IMPACT: Iterative Mask-based Parallel Decoding for Text-to-Audio Generation with Diffusion Modeling

Kuan-Po Huang [1 2 †] Shu-wen Yang [1 2 †] Huy Phan [2] Bo-Ru Lu [2] Byeonggeun Kim [2] Sashank Macha [2]
Qingming Tang [2] Shalini Ghosh [2] Hung-yi Lee [1] Chieh-Chi Kao [2] Chao Wang [2]

ICML 2025

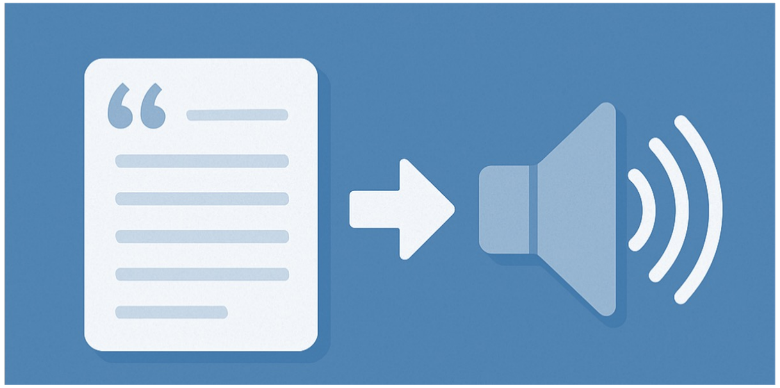
Presenter: Kuan-Po, Huang

amazon

https://arxiv.org/abs/2506.00736

Text-to-audio generation

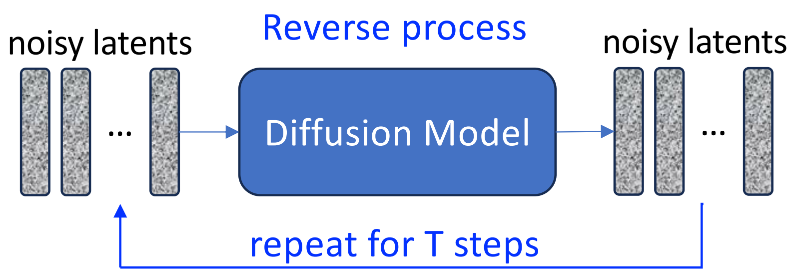
- Converting a written description into a corresponding sound or audio.



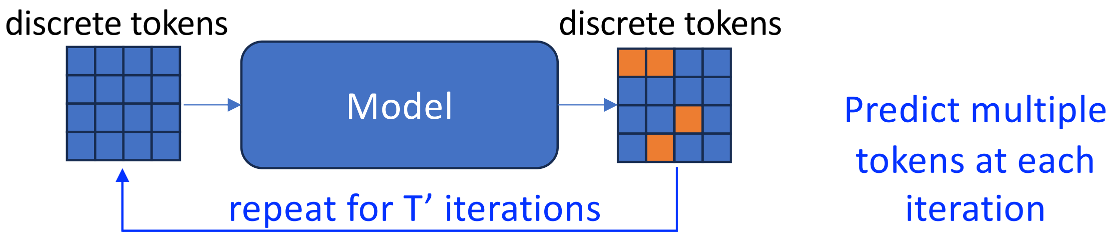
Background

<u>Problems of current Text-to-audio systems:</u>

- High performance on objective metrics, but slow: AudioLDM, Tango, …
  - Heavy-parameterized diffusion-based models



- Fast, but poor performance on objective metrics: MAGNET
  - Iterative parallel decoding discrete tokens



Background

Problems of current Text-to-audio systems:

- High performance on objective metrics, but slow: AudioLDM, Tango, …
  - Heavy-parameterized latent diffusion-based models (LDMs) operating on continuous representations
- Fast, but poor performance on objective metrics: MAGNET
  - Iterative parallel decoding discrete tokens

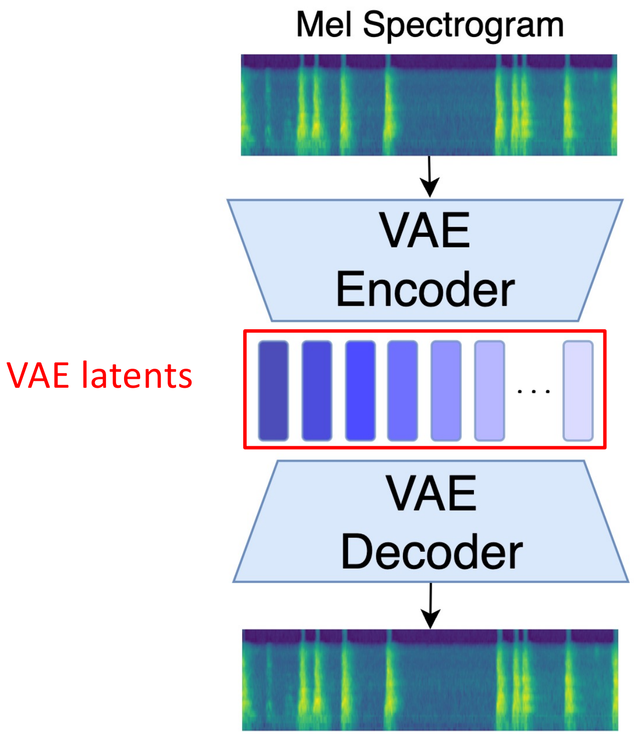
Propose:
Integrate iterative parallel decoding with LDMs operating
on continuous representations using a light-weight diffusion head for text-to-audio.

