IEEE SIGNAL PROCESSING SOCIETY WEBINAR

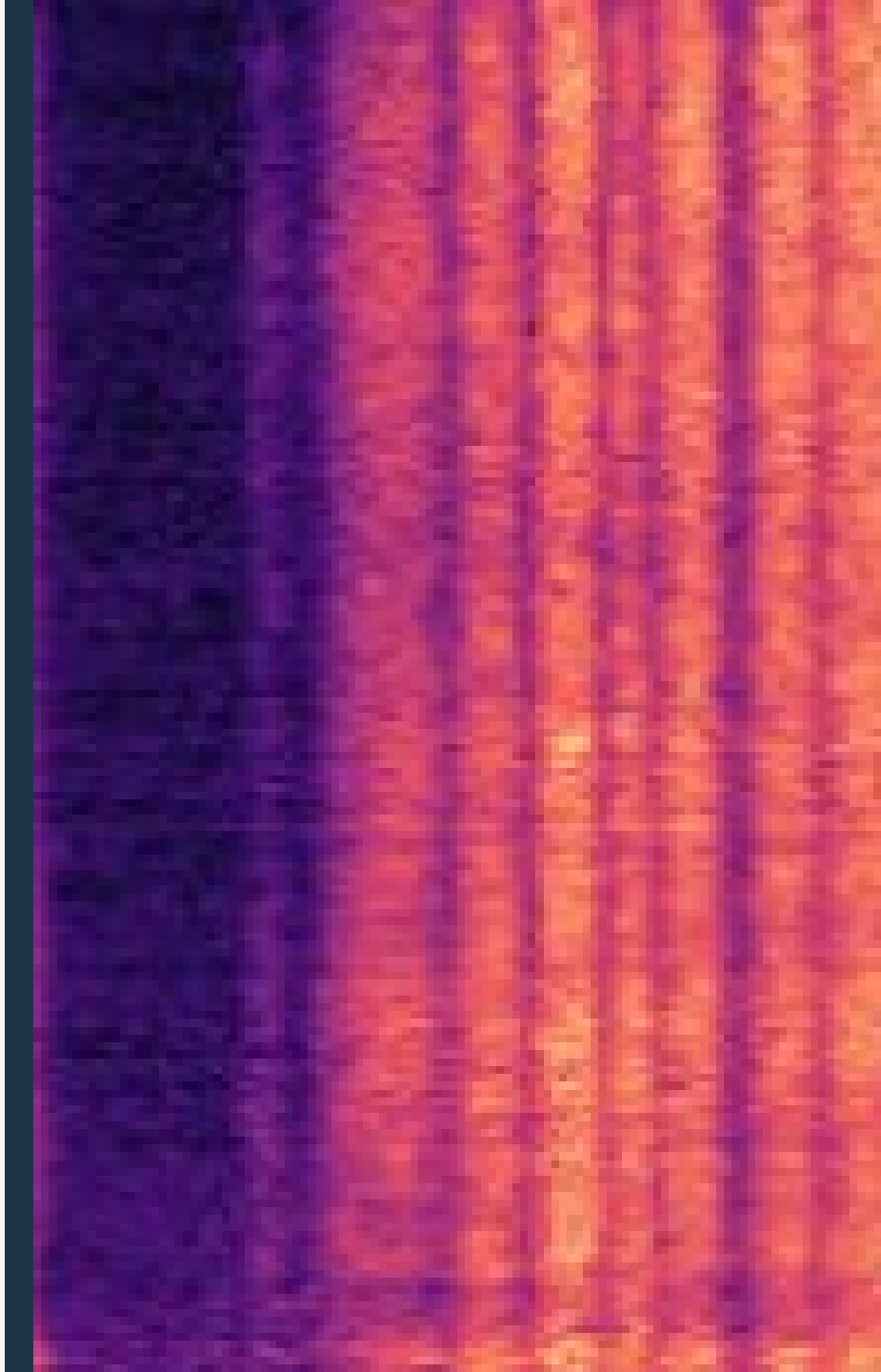
AudioLDM 2

Learning Holistic Audio Generation With Self-Supervised Pretraining

Haohe Liu
Research Scientist
Meta Superintelligence Lab (FAIR)

45-minute talk · 15-minute live Q&A

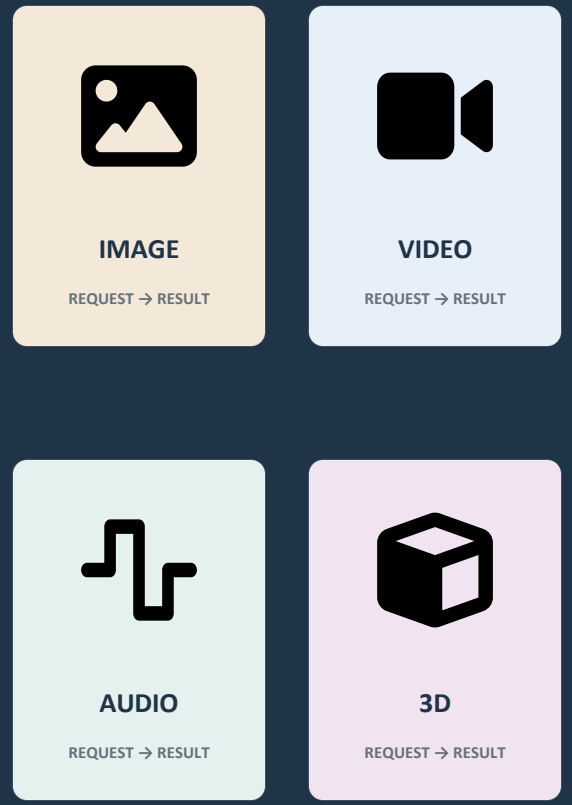
IEEE/ACM TASLP 2024 · DOI 10.1109/TASLP.2024.3399607



GENERATIVE MEDIA · INDUSTRY SHIFT

Media creation is moving from editing assets to describing the result.

Across image, video, audio, and 3D, creators can increasingly start with a request—and then edit the generated result.

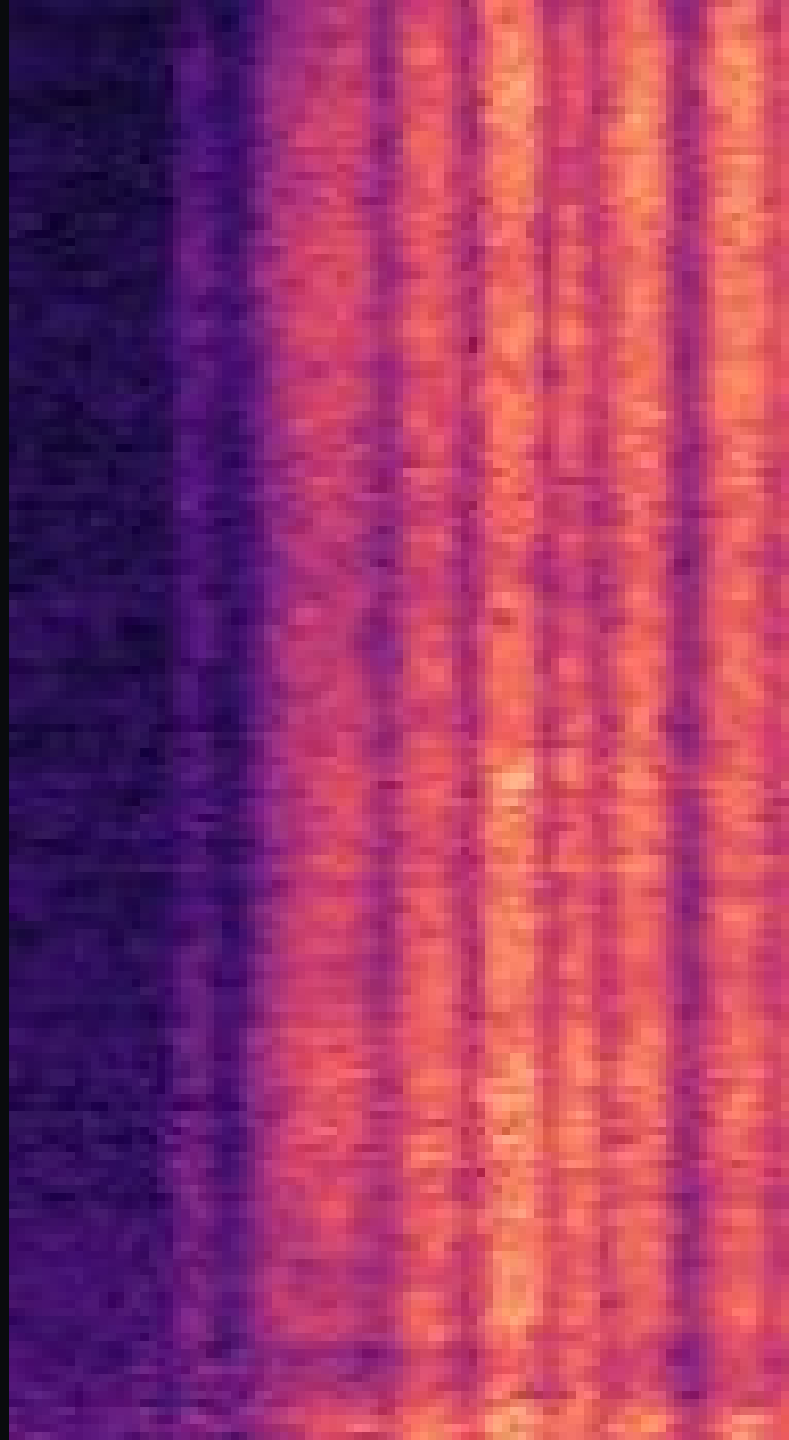


AUDIO IS PART OF THE SAME PRODUCT SHIFT

WHY AUDIO MATTERS

Sound shapes how we experience a story.

In film, games, and radio, audio conveys information, sets the mood, and builds the environment.



