ConsistencyTTA

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

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

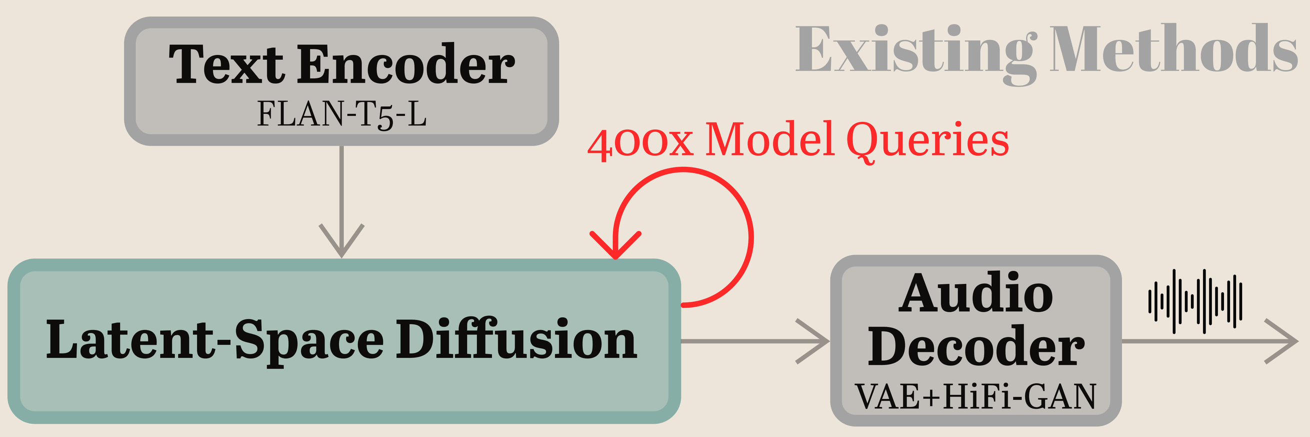
consistency-tta.github.io yatong_bai@berkeley.edu

Project Website 🤗 Live Demo!

