Kan Jen Cheng

Visiting Researcher, UC Berkeley

(+1)650-293-7058
kanjencheng@berkeley.edu
https://github.com/iftrush
https://iftrush.github.io/

Education

University of California, Berkeley — Berkeley, CA
BA in Computer Science — *Dec. 2023*
- Berkeley Speech Group, affiliated with Berkeley Artificial Intelligence Research (BAIR) Lab
- GPA: 3.87/4.0
- Coursework: Audio Signal Processing, Deep Learning, Machine Learning, Digital Signal Processing, Algorithm, Probability and Random Processing

College of San Mateo — San Mateo, CA
Physics — *May. 2020*
- CSM Honors Scholar
- GPA: 3.98/4.0
- Coursework: Calculus, Linear Algebra, Differential Equation, Data Structures, Physics

Research Interests

My research interests center on auditory perception, sound synthesis, texture editing, and computer vision. Recognizing that human perception of the environment relies heavily on the interplay between auditory and visual cues, I aim to develop sophisticated multi-modal systems capable of integrating audio-visual information to enhance human understanding, interpretation, and interaction with the world.
Multi-modal Perception: Audio-Visual Learning, Self-Supervised Learning, Spatial-Temporal Alignment
Generative Modeling: Diffusion Models, Conditional Flow Matching, Multi-modal Diffusion Transformer
Computer Audition: Audio Texture Editing, Spatial Audio, Speech Enhancement, Target Speech/Sound Extraction, Emotion Recognition
Computer Vision: Optical Flow, Segmentation

Publications

Kan Jen Cheng*, Tingle Li*, Gopala Anumanchipalli. "Audio Texture Manipulation by Exemplar-Based Analogy". *In Proc. ICASSP*, 2025.
Jingwen Liu*, **Kan Jen Cheng***, Jiachen Lian, Akshay Anand, Rishi Jain, Faith Qiao, Robin Netzorg, Huang-Cheng Chou, Tingle Li, Guan-Ting Lin, Gopala Anumanchipalli. "EMO-Reasoning: Benchmarking Emotional Reasoning Capabilities in Spoken Dialogue Systems". *In Proc. ASRU*, 2025.

Note: * above denotes equal contribution.

Research Experience

Berkeley Artificial Intelligence Research (BAIR) Lab — UC Berkeley
Advisor : Prof. *Gopala Anumanchipalli* — *Jan. 2023 - Present*

Audio Texture Manipulation by Exemplar-Based Analogy — *ICASSP*, 2025
- Supervisor: Gopala Anumanchipalli
- Designed a latent diffusion, exemplar-based analogy (In-Context Learning) model for audio texture manipulation, which refers to editing the overall perceptual quality of a sound and its interaction with various sound sources (project page).

EMO-Reasoning: Benchmarking Emotional Reasoning Capabilities in Spoken Dialogue Systems *ASRU*, 2025

- Supervisor: Gopala Anumanchipalli
- Designed a holistic benchmark for assessing emotional coherence in spoken dialogue systems through continuous, categorical, and perceptual metrics (project page).

Video to Spatial Audio Generation In Progress

- Supervisor: Gopala Anumanchipalli
- Designed and implemented a binaural audio pretraining foundation model for reconstruction, leveraging a Diffusion Transformer (DiT) backbone with Conditional Flow Matching (CFM) sampling. Fine-tuned on a spatial video–audio dataset conditioned by segmentation masks to generate high-quality spatial audio from video inputs.

Audio-Video Sync Edit Anything In Progress

- Supervisor: Paul Pu Liang, Jim Glass, Yuki Mitsufuji, Takashi Shibuya, Gopala Anumanchipalli
- Develop an audio-visual foundation model for reconstruction, leveraging a Diffusion Transformer (DiT) backbone with Conditional Flow Matching (CFM) sampling. The DiT incorporates synchronized audio-visual inputs and outputs to more effectively learn audio-visual spatial-temporal alignment. Subsequently fine-tuned on a curated dataset for text-guided audio-visual sync editing—enabling the addition and removal of targeted object from both audio and video.
- Collaborate with Sony Research, MIT Media Lab

Acoustically Highlighted, Visually Focused Audio-Video Generation In Progress

- Supervisor: Yuki Mitsufuji, Takashi Shibuya, Gopala Anumanchipalli
- Develop an audio-visual foundation model for reconstruction, leveraging a Diffusion Transformer (DiT) backbone with Conditional Flow Matching (CFM) sampling. The DiT incorporates synchronized audio-visual inputs and outputs to more effectively learn audio-visual spatial-temporal alignment. Subsequently Fine-tuned on a curated photo-realistic audio-visual dataset guided with segmentation-mask, enabling user-driven generation of visually highlighted audio and acoustically focused video pairs, with simultaneous object-specific motion magnification.
- Collaborate with Sony Research

Awards & Achievements

University of California, Berkeley

- Graduation with Distinction in General Scholarship
- Upsilon Pi Epsilon Honors Society
- Honors to Date in all semesters

College of San Mateo

- CSM Honors Scholar
- Dean's List in all semesters

Academic Services

Reviewer: *ICASSP*, 2025