A Multimodal Transformer for Live Streaming Highlight Prediction

Jiaxin Deng[1,2], Shiyao Wang[3], Dong Shen[3], Liqin Zhao[3], Fan Yang[3], Guorui Zhou[3] and Gaofeng Meng[1,2,4]

[1] MAIS, Institute of Automation, Chinese Academy of Science

[2] University of Chinese Academy of Science, [3] Kuaishou Inc.

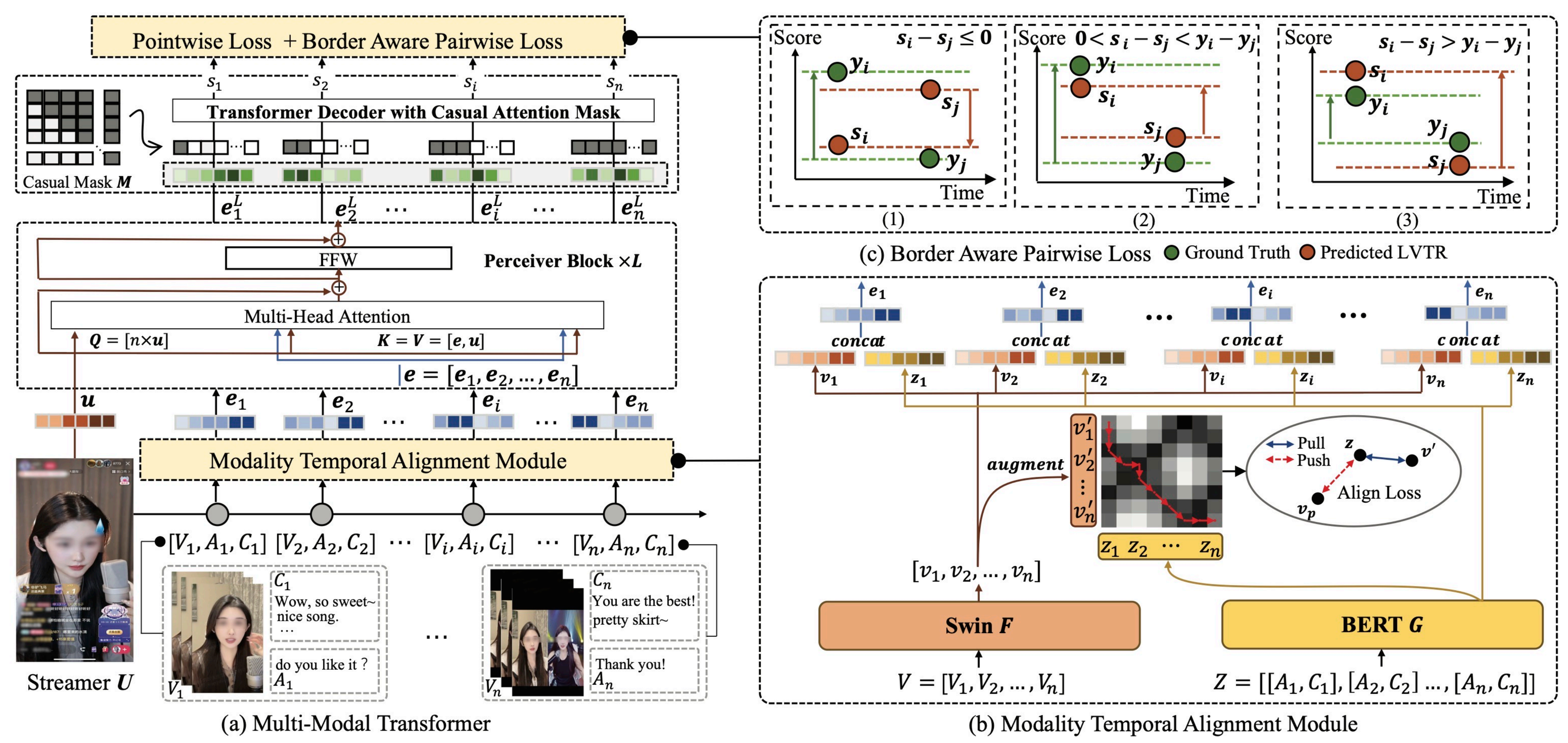
[4] CAIR, HK Institute of Science and Innovation, Chinese Academy of Science

Motivation

(a) Cascading UI (b) Typical Streamers

- Different from traditional video understanding task, live streaming highlight understanding tasks makes predictions only based on information available ***up until that moment***.
- Multimodal information in live streaming videos is usually ***misaligned***. For example, the reaction of hosts and audiences can experience a time lag, so the streamer's speech and audiences' comments may be ambiguous and not sequentially aligned with the visual frames, necessitating a module to mitigate the noise caused by misalignment.
- There is no large-scale public dataset for live streaming highlight detection and a large-scale live streaming dataset with multimodal information is crucial to assessing this topic.

