ICLR Singapore, April 24-28, 2025

MDSGen: Fast and Efficient Masked Diffusion Temporal-Aware Transformers for Open-Domain Sound Generation

Trung X. Pham*, Tri Ton*, Chang D. Yoo

School of Electrical Engineering

KAIST

U-AIM Lab

https://openreview.net/forum?id=yFEqYwgttJ

KAIST

Problems

- **Problems**

❑ Existing approaches used U-Net architecture for denoising diffusion with huge parameter, prone to overfitting and less scalable.

- **→ MDSGen uses Transformers with efficient use of parameters, demonstrating scalability.**

❑ Naïve MDT approach face limitation on VGAG when it uses spatial mask for image NOT for Audio.

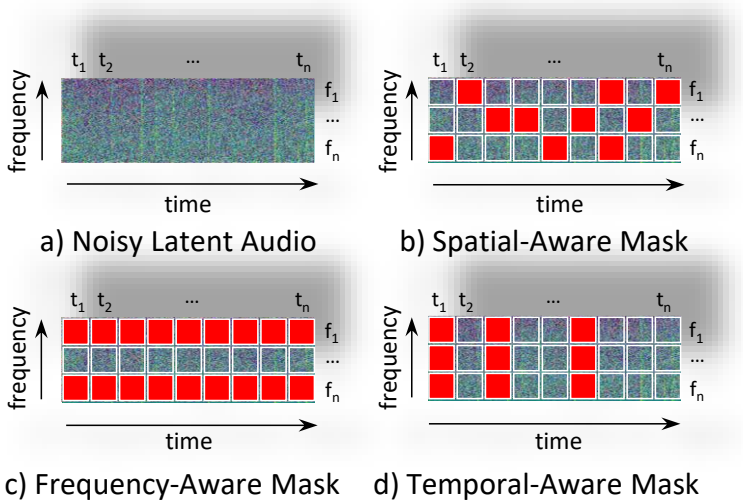
- **→ MDSGen uses temporal-masking for audio guiding better for Transformers.**

❑ Unet-based designs use redundant video features, less efficient, get overfitted**.**

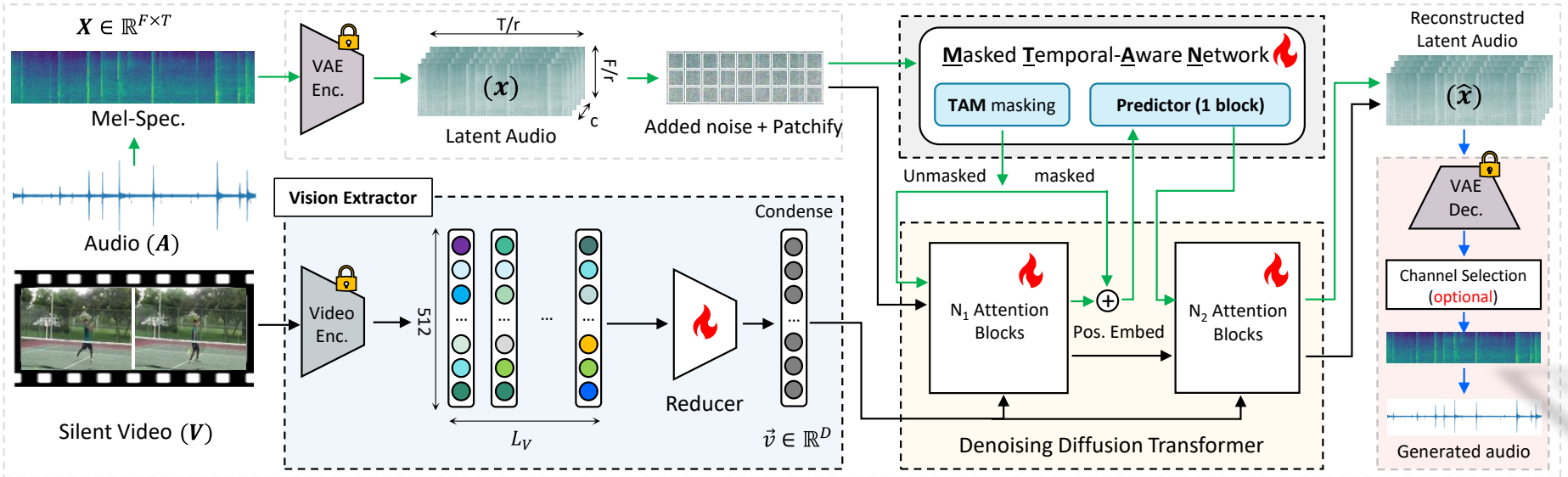
- **→ MDSGen reduces redundant video features, gaining compact supervision, avoid overfitting.**

