ConsistencyTTA

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

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

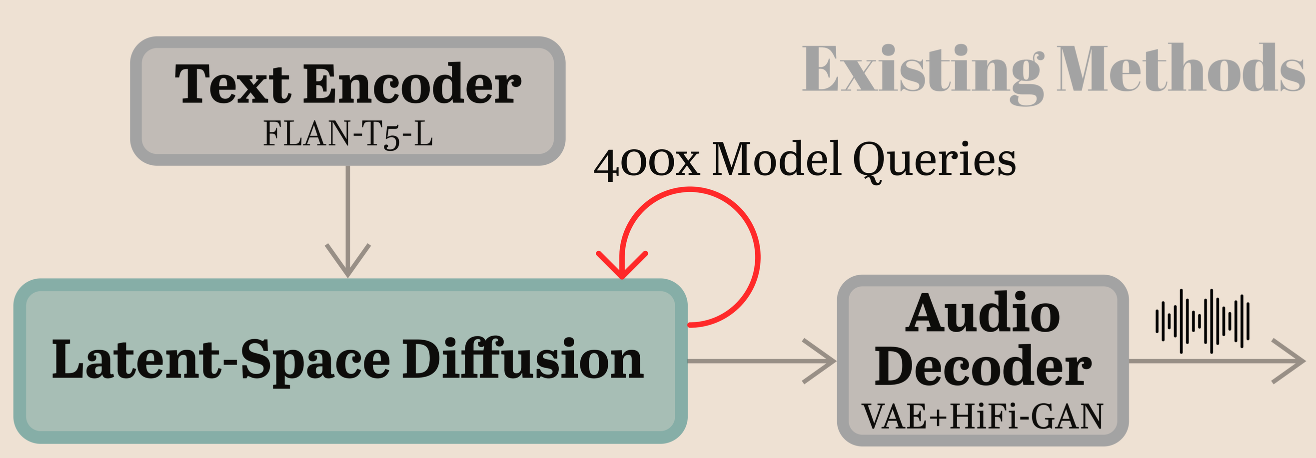
consistency-tta.github.io   yatong_bai@berkeley.edu

Project Website   Live Demo!