Method



(a) Multi-Modal Transformer (b) Modality Temporal Alignment Module

We formulate the task as a prediction task based on historical look-back windows and the casual attention mask is proposed to avoid the information leakage from the future. Second, to alleviate the misalignment between visual and textual modality, we develop a novel ***Modality Temporal Alignment Module (MTAM)*** to address potential temporal discrepancies that may arise during live streaming events. Based on continuous label, we design a novel ***Border Aware Pairwise Loss*** with first-order difference constraints.

Modality Temporal Alignment Module

$$\mathcal{L}_{align} = -\log \frac{\exp\left(d_{\{z,v'\}}/\tau\right)}{\exp\left(d_{\{z,v'\}}/\tau\right) + \sum_{v_p^i \in \omega}^{N} \exp\left(d_{\{z,v_p^i\}}/\tau\right)} \qquad p^{video} = \mathrm{softmax}\left(\frac{D(z,v)_{ij}}{\gamma}\right), (i,j) \in \omega$$

Border Aware Pairwise Loss

$$L^1_{Pair} = \sum_{y_i > y_j} \log\left(1 + e^{-\sigma(s_i - s_j)}\right), (y_i - y_j) - (s_i - s_j) \geqslant 0$$

Experiments

TABLE I: Performances of different methods on KLive and PHD dataset

Methods		KLive Tau $\tau\uparrow$ $\Delta=0$	$\Delta=0.2$	$\Delta=0.4$	$\Delta=0.6$	PHD mAP $\uparrow$
VHD Methods						
Adaptive-H-FCSN [2]	[ECCV'20]	0.5782	0.5707	0.5511	0.5322	15.65
PR-Net [8]	[ICCV'21]	0.5848	0.5818	0.5461	0.5403	18.66
PAC-Net [4]	[ECCV'22]	0.5823	0.5845	0.5537	0.5409	17.51
ShowMe [3]	[MM'22]	0.5798	0.5705	0.5348	0.5407	16.40
LSHD Methods						
AntPivot [5]	[arXiv'22]	0.5818	0.5809	0.5483	0.5421	-
KuaiHL	[Ours]	**0.5961**	**0.5871**	**0.5686**	**0.5563**	**21.89**

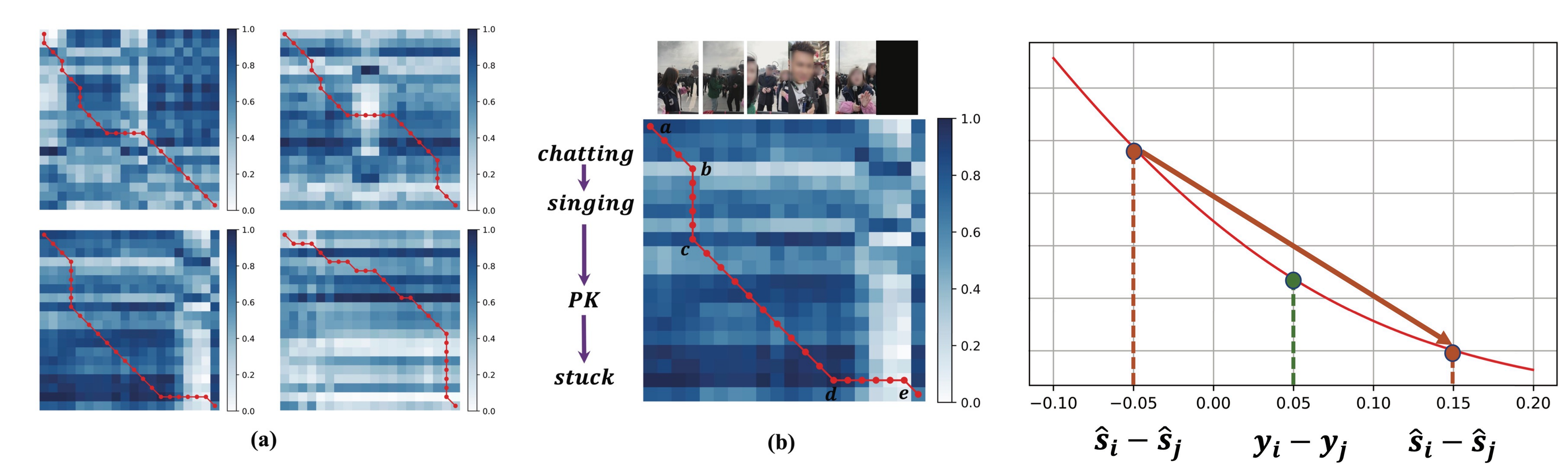
- KuaiHL surpass various strong VHD and LSHD methods.
- Modality Temporal Alignment Module does help train better visual and text encoders that reduce the possible misalignment between the two.

