModel	**0.3375**	**0.3048**	**6.3080**	**1.3870**
CP-360 [4]	0.2761	0.2338	8.3600	0.9515



Resulting Receiver Operating Characteristic (ROC) Curves. Our model produces a curve **closer to the ground truth** than the compared model, achieving **over 90% of recall with only 20% of the most salient regions**.

References

[1] ARNAB A., DEHGHANI M., HEIGOLD G., SUN C., LUČIĆ M., SCHMID C.: ViViT: A Video Vision Transformer. *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021
[2] XU Y., DONG Y., WU J., SUN Z., SHI Z., YU J., GAO S.: Deep 360 Pilot: Learning a Deep Agent for Piloting through 360° Sports Videos. IEEE Conference on Computer Vision and Pattern Recognition. 2017
[3] HU H., LIN Y., LIU M., CHENG H., HANG Y., SUN M.: Gaze prediction in dynamic 360° immersive videos. IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018
[4] CHENG, H., CHAO, C., DONG, J., WEN, H.K., LIU, T.L., and SUN, M.: Cube Padding for Weakly-Supervised Saliency Prediction in 360º Videos. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2018

This work was funded by the Government of Aragon's Departamento de Ciencia, Universidad y Sociedad del Conocimiento through the Reference Research Group "Graphics and Imaging Lab" (ref T34_23R)

Contact:
mvallejo@unizar.es

Graphics and Imaging Lab
Universidad de Zaragoza