ETTA: Elucidating the design space of text-to-audio models

Sang-gil Lee*, Zhifeng Kong*, Arushi Goel, Sungwon Kim, Rafael Valle, Bryan Catanzaro

(*: Equal Contribution)

ICML 2025

Background: Text-to-audio models

Tiny potato kings wearing majestic crowns, sitting on thrones, overseeing their vast potato kingdom filled with potato subjects and potato castles.

Text-to-image models (e.g. DALL-E 3) turn a text description into a high-quality image.

A rolling ball gradually losing momentum as it collides with metal surfaces.

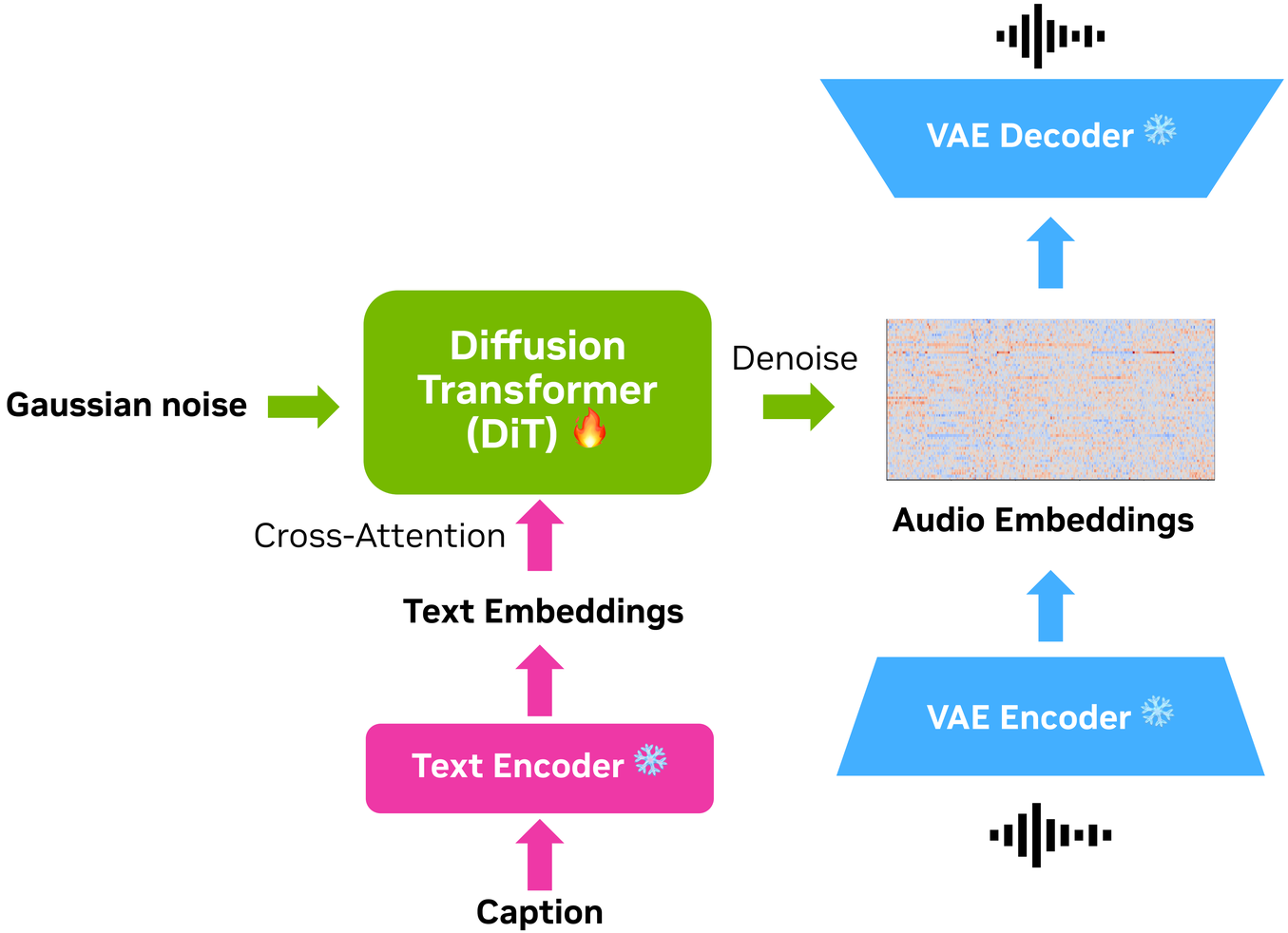
Text-to-audio models, similarly, turn a text description into a high-quality audio.