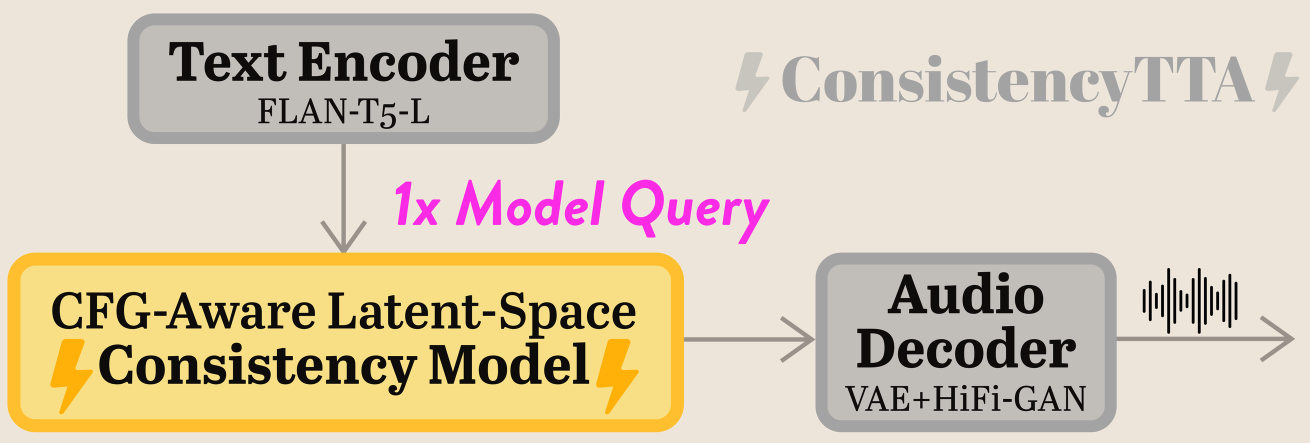
Berkeley UNIVERSITY OF CALIFORNIA

Problem Statement

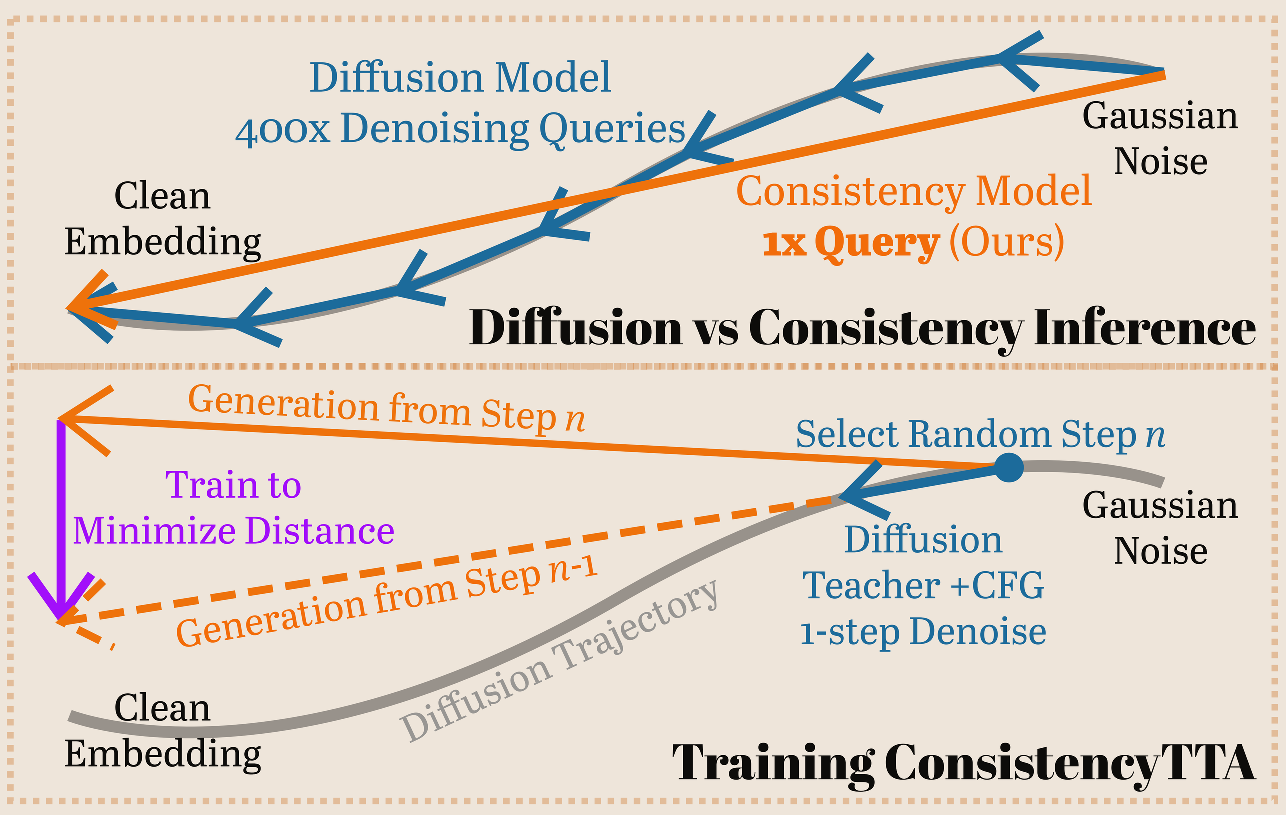
- **Diffusion model** is one of the most popular Text-to-Audio (TTA) methods.
  - **Training:** Add noise and train model to reverse the noise.
  - **Inference:** Start from pure noise and gradually denoise.
  - 400 Model Queries = **SLOW INFERENCE!**



