Highlight Detection from Video, Audio, and Text Prompts

Mathis Doutre, Srikar Viswanatha, Hang Kim, Eugénie Laugier

Motivation

- Video highlights are used by coaches and media
- Soccer matches are long and full of repetitive events

Existing Approaches

- Existing methods are unimodal or weakly multimodal
- No personalization

Research Gap: No prior work of query-based highlights using video, audio and transcript together

Problem Statement

Goal: Retrieve soccer highlights *conditioned on natural-language queries*

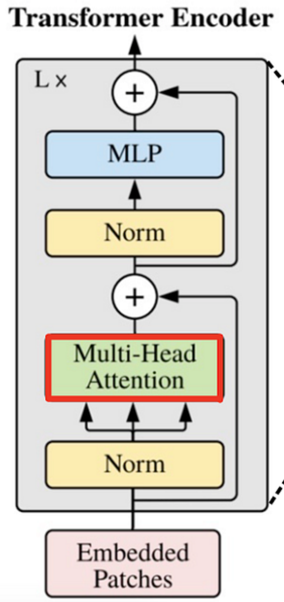
→ Modified endpoint (due to time constraints): **Event classification** as proxy

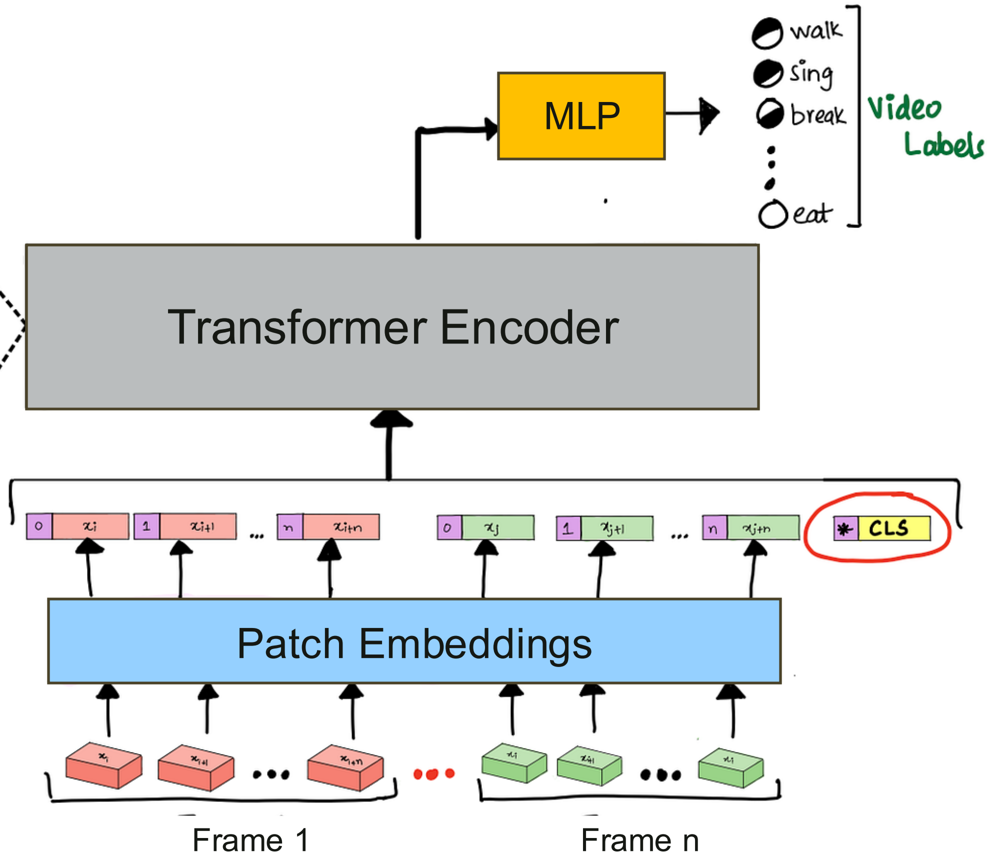
- Dataset: SoccerNet (700h of matches)
- Challenge: Sparse events & multimodal misalignment (video, audio, and transcript)

