Guess Future Anomalies from Normalcy: Forecasting Abnormal Behavior in Real-World Videos

APPENDIX

Snehashis Majhi[1,2,*], Mohammed Guermal[1,2,*], Antitza Dantcheva[1,2], Quan Kong[3], Lorenzo Garattoni[4], Gianpiero Francesca[4], François Brémond[1,2]

[1] INRIA [2] Côte d'Azur University [3] Woven by Toyota [4] Toyota Motor Europe

* Joint first authors.

Table 1. Overview of Appendix

	Content
Section 1	Implementation Details
Section 2	Category Wise Performance Analysis
Figure 1	Object Mask Extraction Pipeline
Figure 3-7	Detailed SoTA Architecture

1. Implementation Details

We implemented our model in pytorch. In order to obtain the pan-optic masks and object categories, we adapted a pipeline described in Figure 1. Next, in CLIP we extract the frame-level class tokens from the output of Image and Text encoders to obtain $512D$ feature vectors for scene, object, and text representatives such as F_S, F_O, F_{txt}. For all usage of functional operators across SIaT, we have the following hyperparameters. In both **f(x)** and **g(x)**, where $x \in \{co., fi., AQ\}$, we have 3 layers of multi-head self attention with 8 parallel heads followed by feed forward network (FFN) 128 output neurons. Similarly, in cross attention (CA) functional operator has 3 layers of multi-head cross attention with 8 parallel heads followed by the feed-forward network (FFN) 128 output neurons. For the uncertainty estimator (α), we use a multi-layer perceptron (MLP), that has input, hidden, and output layers with 128, 64, 1 neuron, and the output is `sigmoid` activated. Next, for the `classifier` in the future decoder is MLP with input and output layer has 256 and 26 neurons, where 26 is the number of abnormal categories in our AHB-F dataset. The proposed framework is end-to-end trainable excluding the backbone CLIP image and text encoders. We train using Adam optimizer with a learning rate of 0.0001. The loss weighting factors are set to $\lambda_1 = \lambda_2 = 1$ and. We also randomly select 8 videos as a mini-batch and compute the gradient. Then the loss is computed and back-propagated for the whole batch. We train the model up to 100 epochs on a single 2080Ti GPU.

Additionally, for the short and long-term evaluation, we have sampled the videos in 32 frames per second to represent 32, 64, 96, 128, 256 frames as 1st, 2nd, 3rd, 4th, and 8th seconds respectively. In our case frames present in 1st-to-3rd seconds account for short-term evaluation and frames present in 4th-to-8th seconds account for long-term evaluation. Next, we have re-implemented and evaluated the state-of-the-art (SoTA) methods like FUTR [2], OADTR [5], LSTR [6], TesTra [8], and JOADAA [3] in our AHB-F dataset and evaluation protocol. Thus, in Figure 3, 7, 5, and 6 we provide the architectural configurations of the SoTAs. Note that, the hyperparameters and model optimizations are kept the same as the original implementations.

2. State-of-the-art Anomaly Category-wise Analysis and Comparison

In this section, we provide an anomaly category-wise performance analysis and comparison of our SIaT with state-of-the-art (SoTA) methods in Long and Short-term future prediction *i.e. 1st second to 8th second*. It can be observed from Figure 2 that the average mAP of our method is best among all other SoTA on Human-to-Human anomalies (such as *fighting, arrest, chasing, shooting*) in both and long-term future prediction. However, on Human-to-Object interaction-based anomalies (such as shoplifting, and stealing) our method is second best after LSTR [6]. This is majorly due to the object of interest getting occluded in many cases and our method is slightly inferior in handling such conditions. Further, on Human-to-(Huamn&Object) interaction-based anomalies (such as Assault, Protest), our method along with others faces difficulties in anticipation. This shows that in the abnormal human behavior anticipation task there exists much room for improvement, thus our datasets and method will serve as a baseline for tasks to promote further research in this direction.

