Research statement

Artificial intelligence (AI) has undergone a paradigm shift over the past years. From training separate models on images or text scraped off the internet, the AI community has started building multimodal models that can combine information from multiple sensory streams, like vision, audio and text. Despite this impressive progress, today's multimodal models are typically limited to learning cues that are directly observable within the relatively narrow field of perception of the sensors and are unable to reason about the larger and often partially observable rich 3D scene in which the streams are captured. Another limitation of these models is that they are designed to learn primarily semantic concepts (*e.g.*, associating the image of a cat with the "meow" sound), but not the spatial relations between the modalities and the underlying scene where the capture takes place (*e.g.*, the sound of a running tap can indicate *where* it is coming from). Additionally, these models are trained with internet-style data, which is usually captured from carefully-chosen locations. Consequently, they often struggle with reasoning about the information asymmetry across different capture locations vis-a-vis understanding 3D scenes and the dynamic activity therein, and predicting the more suitable locations for such capture (*e.g.*, not all camera angles are equally useful for observing a human in a cluttered kitchen).

My research aims to address these limitations and build models that leverage the synergy of vision and other modalities like audio and text to *better* understand a *persistent* 3D scene and the *dynamic* activity happening in the scene, often beyond what is directly captured in the sensors' "field of view". Such models have important applications in robotics (*e.g.*, a scene-aware robot can better navigate a 3D space) and AR/VR (*e.g.*, showing an activity from an optimal camera angle can provide a rich viewing experience to a human user).

In my research, I have developed methods that share two broad themes: 1) multi-modal modeling of a persistent 3D scene and its properties in a resource-efficient manner, and 2) smart sensor placement for high-quality understanding a dynamic activity in a 3D environment. In the following, I summarize these methods and my future research plans.

Learning to move to hear better (*ICCV '21*; *ECCV '22*). Physical factors can either restrict or facilitate our ability to perceive relevant audio-visual events in our daily lives. *E.g.*, a father working upstairs might move near the door to better hear his child calling out from below; a traveler may shift a few places at a busy airport to hear an announcement. These examples show how *controlled sensor movements* can be critical for audio-visual understanding. In terms of audio sensing, a person's nearness and orientation relative to a sound source affects the clarity with which it is heard, especially when there are other competing sounds in the scene. In terms of visual sensing, one must see obstacles to circumvent them, spot desired and distracting sound sources, use visual context to hypothesize an out-of-view sound source's location, and actively look for "sweet spots" in the visible 3D scene that may permit better listening. We are the first to explore how autonomous multi-modal systems might learn to exhibit such intelligent behaviors (*ICCV '21* [7]; *ECCV '22* [9]). Unlike the traditional setup of separating sounds in a passive pre-recorded video [13], we introduce the task of active audio-visual source separation: given a stream of egocentric audio-visual observations from a *novel* 3D scene, an embodied agent must decide how to move within *bounded time* in order to recover the sounds being emitted by some target object. See Fig. 1a. Towards that end, we propose a reinforcement learning framework comprising 1) a policy that temporally aggregates spatial audio-visual observations, to decide where to move, and 2) an audio separator that predicts an estimate of the target audio and refines it over time for better audibility. Please see our example videos[1].

Few-shot estimation of scene acoustics (*NeurIPS '22, IROS '24 Oral*). We now shift our focus from the *object-centric* scene understanding task of active sound separation to the *environment-centric* task of modeling how the physical space in a 3D scene affects how an audio played in it actually sounds. The auditory experience can change drastically from one environment to another—listening to a symphony in a big theater would feel different from listening to it in a cozy bedroom. Conditioned on the scene geometry and materials, and the location of the source relative to the listener, sound undergoes various acoustic phenomena in a 3D scene: direct sound, early reflections, and late reverberations, These factors together comprise the *room impulse response* (RIR)—the transfer function that is commonly used for modeling scene acoustics. Learning to model RIRs can be useful for generating a rich and realistic experience in AR/VR applications and robotics by helping build acoustics-aware audio-visual embodied agents that can better interact with their surrounding world. Stepping away from the tradition of measuring acoustics by densely collecting RIR samples from a previously unseen scene [1], we learn a model that infers RIRs by only using few-shot audio-visual observations from the scene (*NeurIPS '22* [8]). See Fig. 1b. We also show how using a learned RL policy to smartly decide *when* and *where* to sample these sparse observations using a learned RL policy can further improve the acoustic prediction performance (*IROS '24* [16]). Please see our project videos for sample predictions[2].