Our Solutions

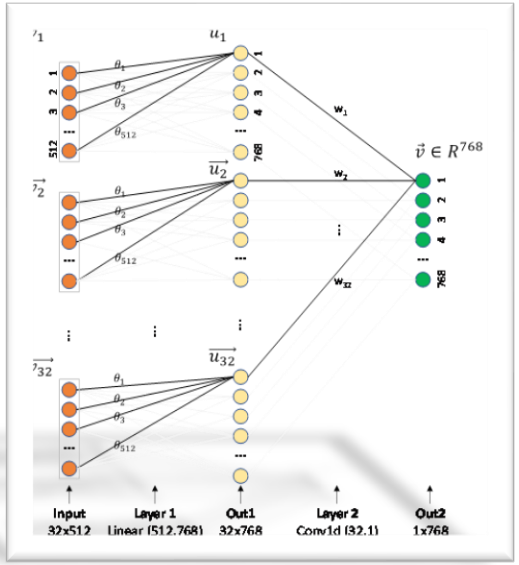
- Utilizing **Masked Diffusion Transformers** with efficient use of parameters for Audio Generation.
- Temporal-Awareness Masking (**TAM**)
- Removing redundant video features with an innovative **Reducer**



Ours masking **TAM** vs others



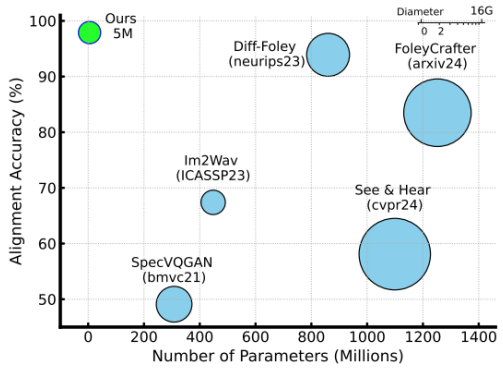
The proposed MDSGen Architecture



Reducer

Evaluations

Video-audio Alignment accuracy



Compact vector
reduced **overfitting**

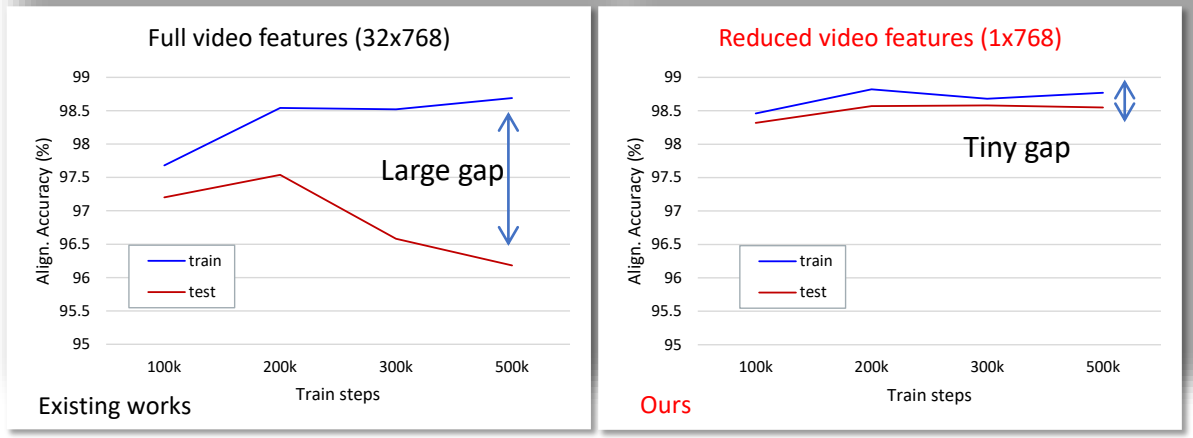


Table 7: **Masking Strategy for Audio Generation.** Comparison of different ways to train diffusion transformer-based models. Masking on temporal audio gives the best performance.

