UALM: Unified Audio Language Model for Understanding, Generation and Reasoning

Jinchuan Tian[1,2*], Sang-gil Lee[2*], Zhifeng Kong[2*], Sreyan Ghosh[2,3], Arushi Goel[2],
Chao-Han Huck Yang[2], Wenliang Dai[2], Zihan Liu[2], Hanrong Ye[2],
Shinji Watanabe[1], Mohammad Shoeybi[2], Bryan Catanzaro[2], Rafael Valle[2], Wei Ping[2]

[1]Carnegie Mellon University [2]NVIDIA [3]University of Maryland

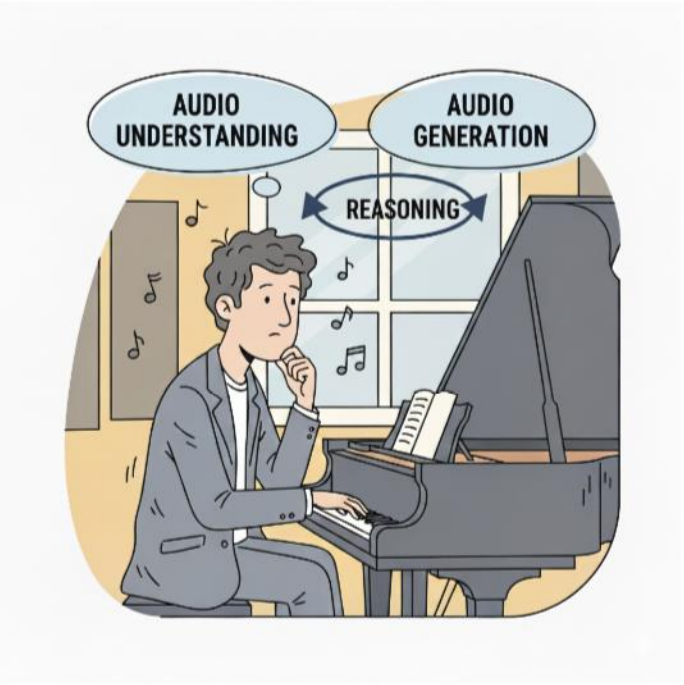
ICLR 2026 Oral

Presenter: Sang-gil Lee | Slides by: Jinchuan Tian

research.nvidia.com/labs/adlr/UALM

* Equal contribution

Why Unified Audio Understanding, Generation, and Reasoning?



A composer creates, listens, and refines -- requiring understanding, generation, and reasoning together

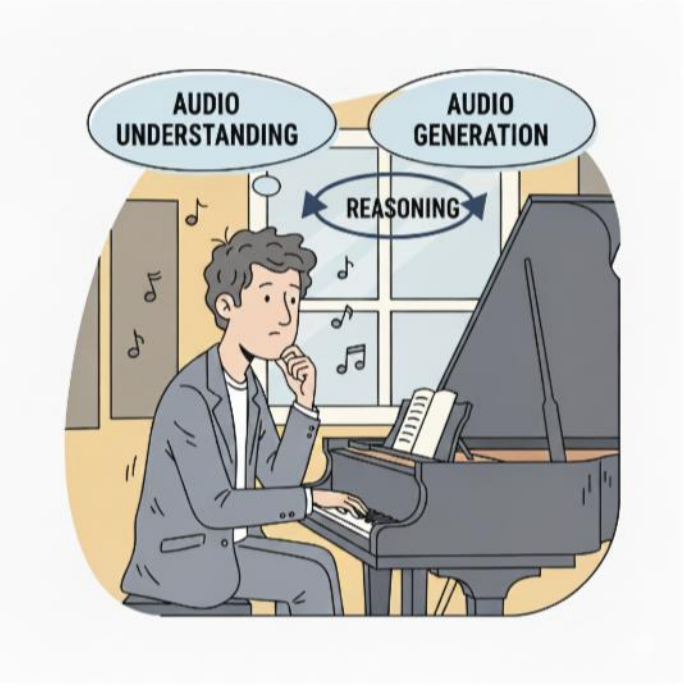
Understanding and generation are deeply coupled

- Neuroscience evidence: impairment in understanding often corresponds to deficits in generation
- Yet current audio AI usually treats them as separate problems

General audio intelligence needs multimodal reasoning

- Complex audio tasks require planning, self-evaluation, and iterative refinement, e.g., music composing
- The model must think iteratively across **both text and audio** modalities to produce high-quality results

Our Approach: Three Stages Toward Unified Audio Intelligence



UALM-Gen

Can language models generate audio as well as diffusion models?