MAR https://arxiv.org/abs/2406.11838

Methodology

Integrate **iterative parallel decoding** with **LDMs** operating on **continuous representations**

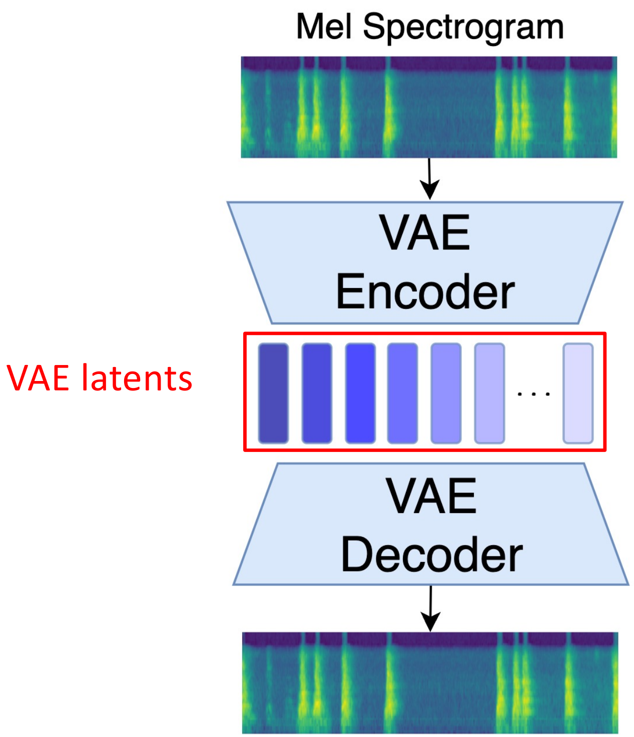
- Continuous representations



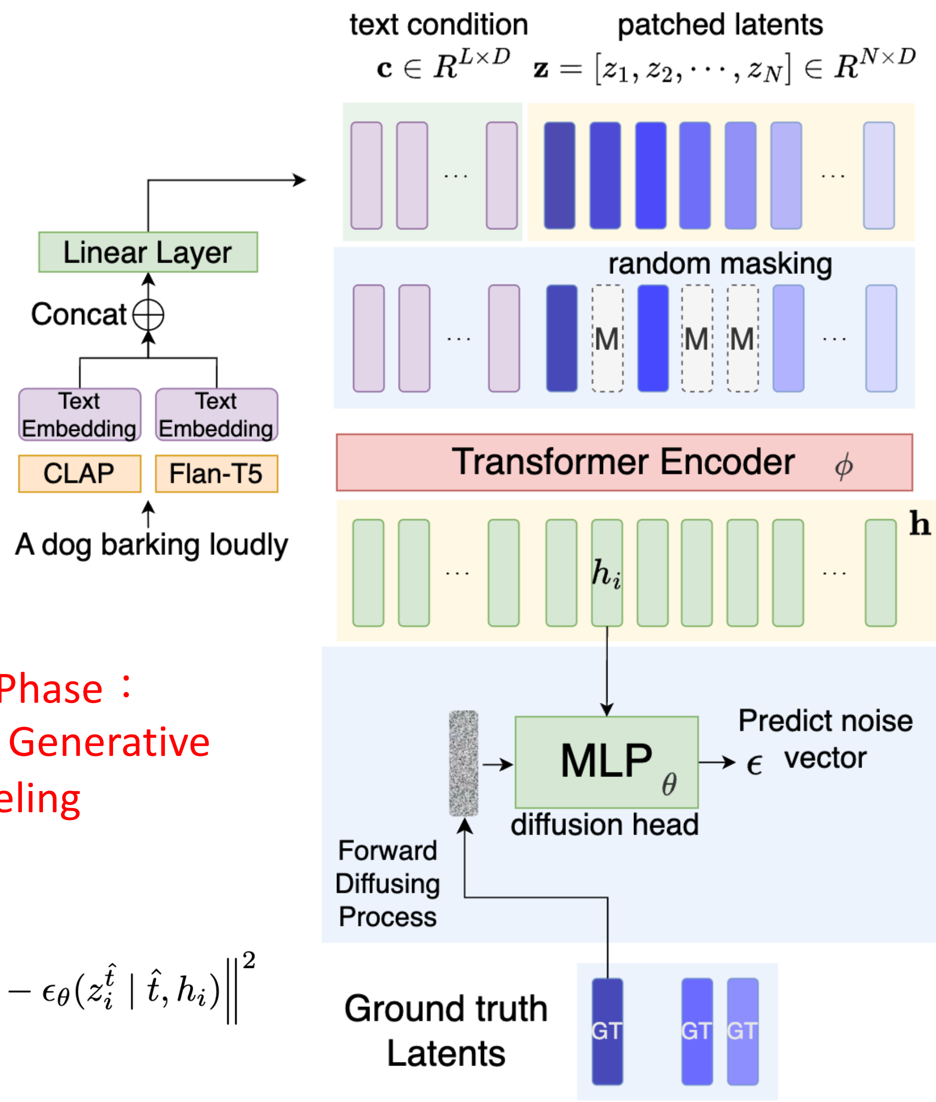
Methodology

Training: Mask generative modeling

- Continuous representations



Training Phase：
Mask-based Generative Modeling



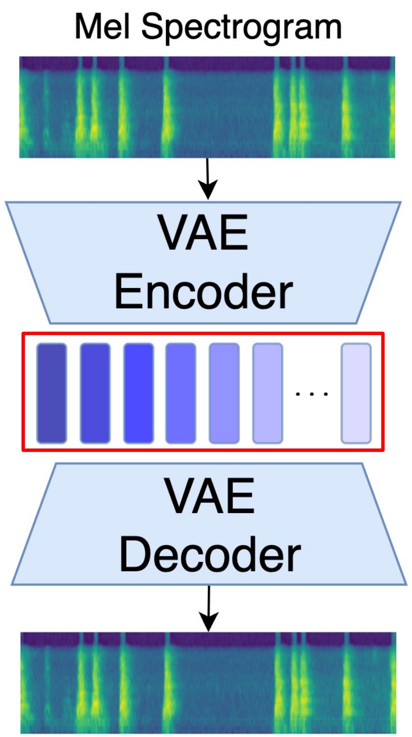
$$\arg\min_{\{\phi,\theta\}} \sum_{\{i|M[i]=1\}} \left\| \epsilon - \epsilon_\theta(z_i^{\hat{t}} \mid \hat{t}, h_i) \right\|^2$$

