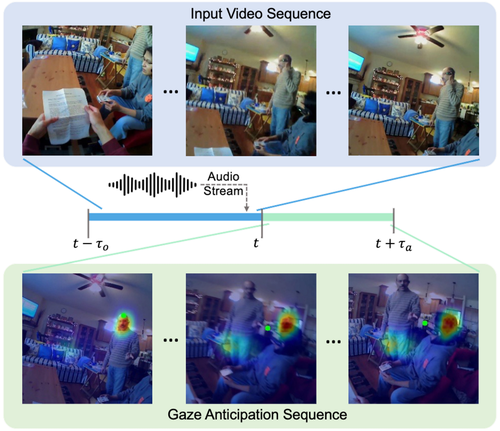
Listen to Look into the Future: Audio-Visual Egocentric Gaze Anticipation

Bolin Lai[1] Fiona Ryan[1] Wenqi Jia[1] Miao Liu[2,*] James M. Rehg[3,*]

[1]Georgia Tech [2]GenAI, Meta [3]UIUC

Problem Definition

- **Input:** video sequence & audio stream
- **Output:** gaze fixation distribution (heatmap) in future frames



We argue that audio signals can serve as a vital auxiliary cue for egocentric gaze forecasting.

Challenges

(1) Viewpoint and scene change due to head movement.

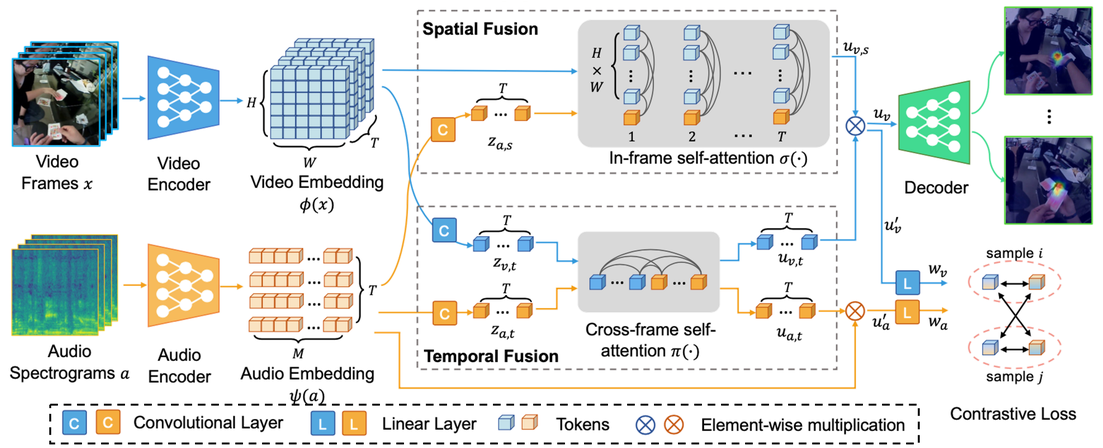
(2) Latency between audio stimulus and gaze reaction.

Demand a model can

(1) learn possible viewpoint and scene change driven by the audio stream over time.

(2) locate potential future gaze target in visual space.

Approach



We propose a Contrastive Spatial-Temporal Separable (CSTS) fusion approach.

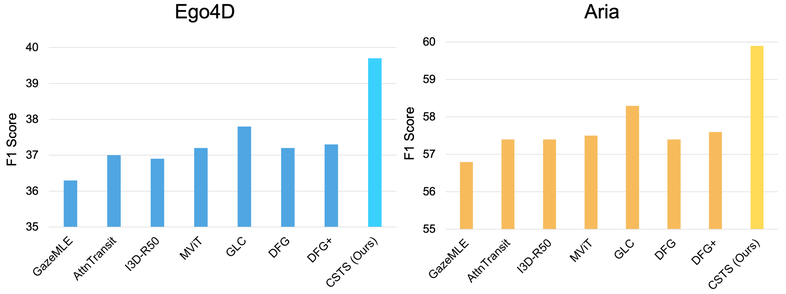
- Spatially (blue arrows), we calculate the correlation of audio token with each visual token in each frame.
- Temporally (orange arrows), we pool each frame into a single token and model the dependency over time.

The visual region (e.g., sound source) that has a stronger correlation with audio signals is more likely to be the potential future gaze target.

Events in the audio signal may drive both egocentric viewpoint change (via head movement) and gaze movements in time.

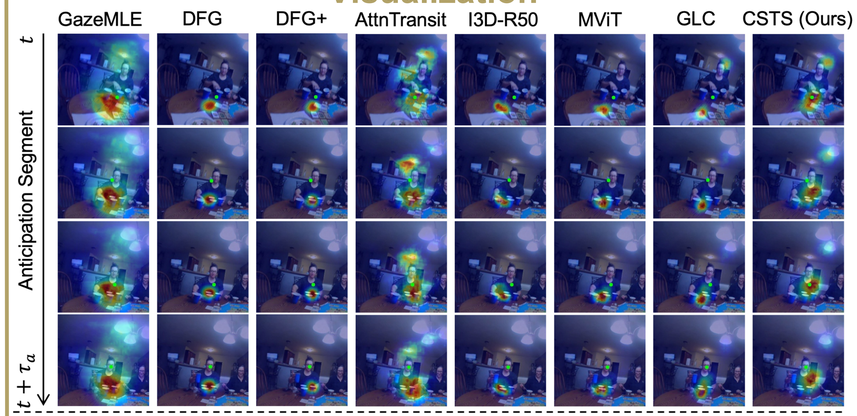
We also introduce a post-fusion contrastive learning scheme on fused modalities to boost performance.

Comparison with SOTA

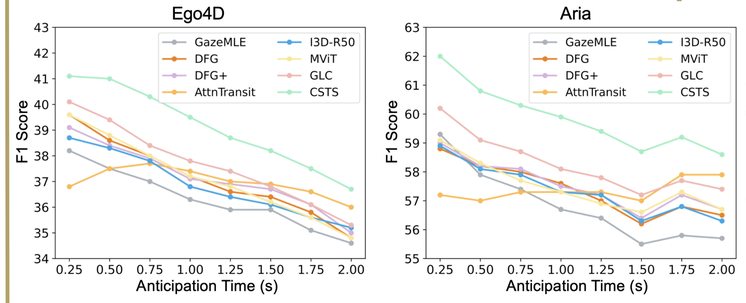


Our model achieves the best performance compared with prior egocentric gaze modeling methods on two datasets.

Visualization



Performance at Each Time Step



Our model outperforms SOTA methods at each time step.

Conclusion

- We introduce the **first audio-visual model** for egocentric gaze anticipation.
- We propose a novel **spatial-temporal separable fusion** module and a **post-fusion** contrastive learning strategy for audio-visual representation learning

Contact Us

Homepage WeChat Twitter Codes