→ Yes, with proper data scaling (10x), CFG, and DPO

UALM

Can one model handle understanding + generation + text reasoning?

→ Yes, with careful data blending

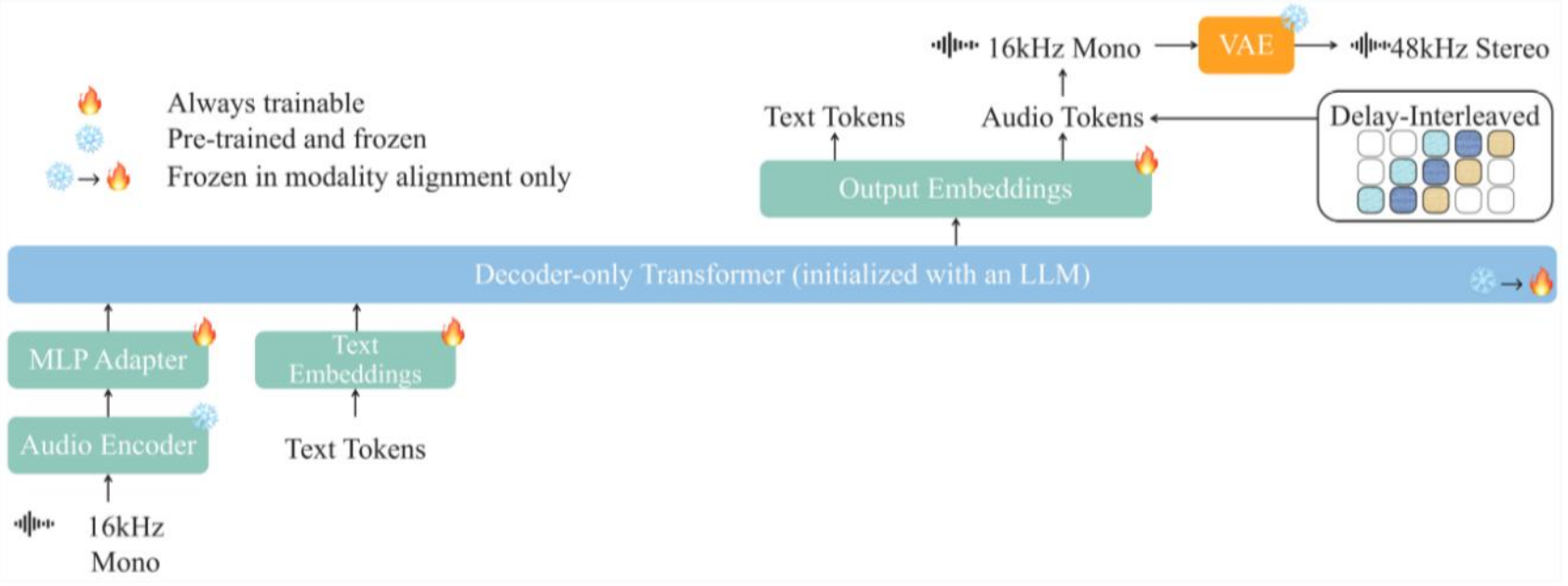
UALM-Reason

Can we achieve multimodal reasoning for audio generation?

→ Yes, via enrichment, dialogue, and self-reflection

UALM Architecture

A single model that can process any audio-text interleaved sequence



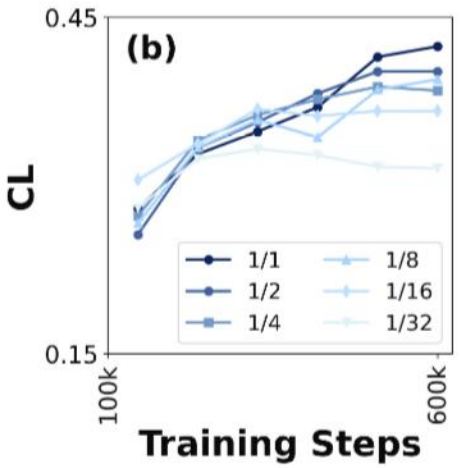
Audio Input — Encoder (25Hz) + MLP Adapter

