Leveraging VLP models for cross-modal video retrieval

Yuqi Liu, Zihao Yue, Qin Jin
AIM3, Renmin University of China

Code is available:
https://github.com/yuqi657/ts2_net

Introduction

- Our method is based on recent image-language pre-trained model CLIP.
- We make adaptions to the origin visual transformer, to leverage VLP for video retrieval tasks.

Results

Main Results:

Team	Run #	xinfAP
C_D_RUCAIM3-Tencent.22	2	**0.175**
C_D_RUCAIM3-Tencent.22	1	0.119
C_D_RUCAIM3-Tencent.22	3	0.109
C_D_RUCAIM3-Tencent.22	4	0.094

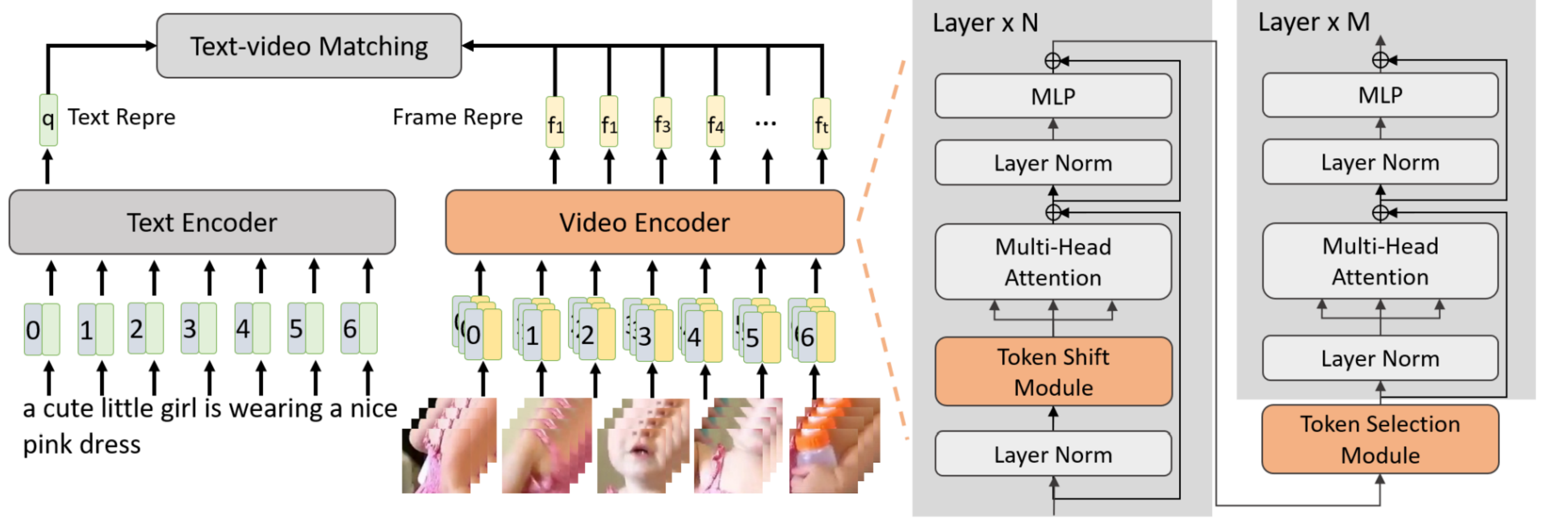
Novelty Score:

Team	Novelty	xinfAP
N_D_VIREO.22_6	38.4	0.088
C_D_RUCAIM3-Tencent.22_2	31.4	0.175
C_D_WasedaMeiseiSoftbank.22_2	26.7	0.282
C_D_kindai_ogu_osaka.22_1	23.1	0.199
C_D_RUCMM.22_2	18.4	0.262
C_D_ITI_CERTH.22_2	15.0	0.210
C_D_CamiloUchile.22_3	5.8	0.002

- Our best xinfAP is 0.175.
- Our method retrieves more unique relevant shots.

Method

Main Architecture:

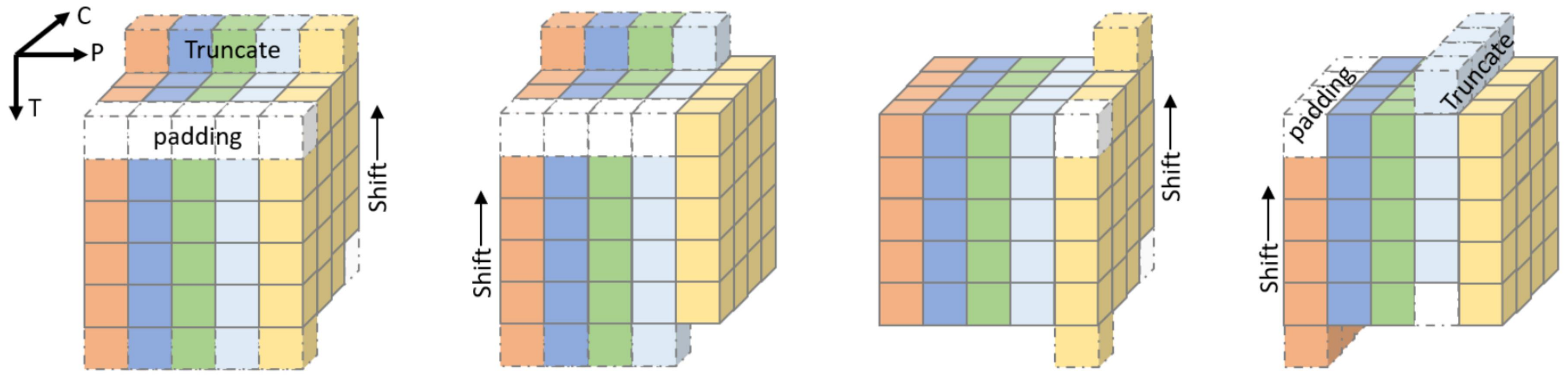


TS2-Net

Token Shift Transformer

Token Selection Transformer

Token Shift Module:



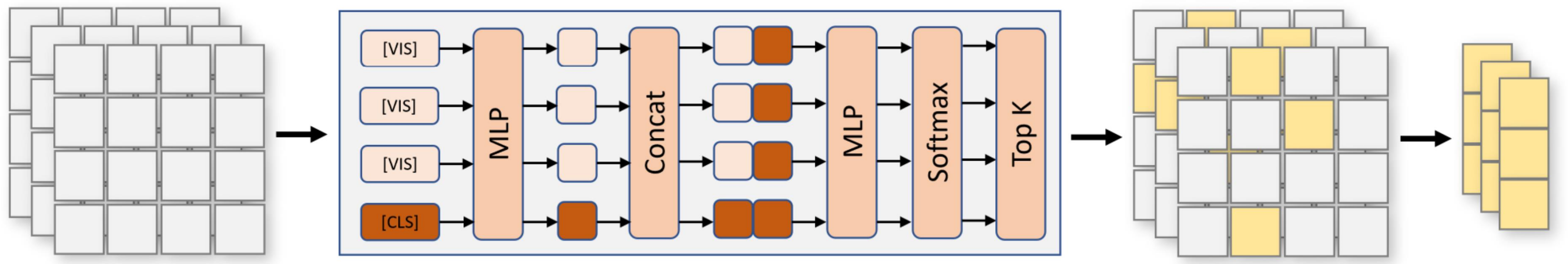
(a) Channel Temporal Shift

(b) [VIS] Channel Temporal Shift

(c) [CLS] Channel Temporal Shift

(d) Token Temporal Shift

Token Selection Module:



Tokens in Video Frame

Token Selection Module

Top K Tokens

Token Repre

Matching, Training and Inference:

Frame level sim: $s_i = \frac{q \cdot f_i}{\|q\| \, \|f_i\|}$, Video level sim: $s = \sum_{i=1}^{n} \alpha_i s_i$

Training Loss:

$$\mathcal{L}_t^{t2v} = -\frac{1}{B}\sum_{i}^{B} \log \frac{\exp\left(\tau \cdot \operatorname{sim}\left(q_i, v_i\right)\right)}{\sum_{j=1}^{B} \exp\left(\tau \cdot \operatorname{sim}\left(q_i, v_j\right)\right)}$$

$$\mathcal{L}_t^{v2t} = -\frac{1}{B}\sum_{i}^{B} \log \frac{\exp\left(\tau \cdot \operatorname{sim}\left(q_i, v_i\right)\right)}{\sum_{j=1}^{B} \exp\left(\tau \cdot \operatorname{sim}\left(q_j, v_i\right)\right)}$$

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{t2v} + \mathcal{L}_{v2t}\right)$$

Analysis

Queries with great results:

704 A parked white car
726 Two teams playing a game where one team have their players wearing white t-shirts

Queries with bad results:

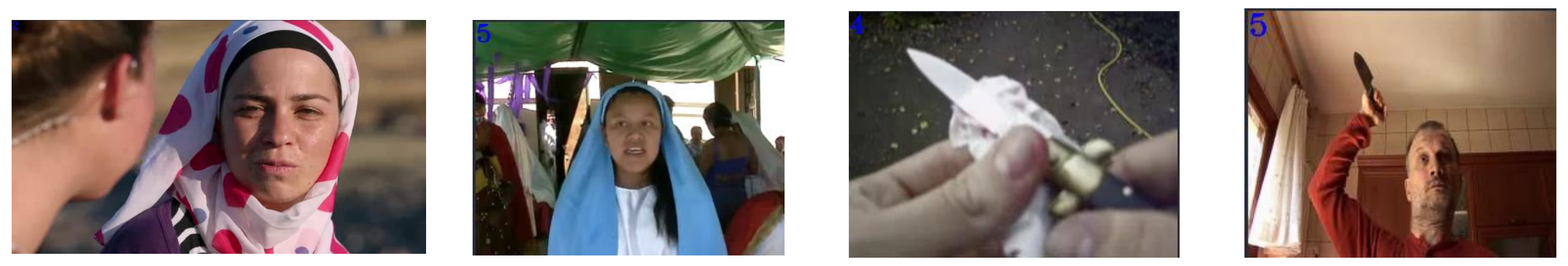
702 A room with blue wall
713 A kneeling man outdoors
728 Two adults are seated in a flying paraglider in the air

- Our method retrievals the correct elements in each query (e.g. blue, wall, paraglider, etc), but fails to model relations between these elements.
- Some elements are difficult to identify (e.g. kneeling). This might be caused by the domain of pre-trained data.

Puzzle

Q1: Are there some annotation errors or misunderstanding of some concepts?

711 A woman wearing a head kerchief (0 in top10)
730 A man is holding a knife in a non-kitchen location (3/10)



Q2: Is there possible to design a metric considering both xinfAP and diversity?