Sound is part of the product experience

FILM & VIDEO

atmosphere · emotion · attention

GAMES & VIRTUAL WORLDS

events · spaces · interaction

RADIO & SPOKEN MEDIA

intelligibility · context

CREATIVE TOOLS

search · generation · revision

Across these products, audio carries information, mood, and a sense of place.

Creating the right sound is still laborious

CREATIVE BRIEF

"A heavy metal door slams once in a large empty warehouse, followed by a long echo."

01

RECORD

Foley, voice, or performance

02

SEARCH

Try keywords across large libraries

03

EDIT & LAYER

Fit timing, distance, and environment

04

LISTEN & REVISE

Compare, adjust, and repeat

A few seconds of finished audio may require several tools—and several rounds of revision.

A sentence can specify what happens, when, and where

EXAMPLE CREATIVE REQUEST

"A heavy metal door slams once in a large empty warehouse, followed by a long echo."

WHAT	WHEN	WHERE	HOW IT SOUNDS
a heavy metal door slams	once … followed by	in a large empty warehouse	a long echo

SEARCH QUERY

"metal door"

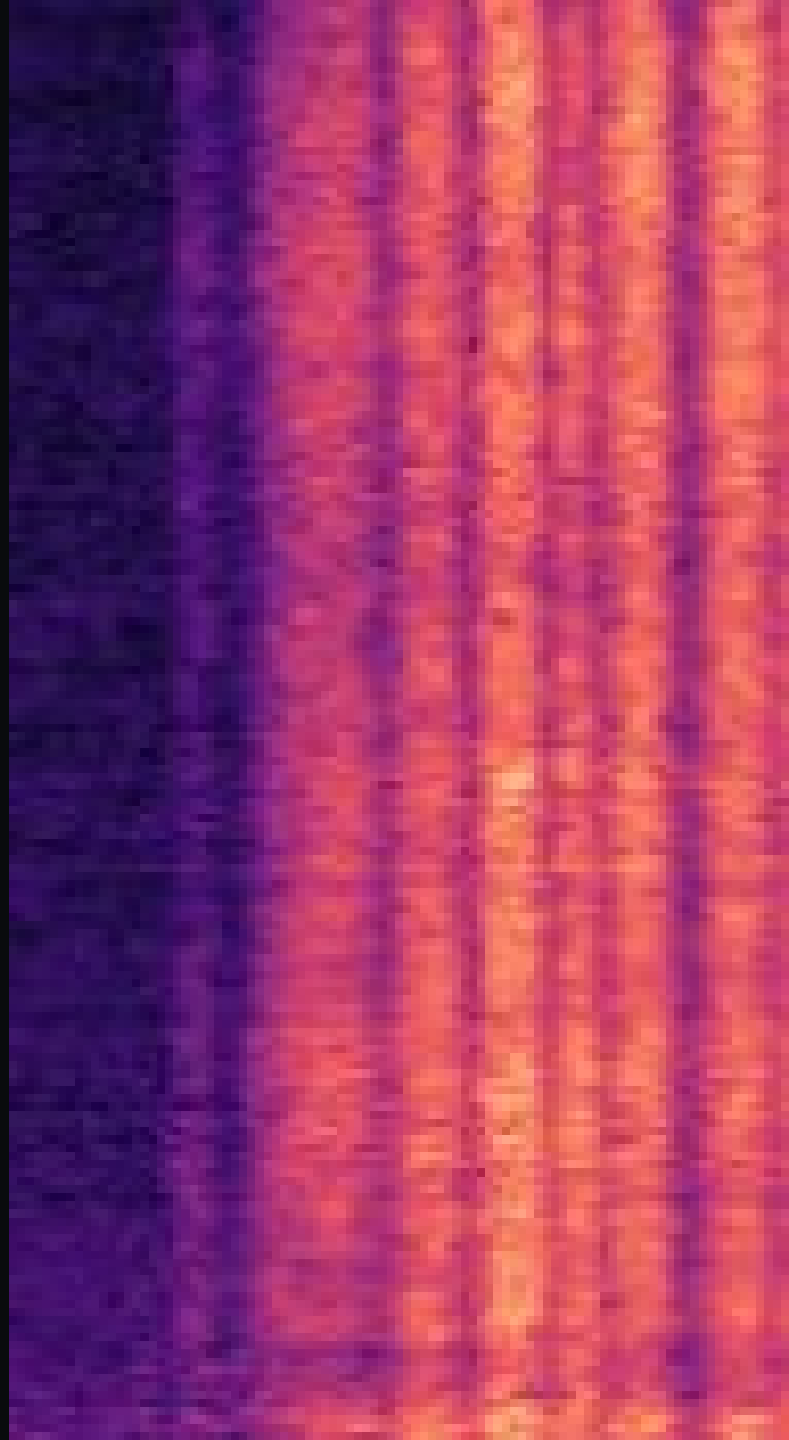
CREATIVE REQUEST

event + timing + place + acoustic detail

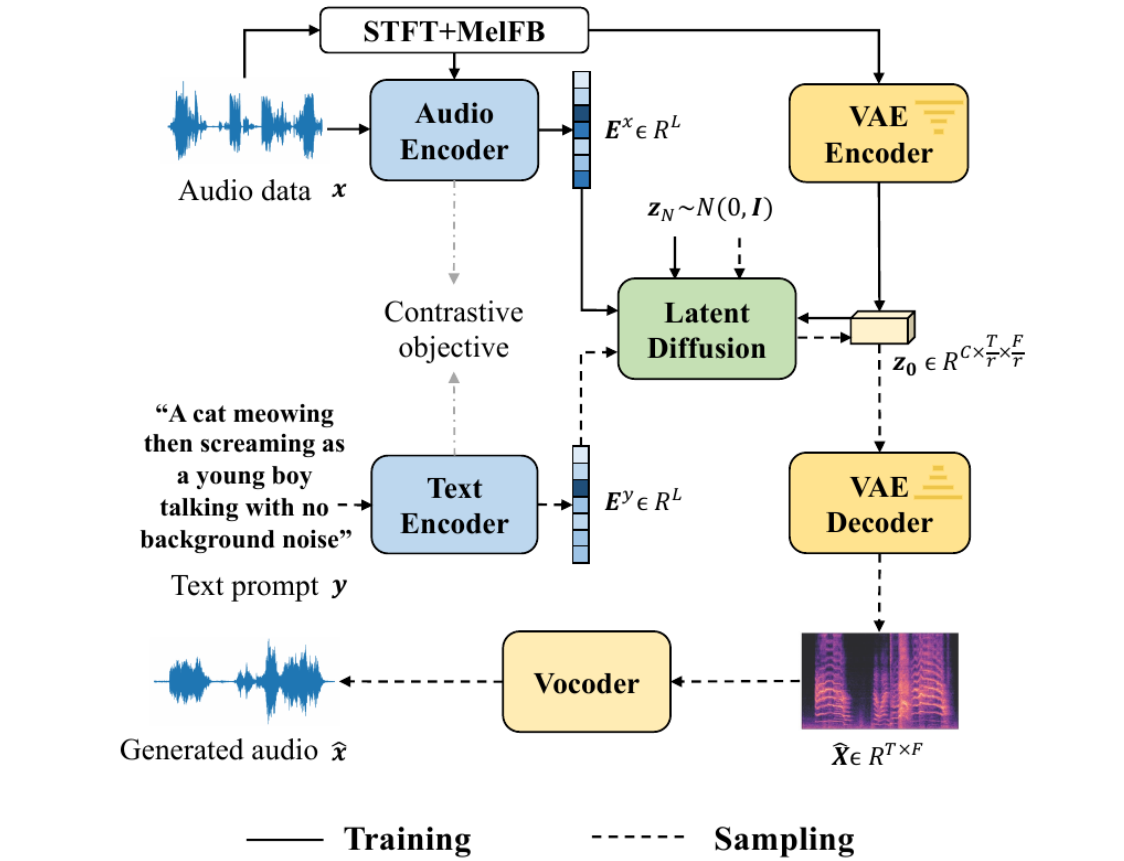
Text is not just a label—it can describe a complete sound scene.

[1] H. Liu, "High Quality Audio Generation with Latent Diffusion Models," PhD thesis, University of Surrey, 2026.

One General, Unified Audio Generator for Sound Effects, Music, and Speech.



AudioLDM 2 retains AudioLDM's VAE, latent diffusion, and vocoder



AudioLDM training and sampling

WHAT CARRIES OVER

VAE latent

latent diffusion

vocoder

WHAT CHANGES

condition → LOA → latent diffusion generator

[1] H. Liu et al., "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models," ICML 2023.

AudioLDM generated general audio—but not intelligible speech

WHAT AUDIOLDM PROVED

01 **Latent diffusion can generate high-quality general audio.**

02 **CLAP enables training with mostly unpaired audio.**

03 **One model covers many environmental sound categories.**

WHAT IT COULD NOT YET DO

Generate intelligible speech with exact words in the right order.

Unintelligible speech showed that global text conditioning was insufficient.

AudioLDM is the foundation; AudioLDM 2 adds LOA as an intermediate audio representation.

[1] H. Liu et al., “AudioLDM: Text-to-Audio Generation with Latent Diffusion Models,” ICML 2023.
[2] Y. Wu et al., “Large-Scale Contrastive Language-Audio Pretraining,” ICASSP 2023.