Audio Output — X-Codec (50Hz, 8 RVQ tokens) + Delay Pattern

Backbone — Qwen2.5-7B

* An Enhancement VAE is used to upsample generated audio into 48kHz

UALM-Gen: LM-Based Text-to-Audio Generation



Data scaling improves LM-based TTA effectively

With the same model and compute, down-sampling the training data by 1/2, 1/4, ..., 1/32 decreases achievable CLAP score (CL) significantly.

Other Findings: Classifier-free guidance (CFG) and direct preference optimization (DPO) also contribute to the TTA performance

* CLAP score (CL) indicates the similarity between input text and generated audio.

UALM-Gen: Results

Model	SongDescriber							AudioCaps						
	FD↓	KL↓	IS↑	CL↑	AES↑	OVL↑	REL↑	FD↓	KL↓	IS↑	CL↑	AES↑	OVL↑	REL↑
Ground Truth	0	0	1.88	0.48	7.20	4.10	4.03	0	0	13.49	0.62	4.50	3.91	3.96
TangoFlux	235.61	0.71	1.70	0.41	6.46	3.80	3.89	103.04	1.02	15.13	0.65	4.42	3.72	3.93
Stable Audio Open	138.58	1.01	2.25	0.42	6.37	3.92	3.97	100.93	2.22	11.80	0.35	4.47	3.81	3.80
ETTA	95.66	0.80	2.15	0.44	6.71	3.92	3.93	80.13	1.22	14.36	0.54	4.51	3.73	3.94
UALM-Gen (Ours)	**74.43**	**0.63**	**1.87**	**0.54**	**7.36**	**4.07**	**3.96**	**75.14**	**1.19**	**14.52**	**0.65**	**5.08**	**3.79**	**3.92**

Key Findings:

- Text-to-audio (TTA) generation has been *previously dominated by diffusion-based methods*
- We show that LM-based generation can *match or exceed diffusion models* with proper data scaling

Metrics: FD = Fréchet Distance (OpenL3, ↓) | KL = KL Divergence (PaSST, ↓) | IS = Inception Score (PANNs, ↑) | CL = CLAP Score (text-audio similarity, ↑) | AES = Aesthetic Score (AudioBox, ↑) | OVL = Overall Quality (human, 5-scale, ↑) | REL = Relevance (human, 5-scale, ↑).
Red bold = best.

UALM: Unified Pre-Training

Context: No unified understanding-generation model existed for general audio before this work.

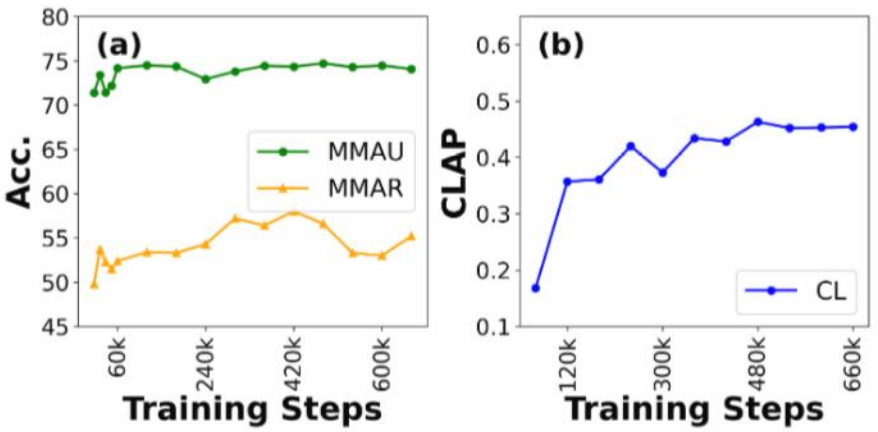
Goal: Unify audio understanding and generation in a single model, while preserving text reasoning from the base LLM.

