COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

# Application of Masked Token Transformer (MaskGIT) in high-quality and efficient audio generation

**Pengyu Zhou, Yiru Gong, Baode Gao, Columbia University**

## Introduction

The limitation in lack of sample diversity, introduced by training instability and mode collapse of the GAN-based models, is always a problem in research. Plus, the best generative transformer model so far, however, still treat an image naively as a sequence of tokens, and decode an image sequentially following the raster scan ordering. Addressing these issues in audio generation and image synthesis still remains open research problems.

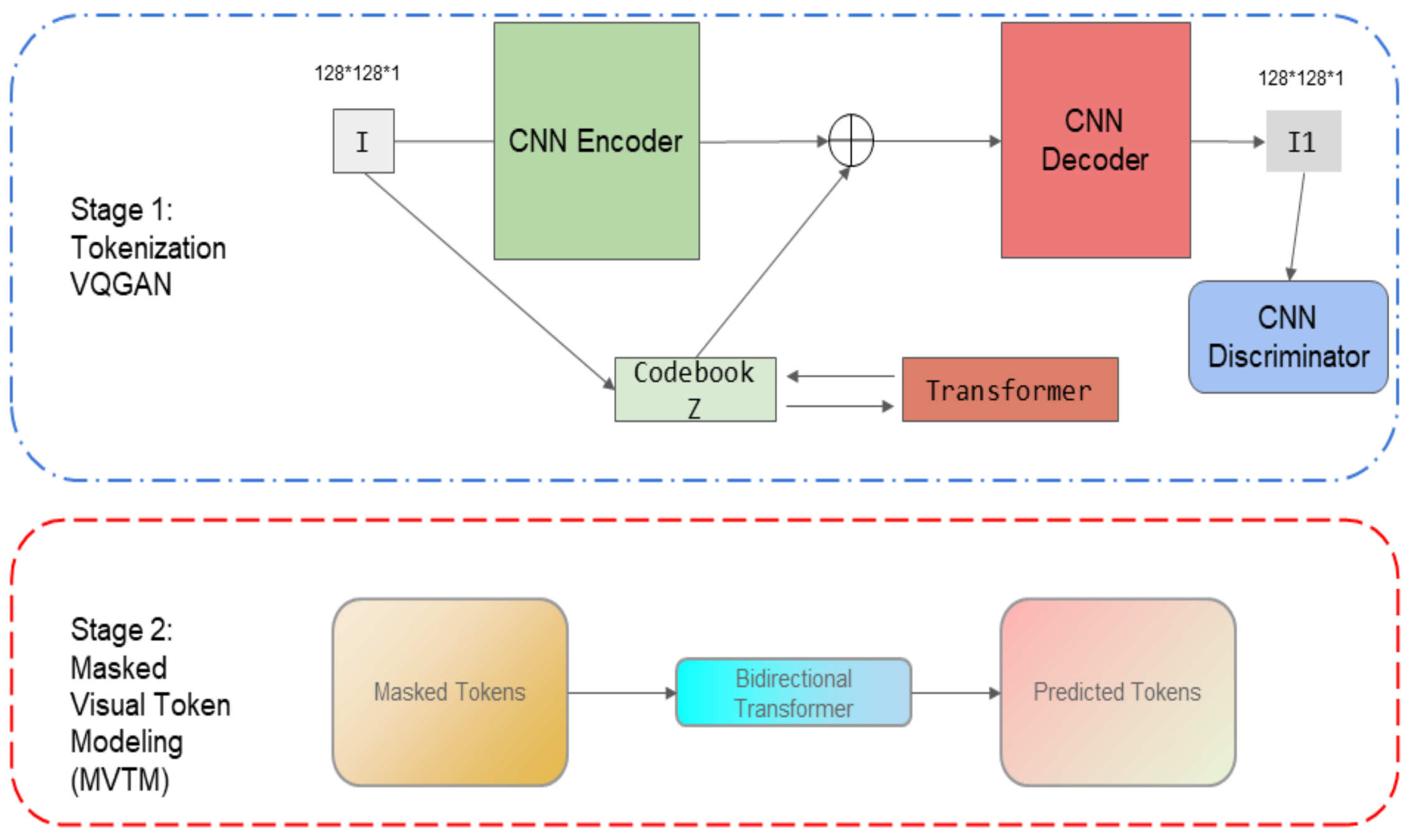
Key contributions:

- Reproducing the MaskGIT code by pre-trained VQGAN and MVTM decoder;
- Combining SpecVQGAN and MVTM to evaluate quality of spectrogram generation for audio.

