The relative contribution of each modality in real-time is dynamically different \([1]\) at a separate time step. So how to learn the fusion dynamically? We summarize the current fusion method for audio-visual fusion in audio-visual navigation and propose our solution in the following figure.

### Contributions

- To realize a more effective (than the existing methods) audio-visual feature fusion strategy during audio-visual embodied navigation, we design a trainable Feature Self-Attention (FSA) module that determines the relative contribution of visual/audio modal in real-time following the ever-changing context.
- We propose an end-to-end framework (FSAAVN: feature self-attention audio-visual navigation) incorporating FSA to train robots to catch up with a moving audio target.
- FSAAVN is easy to train since it requires no extra aid like topology graphs and sound semantics.
- Our comprehensive experiments validate the superior performance (both quantitatively and qualitatively) of FSAAVN compared to the state-of-the-art.
- We also conduct thorough ablation studies on mainstream visual modalities, signal (visual/audio) encoders and audio-visual fusion strategies, providing valuable insights for practitioners and researchers in this field.

### Reference

[1] C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020