Pre-Training Data Mixture

Category	# Samples	%	# Tokens	%
Audio Understanding	59.8M	37.8%	34.4B	39.2%
Audio Generation (2x)	59.0M	37.3%	29.0B	33.1%
Text Reasoning	39.4M	24.9%	24.3B	27.7%

Key Finding: Audio understanding converges much faster than audio generation during unified pre-training.

* Audio generation data is up-sampled 2x due to slower convergence, making it the largest portion by sample count.



(a) MMAU & MMAR: audio understanding accuracy. (b) CLAP: audio generation quality. Both tracked across training steps.

UALM: One Model, Three Capabilities

Audio Generation

Model	SongDescriber							AudioCaps						
	FD↓	KL↓	IS↑	CL↑	AES↑	OVL↑	REL↑	FD↓	KL↓	IS↑	CL↑	AES↑	OVL↑	REL↑
Ground Truth	0	0	1.88	0.48	7.20	4.10	4.03	0	0	13.49	0.62	4.50	3.91	3.96
TangoFlux	235.61	0.71	1.70	0.41	6.46	3.80	3.89	103.04	1.02	15.13	0.65	4.42	3.72	3.93
Stable Audio Open	138.58	1.01	2.25	0.42	6.37	3.92	3.97	100.93	2.22	11.80	0.35	4.47	3.81	3.80
ETTA	95.66	0.80	2.15	0.44	6.71	3.92	3.93	80.13	1.22	14.36	0.54	4.51	3.73	3.94
UALM-Gen	**74.43**	**0.63**	**1.87**	**0.54**	**7.36**	**4.07**	**3.96**	**75.14**	**1.19**	**14.52**	**0.65**	**5.08**	**3.79**	**3.92**
UALM	**83.69**	**0.59**	**2.00**	**0.54**	**7.28**	**3.97**	**3.99**	**65.87**	**1.35**	**15.62**	**0.62**	**4.92**	**3.89**	**3.86**

Audio Understanding

Model	MMAU↑	MMAR↑
Audio Flamingo 3	72.3	58.5
UALM (Ours)	**74.1**	**55.2**

Text Reasoning

Model	MMLU↑	GSM8K↑	HumanEval↑
Qwen2.5-7B-Instruct	74.5	91.6	84.8
UALM (Ours)	**71.6**	**92.1**	**81.1**

Key Finding: Unified pre-training preserves comparable performance across all three tasks compared with specialized models.

* UALM uses the same audio understanding data as Audio Flamingo 3. Our model is initialized from Qwen2.5-7B-Instruct.

UALM-Reason: Rich Caption

Keywords:
Brass band music, Percussion.

Layout:
Brass band music comes first, followed by percussion.

Description:
Brass band music: lively brass band playing an upbeat, rhythmic melody. The music will feature prominent trumpets and trombones.
Percussion: likely a drum kit, providing a steady, driving beat.

Example of a rich caption for audio generation.