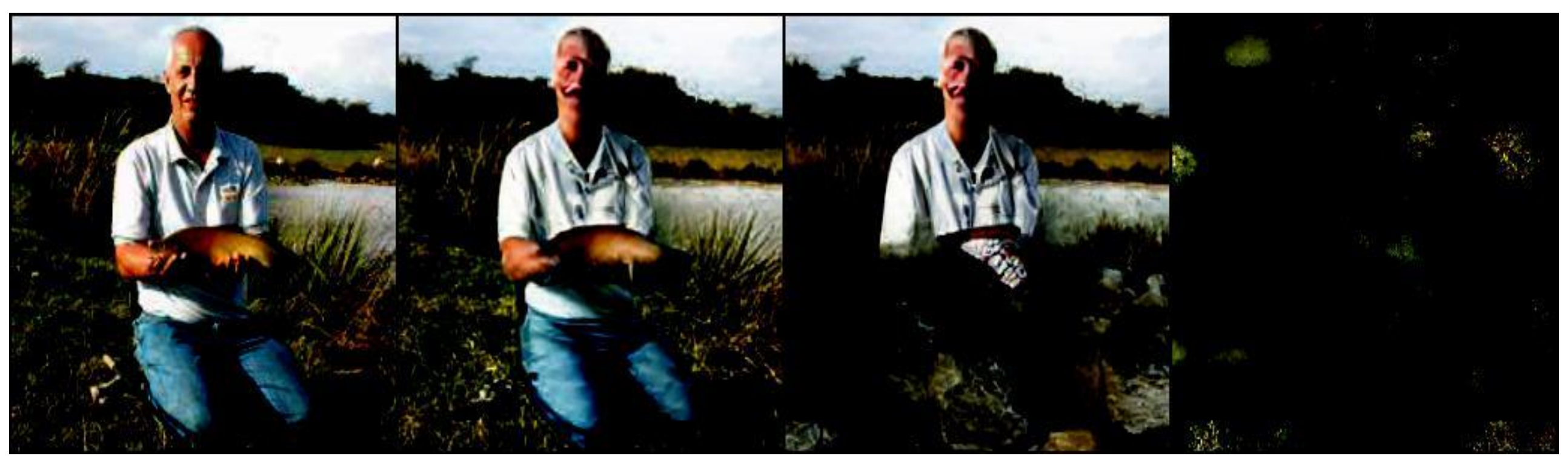
## Methods

The process of generating audio or image can be split into two stages. In the first stage, which is known as tokenization, it tries to compress into discrete latent space. In the second stage, instead of conditioning only on previous tokens in the order of raster scan, bidirectional self-attention allows the model to generate new tokens from generated tokens in all directions, called Masked Visual Token Modeling (MVTM). This MVTM decoding can be implemented in the following steps: 1) predicting the tokens for all the masked location in parallel; 2) Sampling tokens based on the probability of tokens; 3) masking tokens by mask scheduling function; 4) repeating t times.

