11/20/2025

Audio-Guided Dynamic Modality Fusion with Stereo-Aware Attention for Audio-Visual Navigation

This conference presents a novel framework for audio-visual navigation that explicitly addresses the limitations of vision-dominated navigation systems in realistic environments. While many classic navigation tasks rely primarily on visual input, the speaker highlights that vision alone is often insufficient in real-world scenarios such as dark, occluded, or cluttered spaces. Humans naturally exploit auditory cues to infer the direction of unseen targets, motivating the need for navigation agents that can reason jointly over visual and auditory information.

The work focuses on two key observations. First, the reliability of vision and audio varies dynamically depending on environmental factors such as lighting conditions, noise, and echoes. Therefore, the relative importance of each modality should be adapted online rather than being statically fused. Second, binaural audio signals contain critical spatial cues that are essential for accurate sound source localization, yet are often ignored by prior methods. Addressing these observations leads to the central research question: how can an agent effectively leverage both modalities?

To this end, the authors propose an Audio-Guided Dynamic Modality Fusion framework with Stereo-Aware Attention (AGSA). The system consists of three components: feature extraction, multimodal fusion, and policy learning. Visual images and binaural audio signals are processed through separate but structurally identical convolutional encoders, projecting both modalities into a shared latent feature space. These features are then passed to two novel fusion modules.

The first module, Stereo-Aware Attention (SAM), explicitly models left–right auditory differences. Audio features are split into left and right channels, and bidirectional cross-attention is applied so that each channel attends to the other. This design enables the agent to capture spatial relationships, such as the direction from which a sound originates, significantly improving sound localization accuracy.

The second module, Audio-Guided Dynamic Fusion (AGDF), allows audio to dynamically guide how visual and auditory features are combined. Using multi-head attention, the audio representation queries the joint audio-visual feature space to extract the most relevant information. A learnable fusion weight then adaptively balances reliance on vision versus audio, improving robustness when one modality becomes unreliable.

The fused representation is fed into a recurrent policy network with a GRU to capture temporal dependencies, and trained using Proximal Policy Optimization. The proposed method achieves higher success rates, better path efficiency, and stronger generalization to unseen sound categories, including in audio-only settings. Overall, the work demonstrates that stereo-aware auditory modeling and dynamically guided fusion are crucial for robust and efficient audio-visual navigation.