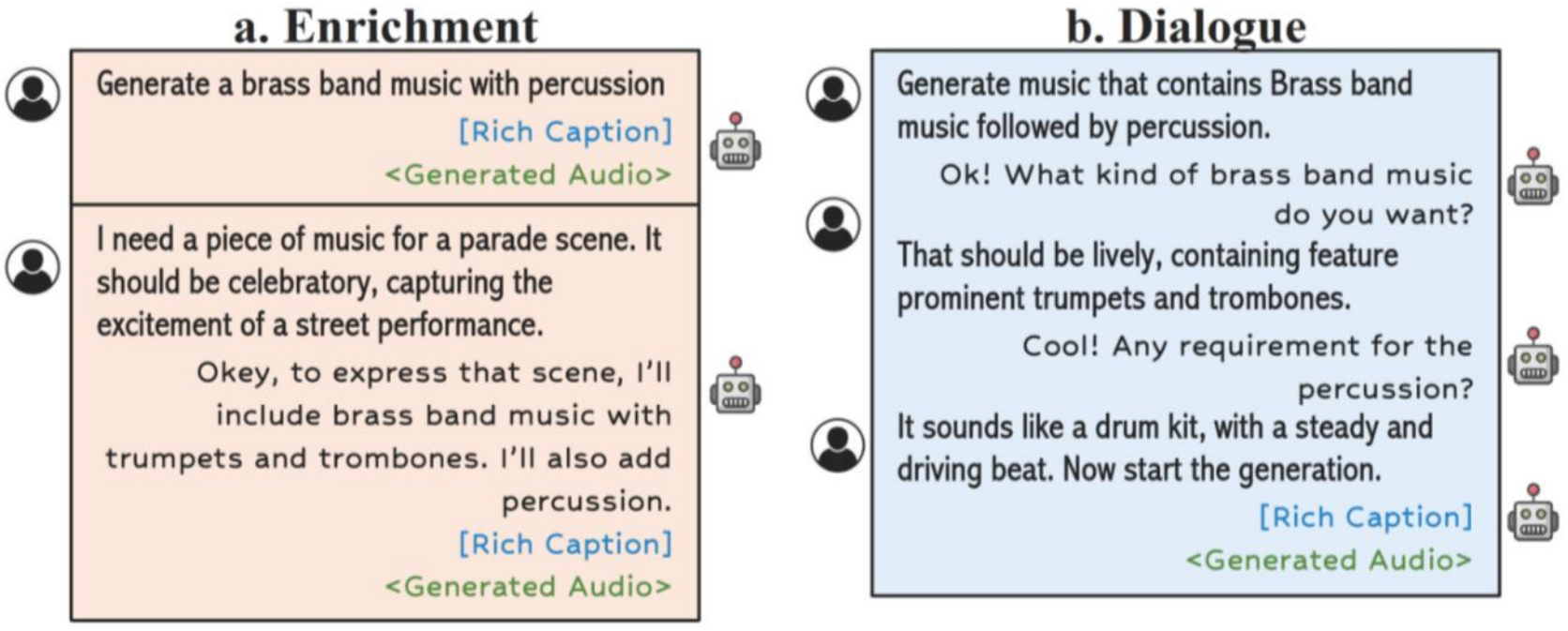
A **rich caption** is a structured and detailed textual blueprint that bridges vague user prompts to high-fidelity audio generation. It consists of three components:

- **Keywords:** Core acoustic events in the target audio
- **Layout:** Temporal arrangement of the events
- **Description:** Detailed characterization of each event's acoustic properties

Rich captions provide nuanced guidance that is critical for controllable audio synthesis — unlike conventional short prompts.

* Rich captions in training data are generated by an external multimodal LLM.

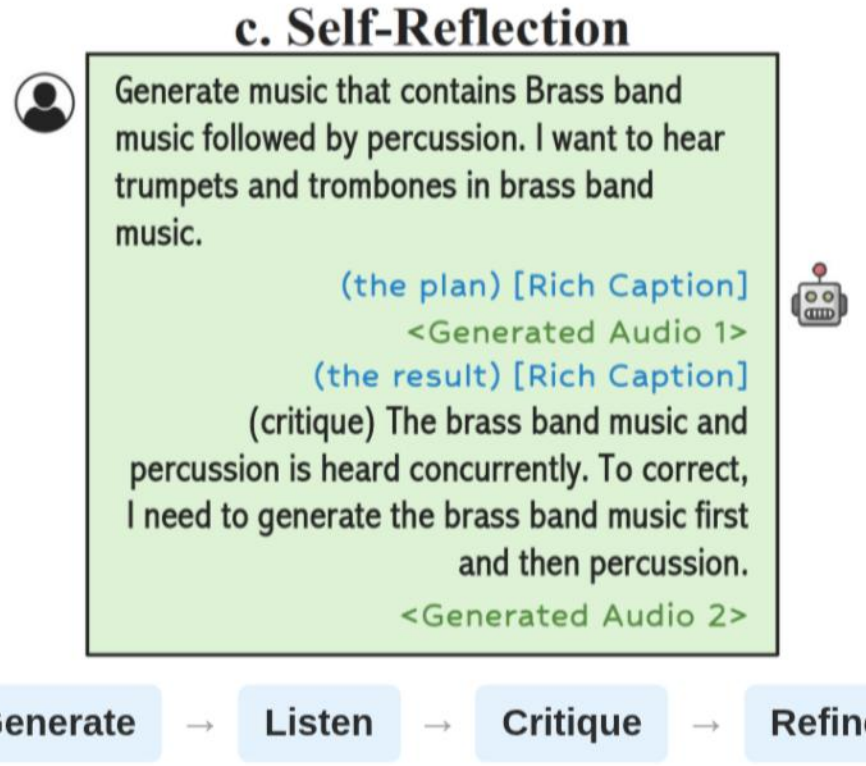
UALM-Reason: Text-Based Reasoning for Audio Generation