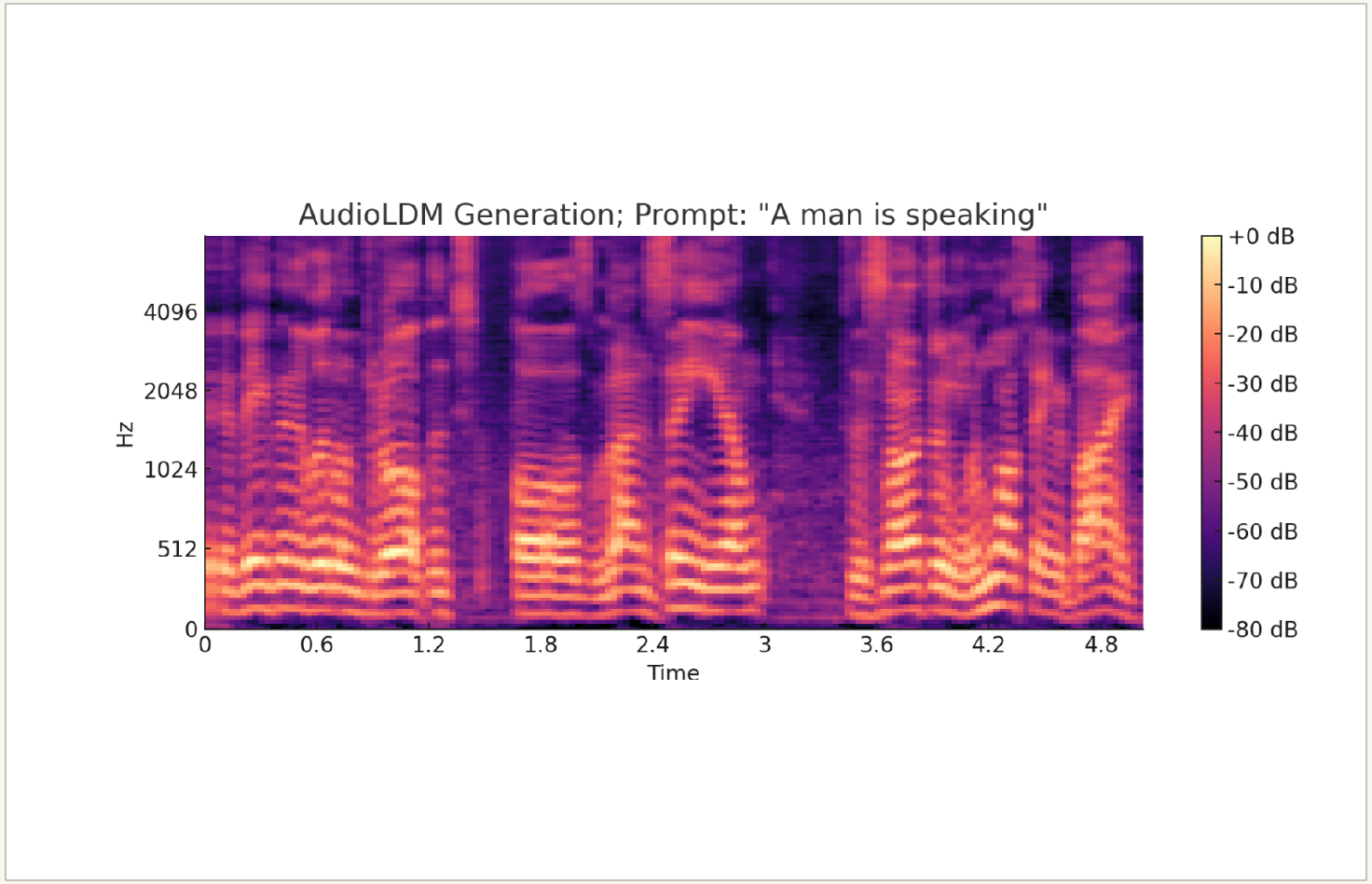
AudioLDM produced speech-like audio, but not intelligible words

PROMPT

“A man is speaking”

The output resembles speech, but the words are uncontrolled and unintelligible.

A conditioning representation was needed that retained linguistic content.

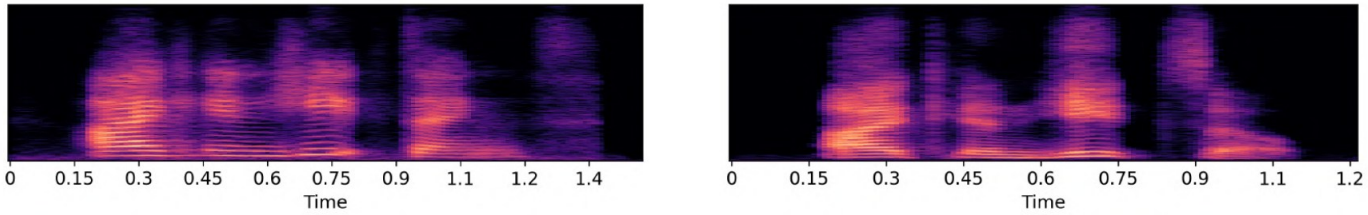


AudioLDM output: speech-like acoustics, unintelligible words

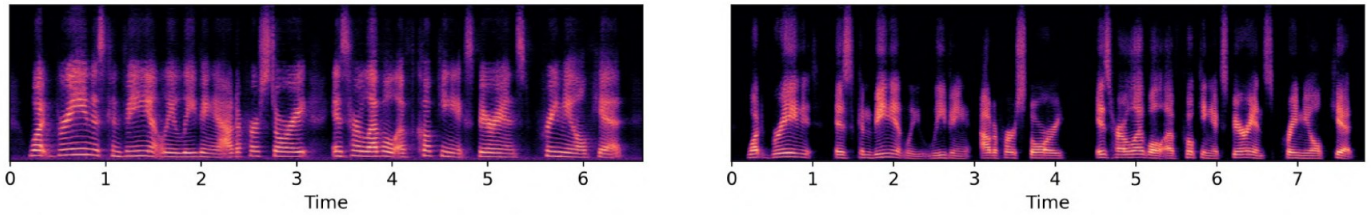
[1] H. Liu et al., “AudioLDM: Text-to-Audio Generation with Latent Diffusion Models,” ICML 2023.

AudioLDM 2 generates intelligible speech with speaker control

Text: I can heat things up.



Text: What green is conveniently leaving out of her story is her level of cooking experience pre-meal kit.



Speaker prompt: A young girl is speaking

Speaker prompt: A young male reporter is speaking

TEXT

“I can heat things up.”

+

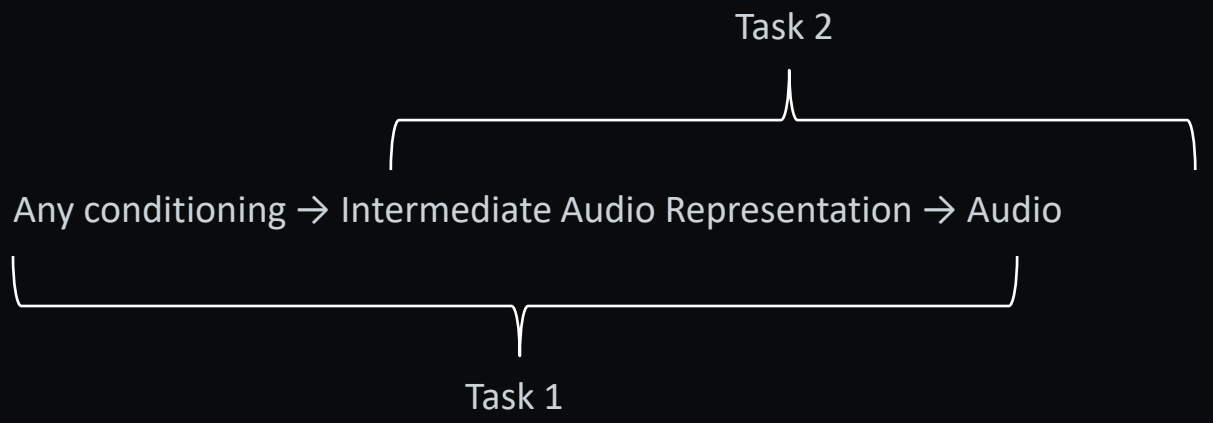
SPEAKER PROMPT

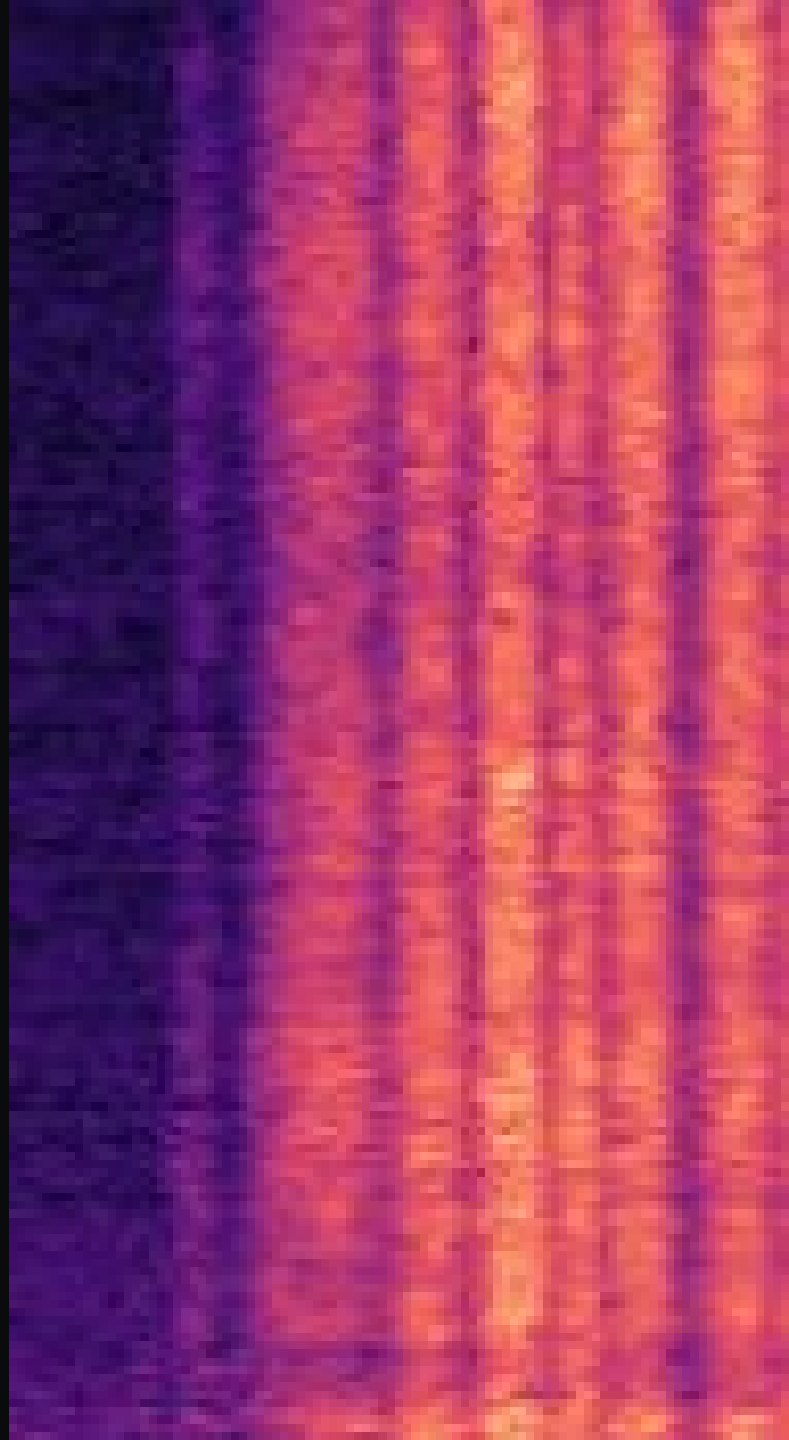
young girl
— or —
young male reporter

LOA conditioning preserves exact words and speaker information.