ETTA: Elucidated Text-to-Audio Model

With a focus on **Latent Diffusion Model / Flow Matching**

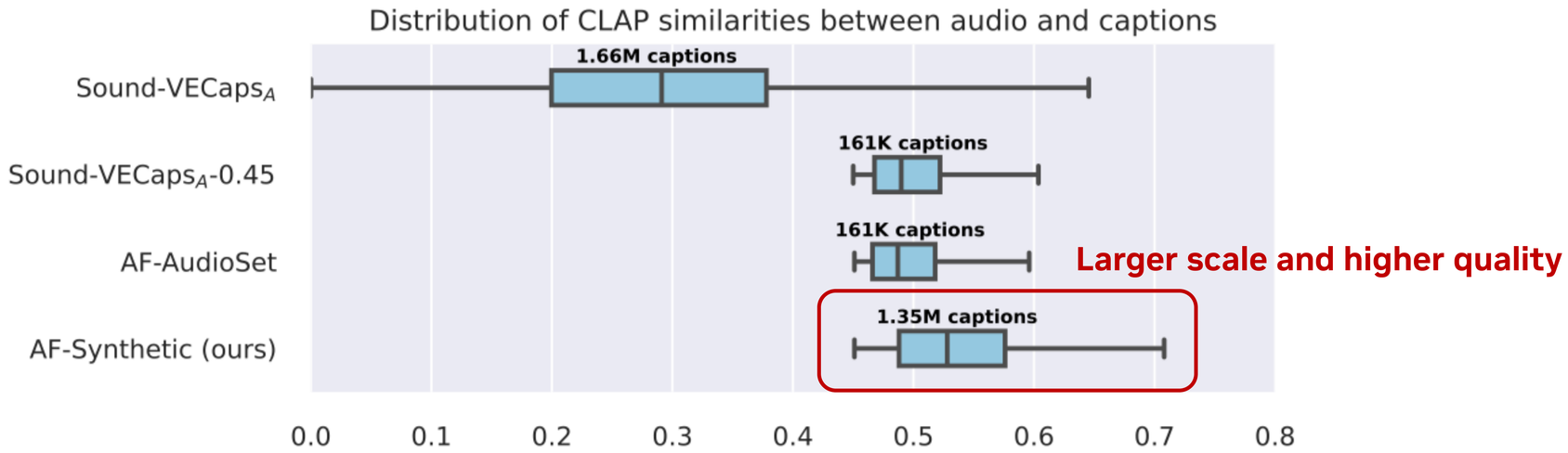
Main Contributions

- **AF-Synthetic**: the first million-scale, high quality, and synthetic caption dataset for TTA.
- We conduct a comprehensive study on **data**, **architectural design**, **training objectives**, and **sampling methods**.
- We implemented a better **DiT**.



AF-Synthetic: Scaling to **>1M** high-quality, synthetic captions

Dataset	Generation Model	Filtering Method	# Hours	# Captions
TangoPromptBank	Collected	None	3.5K	1.21M
Sound-VECaps$_A$	CogVLM + EnClap	Removing visual-only data	14.3K	1.66M
Sound-VECaps$_A$-0.45†	CogVLM + EnClap	CLAP $\geq$ 0.45	448	161K
AutoCap‡	AutoCap	Removing music or speech	8.7K	761K
AF-AudioSet	Audio Flamingo	CLAP $\geq$ 0.45	255	161K
AF-Synthetic (ours)	Audio Flamingo	CLAP $\geq$ 0.45 and others	3.6K	1.35M



Larger scale and higher quality

ETTA shows an emergent ability to generate **creative audio** through **scaling with synthetic data**.