Our Approach



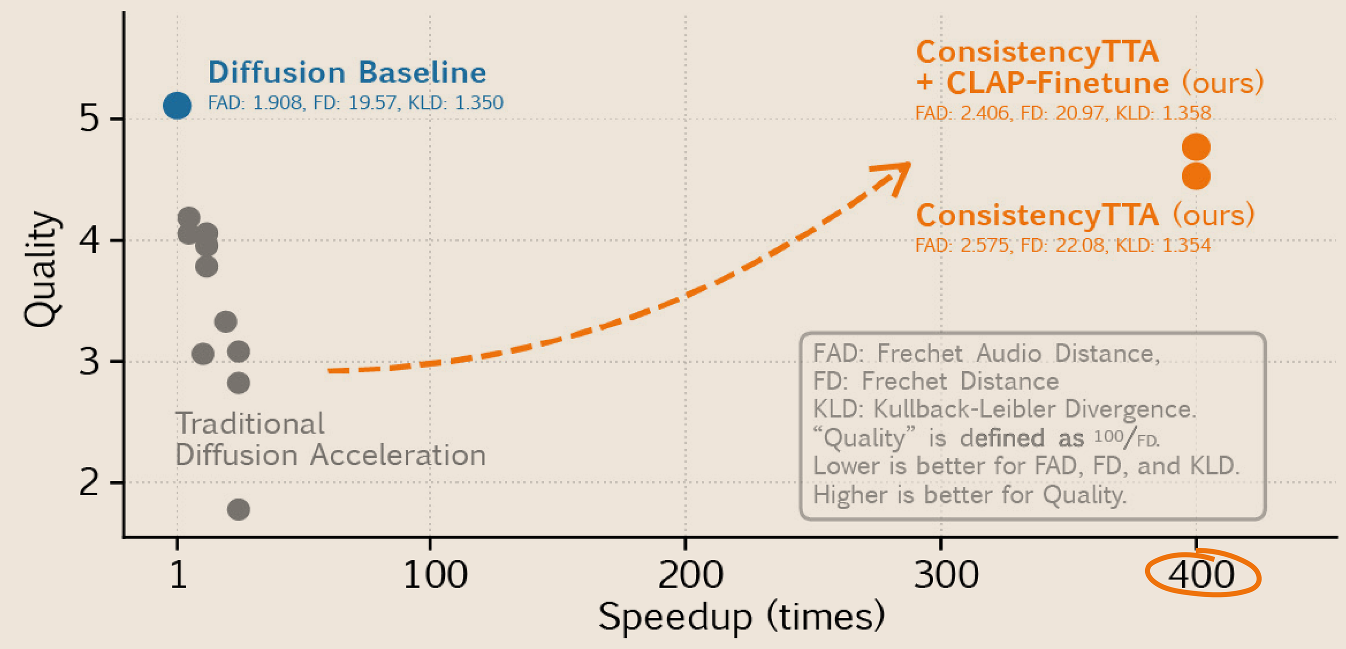
- **Consistency Model**
  - Distilled from a teacher diffusion model (we use TANGO).
  - One-step high-quality generation from anywhere on diffusion trajectory.
- **CFG-Aware Distillation**
  - Classifier-free guidance (CFG):
    - An external operation that strengthens diffusion models.
  - **ConsistencyTTA distills CFG into the model.**
    - Add a CFG weight embedding branch to student neural net, similar to timestep embedding.
    - When querying teacher during distillation, apply a random CFG weight in [0, 6).
    - The same CFG weight is fed into the added student embedding branch.



- **CLAP-Finetuning**
  - Single-step generation means differentiability.
  - Hence, directly optimize generation quality objectives.
    - Cannot be performed directly on diffusion models, thus an advantage of ConsistencyTTA.
  - Finetune ConsistencyTTA to maximize CLAP score.
    - We consider CLAP score w.r.t. ground-truth audio and the CLAP score w.r.t. text prompt.

Results

- **High-quality audio generation with ONE SINGLE MODEL QUERY.**
  - **99.75%** less computation.
  - **98.63%** shorter wall time.
  - **Runs locally on a laptop and still faster** than diffusion model on A100 GPU.



- **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

	Model Queries ↓	Generation Time ↓	Subjective Quality ↑	Subjective Text Align ↑	$CLAP_T$ ↑	$CLAP_A$ ↑	FAD ↓	FD ↓	KLD ↓
AudioLDM-L (Baseline)	400	-	-	-	-	-	**2.08**	27.12	1.86
TANGO (Baseline)	400	168	**4.136**	**4.064**	24.10	**72.85**	**1.631**	**20.11**	1.362
ConsistencyTTA + CLAP-FT	**1**	**2.3**	3.830	**4.064**	**24.69**	72.54	2.406	20.97	1.358
ConsistencyTTA	**1**	**2.3**	3.902	4.010	22.50	72.30	2.575	22.08	**1.354**
Ground Truth	-	-	-	-	26.71	100	-	-	-

Table 1: Main Experiment Results.

- **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**

Table 2. Ablation Studies

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.