Text or phonemes plus a speaker prompt control what is said and how it sounds

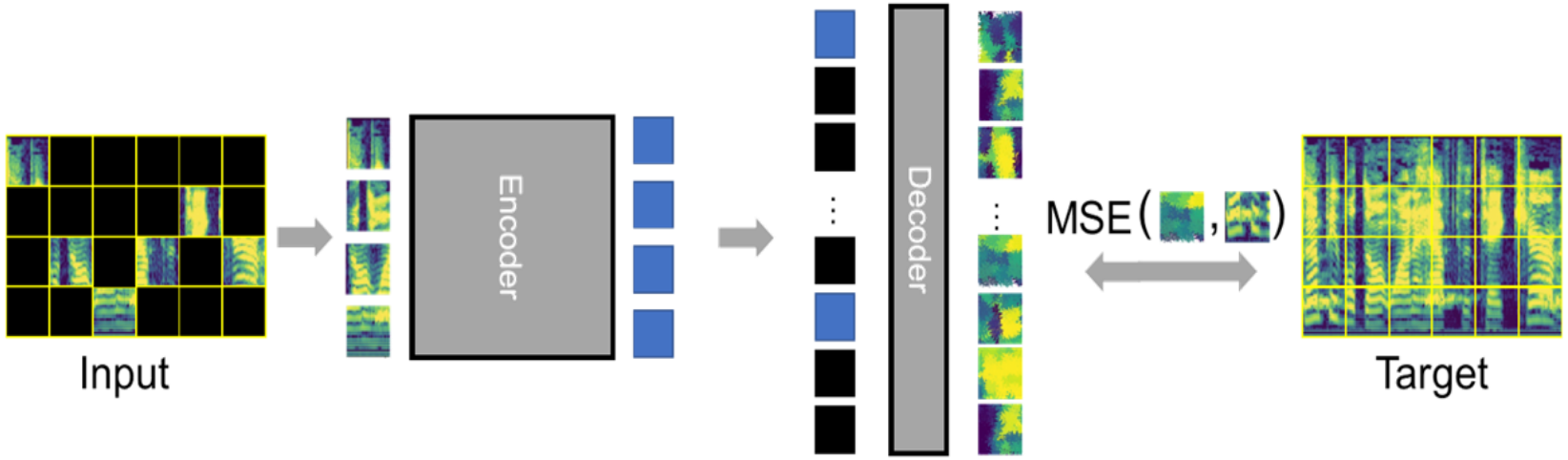
Language of Audio (LOA)





[1] A. Radford et al., “Language Models are Unsupervised Multitask Learners,” OpenAI technical report, 2019.

AudioMAE learns useful features from masked spectrograms



SELF-SUPERVISED

80% masked

the encoder sees only the visible 20% of patches

The decoder reconstructs hidden patches. AudioLDM 2 later uses the encoder features.

Output used later: continuous encoder features

AudioMAE Fig. 1

[1] P.-Y. Huang et al., “Masked Autoencoders that Listen,” NeurIPS 2022.

Why AudioMAE features are used for LOA

01 **Semantic**

Distinguish the intended event, style, and linguistic content.

02 **Temporal**

Preserve an evolving sequence rather than one global summary.

03 **Continuous**

Retain nuance without committing to a hard codebook.

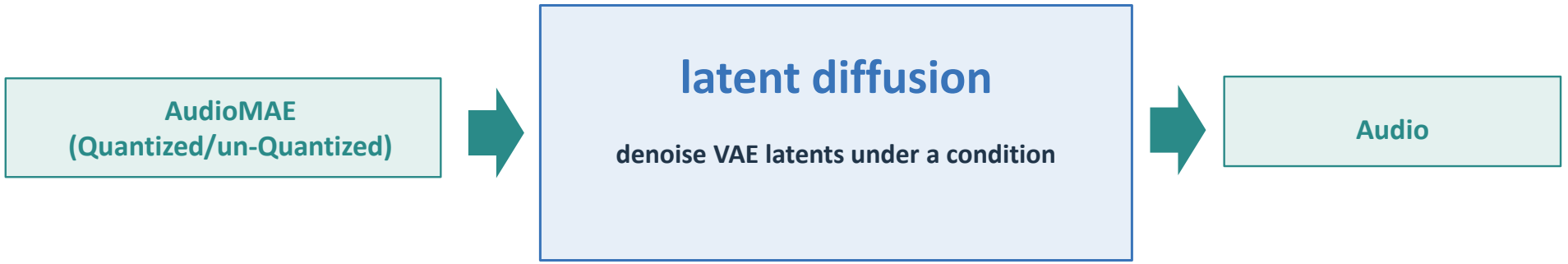
04 **Unlabelled-audio training**

Enable latent-diffusion pretraining without paired annotations.

AudioMAE provides ground-truth LOA; GPT-2 predicts LOA from paired conditions.

[1] P.-Y. Huang et al., “Masked Autoencoders that Listen,” NeurIPS 2022.
[2] A. Radford et al., “Language Models are Unsupervised Multitask Learners,” OpenAI technical report, 2019.

SemantiCodec shows what coarse quantization removes

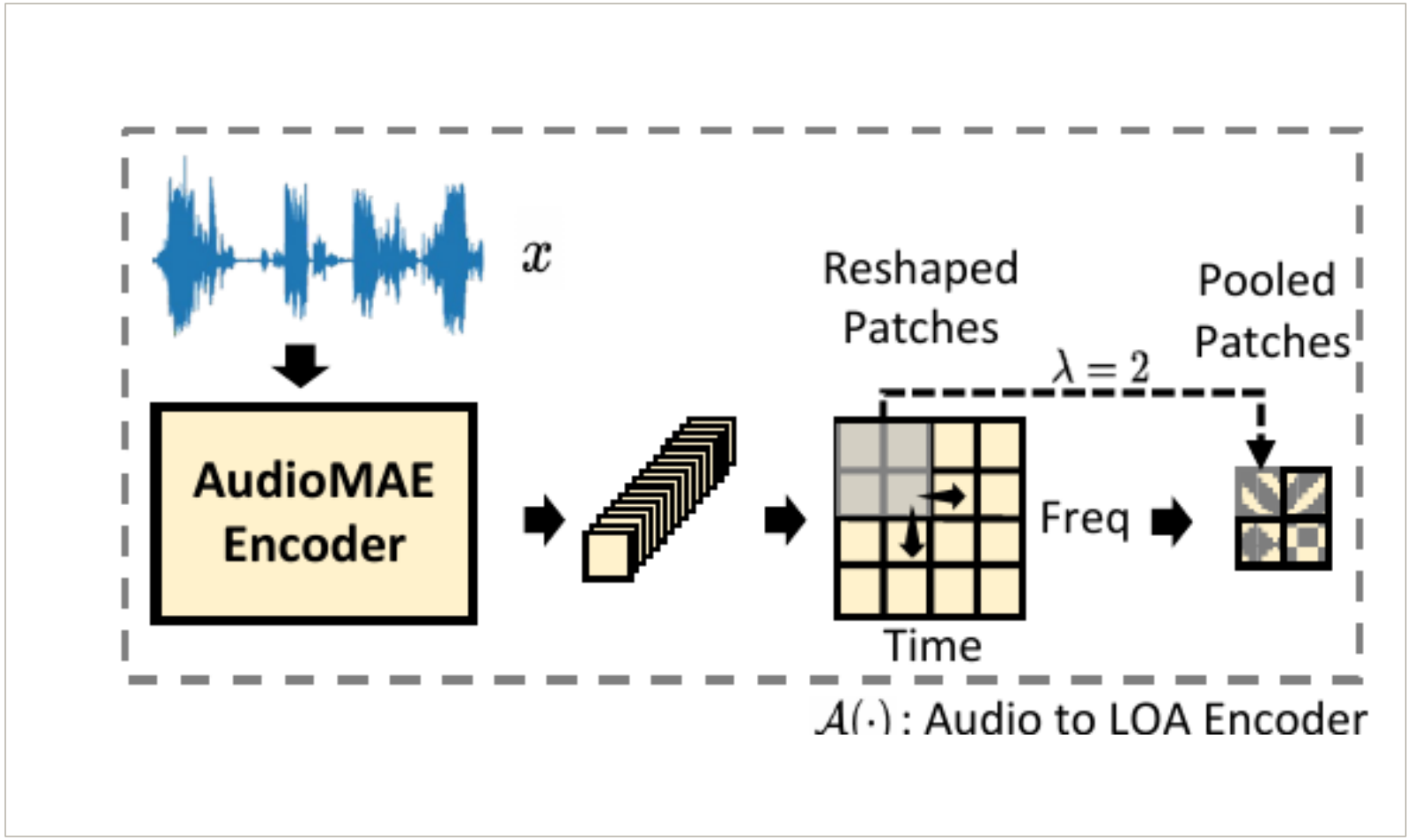


Representation	ViSQOL↑	WER↓	Rate
GT AudioMAE (unquantized)	4.61	2.7	unrestricted
Quantized AudioMAE (no acoustic VQ)	2.61	55.6	0.70 kbps 50 tok/s

Quantizing AudioMAE leads to significantly worse reconstruction performance.