Berkeley UNIVERSITY OF CALIFORNIA

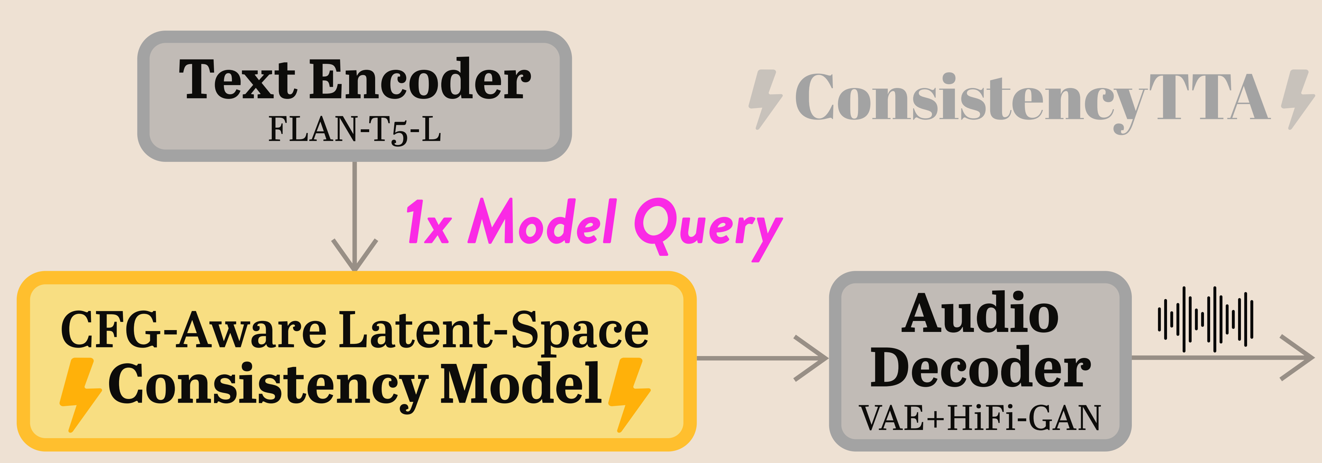
Background

- **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!**



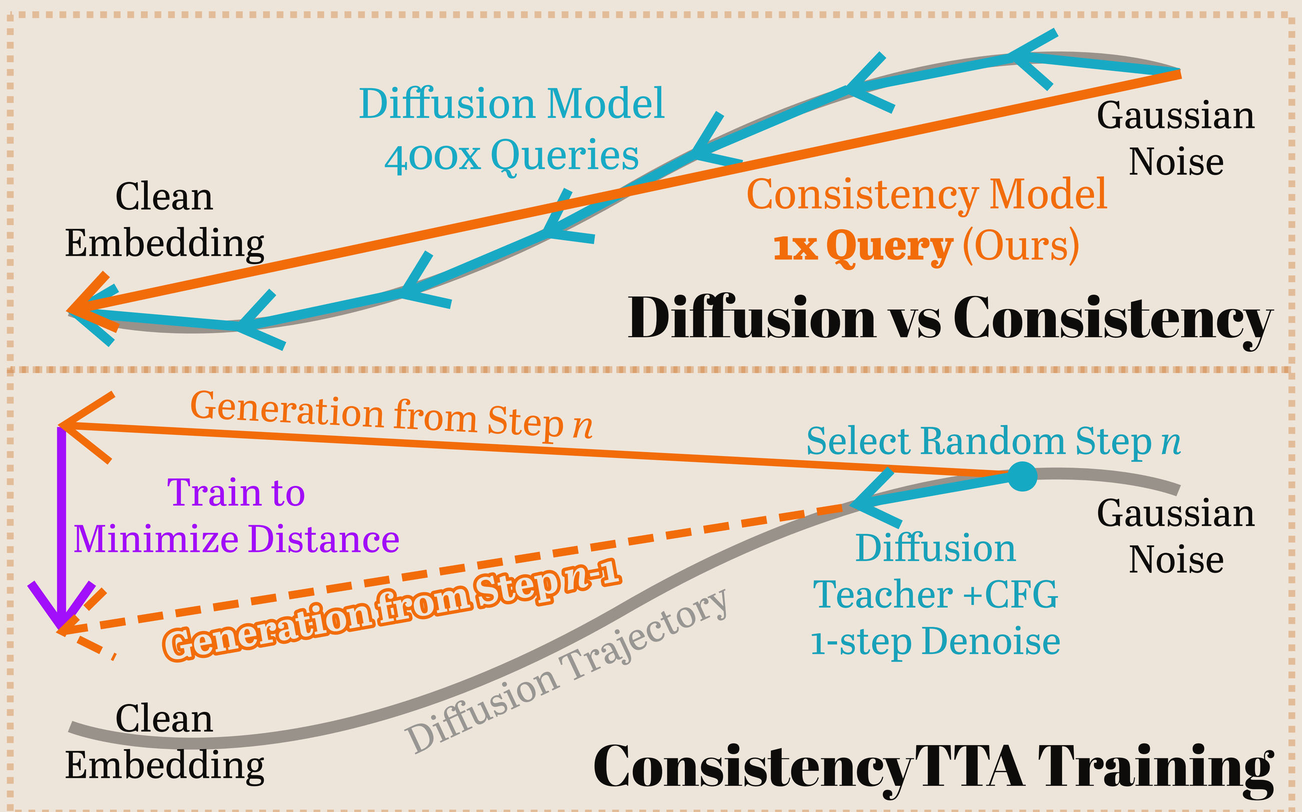
Goal

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



Methods

- **Consistency Model**
  - Distilled from a teacher diffusion model (we use TANGO).
  - 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 CFG into the model.**
    - Add an CFG weight embedding branch to the student neural net, similar to timestep embedding.
    - When querying the teacher during distillation, apply a random CFG weight sampled from [0, 6).
    - The same weight is fed into the added student embedding branch.



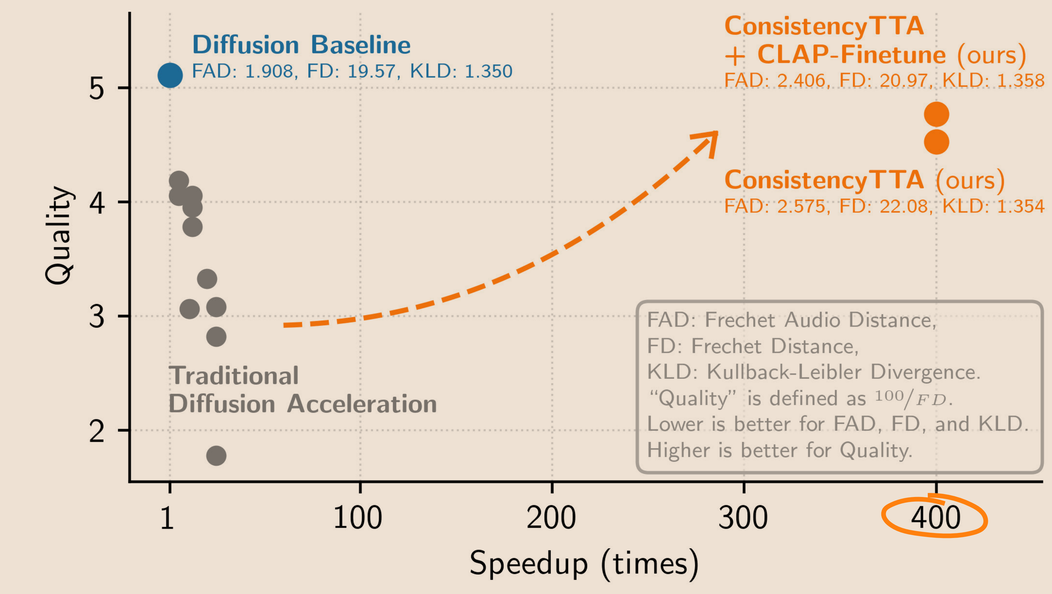
- **CLAP-Finetuning**
  - Single-step generation means differentiability.
  - Hence, directly optimize generation quality objectives, such as CLAP score.