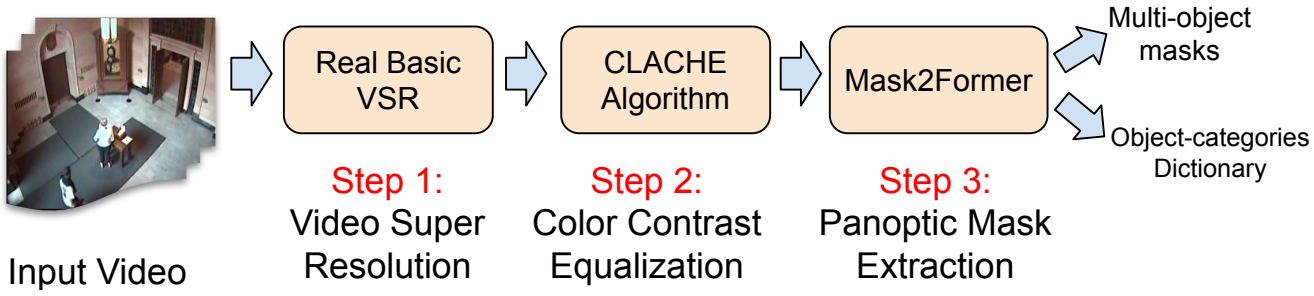

Figure 1. Typical pipeline adopted in our experiments for extracting good quality panoptic masks from low-resolution videos of our AHB-F datasets. Here, each video is sequentially processed to obtain the multi-object masks. In **step-1** a video super-resolution model (RealBasic VSR [1]) is applied on the input video to upscale the video to $4\times$ original resolution. Then in **step-2**, the resultant of step-1 is passed through CLACHE [7] algorithm for color contrast equalization and human body edge sharpening. Thanks to step-1 and step-2, in **step-3** yolov7-pose [4] extracts good quality multi-person key points which are used in SIaT.

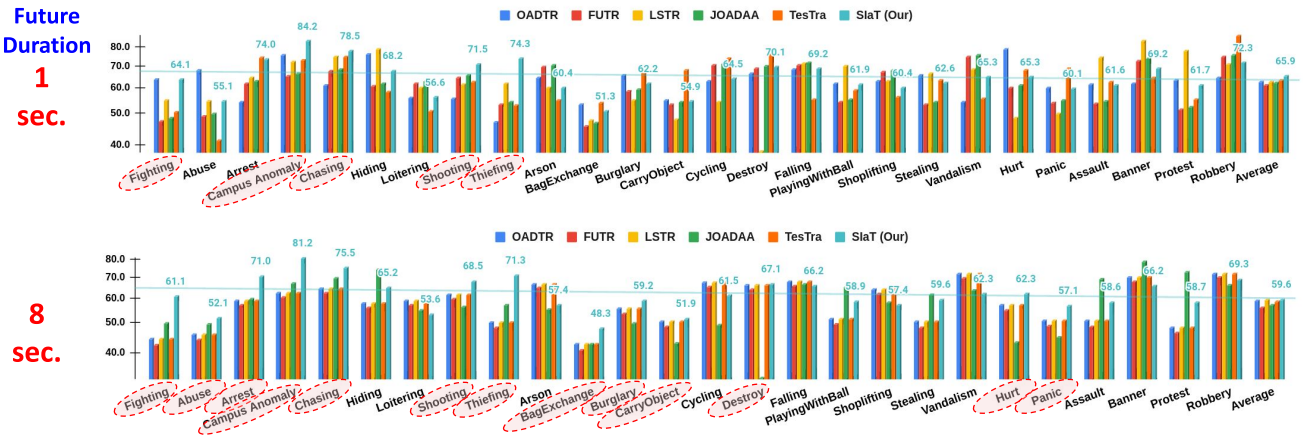

Figure 2. Category-wise anticipation performance on our AHB-F dataset. Categories highlighted with red dotted circles indicate our best performance on those.