Model	AudioLDM2	TANGO2	Stable Audio Open	*ETTA*
OVL↑	3.95 ± 0.05	3.82 ± 0.05	3.94 ± 0.05	$\mathbf{3.99} \pm 0.05$
REL↑	3.79 ± 0.06	3.94 ± 0.05	3.95 ± 0.05	$\mathbf{4.05} \pm 0.05$

MOS Results on Our Creative Audio Generation Benchmark

A saxophone that sounds like meowing of cat.

ETTA is the **SOTA** text-to-**audio** model

Model	Ground Truth	AudioLDM2-Large	TANGO2	Stable Audio Open	*ETTA*	*ETTA*-FT-AC-100k
OVL↑	3.43 ± 0.11	3.00 ± 0.11	3.08 ± 0.10	$\underline{3.29} \pm 0.11$	$\mathbf{3.43} \pm 0.11$	3.26 ± 0.10
REL↑	3.62 ± 0.10	3.11 ± 0.10	3.66 ± 0.09	3.15 ± 0.11	$\underline{3.68} \pm 0.10$	$\mathbf{3.77} \pm 0.10$

AudioCaps MOS Results

A heavy rainstorm with intermittent thunderclaps and raindrops hitting a tin roof.

A rolling ball gradually losing momentum as it collides with metal surfaces.

A construction site alongside a highway, with jackhammer breaking concrete and metal clanging.

ETTA is the **SOTA** text-to-**music** model

Model	Ground Truth	AudioLDM2-Large	TANGO-AF	Stable Audio Open	*ETTA*
OVL↑	3.88 ± 0.10	3.25 ± 0.10	3.38 ± 0.09	$\mathbf{3.92} \pm 0.10$	$\underline{3.53} \pm 0.10$
REL↑	3.90 ± 0.10	3.15 ± 0.10	3.31 ± 0.10	$\underline{3.35} \pm 0.11$	$\mathbf{3.57} \pm 0.10$

MusicCaps MOS Results

This track is a groovy tech house tune with mellow piano melodies. The mood evoked is 'sentimental' and 'dreamy', it follows a 4/4 time signature It's in C minor key, has a tempo of 128 BPM.

The music is an instrumental electronic track with reggae influences in the style of dub. It creates a relaxing mood suitable for a summer day, played at a moderate tempo of 85.7 BPM in G minor key and features a 4/4 time signature.

The music is an instrumental post-rock and ambient drum track in scenes depicting vast landscapes or introspective moments. Ab major with a tempo of around 84 BPM.

This track is an instrumental hip hop loop with melodic violin strings. The key is in D minor with a tempo of 80 BPM. The genre suggests an urban feel with its catchy drum style that harmonizes with violins.

* Sample from a fine-tuned model using Audio Flamingo 2 synthetic captions

Bonus: ETTA Remixing

Studio-quality stereo audio recording of a **cinematic symphonic orchestra**, prominently featuring lush strings including violins, violas, cellos, and double bass, complemented by dynamic, resonant percussion instruments such as timpani, bass drums, cymbals, and orchestral chimes. The composition evokes an **expansive, grandiose, and inspirational mood, characterized by majestic crescendos,** sweeping melodic lines, and deep emotional resonance. Rich stereo imaging with clear spatial separation of instruments, capturing subtle acoustics of a spacious concert hall environment with natural reverb and warm tonal balance.

Input music **ETTA remix**

* Sample from a fine-tuned model using Audio Flamingo 2 synthetic captions

- Paper: https://arxiv.org/abs/2412.19351
- Demo: https://research.nvidia.com/labs/adlr/ETTA/
- Code: https://github.com/NVIDIA/elucidated-text-to-audio