Implementation

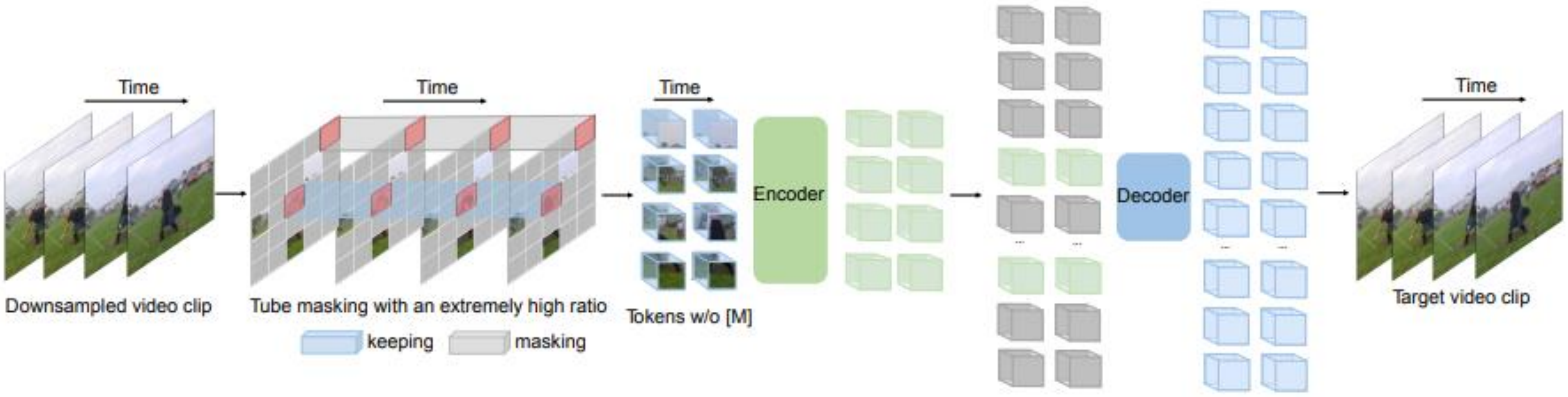
Video Encoder

- Input: short video clip → output: latent embedding
- **Video Transformer ➔** Vision Transformer (ViT) extended to time domain
- Uses spatio-temporal attention to model motion and appearance





Pretraining: VideoMAE

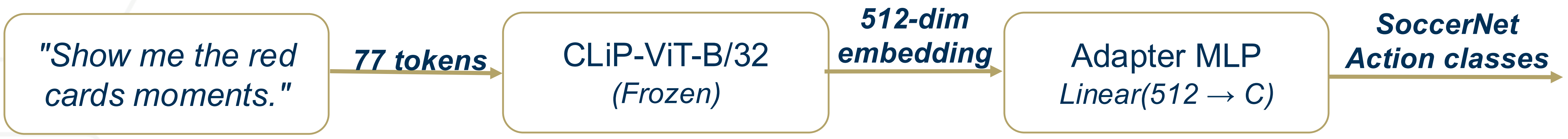


- Self-supervised pretraining (Video MAE) – learns from full video without labels.
- Uses masked autoencoding → model reconstructs missing patches
- Domain adaptation: Pretrain directly on SoccerNet data
- Clustering metrics improved after pretraining:
  Silhouette score 0.42 → 0.58, DB index 2.31 → 1.39

Audio Encoder

- **Wav2Vec2** audio encoder --> 1024-dim embeddings, Whisper transcription --> 382-dim embedding
- Passed these into a combined audio encoder, then trained for classification, 29.6% classification accuracy
- Audio contains meaningful class information, however noisier than video