Methodology

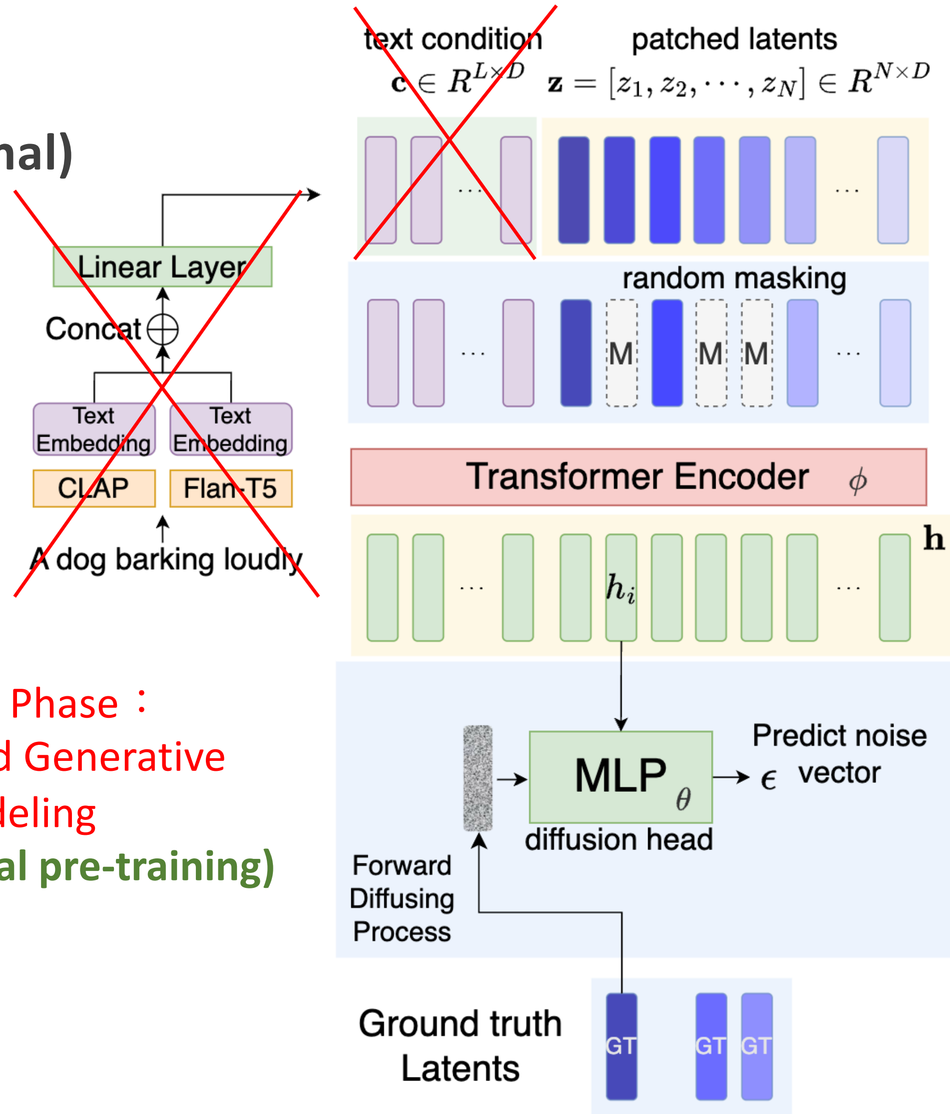
Training: Mask generative modeling (Unconditional)

- Continuous representations

VAE latents



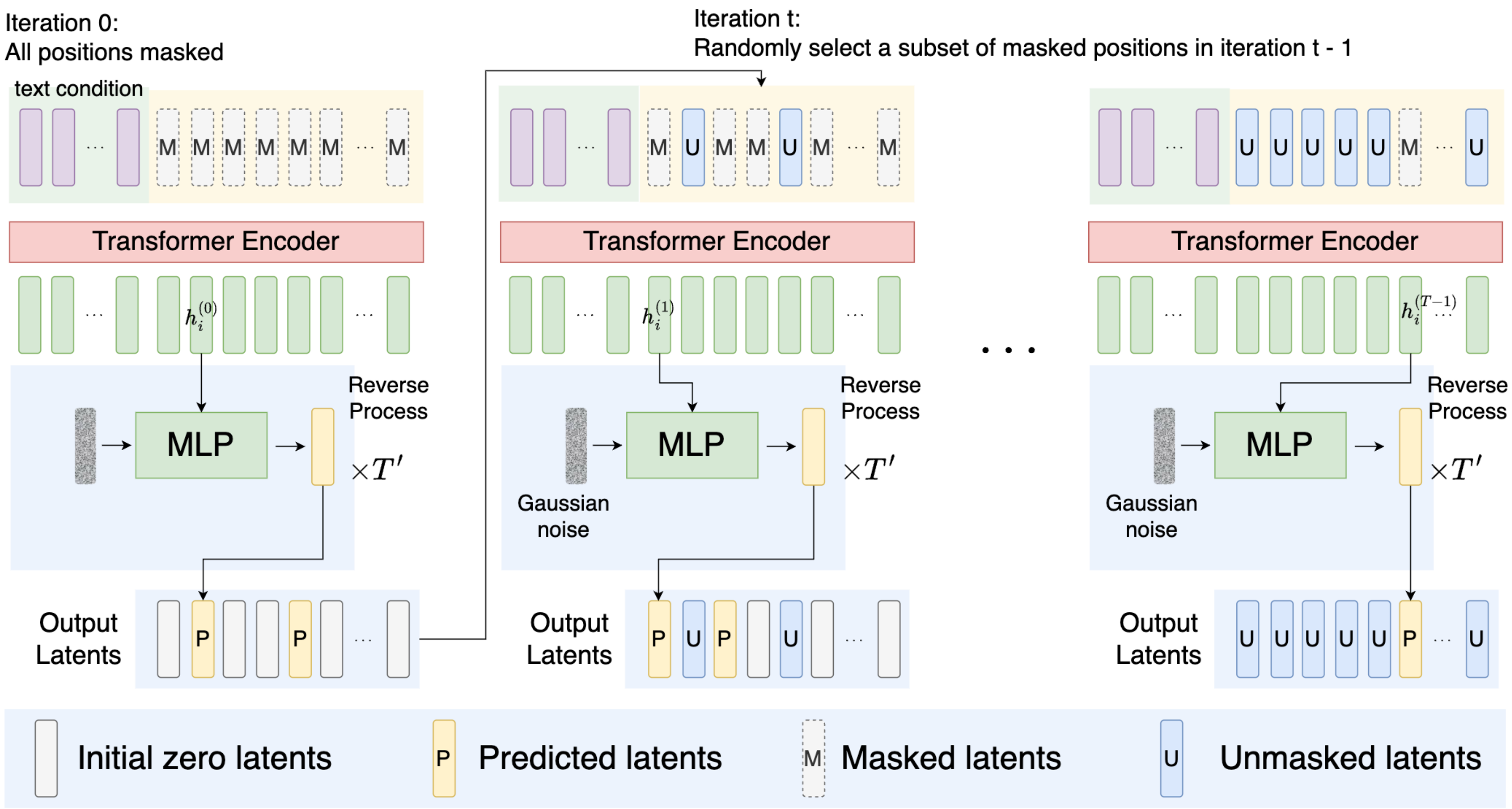
Training Phase：
Mask-based Generative
Modeling
(Unconditional pre-training)