- **Enrichment:** Model autonomously expands a vague prompt into a detailed rich caption, then generates audio
- **Dialogue:** Multi-turn conversation — model actively queries the user to collaboratively build the caption

> Both are *text-based* reasoning — the model reasons in text to plan before generating audio

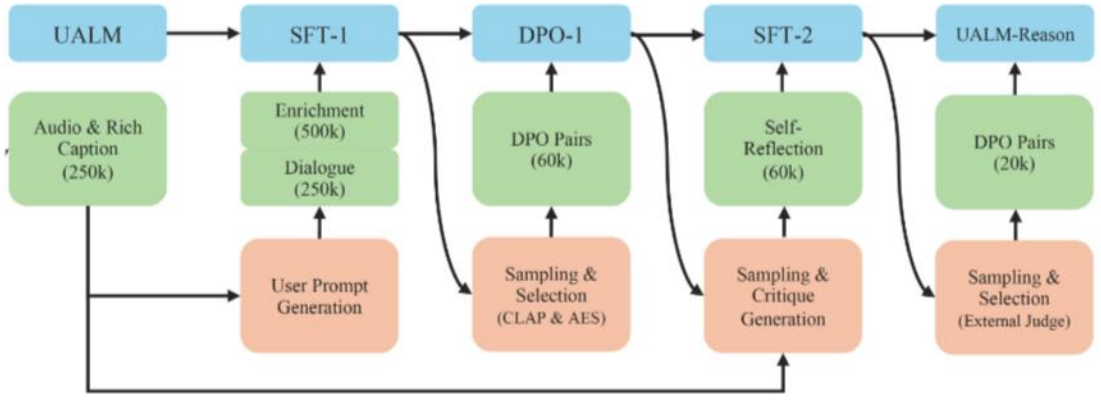
UALM-Reason: Multimodal Reasoning for Audio Generation



- **Generate:** Enrich prompt → rich caption → generate audio clip
- **Listen:** Model listens to its own output, produces a *new* rich caption of what it actually generated
- **Critique:** Compare the plan vs. the result, identify discrepancies in text
- **Refine:** Use critique as feedback to generate an improved audio clip

> Self-Reflection is *multimodal* reasoning — a multimodal CoT spanning both text and audio.

UALM-Reason: Training Recipe



Post-training workflow: two rounds of interleaved SFT + DPO.

Round 1: Enrichment + Dialogue

- 750k SFT samples → SFT-1
- ~60k DPO pairs → DPO-1

Round 2: + Self-Reflection

- Generate audio with DPO-1, create critiques
- Combined SFT → SFT-2
- Targeted DPO on 20k samples → UALM-Reason

UALM-Reason: Results

Subjective Evaluation (5-point MOS)

Model	Enrichment	Dialogue	Self-Reflection
UALM	3.77 ± 0.11	3.92 ± 0.11	3.82 ± 0.11
UALM-Reason	**4.01 ± 0.10**	**4.02 ± 0.10**	**4.04 ± 0.09**

Key Finding: Consistent improvement across all reasoning modes, demonstrating that generation-oriented reasoning from audio understanding and text capabilities effectively advances audio production.

Summary

- **UALM-Gen:** Language models can match diffusion models for audio generation with proper data scaling (10x), CFG, and DPO
- **UALM:** Unified model for audio understanding + generation + text reasoning with minimal degradation vs specialized models
- **UALM-Reason:** First multimodal reasoning for audio generation -- enrichment, dialogue, and self-reflection enable generate-understand-critique-refine cycles

Thank You!

Demo: research.nvidia.com/labs/adlr/UALM

Contact:
Jinchuan Tian (CMU) — jinchuat@andrew.cmu.edu
Sang-gil Lee (NVIDIA) — sanggill@nvidia.com
Zhifeng Kong (NVIDIA) — zkong@nvidia.com
Wei Ping (NVIDIA) — wping@nvidia.com

Questions?