[1] H. Liu et al., "SemantiCodec: An Ultra Low Bitrate Semantic Audio Codec for General Sound," arXiv 2024.
[2] P.-Y. Huang et al., "Masked Autoencoders that Listen," NeurIPS 2022.

AudioMAE provides the ground-truth LOA sequence



10 s audio

512 × 768

pre-pooling AudioMAE feature tensor E

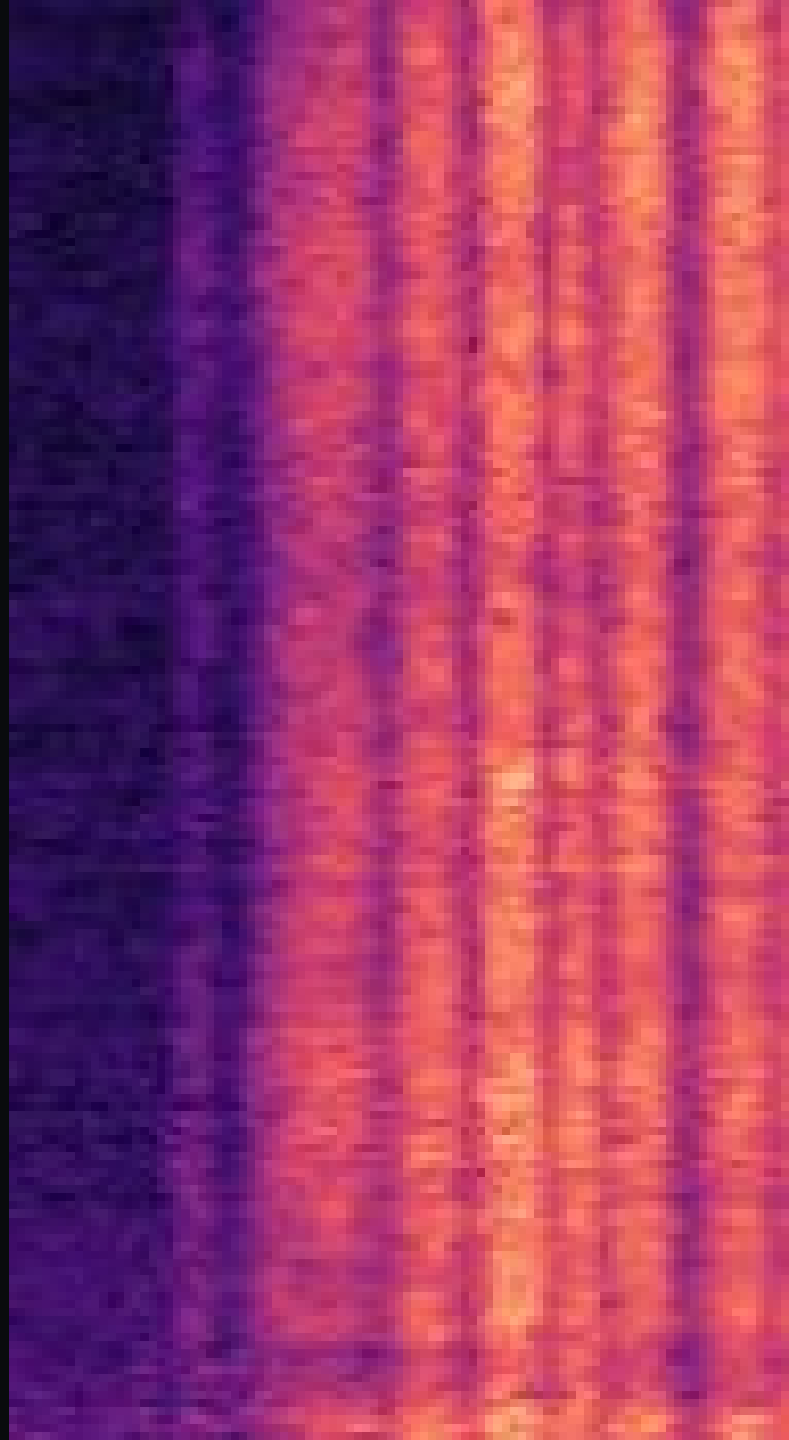
Pooling converts E into the ground-truth LOA sequence Yλ.

AudioMAE extracts the ground-truth LOA sequence

[1] P.-Y. Huang et al., “Masked Autoencoders that Listen,” NeurIPS 2022.

Combining the Pros and Cons from Both AR and Diffusion Models

GPT-2 predicts the Language of Audio; a latent diffusion model generates the waveform.



[1] A. Radford et al., “Language Models are Unsupervised Multitask Learners,” OpenAI technical report, 2019.

LM+Diffusion > Diffusion/LM?

Auto-regressive modeling:

- ✅ **Explicit modeling of temporal dependencies.**
- ✅ **Enjoy the advance of recent LLM development.**
- ✅ **Good in-context learning performance.**
- ❌ Long generation sequence/ lack of parallelism
- ❌ Long range dependencies
- ❌ Error propagation

Diffusion-based approach:

- ✅ **Stable**
- ✅ **State-of-the-art generation quality**
- ✅ **Flexible formulation for manipulation, interpolation, etc.**
- ❌ Do not explicitly model temporal dependencies
- ❌ Less flexible on duration

AUTOREGRESSIVE MODELS

AudioGen · general sound
MusicGen · music

DIFFUSION / FLOW MODELS

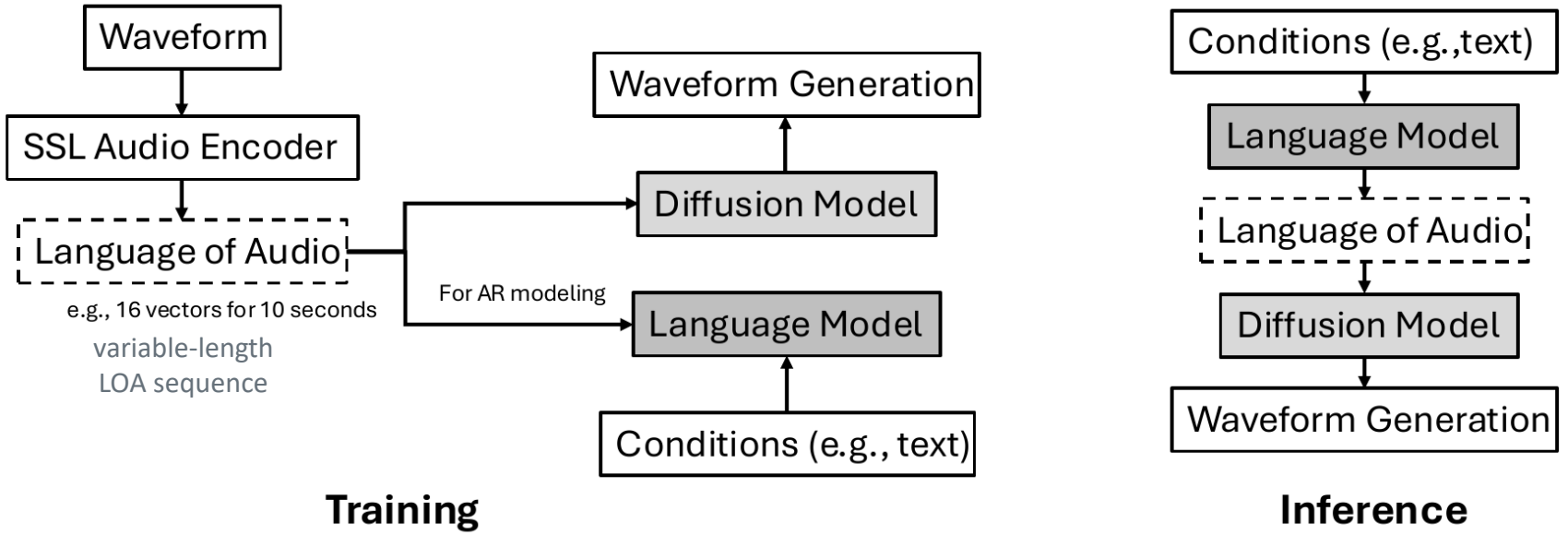
DiffSound · discrete diffusion
Make-An-Audio · AudioLDM · latent diffusion
Voicebox · flow matching for speech