Methodology

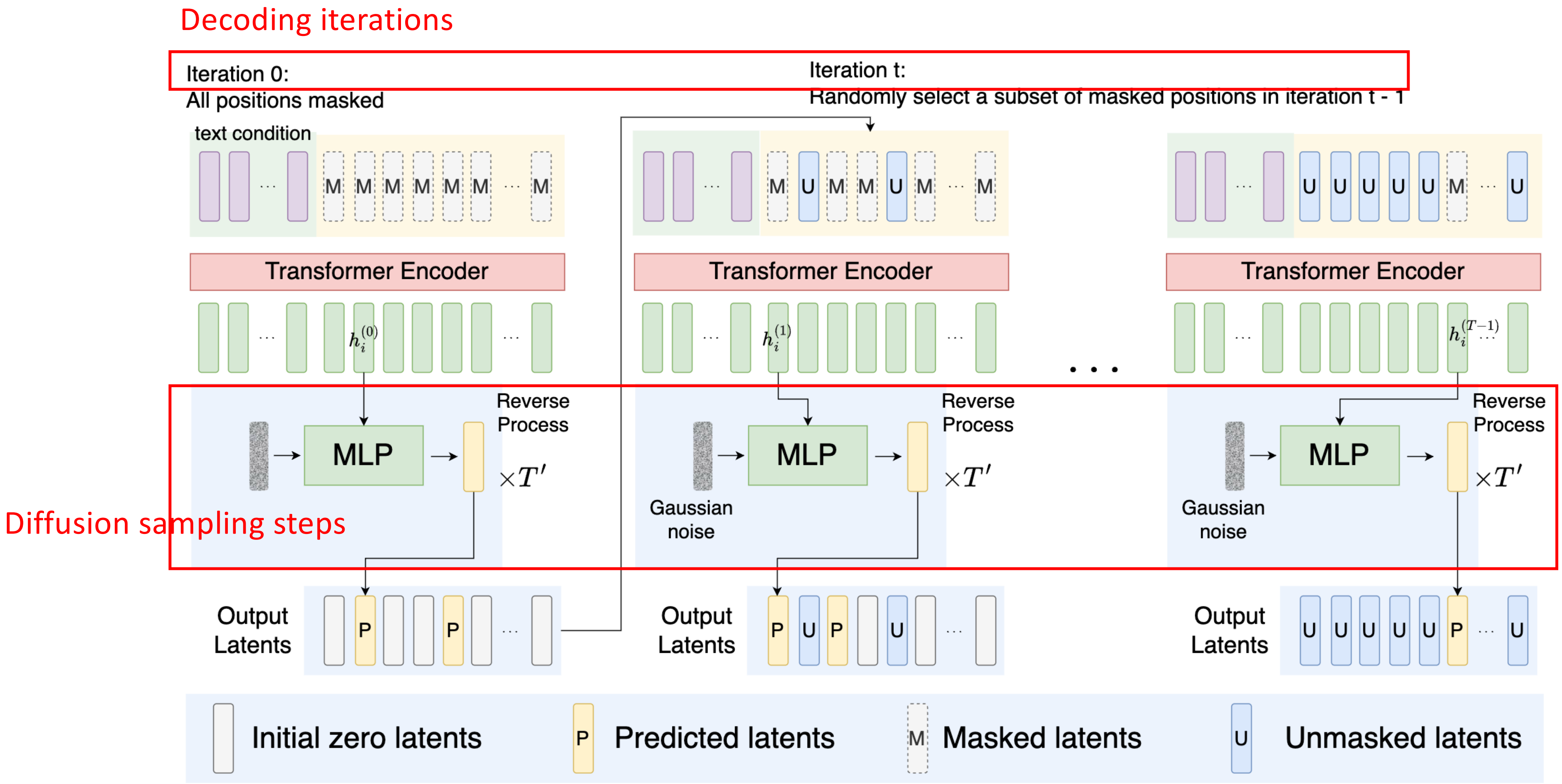
- **Inference: Iterative parallel decoding**