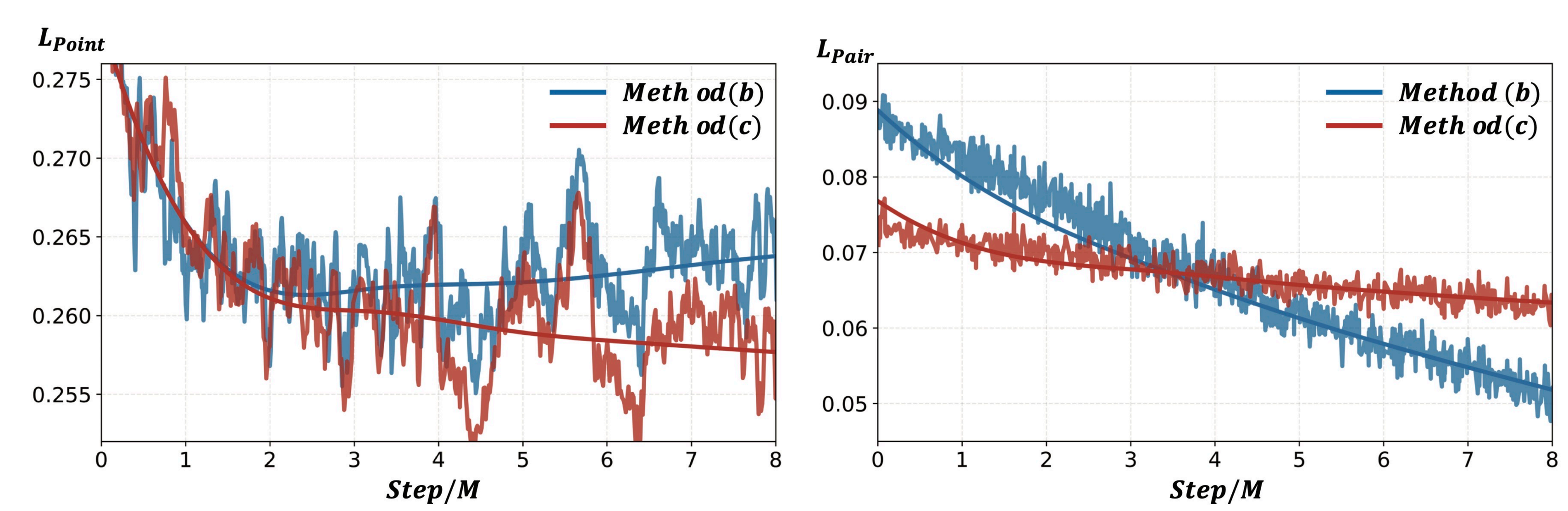
TABLE II: Ablation study of KuaiHL with different loss functions on KLive dataset.

Methods	L_{Point}	L^0_{Pair}	L^1_{Pair}	L^2_{Pair}	L^3_{Pair}	L'_{align}	L_{align}	Tau τ
(a)	✓	-	-	-	-	-	-	0.5761
(b)	✓	✓	-	-	-	-	-	0.5857 ↑ 0.96%
(c)	✓	-	✓	-	-	-	-	0.5872 ↑ 1.11%
(d)	✓	-	-	✓	-	-	-	0.5256 ↓ 5.05%
(e)	✓	-	-	-	✓	-	-	0.5824 ↑ 0.66%
(f)	✓	-	✓	-	-	✓	-	0.5919 ↑ 1.58%
(g)	✓	-	✓	-	-	-	✓	**0.5961** ↑ 2.00%

TABLE III: Ablation study on different modality impact.

Model	v	a	x	u	c	Tau τ	mAP(%)
KLive dataset							
KuaiHL	✓	✓	✓	✓	-	**0.5961**	-
KuaiHL w/o item	✓	✓	✓	-	-	0.5910 ↓ 0.51%	-
KuaiHL w/o text	✓	-	-	✓	-	0.5760 ↓ 2.01%	-
KuaiHL w/o visual	-	✓	✓	✓	-	0.5489 ↓ 4.72%	-
PHD dataset							
KuaiHL	✓	-	-	-	✓	-	21.89
KuaiHL w/o visual	-	-	-	-	✓	-	19.55 ↓ 2.34%
KuaiHL w/o caption	✓	-	-	-	-	-	20.06 ↓ 1.11%





- Border Aware Pairwise Loss helps model to effectively exploit the contrastive information between the highlight and no-highlight frames and avoids the collapse due to the over optimization.