AUDIOLDM 2'S ANSWER

Language modeling organizes content; latent diffusion renders high-quality audio

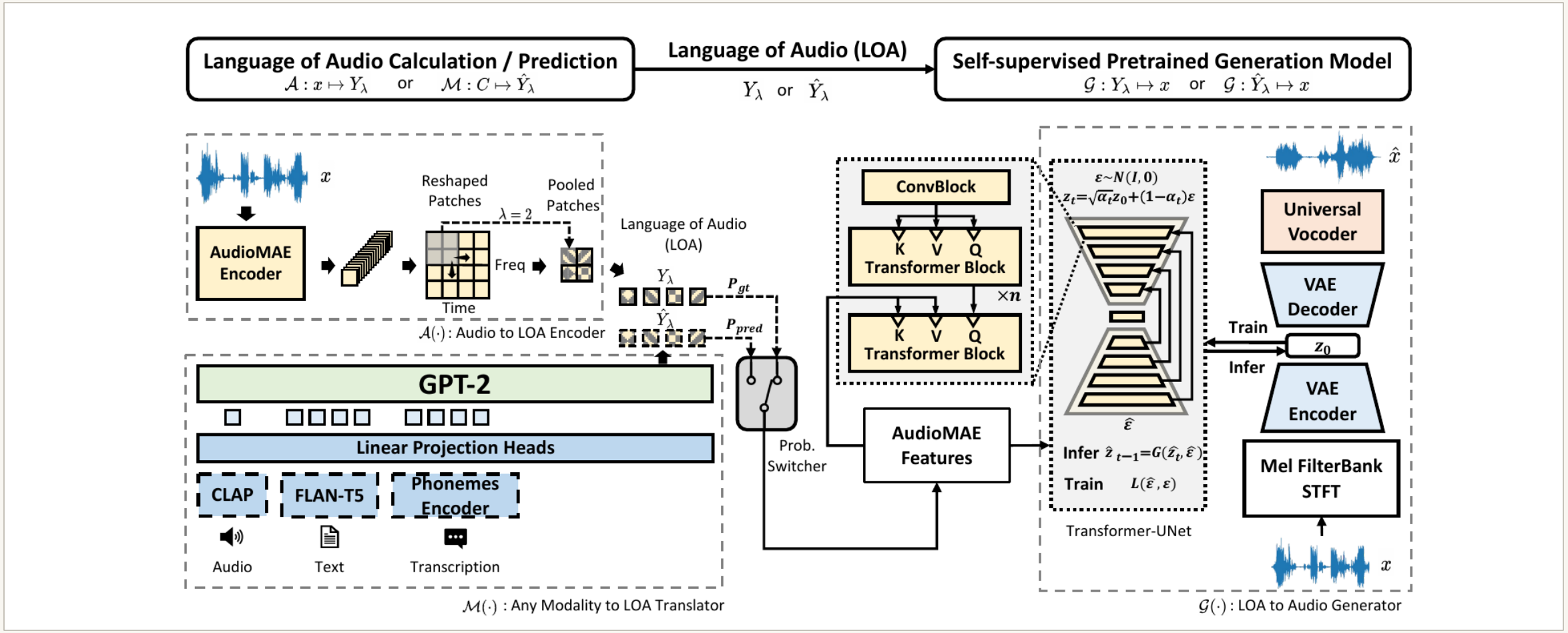
[1] F. Kreuk et al., "AudioGen," ICLR 2023. [2] J. Copet et al., "MusicGen," NeurIPS 2023.
[3] D. Yang et al., "DiffSound," IEEE/ACM TASLP 2023. [4] R. Huang et al., "Make-An-Audio," ICML 2023.
[5] H. Liu et al., "AudioLDM," ICML 2023. [6] M. Le et al., "Voicebox," NeurIPS 2023.

The same LOA representation connects training and inference



Training uses ground-truth LOA from audio; inference uses LOA predicted from conditions

AudioLDM 2 translates conditions into LOA, then LOA into audio

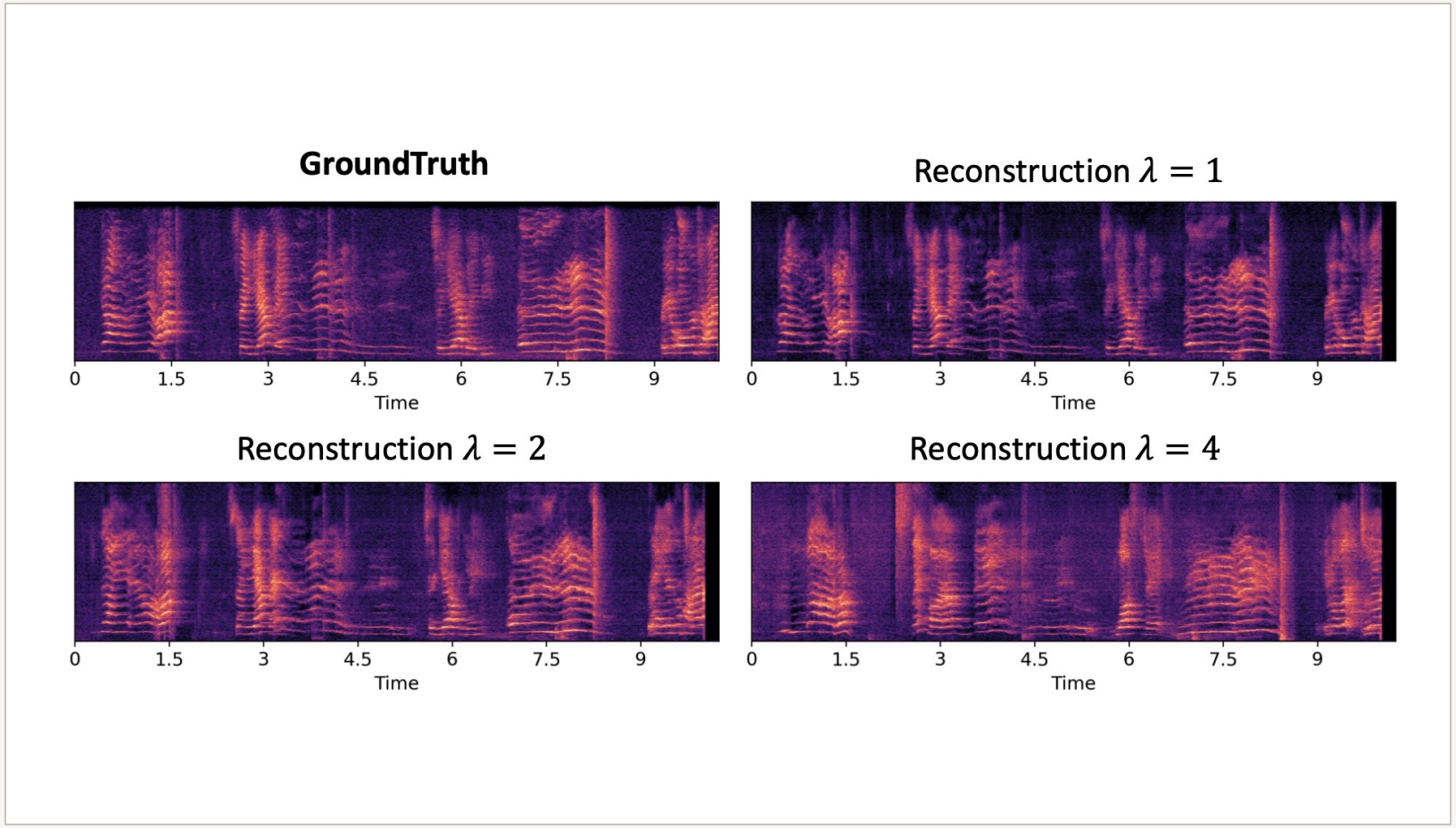


Condition-to-LOA GPT-2 → probabilistic switcher → LOA-to-audio latent diffusion

[1] P.-Y. Huang et al., "Masked Autoencoders that Listen," NeurIPS 2022.

[3] Y. Wu et al., "Large-Scale Contrastive Language-Audio Pretraining," ICASSP 2023.

More pooling shortens LOA but reduces reconstruction detail



AUDIO + MUSIC

λ = 8

8 LOA vectors

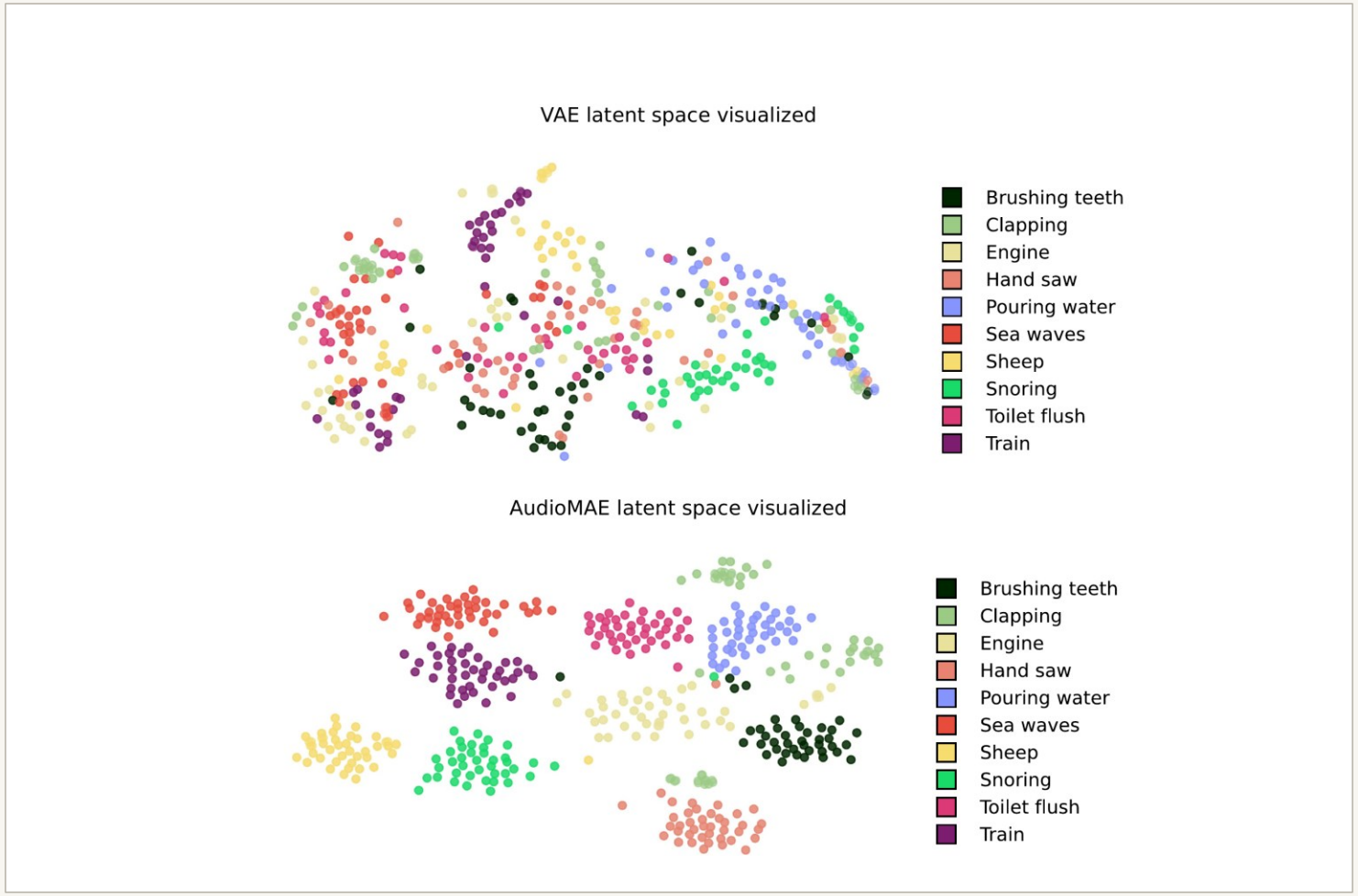
SPEECH

λ = 1

512 LOA vectors

A larger pooling factor λ shortens LOA but reduces reconstruction detail

AudioMAE separates ESC-50 sound classes more clearly than the VAE latent



Ten ESC-50 classes: VAE overlap versus AudioMAE clustering

VAE

optimized for mel reconstruction

AudioMAE

same-class samples cluster together

VAE and AudioMAE represent the same audio for different objectives.

