

Ami Baid

amibaid@stanford.edu | amibaid.github.io | Google Scholar | GitHub

EDUCATION

Stanford University
M.S. Computer Science

September 2026 – June 2028 (Expected)
Stanford, CA

The University of Texas at Austin
B.S. Computer Science (Turing Scholars Honors) & Mathematics; GPA: 3.87

August 2022 – May 2026
Austin, TX

RESEARCH

UT Austin Computer Vision Group

March 2023 – May 2026

Advised by Prof. Kristen Grauman; research in audio-visual multimodal learning.

Don't Let the Video Speak: Audio-Contrastive Preference Optimization for AVLMS

ECCV, 2026

A. Baid, Z. Xue, K. Grauman [\[project page\]](#)

- Combats video-driven audio hallucination, a systematic failure mode where audio-visual language models (AVLMS) exploit visual shortcuts and describe the sounds implied by a scene rather than the audio events actually present
- Proposes a lightweight preference optimization framework: output-contrastive pairs penalize visual descriptions masquerading as audio facts, and input-contrastive pairs swap audio tracks to penalize audio-invariant generation
- Demonstrates improved audio hallucination resistance (64.5 $\rightarrow$ 82.0 on CMM, 74.0 $\rightarrow$ 79.9 on AVHBench), stronger audio captioning on a new benchmark (CIDEr 27.9 $\rightarrow$ 43.6), and generalization across multiple AVLMS backbones

Personal Visual Context Learning in Large Multimodal Models

arXiv Preprint, 2026

Z. Xue, A. Baid, S. Kim, M. Luo, K. Grauman [\[project page\]](#)

- Formalizes personalized question answering from a user's egocentric context and introduces Personal-VCL-Bench (2,255 instances, 7 tasks), showing frontier models fail to synthesize multiple observations into personal patterns

Action2Sound: Ambient-Aware Generation of Action Sounds

ECCV, 2024 (Oral)

C. Chen, P. Peng, A. Baid, Z. Xue, W. Hsu, D. Harwath, K. Grauman [\[project page\]](#)

- Presents AV-LDM, a latent diffusion model that generates action sounds from egocentric video while disentangling them from ambient noise, and Ego4D-Sounds, a curated 1.2M-clip dataset with high action-audio correspondence

Self-Supervised Visual Acoustic Matching (*acknowledged contributor*)

NeurIPS, 2023

- Conditional GAN for acoustic matching, trained without the need for paired source-target audio; contributed a ViLT-based pipeline to curate training samples with high audio-visual correspondence from web videos ([GitHub repo](#))

EXPERIENCE

Stripe

May 2025 – August 2025

Software Engineering Intern

South San Francisco, CA

- Added a visual modality to Stripe's LLM-based Terms-of-Service (ToS) evaluation system, enabling detection of violations that are only visible in merchant website imagery, such as the sale of counterfeit goods and weapons
- Designed the image representation strategy and engineered the production pipeline: selecting images, generating text descriptions using a VLM, and integrating them into the existing system; ablated design choices at each stage
- Deployed the best configuration to production; offline evaluations showed a 4-point recall and 1-point precision gain

Salesforce

May 2024 – August 2024

Software Engineering Intern

San Francisco, CA

- Built infrastructure-as-code automation for creating custom search attributes on Salesforce's internal Temporal.io workflow orchestration platform (1M+ workflows monthly), cutting creation time from 2 weeks to 30 minutes
- Contributed the custom search attribute resource, written in Go, to the open-source [Terraform Temporal provider](#)

TECHNICAL SKILLS

Languages: Python, C, C++, Go, Java, Bash

Tools and Frameworks: PyTorch, NumPy, scikit-learn, Hugging Face, Weights & Biases, Slurm, FFmpeg, Terraform