Masking Method	Mask	Temporal	FID↓	IS↑	KL↓	Align. Acc.↑ (%)
DiT (Peebles & Xie, 2023)	×	×	14.55	46.11	6.51	97.12
Random, SAM (Gao et al., 2023)	✓	×	12.44	48.66	6.30	98.15
Frequency, FAM	✓	×	12.79	46.33	6.41	97.58
Temporal, TAM (Ours)	✓	✓	**11.19**	**52.77**	**6.27**	**98.55**

Table 1: **Benchmark on VGGSound test.** Generation quality comparison of different approaches. † gets from (Luo et al., 2023), ‡ denotes without pre-trained SDv1.4. * denotes results with pre-trained SDv1.4, we reproduce it using the public checkpoint. "()" in FAD denotes AudioVAE (ImageVAE).

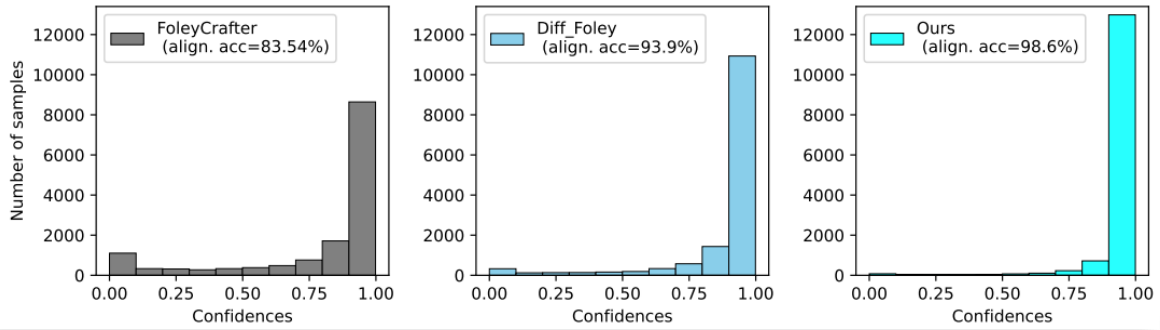
Method	FAD↓	FID↓	IS↑	KL↓	Align. Acc.↑	Time↓ (s)	#Params↓	Cost↓
SpecVQGAN (Iashin & Rahtu, 2021)†	-	**9.70**	30.80	7.03	49.19	5.47	308M	61×
Im2Wav (Sheffer & Adi, 2023) †	-	11.44	39.30	**5.20**	67.40	6.41	448M	90×
Diff-Foley (Luo et al., 2023) †‡	-	16.98	24.91	6.05	92.61	0.38	860M	172×
Diff-Foley (Luo et al., 2023) *	4.71	10.55	56.67	6.49	93.92	0.36	860M	172×
See and Hear (Xing et al., 2024)	5.55	21.35	19.23	6.94	58.14	18.25	1099M	220×
FoleyCrafter (Zhang et al., 2024)	2.45	12.07	42.06	5.67	83.54	2.96	1252M	250×
MDSGen-T (Ours) 500k	**2.21 (3.69)**	14.18	37.51	6.25	**97.91**	**0.01**	**5M**	1.0×
MDSGen-S (Ours) 500k	**1.66 (2.75)**	12.92	44.38	6.29	**98.32**	**0.02**	**33M**	6.6×
MDSGen-B (Ours) 500k	**1.34 (2.16)**	11.19	52.77	6.27	**98.55**	**0.05**	**131M**	26.2×
MDSGen-B (Ours) 800k	-	12.29	**57.12**	6.43	91.62	**0.05**	**131M**	26.2×

Benchmarking VGGSound

#Params: 1105 M Existing method 1 Infer time: 12.52s

#Params: 1250M Existing method 2 Infer time: 2.96s

#Params: 131M **Ours** Infer time: 0.05s



Our methods produces much higher samples with high confidence scores

Takeaways

- U-Net-based video-to-audio frameworks with **Stable Diffusion** are **too heavy and inefficient**.
- **Masked Diffusion Transformers** provide a scalable, efficient alternative.
- The **Innovative Reducer** compresses video into a **Compact Vector**, improving efficiency and reducing overfitting.
- **Temporal-Aware Masking** enhances audio generation quality.
- **MDSGen** achieves state-of-the-art **FAD (1.34)** and **~99% alignment accuracy** with minimal parameters.

More information: https://trungpx.github.io/