Omnimodal Encoders

We are heading towards a near future where we use multimodal large language models (LLMs) to do most tasks involving a variety of modalities -- written language, spoken language, general audio, images and video, and beyond. Today's multimodal LLMs typically involve a combination of pre-trained text LLMs and modality-specific encoders for the various input modalities. This approach places a large burden on the encoders, and raises several natural questions:

- For this paradigm to work well, the encoders must be *task-universal* -- i.e. a single set of encoders should provide the needed information for all of the tasks the multimodal LLM might be used for. How should we train encoders to satisfy this goal? Can we learn deep semantic representations that are task-universal, or only low-level ones (leaving the deeper work to downstream models)?
- How *modality-universal* can the encoders and their learning techniques be? That is, can we jointly learn encoders for multiple modalities, or even share much of the model across modalities? How do we account for modality-specific information vs. shared information?
- Can we reliably evaluate encoder quality in a more efficient way than plugging them into LLMs and using large LLM benchmarks? What is a necessary and sufficient set of *intrinsic encoder evaluation tasks*?

Existing work has begun to address some of these questions, but more of the space is still unexplored. Our proposed workshop project will address these challenges by: (1) developing techniques for learning task- and modality-universal encoders and (2) establishing benchmark tasks for efficient intrinsic encoder quality evaluation that correlate well with downstream performance.

Thrust 1: Developing New Task- and Modality- Universal Encoders for LLMs

Task Universality: Current multimodal LLMs typically utilize off-the-shelf encoders such as Whisper, HuBERT, CLIP, CLAP, etc. which are locally optimized for pretext tasks which may cause them to discard some information and learn representations that do not generalize well to all tasks that the multimodal LLM should handle. We will investigate how downstream performance across a range of tasks depends on various encoder design decisions, such as the pretext training task (supervised learning vs. contrastive learning vs. autoencoders vs. self-supervised learning), pre-training data (e.g. speech vs. non-speech audio), and architectural design (e.g. convolutional waveform encoders vs. spectrogram transformers, etc.)

Modality Universality: Current LLMs use a separate encoder and set of weights for every different input modality. These encoders often have billions of parameters each, making the full LLM more complex and expensive. We will investigate to what extent these encoders can be merged into a shared model using techniques such as knowledge distillation from existing pre-trained encoders, multi-task joint training, and model stitching of pre-trained encoders. We will also investigate how much the performance on downstream LLM tasks depends upon various properties of the multimodal encoders, such as how well their representations are semantically aligned (e.g. do image embeddings of dogs live nearby to audio embeddings of dogs barking or word embeddings of the word "dog", etc.) and to what extent the encoder's embeddings capture and/or disentangle the "shared" information between modalities and the "private" information that is specific to each modality.

Thrust 2: Intrinsic Evaluation of Our Task- and Modality-Universal Encoders

A key outcome for our project will be the establishment of an intrinsic evaluation framework that enables us to predict how well a set of encoders will function when integrated into an LLM, ***without having to actually train the LLM.*** To this end, we will need to 1) curate a set of encoder evaluations that span a range of tasks and modalities 2) determine which tasks from the encoder evaluation are predictive of performance on multimodal LLM benchmarks.

For the encoder evaluation, we will utilize a set of benchmarks such as MSEB[1], HEAR[2] and SUPERB[3] for speech, MTEB[4] for text and MIEB[5] for vision. Massive Sound Embedding Benchmark (MSEB) [1] is a

model-agnostic framework designed to assess the auditory capabilities of multimodal encoders that is readily extensible to incorporate relevant tasks from other benchmarks, such as MMAU [6] and MAEB [7].

A key objective of our proposal is to demonstrate that the performance of encoder models and tokenizers by themselves on MSEB is strongly indicative of their final performance after integration into a larger LLM. A strong indication will allow us to pursue quicker development cycles focused on optimizing the lightweight sound components before the expensive LLM integration.

Senior members:
David Harwath (UT Austin) - tentative team leader
Karen Livescu (TTIC) - full time
Georg Heigold (Google) - part-time (1 week)
Shankar Kumar (Google) - part-time (1 week)

Potential senior members:
Ehsan Variani (Google)
David Chan (UC Berkeley)
Shiry Ginosar (TTIC)
Andrew Owens (Cornell Tech)
Sidd Karamcheti (Stanford)
Ruohan Gao (UMD)
Bowen Shi (Meta)

References:
[1] G. Heigold, E. Variani, T. Bagby, C. Allauzen, J. Ma, S. Kumar, M. Riley. Massive Sound Embedding Benchmark (MSEB). NeurIPS, 2025.
[2] J. Turian et al. HEAR: Holistic Evaluation of Audio Representations, In NeurIPS 2021 Competitions and Demonstrations Track, 2022.
[3] S. Yang et al. SUPERB: Speech processing Universal PERformance Benchmark, Interspeech, 2021.
[4] N. Muennighoff, N. Tazi, L. Magne, N. Reimers, MTEB: Massive Text Embedding Benchmark, EACL, 2023
[5] C. Xiao et al.. MIEB: Massive image embedding benchmark. arXiv preprint arXiv:2504.10471, 2025. doi: 10.48550/ARXIV.2504.10471. URL https: //arxiv.org/abs/2504.10471.
[6] S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. *ICLR*, 2025.
[7] MAEB. Coming soon.