

Stage 1: Semantic Modeling (VAE)

Latent Space
 $z \sim N(0, I)$

Sample

VAE Decoder

Generate

Attribute List
[folk, guitar, upbeat]

Conditioning

Stage 2: Text Generation (LLM)

Fine-Tuned LLM
(*Llama 3.1 8B*)

Translation

Generated Caption
"An upbeat folk song
featuring acoustic guitar..."

Prompting

Stage 3: Audio Synthesis

Music Generation Model
(*MusicGen Large*)

Synthesis

Synthetic Audio
(Copyright-Free)