Gradually generate the latent sequence



Methodology

- **Iterative parallel decoding** (Inference Phase): Gradually generate the latent sequence



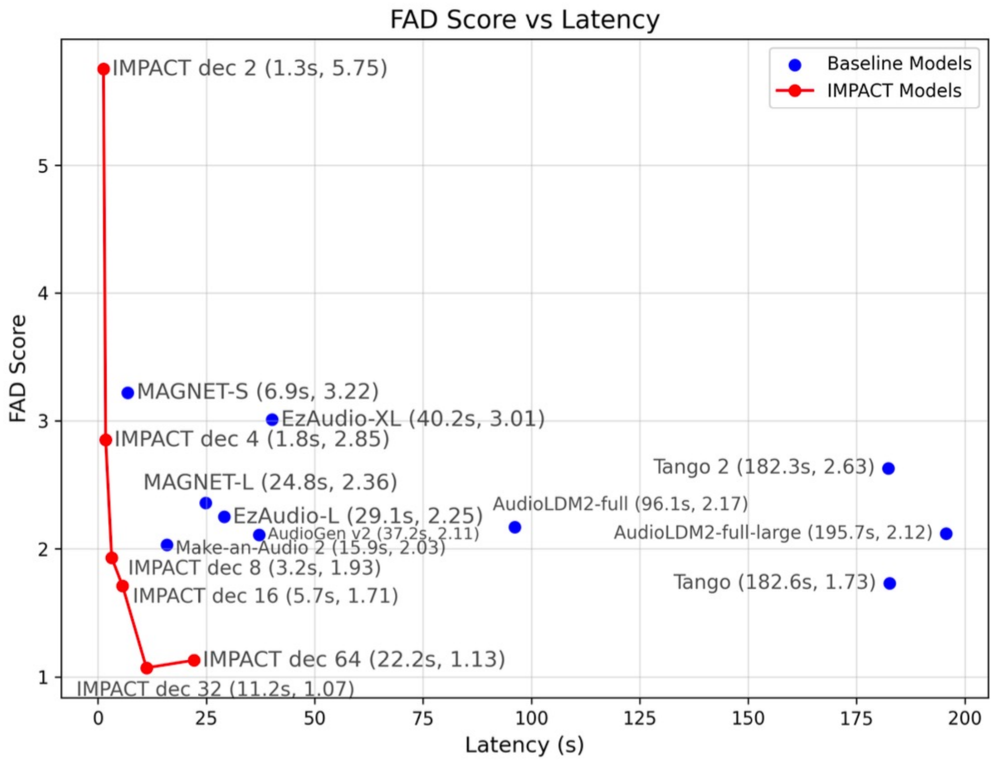
Results on objective and subjective metrics

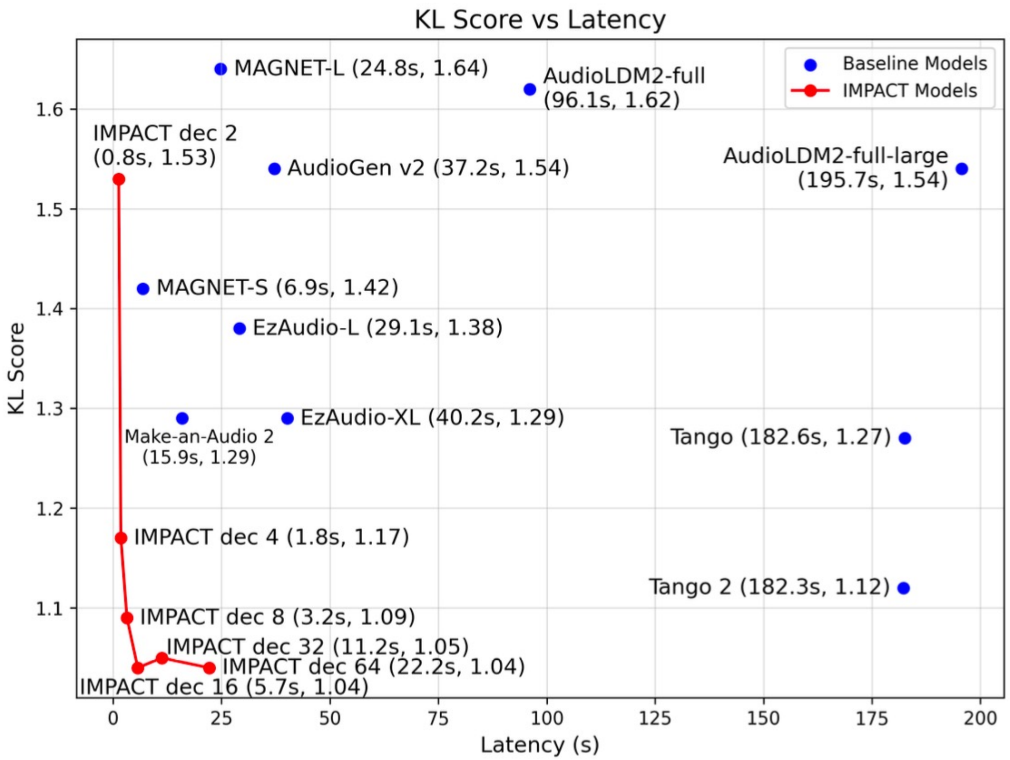
Latency: Required time for generating a batch of 8 audios. (measured in seconds)