Query Encoder



- **CLiP-ViT-B/32** text encoder --> 512-dim embeddings
- Few shot fine-tuning via a MLP layer on synthetic prompts for each SoccerNet class
- Top-1 accuracy improved: **68%** with raw CLIP --> **77%** after fine-tuning

Georgia Tech

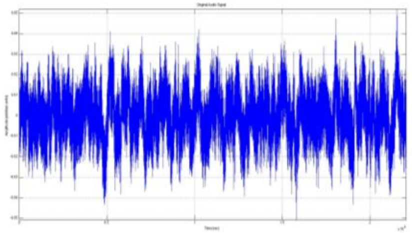
Fusion Transformer Architecture

- Overall Multimodal Pipeline

VideoMAE embeddings
768-D sequences

Not this. It's back to Raish now. He's got Shiru in support. Lovely back heel. Oh, and Olivia Giru with a very, very classy finish.
Well, that was a goal that just got Marco straight into the wall once again.
Marco wrestle.
Wow.
[Applause]
Well, we talked about the shot

MPNet 's SentenceTransformer
768-D



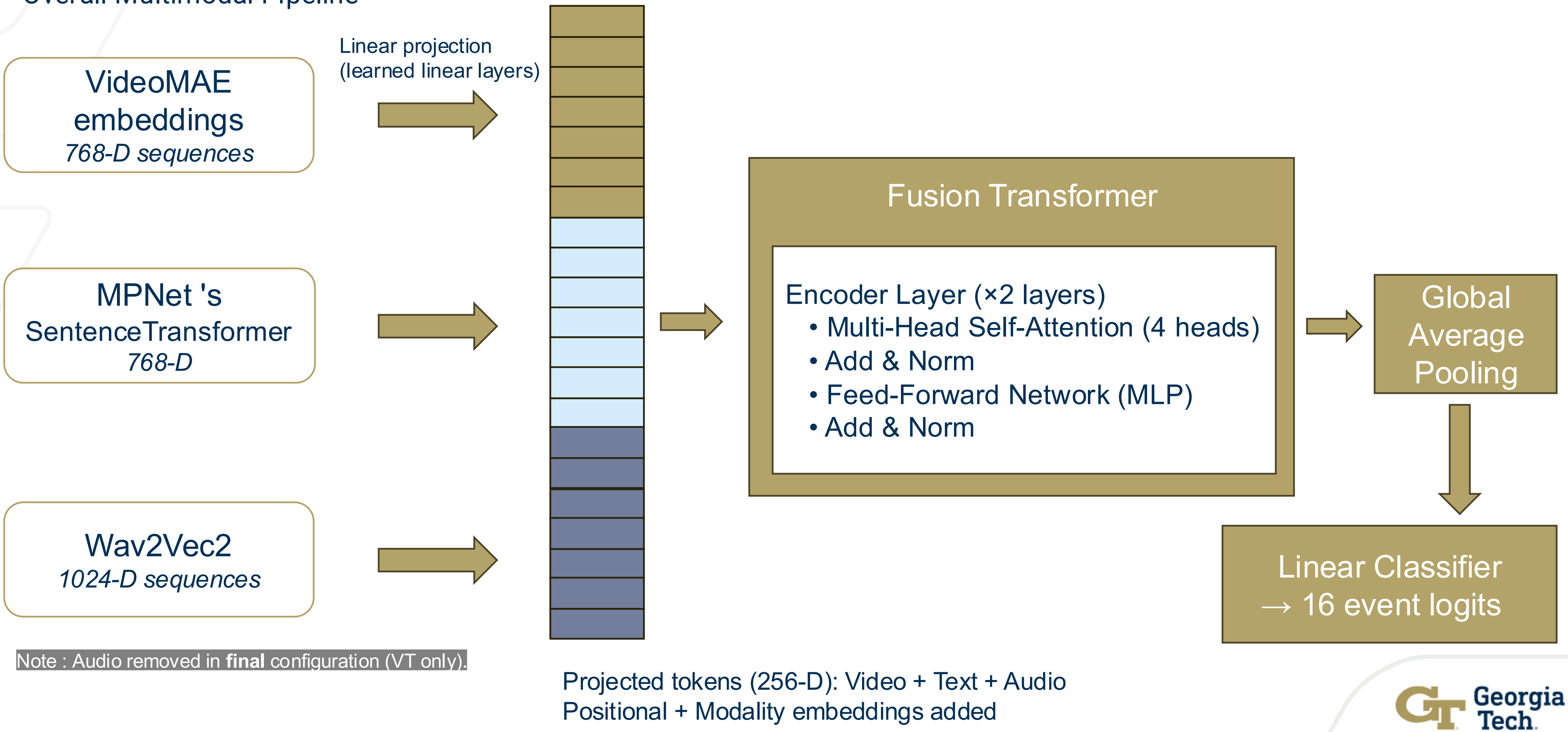
Wav2Vec2
1024-D sequences

Frozen

Encoders are frozen → only projection layers + fusion transformer are trained

Fusion Transformer Architecture

- Overall Multimodal Pipeline



Fusion Transformer : Experiments and Results

- **We compare all modality combinations:**
  - **3 modalities** available : T = Transcript (MPNet), V = Video (VideoMAE), A = Audio (Wav2Vec2)
  - We test all combinations to understand **which modalities actually contribute** to event classification.
  - We evaluate :
    - Unimodal: T, V, A
    - Bimodal: VT, VA, TA
    - Trimodal: VTA
  - The goal is to detect whether adding modalities improves or hurts accuracy
- **Results :**

Model	T	V	A	VT	VA	TA	VTA
Accuracy	0.527	0.658	0.293	**0.747**	0.662	0.516	0.740

  - V > T > A as unimodal baselines
  - VT is the best overall (0.747) -- Adding Audio hurts performance (VT > VTA)
  - Audio (A) is weaker than the transcript (T) (0.293 < 0.658)
    - Wav2Vec2 only captures noisy acoustics (crowd noise, commentary variation)