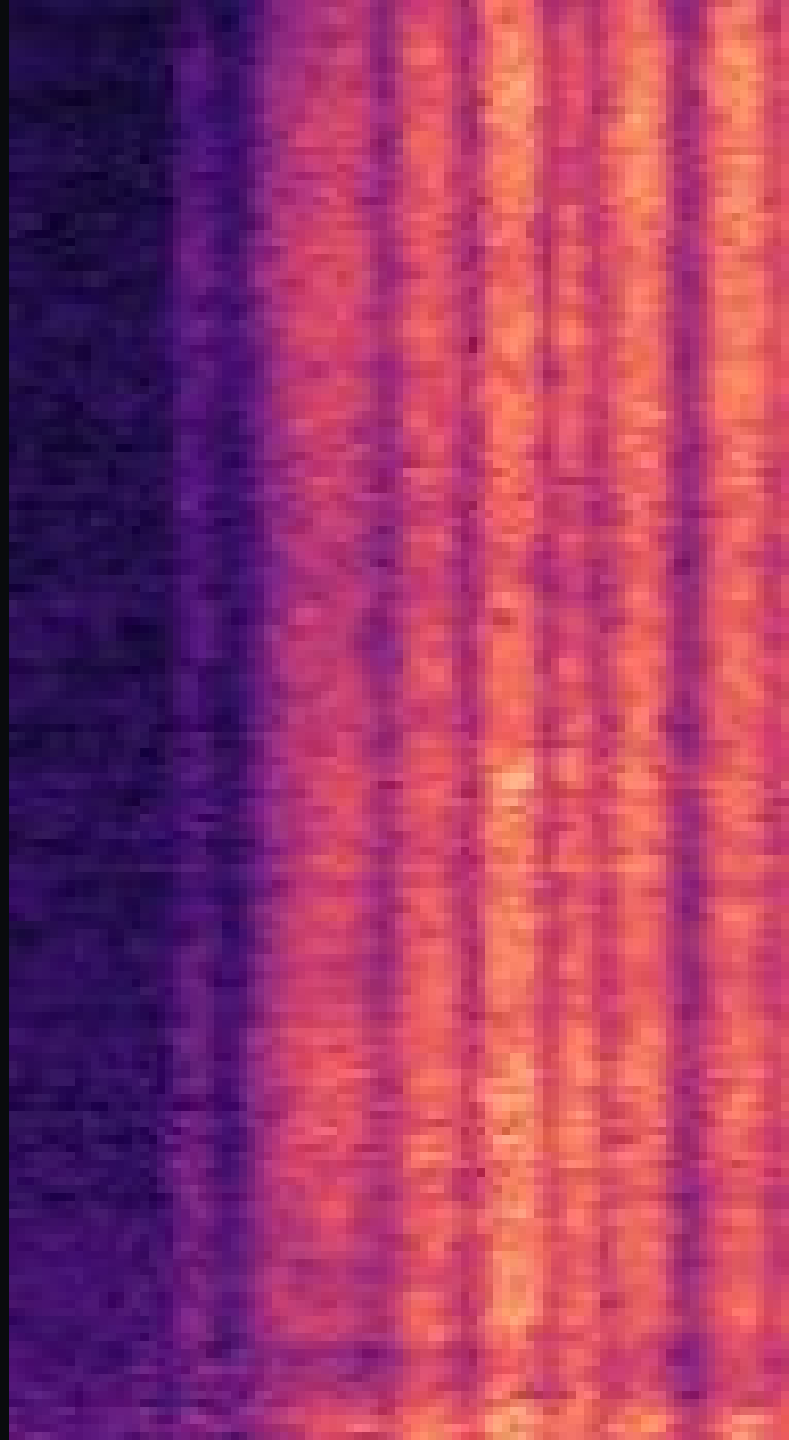
[1] P.-Y. Huang et al., “Masked Autoencoders that Listen,” NeurIPS 2022.
[2] K. Piczak, “ESC: Dataset for Environmental Sound Classification,” ACM Multimedia 2015.

The official one-minute demo spans sound effects, music, and speech

SECTION 2 · RESULTS

How well does AudioLDM 2 perform across tasks?

Objective fidelity, text alignment, and listener ratings measure different properties.



No single metric captures every aspect of audio quality

FAD / KL
Fréchet Audio Distance / KL divergence

Does the generated distribution resemble real audio?

LOWER IS BETTER

CLAP
audio–text alignment score

Does the sound match the text instruction?

HIGHER IS BETTER

OVL / REL / MOS
overall quality / relevance / mean opinion score

What do listeners judge as natural and relevant?

HIGHER IS BETTER

Fidelity, text relevance, and listener ratings answer different questions.

[1] Y. Wu et al., “Large-Scale Contrastive Language-Audio Pretraining,” ICASSP 2023.
[2] K. Kilgour et al., “Fréchet Audio Distance,” Interspeech 2019.

AudioLDM 2 checkpoints trade off AudioCaps metrics

Model	Train h	Params	FAD↓	KL↓	CLAP↑	OVL↑	REL↑
AudioLDM-M	9,031	416M	4.53	1.99	.141	3.61	3.55
Make-an-Audio 2	3,700	937M	2.05	1.27	.173	3.68	3.62
TANGO*	145	866M	1.73	1.27	.176	3.75	3.72
AudioLDM 2-AC*	145	346M	1.67	1.01	**.249**	3.88	**3.90**
AudioLDM 2-Full	29,510	346M	1.78	1.60	.191	3.83	3.77
AudioLDM 2-AC-Large*	145	712M	**1.42**	**0.98**	.243	**3.89**	3.87
AudioLDM 2-Full-Large	29,510	712M	1.86	1.64	.182	3.79	3.80

AC-Large leads FAD, KL, and OVL. AC leads CLAP and REL.

Both are AudioCaps-specialized checkpoints trained on the 145-hour subset.

* trained exclusively on AudioCaps

[1] H. Liu et al., "AudioLDM," ICML 2023.
[2] R. Huang et al., "Make-An-Audio 2," arXiv 2023.
[3] D. Ghosal et al., "TANGO: Text-to-Audio Generation Using Instruction-Guided Latent Diffusion," ACM Multimedia 2023.

More diverse training data did not improve AudioCaps results

Checkpoint	Train h	FAD↓	CLAP↑	What changed?
AC	145	**1.67**	**.249**	AudioCaps only
Full	29,510	1.78	.191	broader audio + music
AC-Large	145	**1.42**	**.243**	larger latent diffusion model
Full-Large	29,510	1.86	.182	larger + broader data

145 h
scores better on AudioCaps

≠

29,510 h
guaranteed AudioCaps gain

AudioLDM 2-Full leads 3 of 5 MusicCaps metrics

Model	FAD↓	KL↓	CLAP↑	OVL↑	REL↑
MusicGen-Medium (reported)	3.40	1.23	.320	—	—
MusicGen-Medium*	4.89	1.35	.291	3.37	3.38
AudioLDM-M*	3.20	1.29	**.360**	3.03	3.25
AudioLDM 2-MSD	4.47	1.32	.294	**3.41**	3.30
AudioLDM 2-Full	**3.13**	**1.20**	.301	3.34	**3.54**

Full leads FAD, KL, and REL; AudioLDM-M leads CLAP; MSD leads OVL.

Reported MusicGen score shown separately; * denotes reruns with publicly available implementations.

[1] J. Copet et al., “MusicGen: Simple and Controllable Music Generation,” NeurIPS 2023.
[2] H. Liu et al., “AudioLDM,” ICML 2023.

Speech pretraining improves MOS, but not to recorded-speech quality

System	MOS↑
Ground truth	4.63 ± 0.08
GT-AudioMAE	4.14 ± 0.13
FastSpeech2	3.78 ± 0.15
AudioLDM 2-LJS	3.65 ± 0.21
AudioLDM 2-LJS-Pretrained	4.00 ± 0.13

3.65 → 4.00

PRETRAINING EFFECT · GigaSpeech improves the condition-to-LOA GPT-2 model

Predicted LOA remains below ground-truth LOA and recorded speech.

One prompted-speech spectrogram crop

Thesis Fig. 4.7

[1] P.-Y. Huang et al., “Masked Autoencoders that Listen,” NeurIPS 2022.
[2] Y. Ren et al., “FastSpeech 2,” ICLR 2021.
[3] G. Chen et al., “GigaSpeech,” Interspeech 2021.
[4] A. Radford et al., “Language Models are Unsupervised Multitask Learners,” OpenAI technical report, 2019.

Joint fine-tuning improves FAD, KL, and CLAP score

Setting	FAD↓	KL↓	CLAP↑
AudioLDM 2-AC (joint-tuned)	**1.67**	**1.01**	**.249**
without joint fine-tuning	**2.24**	1.07	**.234**
without CLAP embedding in GPT	2.48	1.07	.245
without FLAN-T5 embedding in GPT	2.73	1.05	.250
without CLAP + FLAN-T5 in GPT	2.11	1.06	.213

Jointly fine-tuning GPT-2 and latent diffusion improves all three metrics.