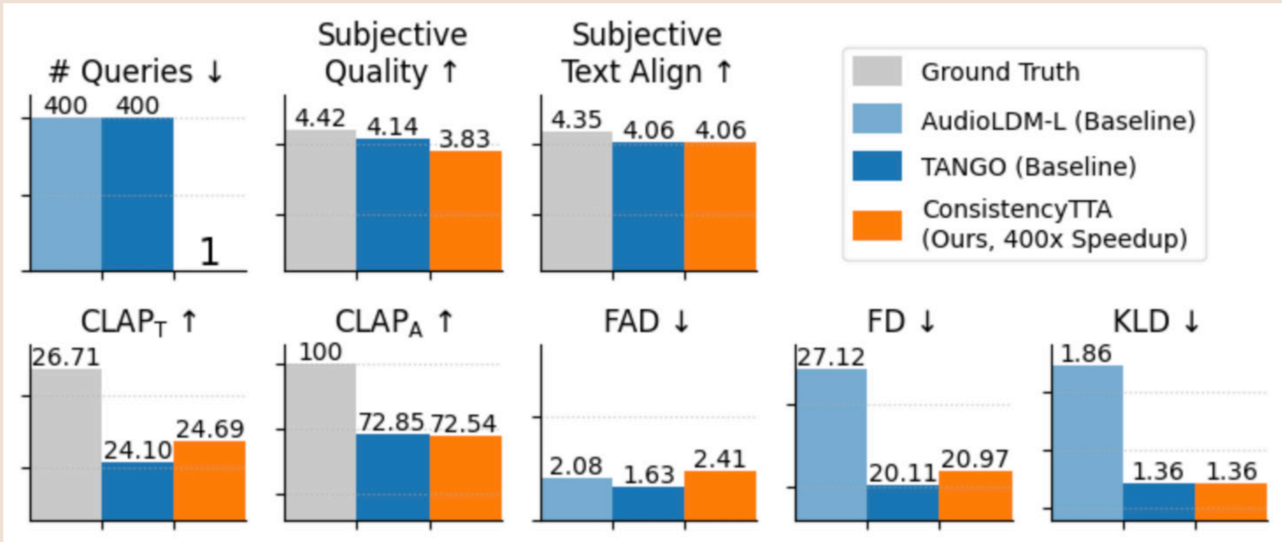
Data

- **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.

Results

- Evaluation Metrics:
  - Frechet Audio Distance (FAD) w/ VGG-ish embeddings
  - Frechet Distance (FD) w/ PANN embeddings
  - KL Divergence (KLD) w/ PANN embeddings
  - CLAP Score w.r.t. Prompt ($CLAP_T$)
  - CLAP Score w.r.t. Ground-Truth Audio ($CLAP_A$)
  - Human Subjective Quality & Prompt Alignment





- **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.

- **Ablation Studies** with short training runs:
  - Distilling CFG into the model outperforms external CFG.
  - Training with random CFG weight is better than fixed.
  - Using Heun solver to query teacher is better than DDIM.
  - Uniform noise schedule is preferred over Karras.

Guidance Method	Solver	Noise Schedule	CFG w	# Queries (↓)	FAD (↓)	FD (↓)	KLD (↓)
Unguided	DDIM	Uniform	1	1	13.48	45.75	2.409
Direct Guidance	DDIM Heun	Uniform Karras	3	2	8.565 7.421	38.67 39.36	2.015 1.976
Fixed Guidance Distillation	Heun	Karras Uniform Uniform	3	1	5.702 4.168 3.859	33.18 28.54 **27.79**	1.494 1.384 1.421
Variable Guidance Distillation	Heun	Uniform	4 6	1	3.180 **2.975**	27.92 28.63	1.394 **1.378**