DEL: Dense Event Localization for Multi-modal Audio-Visual Understanding

Mona Ahmadian, Amir Shirian, Frank Guerin, Andrew Gilbert

UNIVERSITY OF SURREY

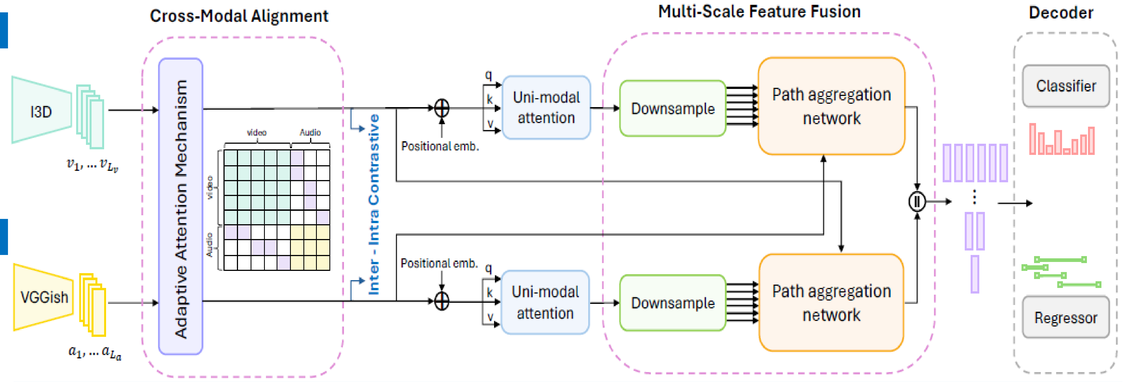
ICCV OCT 19-23, 2025 HONOLULU HAWAII

Motivation

- Real videos contain dense, overlapping and concurrent multimodal events
- Audio–visual cues are often misaligned
- Need to capture intra-video temporal coherence
- Discover fine-grained cross-modal temporal dependencies

Contributions

- **Unified framework** for dense multimodal event localization
- **Adaptive attention mechanism** for audio-visual alignment, ensuring event-synchronous and noise-robust interactions.
- **Score-based contrastive learning** with token-level supervision for dynamic selection of positives and hard negatives, enhancing intra-video discrimination.
- **Path Aggregation Network (PAN)** for multi-scale audio-visual fusion, preserving fine-grained temporal details while integrating high-level semantics.
- **State-of-the-art performance** onTHUMOS14, ActivityNet, EPIC-Kitchens and UnAV-100. consistently surpassing prior methods using the same encoders.



Method

- **Cross-Modal Alignment**

  Apply a learned event-aligned mask to the attention weights, ensuring temporal coherence and intra-modal consistency

  This guides feature interactions within and across modalities. Unlike standard self-attention, a learned mask matrix that constrains attention based on temporal alignment,

  Emphasising interactions likely to represent the same event while suppressing irrelevant ones.

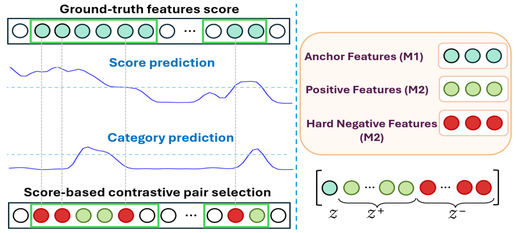
- **Score-based Contrastive Learning**

  Standard contrastive learning struggles with fine-grained temporal dynamics in untrimmed videos.

  Heuristic sampling often overlooks concurrent events and intra-video relationships.

  Need for a more adaptive method to align audio-visual features within and across samples.

  - We introduce a score-based contrastive learning approach that dynamically selects contrastive pairs using a learned similarity function.
  - Combines inter-sample and intra-sample objectives to refine both global and fine-grained audio-visual alignment.
  - Capturing contextual similarity and temporal alignment across modalities.



- **Multi-Scale Feature Fusion**

  Events span short to long durations → multi-scale fusion preserves details and context while aggregating high-level semantics.

  Enriches features across both modalities at multiple scale using attention-based fusion.

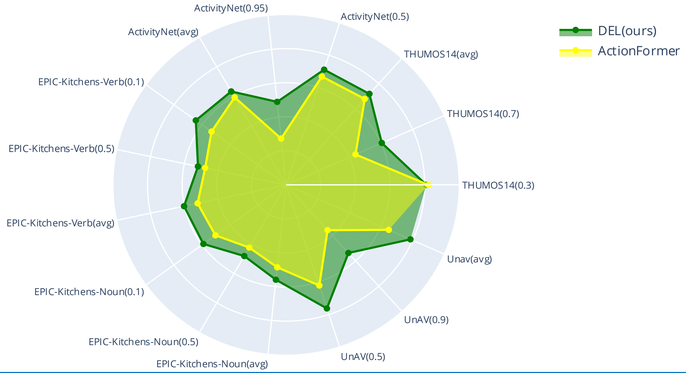
  Dynamically weight temporal features across scales for better event discrimination.

  Combines top-down and bottom-up pathways to preserve fine details and aggregate context.

Ablation studies, results, and implementation details are provided in our arXiv paper. The DEL code and pretrained models are publicly available on GitHub.

Code arXiv

Results



Component Ablation

Evaluating the individual contributions of our proposed modules on UnAV reporting mAP across tIoU thresholds and their average:

AAT	SCL	PAN	0.5	0.6	0.7	0.8	0.9	Avg
×	✓	✓	51.1	45.7	41.0	34.5	25.8	49.6
✓	×	✓	51.5	44.7	38.7	33.3	25.8	49.7
✓	✓	×	51.3	45.0	39.4	33.8	25.3	49.5
✓	✓	✓	53.4	48.1	42.6	35.6	26.9	51.1