[1] Y. Wu et al., "Large-Scale Contrastive Language-Audio Pretraining," ICASSP 2023.
[2] H. W. Chung et al., "Scaling Instruction-Finetuned Language Models," JMLR 2024.
[3] A. Radford et al., "Language Models are Unsupervised Multitask Learners," OpenAI technical report, 2019.

A lower FAD score can still mean worse text alignment

Setting	FAD↓	KL↓	CLAP↑
AudioLDM 2	1.67	1.01	.249
without FLAN-T5 cross-attention	1.38	1.30	.211

FAD

1.67 → 1.38

lower FAD

but

CLAP

.249 → .211

lower CLAP score

FAD and CLAP measure different properties.

[1] Y. Wu et al., “Large-Scale Contrastive Language-Audio Pretraining,” ICASSP 2023.
[2] H. W. Chung et al., “Scaling Instruction-Finetuned Language Models,” JMLR 2024.

AudioLDM 2 influenced later work through direct reuse and design influence.

Some systems reuse AudioLDM 2 weights or components. Others build on the broader two-stage architecture.

MODEL / COMPONENT REUSE

released checkpoints and components are reused for new tasks

ARCHITECTURAL INFLUENCE

predict an intermediate representation, then generate detailed audio

AudioLDM established a strong open baseline before AudioLDM 2

1,304

GOOGLE SCHOLAR CITATIONS

The ICML 2023 paper became a widely used reference for text-to-audio generation.

≈2,900

GITHUB STARS

The original open-source repository built a large developer community.

93,445

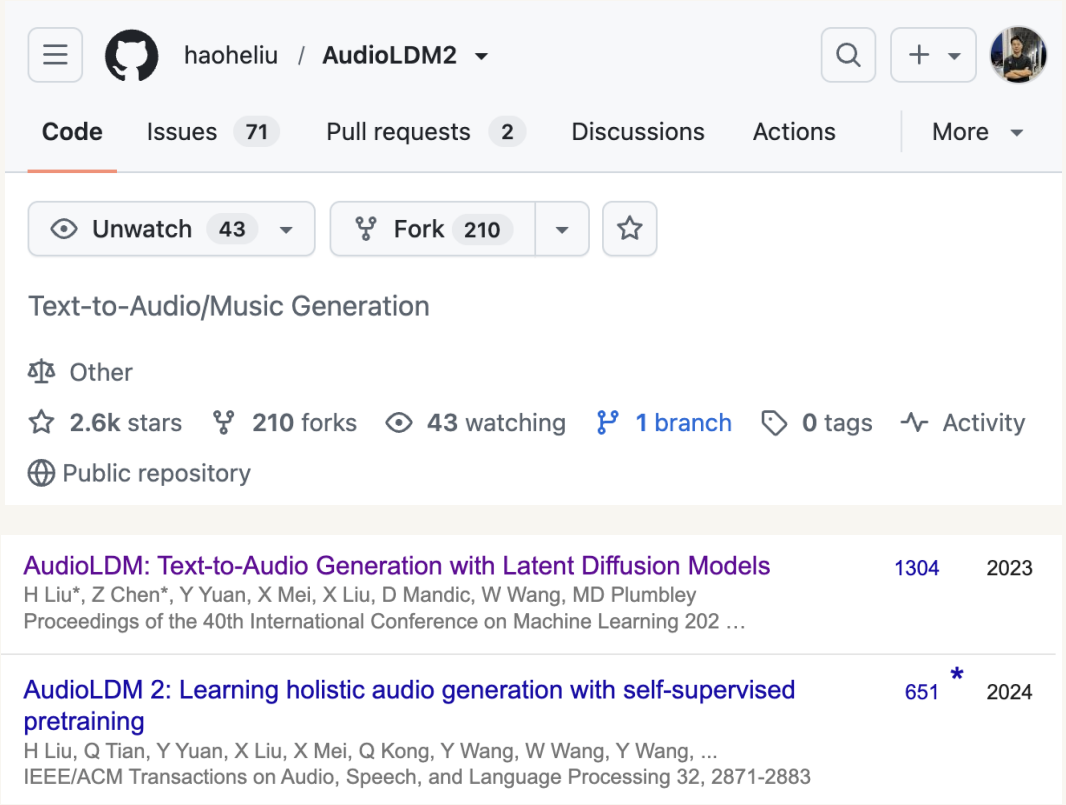
HUGGING FACE DEMO USES

The public demo made the predecessor accessible beyond paper readers.

AudioLDM 2 builds on a predecessor that already reached both research and developer communities.

[1] H. Liu et al., “AudioLDM: Text-to-Audio Generation with Latent Diffusion Models,” ICML 2023.

Public code and checkpoints lowered the barrier to reuse



PUBLIC SNAPSHOT · 10 AUG 2026 · SCREENSHOT ROUNDS STARS

2,638

GitHub stars

1.8M

Hugging Face download events
3 official model repos

AudioLDM2Pipeline

Diffusers implementation for inference

audioldm_eval

shared evaluation toolkit

These numbers show public availability and frequent downloads—not unique users or model quality.

From Official GitHub, Hugging Face, Diffusers, and audioldm_eval pages, accessed 10 Aug. 2026.

Follow-up systems relate to AudioLDM 2 in three different ways

1 USES AUDIOLDM 2 WEIGHTS

AudioLDM 2 checkpoints or denoising weights remain in the follow-up system.

2 USES A COMPONENT OR OUTPUT

A released component or AudioLDM 2-generated output is reused.

3 BUILDS ON THE TWO-STAGE IDEA

A later system predicts an intermediate representation before continuous audio generation.

These categories distinguish direct reuse from broader architectural similarity.

AudioLDM 2 became a backbone, a component, and an architectural reference

GROUPED IMPACT MAP · EVIDENCE-BOUNDED

RELATIONSHIP	REPRESENTATIVE WORKS	WHAT AUDIOLDM 2 CONTRIBUTES	WHAT FOLLOW-UPS ADD
DIRECT BACKBONE / WEIGHT REUSE	M²UGen [4] · MusicMagus [5] ZETA [6] · AP-Adapter [7] MEDIC [8] · SteerMusic [9] · ZeroSep [10]	released checkpoints, frozen denoiser / score model, or full generator	multimodal conditioning, editing + preservation, and source separation DIRECT REUSE · STRONGEST EVIDENCE
COMPONENT / OUTPUT REUSE	JavisDiT [11] XAttnMark [12]	frozen VAE or AudioLDM 2-generated edits	joint audio–video diffusion or watermark-robustness testing NARROWER REUSE
EXPLICIT EXTENSION / RELATED DIRECTION	LeVo [13] InspireMusic [14] AudioCALM [15]	the LM / autoregressive → continuous-generation division of labor	song tokens + diffusion codec, super-resolution flow, or continuous latent flow EXPLICIT FOR LEVO; RELATED FOR OTHERS

Direct reuse is the strongest evidence. "Related direction" does not mean weight reuse. Primary sources [4]–[15] are listed next.

Backbone reuse: AudioLDM 2 still does the core audio work

DIRECT BACKBONE REUSE

M²UGEN [4]

REUSES
AudioLDM 2 as a selectable audio generator

ADDS
multimodal LLM + output projector

Multimodal-to-audio generation using AudioLDM 2 as the audio generator.

AP-ADAPTER [7]

REUSES
frozen AudioLDM 2-Large + AudioMAE

ADDS
22M-parameter reference-audio adapter

Reference-guided audio generation using a frozen AudioLDM 2 backbone.

ZEROSEP [10]

REUSES
pretrained AudioLDM 2 variants

ADDS
mixture inversion + text-guided denoising

Text-guided source separation using AudioLDM 2 as a generative prior.

AudioLDM 2 stays in the system; each follow-up changes the input, control method, or task.

[4] S. Liu et al., "M²UGen," arXiv 2023.
[7] F.-D. Tsai et al., "Audio Prompt Adapter," ISMIR 2024.
[10] C. Huang et al., "ZeroSep," NeurIPS 2025.

Component reuse: only a part—or an output—is carried forward

NOT FULL-MODEL REUSE

JAVISDIT [11]

AUDIOLDM 2 CONTRIBUTES

frozen VAE encoder + decoder

FOLLOW-UP BUILDS

joint audio–video diffusion transformer

reused component

XATTNMARK [12]

AUDIOLDM 2 CONTRIBUTES

generated audio edits used as stress tests

FOLLOW-UP BUILDS

watermark model robust to edits

reused output

AudioLDM 2 is not the generator inside either new system: only its VAE or generated data is reused.

[11] K. Liu et al., “JavisDiT,” ICLR 2026.
[12] Y. Liu et al., “XAttnMark,” ICML 2025.

Autoregressive-plus-flow systems now span speech, music, and audio

RELATED ARCHITECTURES · NO WEIGHT REUSE

DOMAIN	SYSTEM	AUTOREGRESSIVE REPRESENTATION	FLOW-MATCHING GENERATOR
SPEECH	CosyVoice 2 2024	semantic speech tokens	causal flow matching → mel related architecture · no direct extension claim
MUSIC + AUDIO	InspireMusic 2025	Qwen2.5 → coarse audio tokens	super-resolution flow → 48 kHz cites AudioLDM 2 · independent weights
SPEECH + SOUND + MUSIC	AudioCALM 2026 preprint	continuous latent blocks in one causal LM	flow head generates each continuous latent block cites AudioLDM 2 · independent weights

The intermediate representations, training objectives, and model weights are new.

[16] Z. Du et al., "CosyVoice 2," arXiv 2024.
[14] C. Zhang et al., "InspireMusic," technical report 2025.
[15] H. Liu et al., "AudioCALM," preprint 2026.

LIVE Q&A

Thank you!

PAPER

doi.org/10.1109/TASLP.2024.3399607

DEMOS

audioldm.github.io/audioldm2

CODE

github.com/haoheliu/AudioLDM2