MTLLFM: Multimodal-Temporal Laughter Localization: UR-FUNNY-Temporal and SMILE-Temporal Benchmarks with an Adaptive Multimodal Fusion Model

Eyal Hanania, Nadav Kirsch, Daniel Arkushin, Jonathan Benvenisti, Amos Bercovich, Elie Zemmour, Sahar Froim

WSC-Sports

CVPR JUNE 3-7, 2026 DENVER COLORADO

Introduction & Motivation

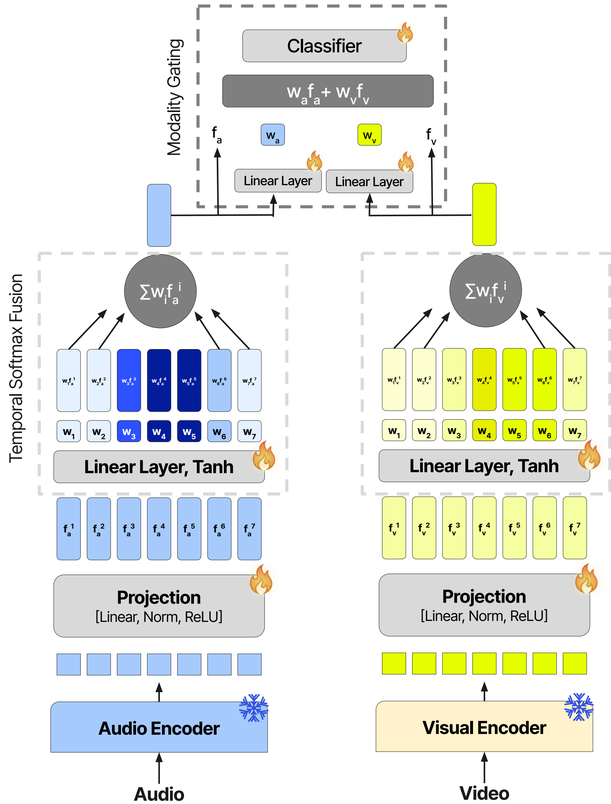
- **The Problem:** Laughter is a fundamental social signal, yet most current datasets and models treat detection as a coarse, clip-level classification task.
- **The Consequence:** This formulation fails to capture the precise **temporal boundaries** (onset and offset) of brief, sporadic laughter events.
- **The Solution:** We introduce **MTLLFM**, a lightweight framework designed for fine-grained temporal grounding using complementary audio and visual cues.
- **Weak Supervision:** Our model learns precise temporal structure from **clip-level labels alone**, requiring no frame-level annotations during training.
- **New Benchmarks:** To support this task, we release **UR-FUNNY-Temporal** and **SMILE-Temporal**, featuring manual onset/offset annotations for 11,053 videos (78.8 hours).

MTLLFM Methodology

- **Multimodal Encoders:** Our architecture leverages frozen, pre-trained audio and vision encoders to extract robust representations.
- **Temporal Softmax Pooling:** We employ a lightweight, saliency-driven aggregation mechanism that computes a learned attention distribution over timesteps. This allows the model to isolate brief "affective peaks" from neutral background context under weak supervision.
- **Adaptive Modality Gating:** To resolve modality conflicts - such as audible laughter without facial expression or silent chuckles- our gating mechanism dynamically computes modality weights (wa, wv) based on per-instance reliability.
- **Computational Efficiency:** The framework operates in O(T) time complexity, avoiding the O(T^2) cost of pairwise self-attention, making it ideal for continuous, large-scale video analysis.

Experimental Results

- **SOTA Performance**: MTLLFM achieves **99% F1** and **68.1% precision** (IoU=0.5) on sports broadcast data.



- Vs. Foundation Models: Substantially outperforms **Gemini 3 Flash** and **Qwen2.5 Omni** in precise temporal grounding.
- **Multimodal Synergy**: Adaptive gating provides a **+7.7 point gain** over audio-only variants.
- **Attention Stability**: **Tanh Gating** prevents saturation in brief events, adding **+4.2 points** to localization precision.

Downstream Impact & Conclusion

- **Reasoning Boost**: Precise markers enable **GPT-3.5 to outperform the GPT-4.0** baseline on Video Laugh Reasoning.
- **Quantitative Jump**: MTLLFM temporary tags yielded a **227.2%** CIDEr improvement.
- **Grounding vs. Scaling**: Task-specific temporal grounding is a superior, lower-cost alternative to massive model scaling for social signals.
- **Conclusion**: Specialized modeling is essential for sub-second event localization.
- The code and datasets are publicly available at https://github.com/WSCSports/MTLLFM-temporal-laughter-localization

Contact Eyal Hanania, WSC-Sports
eyal.hanania@wsc-sports.com

10th Workshop and Competition on Affective & Behavior Analysis in-the-wild (ABAW).
Paper ID: 57