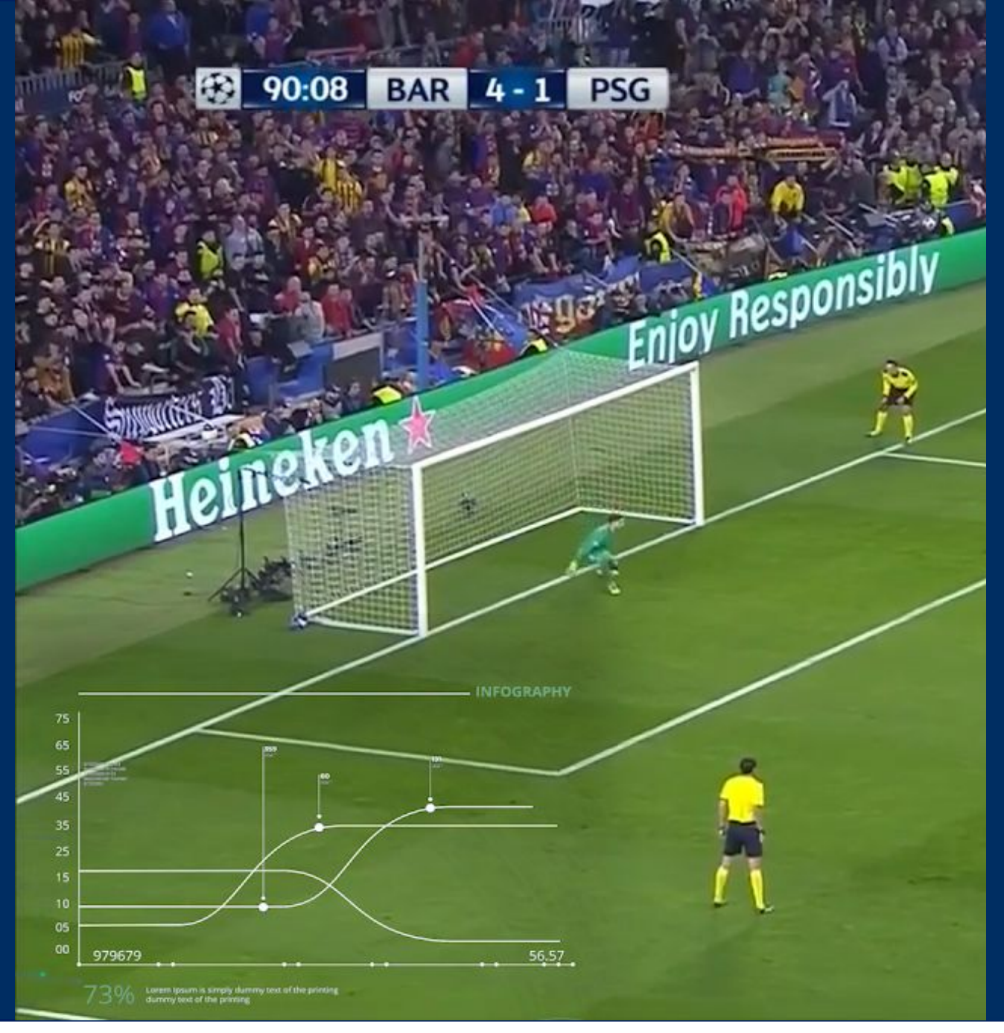
A Small Video-Language Model for Sparse Action Spotting

Evan Zimmerman, Naveen Unnikrishnan, Nicholas Mellon, Pranav Mallela, Vishal Chandra

