



(a) Train a text-to-token LLM using TTS data

All NLP tasks are generation tasks.

(b) Sample spans from the input text

All NLP tasks are generation tasks.

(c) Generate audio token for sampled spans

Interleaved Speech-Text

All NLP tasks are generation tasks.

Unsupervised Speech

Hey, how can I help you?

Speech-Text Pair (ASR, TTS)

Hello

Hello



(d) Interleaved Speech-Text Pre-training