References

[1] Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Investigating tradeoffs in real-world video super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[2] Dayoung Gong, Joonseok Lee, Manjin Kim, Seong Jong Ha, and Minsu Cho. Future transformer for long-term action anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3052–3061, 2022. 1

[3] Mohammed Guermal, Abid Ali, Rui Dai, and François Brémond. Joadaa: Joint online action detection and action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6889–6898, January 2024. 1

[4] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475, 2023. 2

[5] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Zhengrong Zuo, Changxin Gao, and Nong Sang. Oadtr: Online action detection with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7565–7575, 2021. 1

[6] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. *Advances in Neural Information Processing Systems*, 34:1086–1099, 2021. 1

[7] Garima Yadav, Saurabh Maheshwari, and Anjali Agarwal. Contrast limited adaptive histogram equalization based enhancement for real time video system. In *2014 international conference on advances in computing, communications and informatics (ICACCI)*, pages 2392–2397. IEEE, 2014. 2

[8] Yue Zhao and Philipp Krähenbühl. Real-time online video detection with temporal smoothing transformers. In *European Conference on Computer Vision*, pages 485–502. Springer, 2022. 1

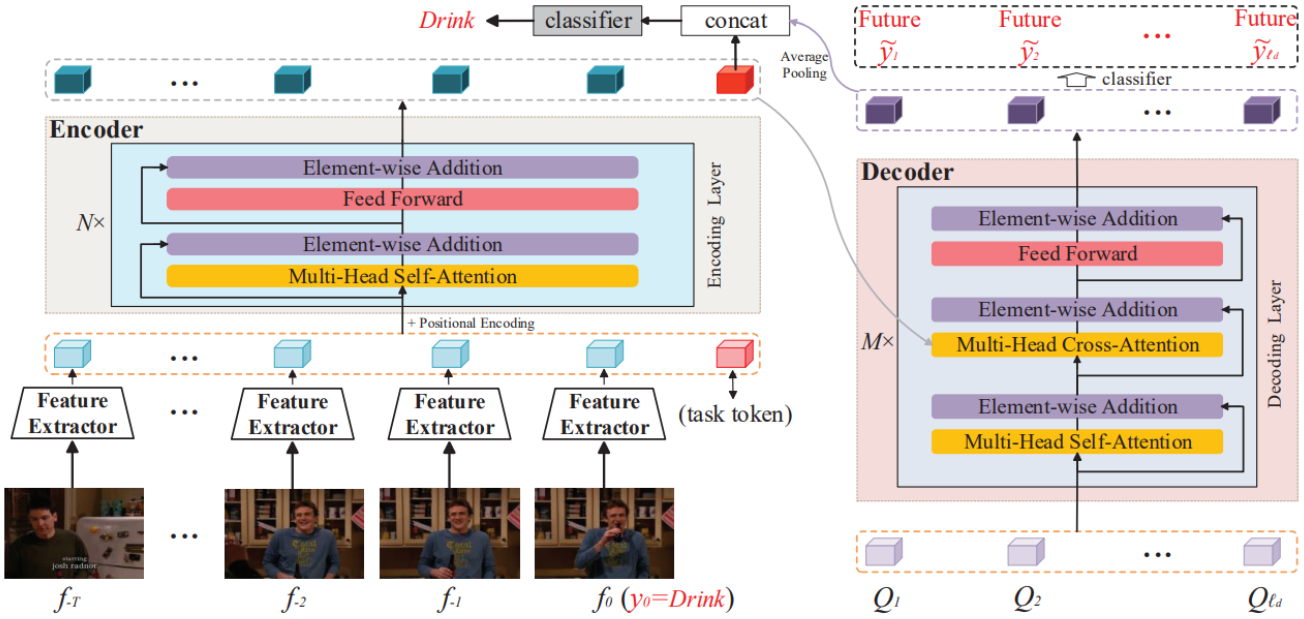

Figure 3. **OADTR:** Given an input streaming video $\mathbf{V} = \{f_t\}_{t=-T}^{0}$, a task token is attached to the visual features output by the feature extraction network. Then the token feature sequence is input into the standard Transformer's encoder to model long-range historical temporal dependencies. Afterward, the decoder of OADTR anticipates the future context information in parallel.. Note that OADTR, including the encoder and decoder, is an end-to-end parallel framework.

