ConsistencyTTA

Accelerating Diffusion-Based Text-to-Audio Generation with Consistency Distillation

Yatong Bai, Trung Dang, Dung Tran, Kazuhito Koishida, Somayeh Sojoudi

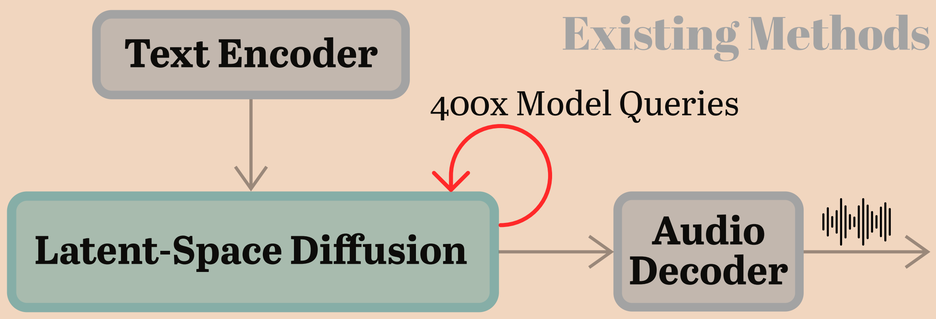
consistency-tta.github.io

ConsistencyTTA

Live Demo!

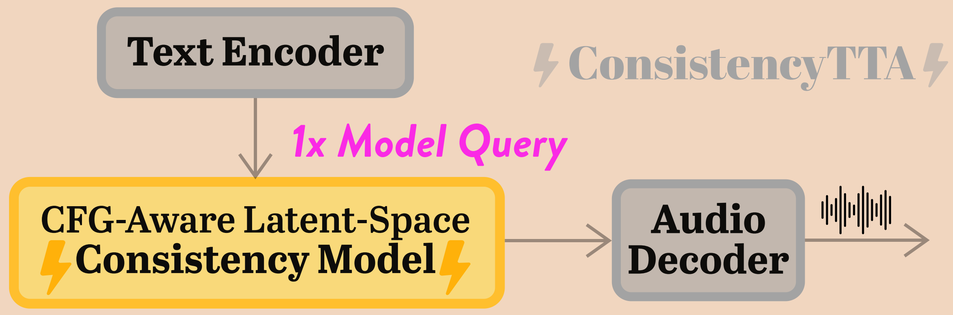
Background

- **Diffusion model** is one of the most popular Text-to-Audio (TTA) methods.
  - **Training:**
    - Add noise and train model to reverse the noise.
  - **Inference:**
    - Start from pure noise and gradually denoise.
  - 400 Model Queries = **SLOW INFERENCE!**