Efficient scene mapping from multi-ego conversations (*CVPR '23*). Whereas scene acoustics is crucial for revealing an environment's auditory context, the spatial layout of the scene is fundamental to understanding its physical context. By representing the walls, furniture, and other major structures in a space, scene maps ground activity and objects in a persistent frame of reference. This can facilitate high-level reasoning for many downstream applications in AR (*e.g.*, floorplan estimation, finding objects in a video walkthrough of a scene a.k.a episodic memory) and robotics

[1] https://vision.cs.utexas.edu/projects/active-av-dynamic-separation, https://vision.cs.utexas.edu/projects/move2hear/
[2] https://vision.cs.utexas.edu/projects/fs_rir, https://vision.cs.utexas.edu/projects/active_rir/

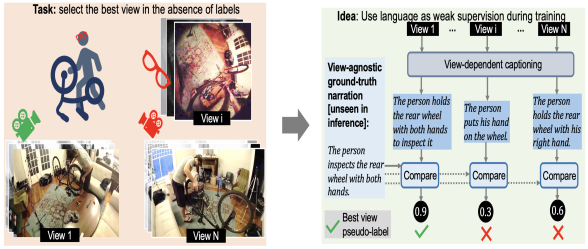

Figure 1a. LangView [12]: use view-agnostic captions in multi-view videos to gauge view quality and use this measure train a view selector without any manual labels.

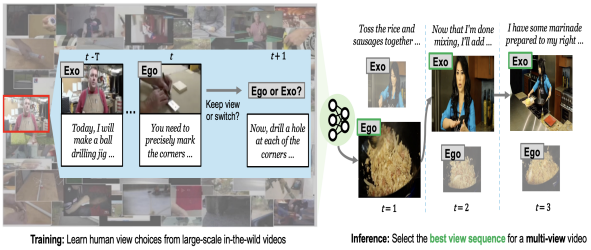

Figure 1b. SWITCH-A-VIEW [11]: given unlabeled in-the-wild varying-view videos, pretrain a view-switch detector and finetune it with limited labels for view selection

(*e.g.*, indoor navigation of mobile robots). The status quo in topdown mapping is to either do a *dense* visual scan of the scene [14] or use sounds (*e.g.*, broad-range frequency sweep) [15], which are often intrusive in nature when emitted around humans, coupled with a *continuous* camera stream, making the process energy-intensive and time-consuming. We introduce a new scene mapping task aimed at mitigating these challenges (*CVPR '23* [10]). In our proposed setting, multiple people converse as they move casually through the scene while wearing AR glasses equipped with an egocentric camera, microphones, and potentially other sensors (e.g., for odometry). Given their egocentric audio-visual data streams, the goal is to infer the ground-plane occupancy map for the larger environment around them, while satisfying a pre-specified *visual budget*. To realize this vision, we propose a novel framework comprising 1) an audio-visual mapper that predicts a shared map for the egos, and 2) a visual sampling policy that samples visual frames only when they are deemed absolutely necessary for mapping. Please watch our project video for qualitative examples[3].

Learning to select camera viewpoints in instructional videos (*CVPR '25 Highlight, ICCV '25*). Not all camera *viewpoints* (views) are created equal when it comes to observing an intricate activity in a geometrically complex 3D scene. That is, our ability to observe such activities depends on the chosen view. Solving this problem of camera view selection is especially important for making instructional (a.k.a "how-to") videos, where the goal is to produce a *varying-view* video from a multi-camera capture, such that each and every video segment is shown from an informative viewpoint. For example, in a typical high-quality how-to video, a close-up view of the hands is desirable when a knitter shows how to add stitches of yarn to a needle, or when a rock-climber demonstrates a particular hold—whereas a view from afar may be preferable when the knitter shows the knitted sweater being worn, or the climber shows their selected path up the wall. Unfortunately, today's view selectors are primarily designed to provide a high-level understanding of the video scene and optimize for viewing pleasure [6, 17, 2]. Besides, existing work is limited by relying on hand-coded heuristics [3, 4] or assuming access to manual labels indicating the favored views for training [5, 6, 18]. Such labels are expensive and quite special purpose. In our research, we address these limitations and design view selectors specifically for instructional videos (*CVPR '25* [12], *ICCV '25* [11]). Importantly, our methods can learn to choose informative views *without* access to large-scale manual best view labels. To this end, we turn to two alternate sources of weak supervision during training: 1) viewpoint-agnostic natural language captions of multi-view instructional videos, for indicating which view better captures the important details of the activity, as mentioned in the caption [12], and 2) auto-detected viewpoint switches in edited in-the-wild how-to videos, for revealing human view preferences when filming instructional videos [11]. Please see our project videos[4] for our models' view selection examples.