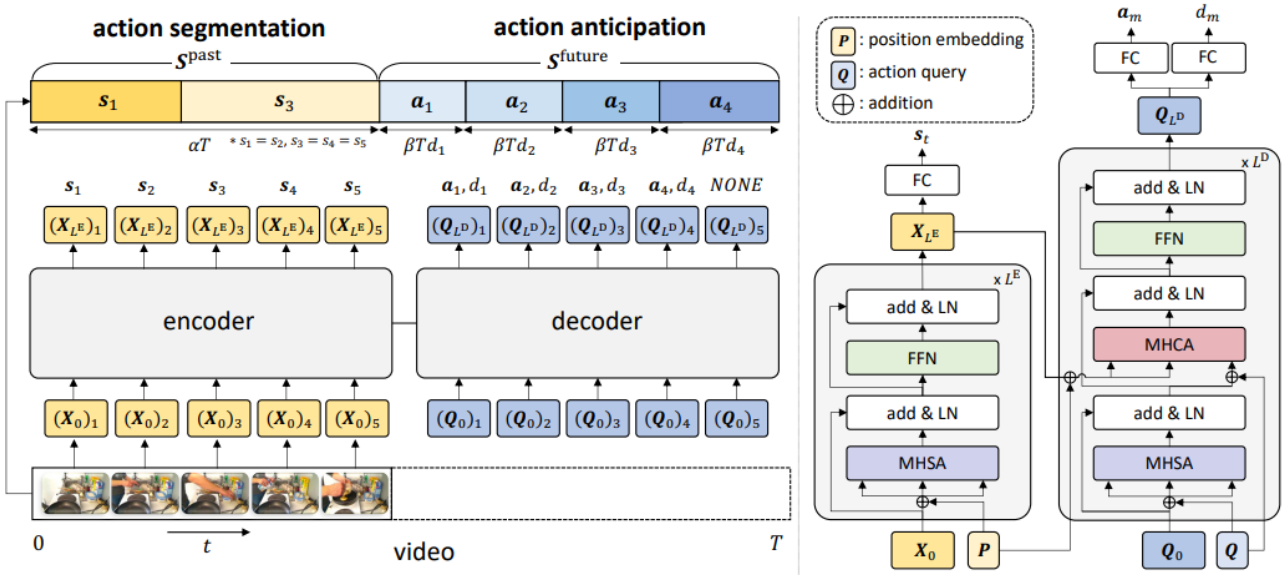

Figure 4. **FUTR:** It is an an end-to-end attention neural network to anticipate actions in parallel decoding, leveraging global interactions between past and future actions for long-term anticipation. FUTR is composed of an encoder and a decoder; each classifies action labels of past frames (action segmentation) and anticipates future action labels and corresponding durations (action anticipation), respectively. The encoder learns distinctive feature representation from past actions via self-attention, and the decoder learns long-term relations between past and future actions via self-attention and cross-attention. For simplicity, FUTR set the number of past frames αT as 5 and the number of object queries M as 5 in this figure. Note that $(X_l)_i$ and $(Q_l)_i$ indicate i^{th} index of X_l and Q_l, respectively.