AudioCaps	# para	FD ↓	FAD ↓	KL ↓	IS ↑	CLAP ↑	REL ↑	OVL ↑	diff.	Lat. ↓
Ground Truth	-	-	-	-	-	0.373	4.43	3.57	-	-
AudioGen	1.5B	16.51	2.11	1.54	9.64	0.315	-	-	-	37.2
Tango	866M	24.42	1.73	1.27	7.70	0.313	-	-	200	182.6
Tango-full-ft	866M	18.93	2.19	1.12	8.80	0.340	-	-	200	181.6
Tango-AF&AC-FT-AC	866M	21.84	2.35	1.32	9.59	0.343	-	-	200	182.6
Tango 2	866M	20.66	2.63	1.12	9.09	0.375	4.13	3.37	200	182.3
EzAudio-L (24kHz)	596M	15.59	2.25	1.38	11.35	**0.391**	4.05	3.44	50	29.1
EzAudio-XL (24kHz)	874M	14.98	3.01	1.29	**11.38**	0.387	4.00	3.35	50	40.2
MAGNET-S	300M	23.02	3.22	1.42	9.72	0.287	3.83	2.84	-	**6.9**
MAGNET-L	1.5B	26.19	2.36	1.64	9.10	0.253	-	-	-	24.8
Make-an-Audio 2	160M	16.23	2.03	1.29	9.95	0.345	-	-	100	15.9
AudioLDM2-full	346M	32.14	2.17	1.62	6.92	0.273	3.74	3.19	200	96.1
AudioLDM2-full-large	712M	33.18	2.12	1.54	8.29	0.281	-	-	200	195.7
IMPACT base, dec iter 32	193M	14.90	**1.07**	**1.05**	10.06	0.364	4.20	3.46	100	11.2
IMPACT base, dec iter 64	193M	**14.72**	1.13	1.09	10.03	0.353	**4.31**	**3.51**	100	22.2

Latency analyses

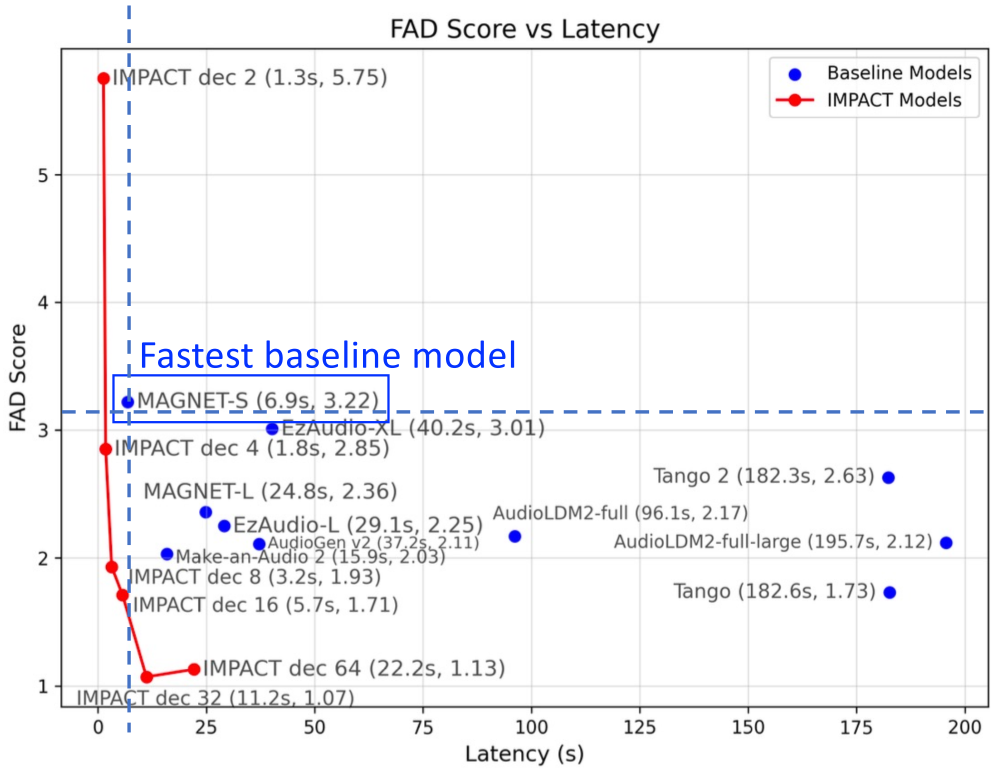
Latency: Required time for generating a batch of 8 audios. (measured in seconds)