Conclusion and future work. My research so far has shown how we can build multi-modal methods that can reason about a 3D scene and the dynamic activity therein, and solve both scene-centric (estimation of environment acoustics and topdown maps) and activity-centric (audio source separation and camera view selection in instructional videos) tasks. My ultimate goal is to design unified but efficient models that can be not only be used to tackle multiple such tasks at once, but can also be deployed on edge computing devices with compromising on performance. Towards this goal, I am excited to see my research evolve in two broad directions in the future.

First, I want my future research to go hand-in-hand with the development of state-of-the-art edge computing hardware. Specifically, I plan to benchmark my methods on wearables like Meta's Aria glasses and Apple VisionPro, and continue to work on enhancing them while also meeting the compute and energy requirements of such devices. I believe that such efforts can also contribute to improvements in wearable hardware design, thereby creating a tight feedback loop between the field of building energy-efficient AI tools and that of designing edge computing devices.

Second, my approach so far has involved designing individual methods that tackle different scene and activity understanding tasks separately. However, many of these methods inherently rely on learning similar features (*e.g.*, active audio separation and camera view selection require precise understanding of scene occlusions; audio-visual scene mapping and acoustics estimation require high-level reasoning about scene layout and materials), and consequently, can

[3] https://vision.cs.utexas.edu/projects/chat2map

[4] https://vision.cs.utexas.edu/projects/which-view-shows-it-best/, https://vision.cs.utexas.edu/projects/switch_a_view/

potentially benefit from shared representations. I plan to explore this avenue in future work and design models that learn shared features through multi-task learning or self-supervision, which can be used for solving multiple such tasks.

References

[1] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 17–36. Springer, 2020. 1

[2] J. Chen, K. Lu, S. Tian, and J. Little. Learning sports camera selection from internet videos. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1682–1691. IEEE, 2019. 2

[3] D. Elson and M. Riedl. A lightweight intelligent virtual cinematography system for machinima production. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 3(1):8–13, Sep. 2021. 2

[4] L.-w. He, M. F. Cohen, and D. H. Salesin. *The Virtual Cinematographer: A Paradigm for Automatic Real-Time Camera Control and Directing*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. 2

[5] H.-N. Hu, Y.-C. Lin, M.-Y. Liu, H.-T. Cheng, Y.-J. Chang, and M. Sun. Deep 360 pilot: Learning a deep agent for piloting through 360deg sports videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3451–3460, 2017. 2

[6] S. Lee, J. Sung, Y. Yu, and G. Kim. A memory network approach for story-based temporal summarization of 360 {\deg} videos. *arXiv preprint arXiv:1805.02838*, 2018. 2

[7] S. Majumder, Z. Al-Halah, and K. Grauman. Move2hear: Active audio-visual source separation. In *ICCV*, 2021. 1

[8] S. Majumder, C. Chen, Z. Al-Halah, and K. Grauman. Few-shot audio-visual learning of environment acoustics. *NeurIPS*, 2022. 1

[9] S. Majumder and K. Grauman. Active audio-visual separation of dynamic sound sources. In *ECCV*, 2022. 1

[10] S. Majumder, H. Jiang, P. Moulon, E. Henderson, P. Calamia, K. Grauman, and V. K. Ithapu. Chat2map: Efficient scene mapping from multi-ego conversations. In *CVPR*, 2023. 2

[11] S. Majumder, T. Nagarajan, Z. Al-Halah, and K. Grauman. Switch-a-view: View selection learned from unlabeled in-the-wild videos. *arXiv preprint arXiv:2412.18386*, 2024. 2

[12] S. Majumder, T. Nagarajan, Z. Al-Halah, R. Pradhan, and K. Grauman. Which viewpoint shows it best? language for weakly supervising view selection in multi-view instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29016–29028, June 2025. 2

[13] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen. An overview of deep-learning-based audio-visual speech enhancement and separation. *Transactions on Audio, Speech, and Language Processing*, 2021. 1

[14] J. A. Placed, J. Strader, H. Carrillo, N. Atanasov, V. Indelman, L. Carlone, and J. A. Castellanos. A survey on active simultaneous localization and mapping: State of the art and new frontiers. *Transactions on Robotics*, 2023. 2

[15] S. Purushwalkam, S. V. A. Gari, V. K. Ithapu, C. Schissler, P. Robinson, A. Gupta, and K. Grauman. Audio-visual floorplan reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1183–1192, 2021. 2

[16] A. Somayazulu, S. Majumder, C. Chen, and K. Grauman. Activerir: Active audio-visual exploration for acoustic environment modeling. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13830–13836, 2024. 1

[17] Y.-C. Su, D. Jayaraman, and K. Grauman. Pano2vid: Automatic cinematography for watching 360 videos. In *Asian Conference on Computer Vision*, pages 154–171. Springer, 2016. 2

[18] B. Xiong and K. Grauman. Snap angle prediction for 360 panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–18, 2018. 2