    - Whisper transcripts encode clear semantic cues, aligned with event labels (more discriminative)

Georgia Tech

Fusion Transformer : Experiments and Results

- **Transcript + Video is the best multimodal pair** → We therefore **drop A** and focus on **VT only**
  - **Strong class imbalance (e.g., 1676 BOO vs 118 Goals)**
  - **Video** resolves visually distinct events: corners, throw-ins
  - **Transcript** provides explicit semantics: foul, ball out of play, goals

- Counts show strong class imbalance (e.g., 1676 BOO vs 118 Goals), which impacts precision.

→ Classes with weak transcript alignment (Goal, Kick-off) degrade under fusion.

- → confirm complementary roles of semantics & vision

Class	# of events	Text	Video	Fusion
Ball out of play	1676	0.809	0.834	**0.868**
Throw-in	1017	0.512	0.808	**0.837**
Foul	611	0.504	0.568	**0.799**
Indirect free-kick	487	0.260	0.375	**0.583**
Clearance	433	0.247	0.588	**0.600**
Shots on target	299	0.311	**0.426**	0.279
Shots off target	265	0.471	0.412	**0.686**
Corner	249	0.312	0.729	**0.833**
Goal	118	0.333	**0.500**	0.292
Kick-off	126	0.125	**0.469**	0.344
Total / Avg accuracy	5816	0.527	0.658	**0.747**
Δ to *Fusion*		-0.220	-0.089	**-**

Per-class results

Georgia Tech

Discussion & Conclusion

- Video + Audio transcripts provided best results
- CLIP had good accuracy, not integrated into Fusion model
- Fusion improved most event classes compared to unimodal baselines

- Limitations
  - CLIP embeddings not used, no full query-conditioned retrieval
  - Slow embedding extraction, class imbalance limited experiment scope

- Future Work
  - Integrate CLIP to Fusion Transformer
  - Improve speed of audio + video embedding extraction
  - Extend from event classification to full highlight retrieval