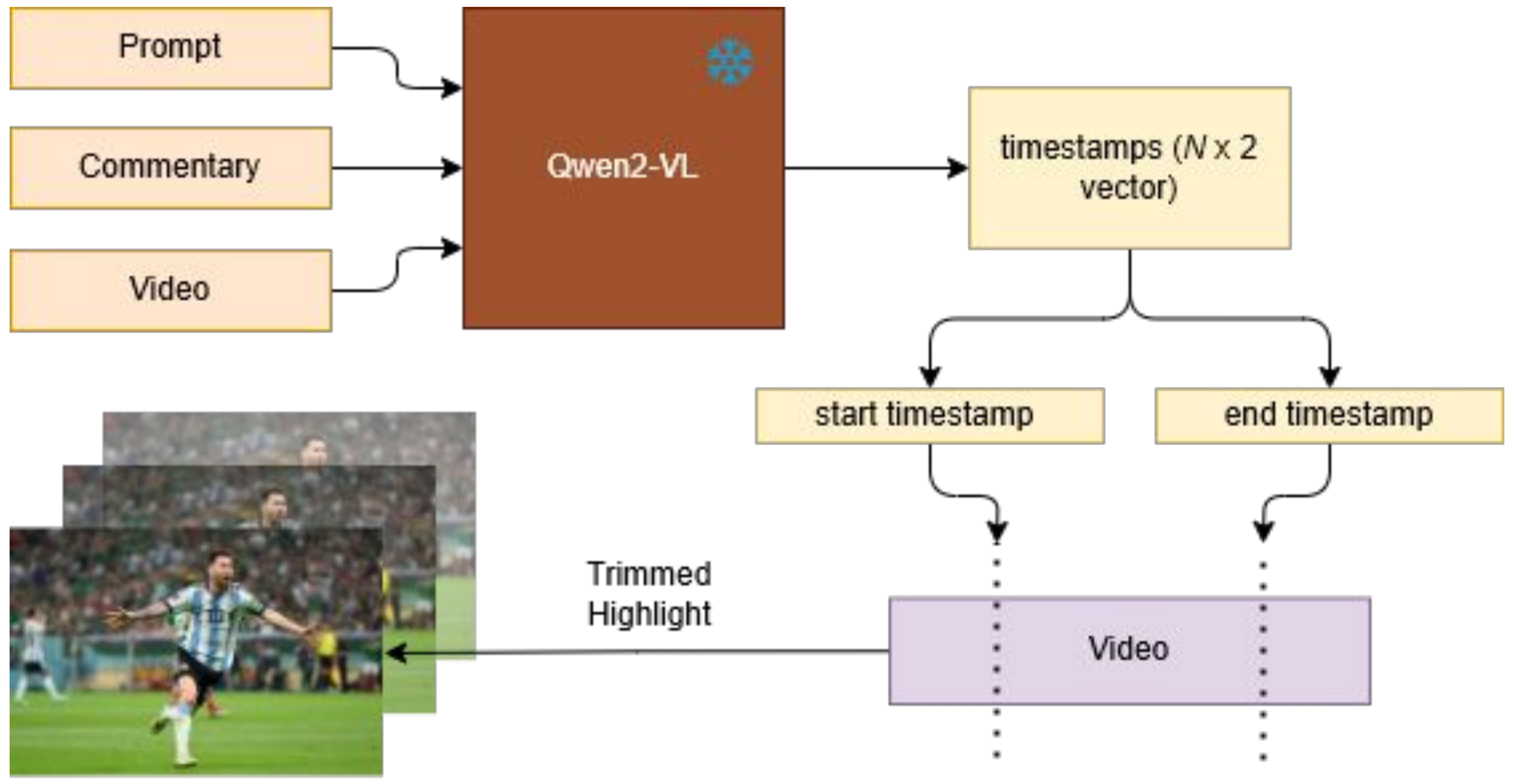
EECS 545 | Machine Learning | Poster 7



UNIVERSITY OF MICHIGAN

Motivation

- Sparse action spotting methods in sports have traditionally been CNN or feature-extractor based
- We introduce a purely token-based pipeline that, when given a video of a sports game, can extract desirable highlights.
- A user should be able to query our system: *"Give me all the highlights of free-kicks."*, and receive a response: *"A free-kick occurs at (36:09, 36:24)"*.



Data

- The **SoccerNet v2 Action Spotting Dataset**:
  - **500 videos** from past soccer games.
  - Annotations for **17** common soccer actions such as **Penalty, Goal, Substitution, and Offside**.
  - Each action annotated with **in-game timestamps**.
- **SoccerNet Echoes commentary** transcript dataset.
- Split into **1-minute video chunks** with attached **commentary** and ground-truth actions:

Train Split	23151 action-clip pairs
Test Split	5788 action-clip pairs

Methods

Approach 1 : VLM Fine-Tuning

1. Load `Qwen2-VL-2B-Instruct-AWQ`
2. Fine-tune (train) using ~8000 training videos
   - Format a prompt with video and timestamps
3. For each example in the test set, we:
   - Format a prompt with video & no timestamps
     - The VLM generates the timestamps for us after learning from the examples in the fine-tuning

Approach 2: LLM Out-of-the-box

1. Load `Mistral-7B-v0.3`
2. For each example in the test set, we:
   - Set a system role (ie. Assistant)
   - Set a user prompt
     - Zero-shot or few-shot examples
     - Instructions for extracting timestamps
     - Commentary for the selected game segment

Approach 3: LLM Fine-tuning

1. Load `Mistral-7B-v0.3`
2. Fine-tune (train) using ~8000 training examples
   - Format a prompt - commentary and timestamps
3. For each example in the test set, we:
   - Format a prompt - commentary & no timestamps
     - The LLM generates the timestamps for us after learning from the examples in the fine-tuning

Approach 4 : Open AI Reasoning API

1. Use the `o4-mini` reasoning API
2. For each example in the test set, we:
   - Format a prompt with commentary & no timestamps
     - The reasoning model generates timestamps

Results

Model	Precision	Recall	F1 Score
`Qwen2-VL-2B`	**0.0662**	**0.4045**	**0.1138**
`Mistral-7B`	0.0551	0.3364	0.0947
`o4-mini`	0.0345	0.1887	0.0583

Model	Action 1 (F1)	Action 2 (F1)	Action 3 (F1)
`Qwen2-VL-2B`	shots on target (0.1792)	foul (0.1572)	free-kick (0.1201)
`Mistral-7B`	goal (0.1778)	clearance (0.1425)	corner (0.1085)
`o4-mini`	goal (0.3333)	substitution (0.2667)	corner (0.1714)

Metric: Intersection over Union (IoU)

- The best IoU between ground truth time intervals and predicted time intervals is taken.

Future Work

- Integrate context-aware losses used in previous SOTA CNN-based methods
- Investigate the role of ResNet features in current transformer-based SOTA method.
- Understand token-based versus CLIP-based approaches to sparse action spotting (LG AI)