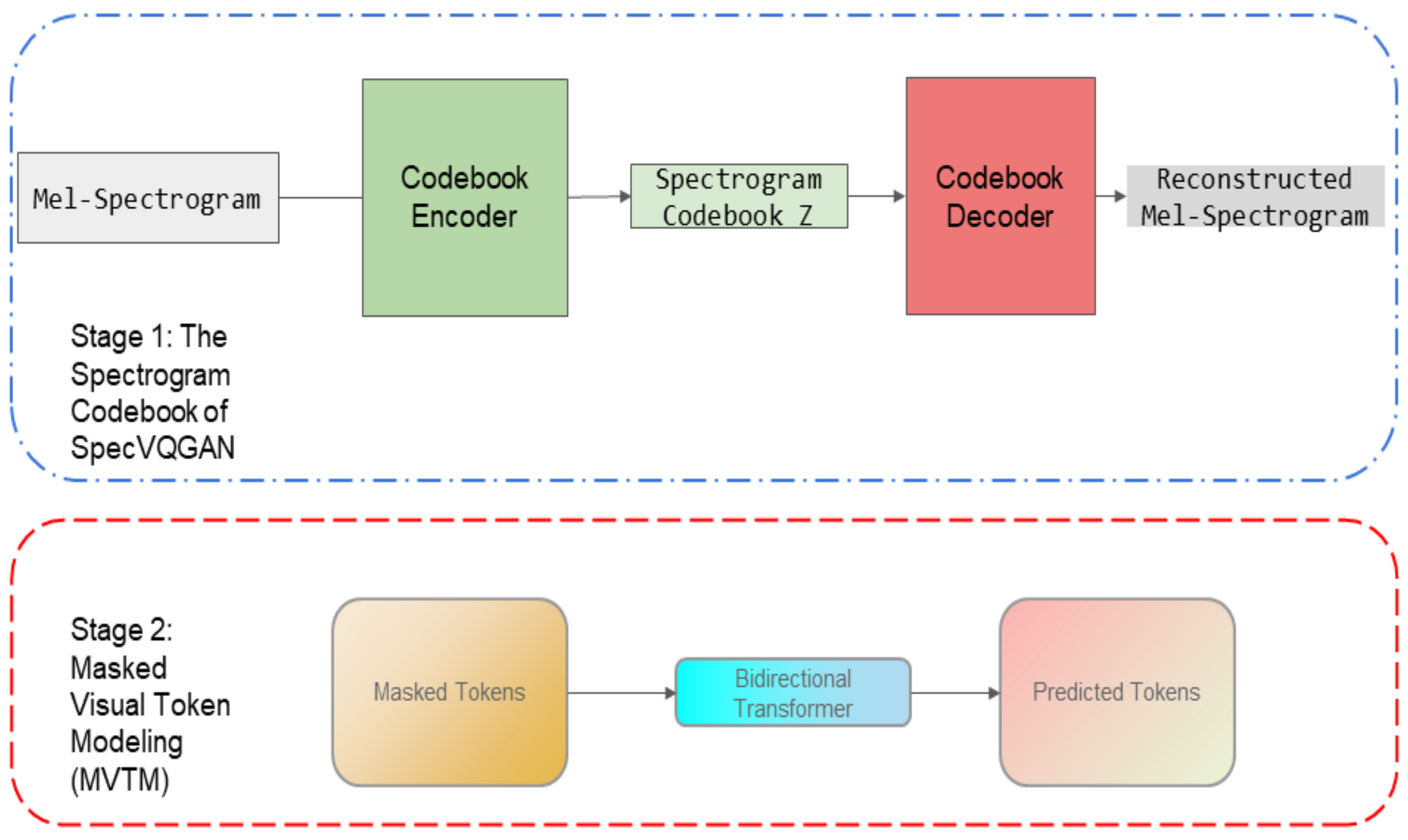
## Results



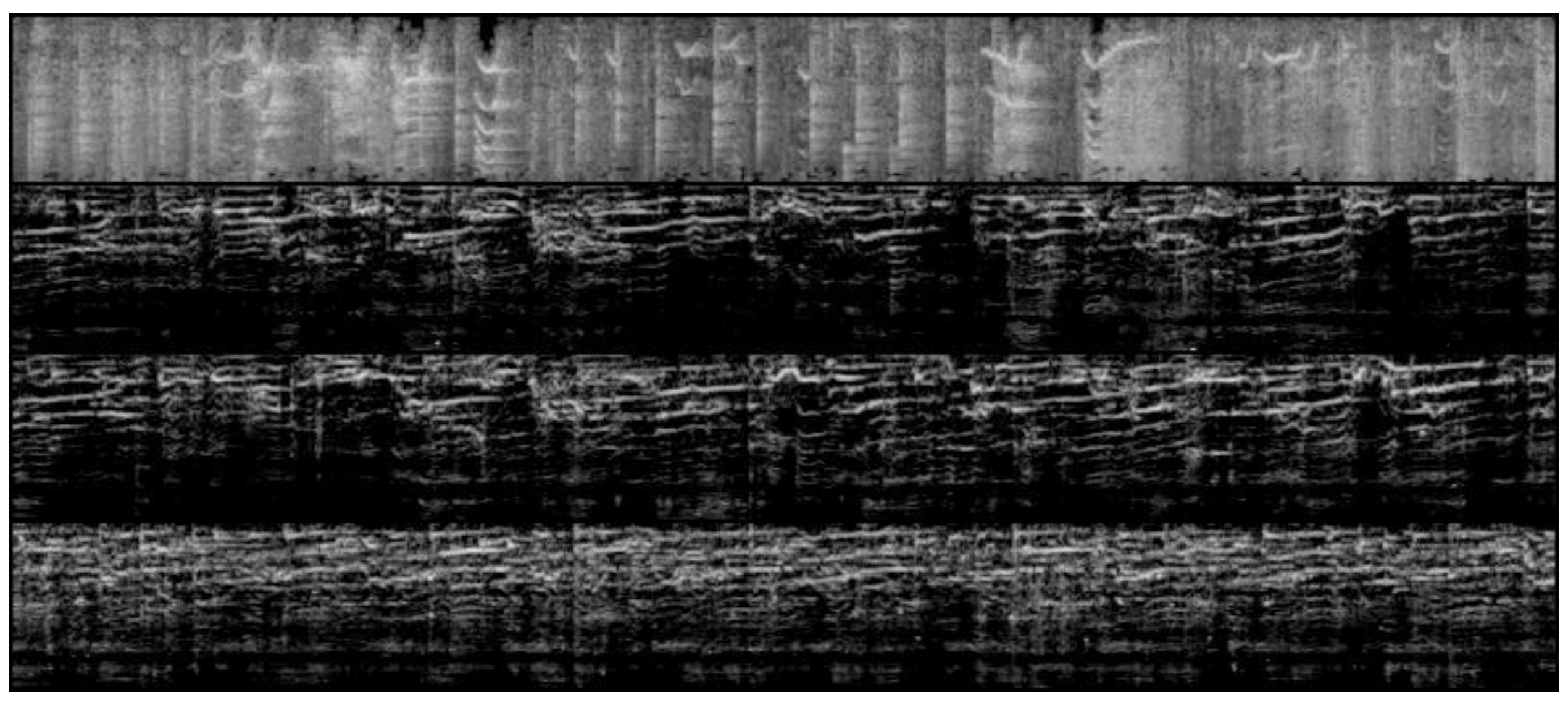
MaskGIT Pipeline Overview



Synthesized Image
(input, reconstruction, half-masked, new)



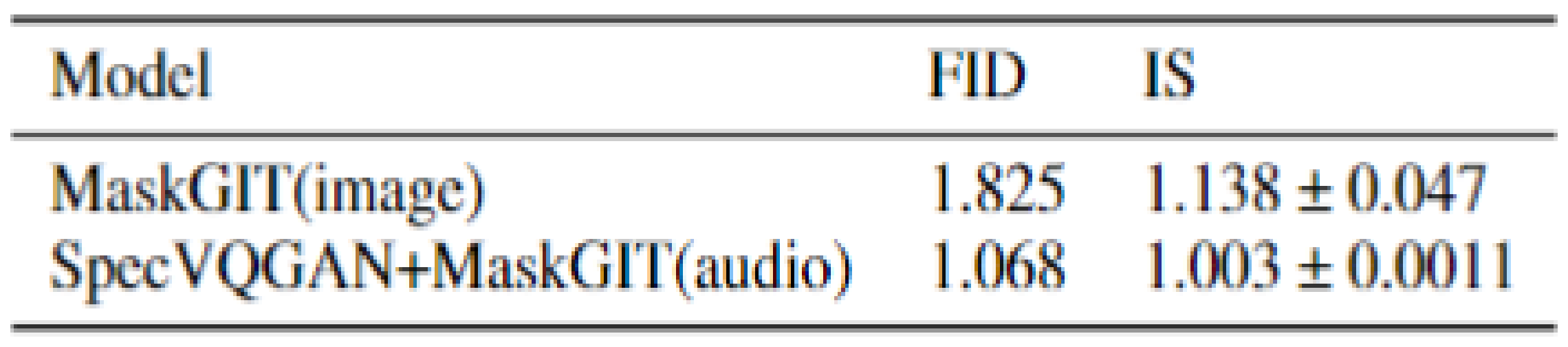
SpecVQGAN+MVTM



Synthesized spectrogram with log transformation
(input, reconstruction, half-masked, new)

Training the model on the entire ImageNet data requires 170 hours per epoch, and 50 epochs in total are set for model training. Due to the time and GPU resource limitation, we decided to train on a subset of ImageNet (3900 pictures) and baby VAS audio (2061 audio). We show our results of FID and IS with 500 test size. The future implementation thus lies in completing the training process on the entire ImageNet and VAS data. Meanwhile, the MaskGIT model is also flexible for adopting video, image, and word keys to facilitate audio generation. In general, we showed a new application of MVTM masking transformers in the audio generation field.

Table 1: Evaluation Score



| Model | FID | IS |
|---|---|---|
| MaskGIT(image) | 1.825 | 1.138 ± 0.047 |
| SpecVQGAN+MaskGIT(audio) | 1.068 | 1.003 ± 0.0011 |

## Reference

1. Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. arXiv preprint arXiv:2202.04200, 2022
2. Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. arXiv preprint arXiv:2110.08791, 2021.
3. MaskGIT basic code: https://github.com/dome272/MaskGIT-pytorch