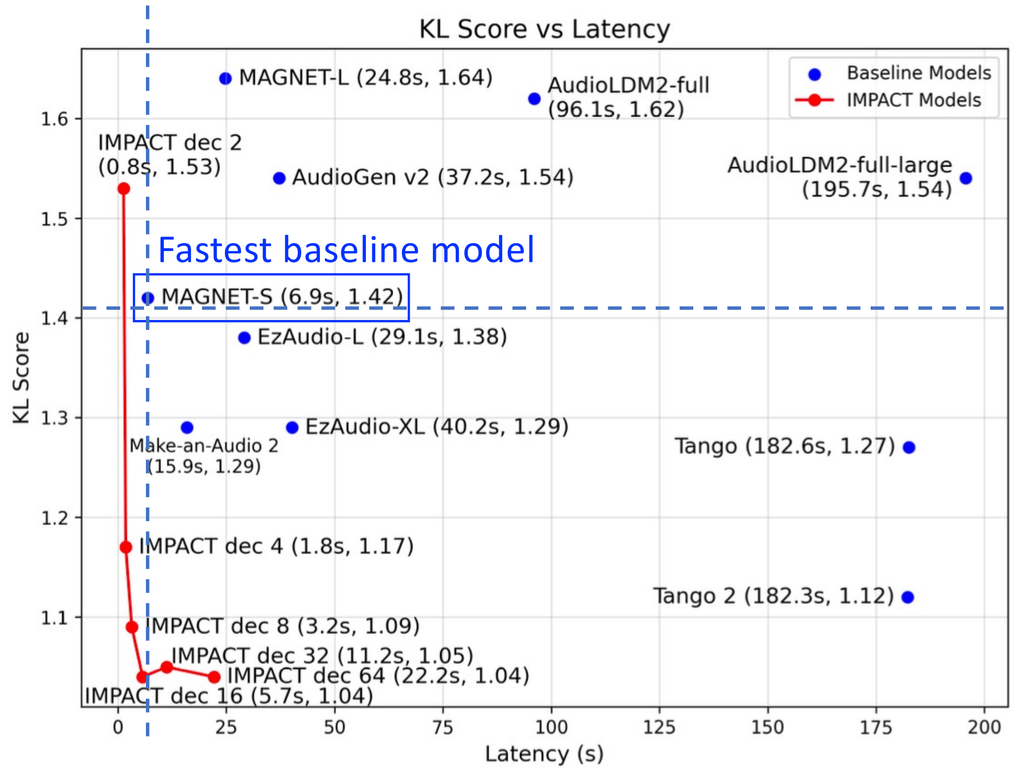




Latency analyses

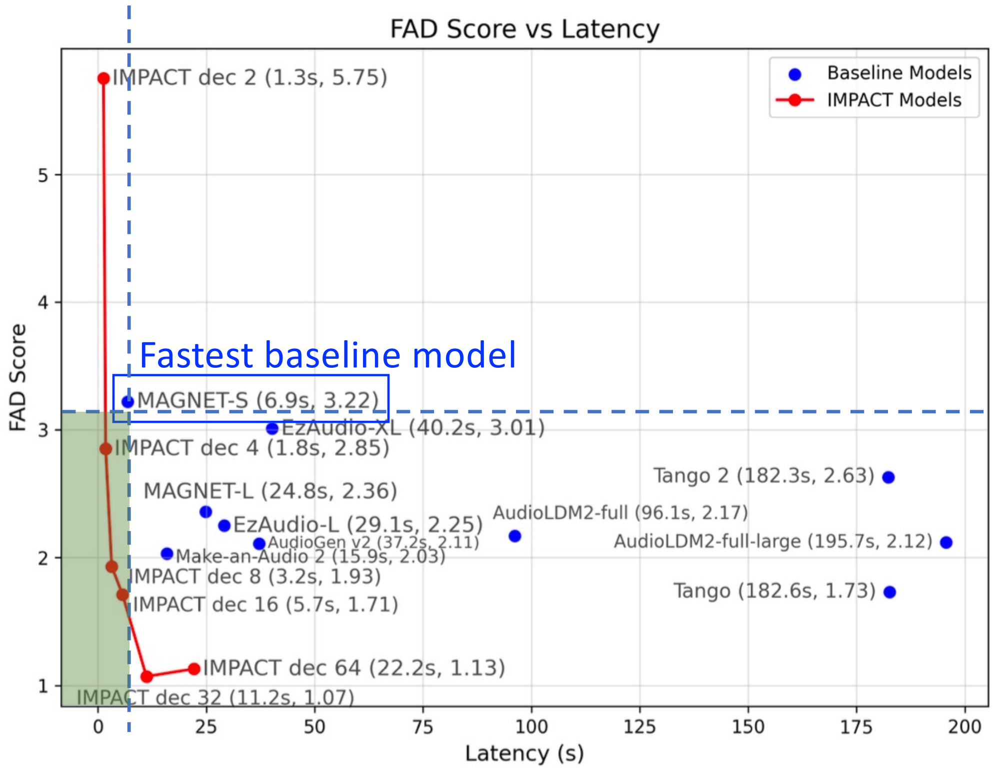
IMPACT using 16 decoding iterations (5.7s) is faster than MAGNET-S (6.9s), while having better FAD, KL, IS, and CLAP score.