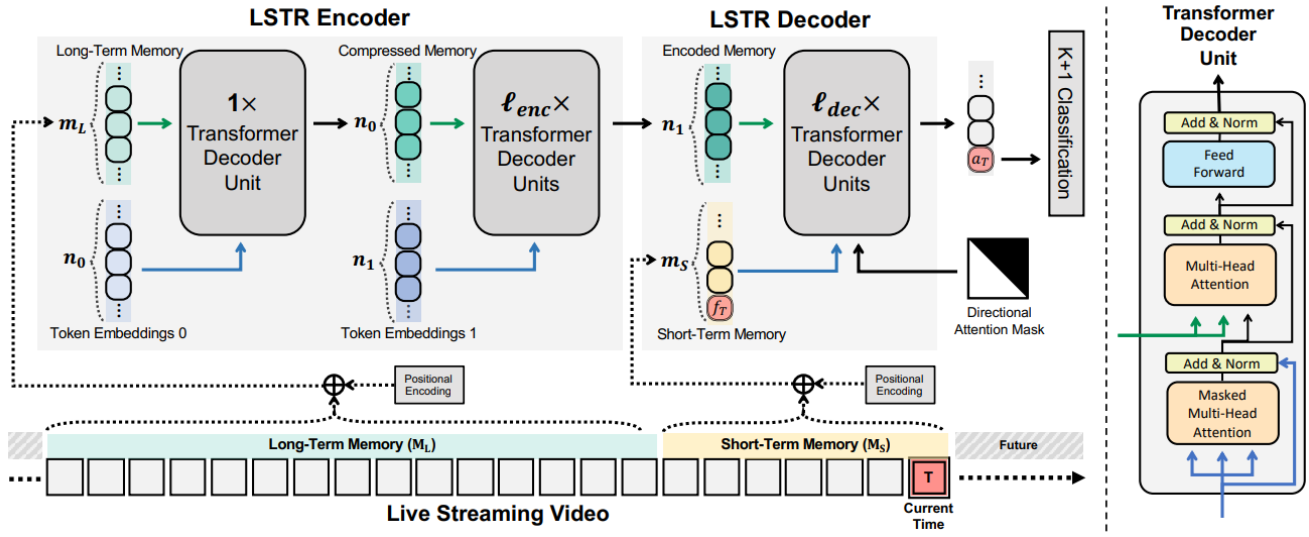

Figure 5. **LSTR:** It is formulated in an encoder-decoder manner. Specifically, the LSTR encoder compresses the long-term memory of size m_L to n_1 encoded latent features, and the LSTR decoder references related context information from the encoded memory with the short-term memory of size m_S for action recognition of the present. The LSTR encoder and decoder are built with Transformer decoder units which take the input tokens (dark green arrows) and output tokens (dark blue arrows) as inputs. During inference, LSTR processes every incoming frame in an online manner, absent future context.

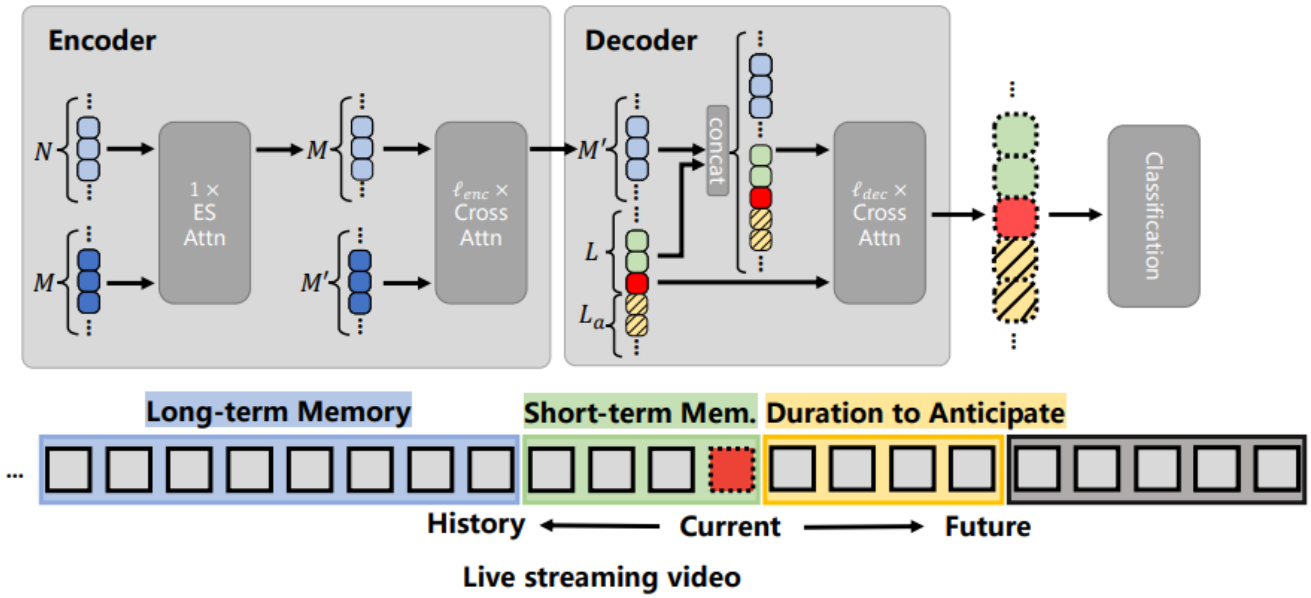

Figure 6. **TesTra:** The basic setup follows LSTR. A long-term memory compresses a long temporal history into M representative queries. A short-term attention mechanism uses compressed memory and a short history of frames to compute current and future actions. The main advantage of TeSTra is that the long-memory incurs only constant cost, and thus allows for much more efficient long-term reasoning.