Goal

- **High-quality audio generation with ONE SINGLE MODEL QUERY.**

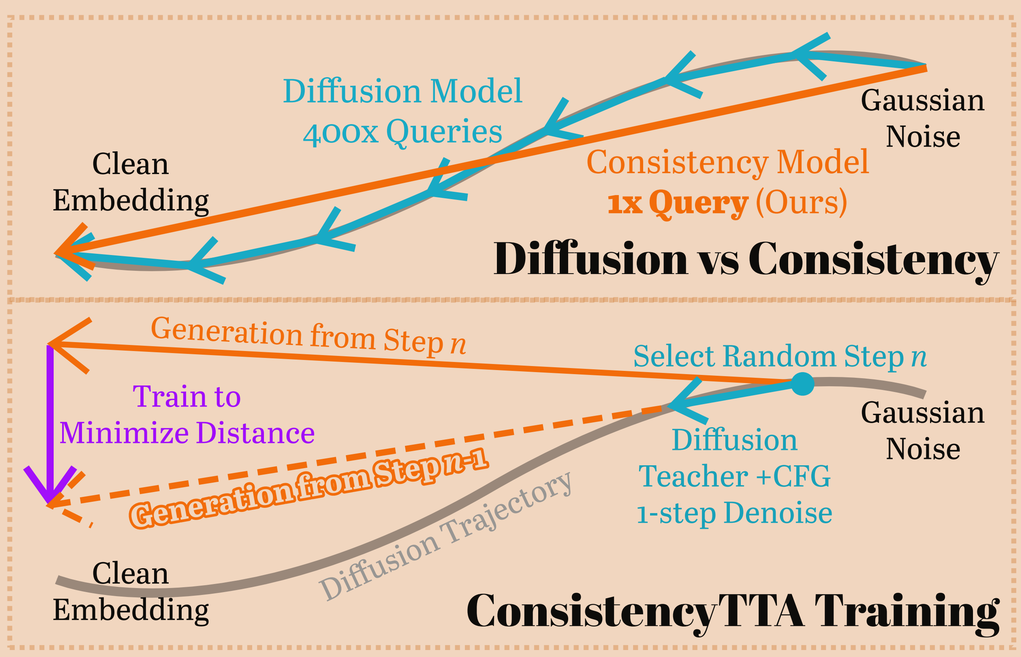


Data

- **In-the-wild audio generation.**
  - **AudioCaps** (YouTube video soundtracks + captions).
  - 45,260 training audio clips (10s); 882 validation clips.
- **Example prompts:**
  - A telephone ringing with loud echo.
  - A horn and then an engine revving.

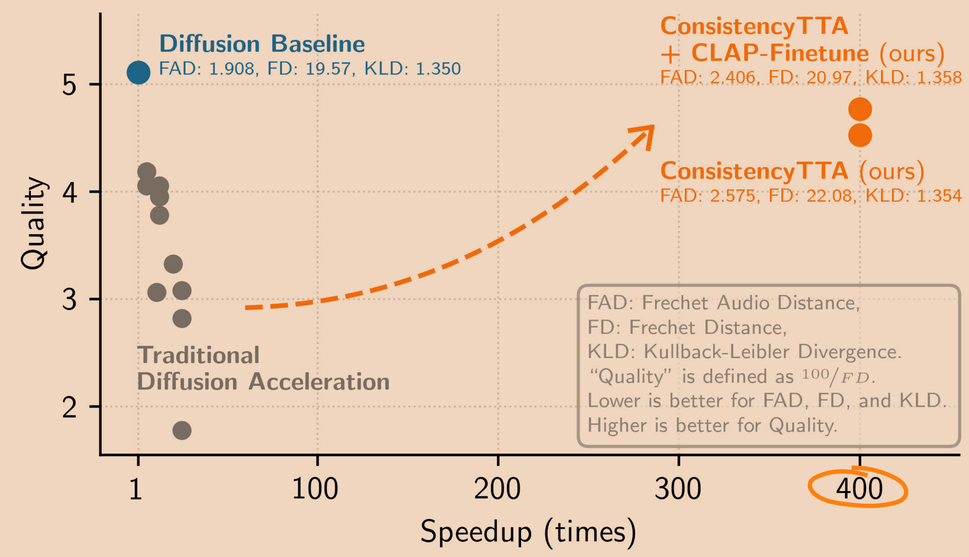
Methods

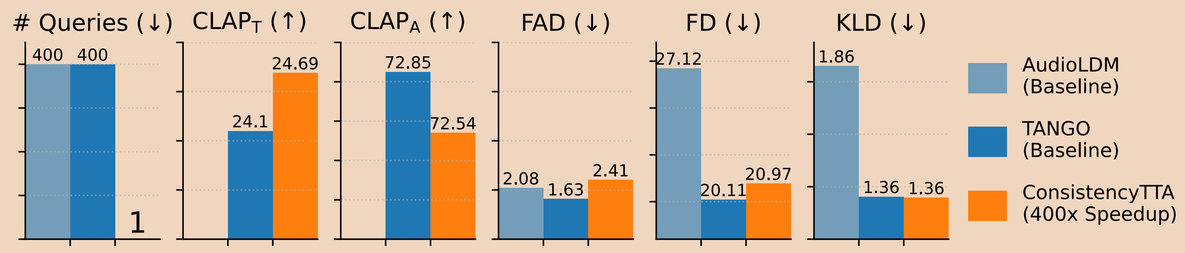
- **Consistency Model**
  - Distilled from a teacher diffusion model
  - Single-step high-quality generation from anywhere on the diffusion trajectory.
- **CFG-Aware Distillation**
  - Classifier-free guidance (CFG): an external operation that strengthens diffusion models.
  - ConsistencyTTA distills with CFG and absorbs it.



- **CLAP-Finetuning**
  - Single-step generation means differentiability.
  - Hence, directly optimize generation quality objectives, such as CLAP score.

Results





- **Uncompromised generation quality in a single step.**
  - **99.75% computation reduction.**
  - **98.63% wall time reduction.**
  - **Runs locally on a laptop and still faster** than a diffusion model on A100 GPU.