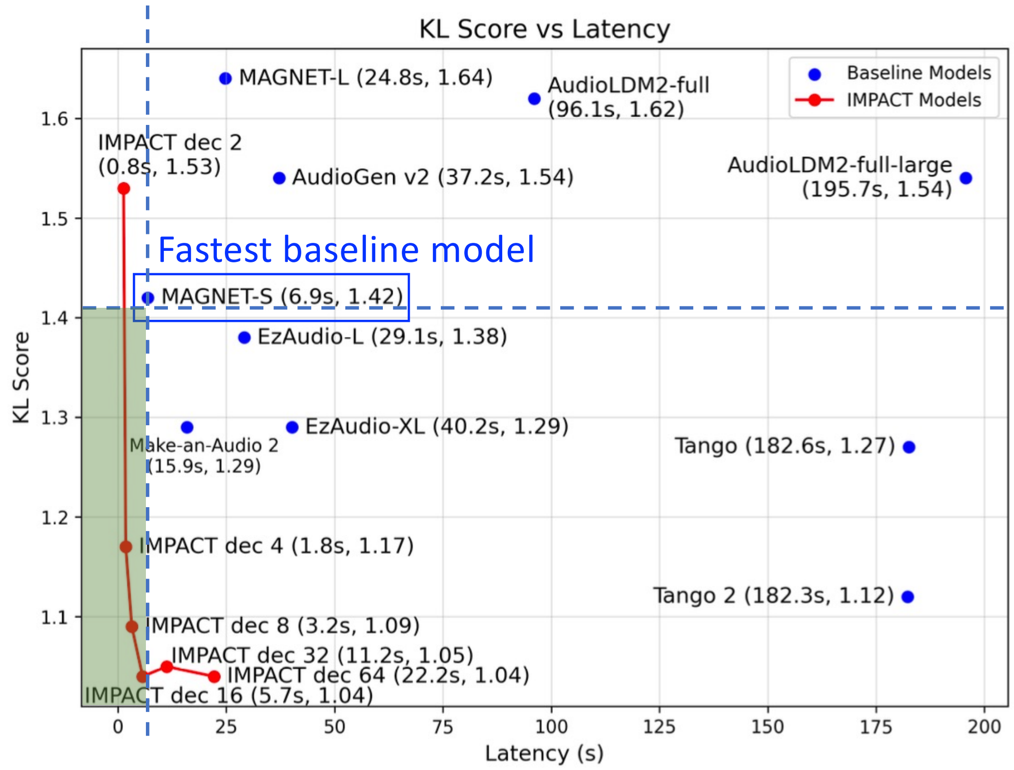




Latency analyses

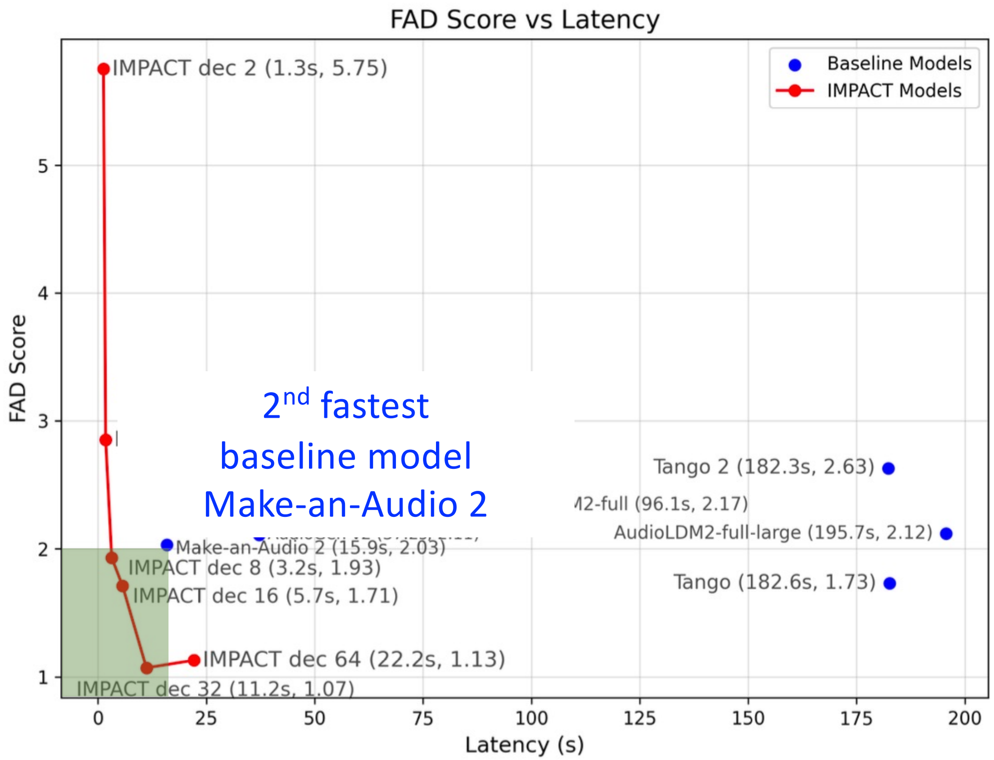
IMPACT using 16 decoding iterations (5.7s) is faster than MAGNET-S (6.9s), while having better FAD, KL, IS, and CLAP score.

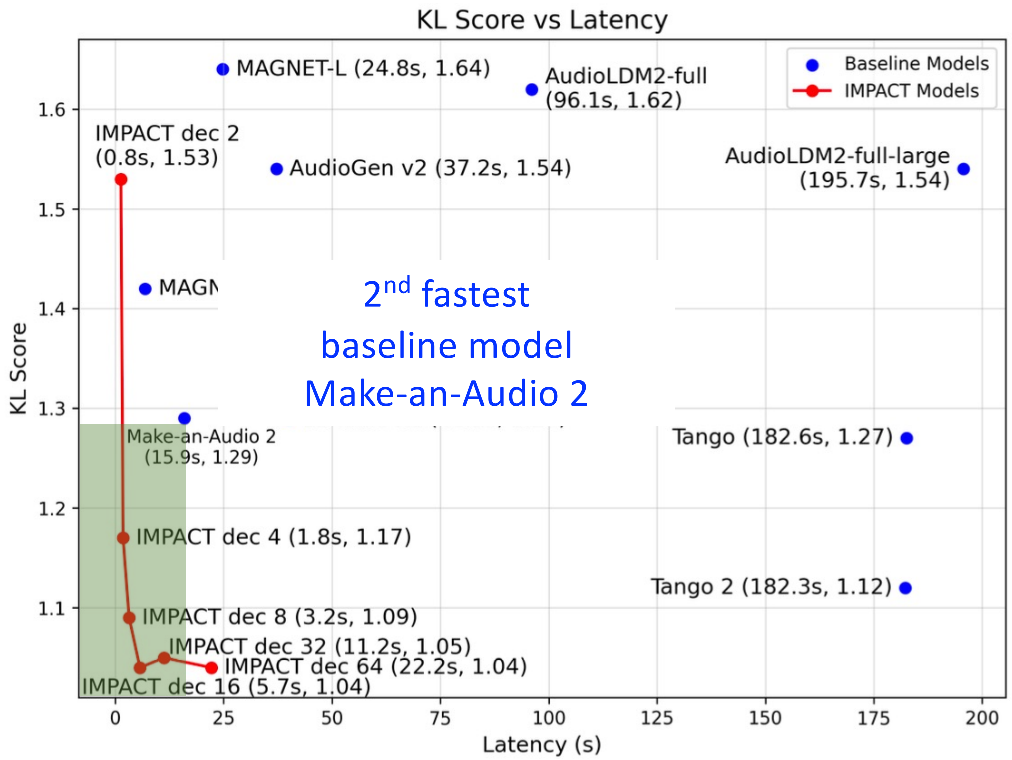




Latency analyses

IMPACT using 16 decoding iterations (5.7s) is faster than MAGNET-S (6.9s), while having better FAD, KL, IS, and CLAP score.





Conclusions

- State-of-the-art performance on objective metrics FD and FAD.
- State-of-the-art performance on subjective metrics for overall audio quality and text-relevancy.

- Faster than current fastest Text-to-audio model, MAGNET.

amazon