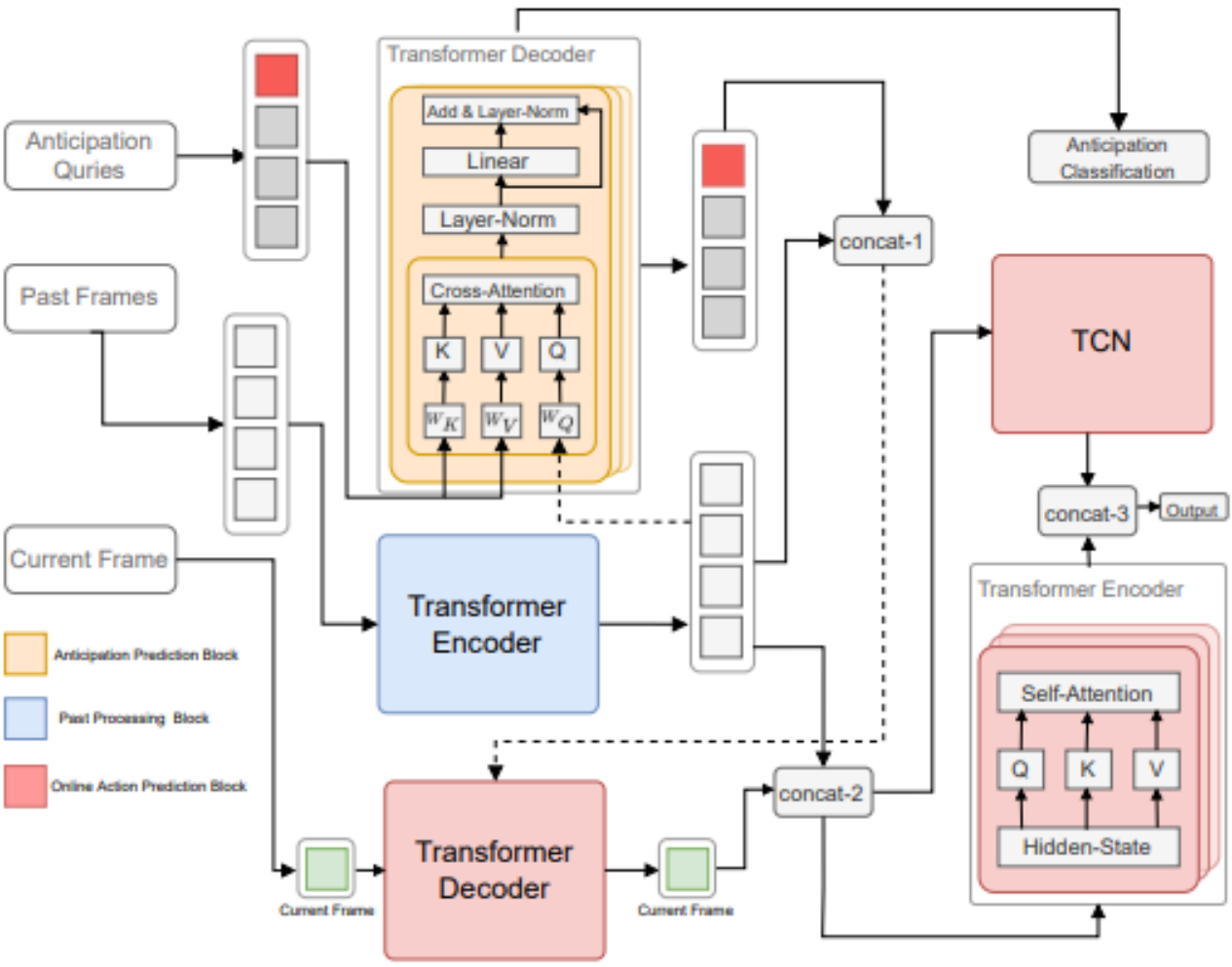

Figure 7. **JOADAA:** This architecture has three units i) Past processing: a short-term past transformer-encoder that enhances observation features, ii) Anticipation prediction:a transformer-decoder that anticipates the upcoming actions in the future frames, using embedding output from the past processing block and a set of learnable queries, and iii) Online Action prediction, to detect current ongoing action.