15th Israel Machine Vision Conference (IMVC) 2026
Monday | April 27, 2026 | Pavilion 10, EXPO Tel Aviv

Capturing the Moment: A Multimodal Framework for Precise Temporal Localization of Emotions in Live Broadcast Streams

Eyal Hanania, WSC-Sports

Laughter is a social signal, yet most datasets and models treat its detection as a coarse, clip-level classification task. This fails to capture the precise temporal boundaries (the onset and offset) of brief, sporadic laughter events. We address this with a lightweight framework for fine-grained temporal localization of laughter using complementary audio and visual cues, learning precise temporal structure from clip-level labels alone, with no frame-level annotations during training.

The architecture leverages frozen, pre-trained audio (HuBERT) and visual (MAE) encoders. A temporal softmax pooling mechanism computes a learned attention distribution over timesteps, isolating brief affective peaks from neutral context under weak supervision. To resolve modality conflicts - such as audible laughter without facial expression, or silent chuckles - an adaptive gating mechanism dynamically weights each modality by per-instance reliability. The framework operates in linear time, avoiding the quadratic cost of self-attention and scaling to continuous, large-scale video.

To support this task, we release two benchmarks, UR-FUNNY-Temporal and SMILE-Temporal, with manual onset/offset annotations for 11,053 videos (78.8 hours). Our method substantially outperforms multimodal foundation models, including Gemini and Qwen2.5-Omni, on precise temporal grounding. Beyond detection, precise temporal markers yield a 227% CIDEr improvement on downstream video laugh reasoning and let a smaller language model surpass a larger baseline. These results show that task-specific temporal grounding is a superior alternative to model scaling for understanding social signals.

Code and data: https://github.com/WSCSports/MTLLFM-temporal-laughter-localization