NEURAL INFORMATION PROCESSING SYSTEMS

Large-Scale Training Data Attribution for Music Generative Models via Unlearning

Sony AI SONY

*Woosung Choi[1], *Junghyun Koo[1], *Kin Wai Cheuk[1], Joan Serrà[1], Marco A Martínez-Ramírez[1], Yukara Ikemiya[1], Naoki Murata[1], Yuhta Takida[1], Wei-Hsiang Liao[1], Yuki Mitsufuji[1,2] , 1 Sony AI 2 Sony Group Corporation

Background & Motivation

Existing efforts to address attribution in music generation, often using white-box or black-box methods:

- **White-Box Methods:** initial work used the Influence Function on Music Transformer, trained on MAESTRO dataset (~ 200 hours of piano music).
- **Black-Box Methods:** they rely on external encoders, but the resulting embeddings may not reflect the model's actual internal perspective.
- **This work** pioneers machine unlearning, directly emulating the counterfactual state to offer large-scale TDA on a text-to-music DiT.

Methodology

- We approximate the attribution score $\tau(\hat{z}, z_i)$ via unlearning and the Fisher information matrix (FIM).

$$\tau(\hat{z}, z_i) = \boldsymbol{\mathcal{L}}(z_i, \Theta_{\setminus\hat{z}}) - \boldsymbol{\mathcal{L}}(z_i, \Theta_0)$$

Unlearning Algorithm

- We apply the *mirrored influence hypothesis*, which enables attribution by unlearning the generated sample instead of each training sample, assuming that related training examples will also be unlearned.

$$\boldsymbol{\mathcal{L}}^{\hat{z}}_{unlearn}(\Theta) = \underbrace{-\boldsymbol{\mathcal{L}}(\hat{z},\Theta)}_{\text{Maximizing objective (for unlearning)}} + \underbrace{\frac{N}{2}(\Theta - \Theta_0)^{\top}\mathbf{F}(\Theta - \Theta_0)}_{\text{Regularization term}}$$

- The objective balances forgetting the generated sample with retaining overall model performance, using a regularization term based on Fisher Information Matrix (FIM) to prevent catastrophic forgetting.

Masking Silence

- For unlearning in Diffusion Transformer (DiT)-based music models, we optionally apply a mask M_U to ignore silent regions during unlearning, although the generative model was trained without masking.
- During attribution, a separate mask M_L can be applied with computing loss to exclude padded regions. We omit M_L to stay consistent with the original training setup.

Experiment Setup

- **Model**: Latent DiT text-to-music model using Stable Audio Framework.
- **Dataset**: 115K (4,356 hours) in-house music tracks across diverse genres.
  - *Train-to-Train*: unlearns training samples for validation (40 samples)
  - *Test-to-Train*: generates 16 tracks to assess attribution to training data.

Results

Self-Influence Experiment and Tuning

Table 1: Grid search results for optimal unlearning hyperparameters. FDopenl3 is 110.5 for the original checkpoint.

Target Layer	$\mathrm{M_U}$	$\mathrm{M_L}$	$\mathbf{R(z_{tar})}$	$\mathbf{CLAP_{topk}}$	$\mathbf{CLAP_{botk}}$	$\mathbf{FD_{openl3}}$
Cross-Attention's *to_kv* weights	✓		103.2	0.38	0.35	110.5
Cross-Attention Layers	✓		1.4	0.60	0.32	110.4
Self-Attention Layers	✓		1.1	0.63	0.30	110.5
All the Transformer Layers	✓	✓	**1.0**	0.80	0.38	110.5
All the Transformer Layers			6615.7	**0.82**	0.42	110.5
All the Transformer Layers	✓		**1.0**	0.66	**0.26**	110.5

Comparison with Non-counterfactual Methods

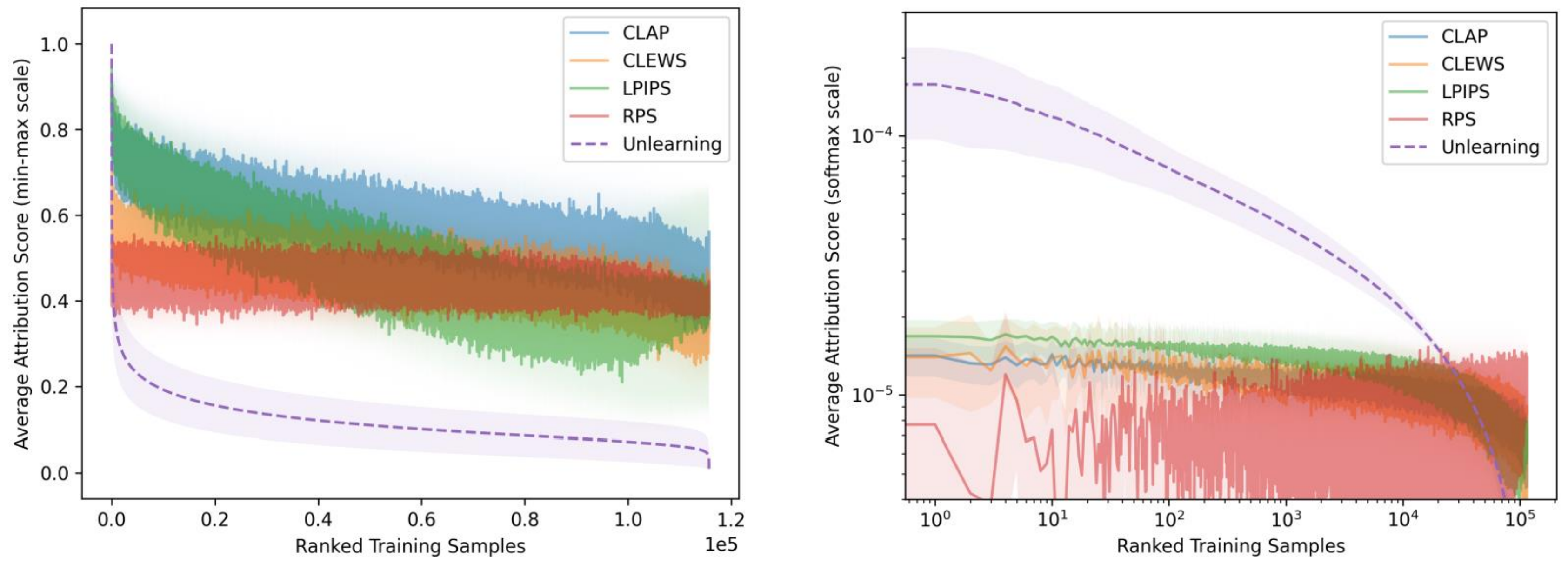

Figure 1: Comparison of attribution scores from unlearning- and similarity-based methods. Mean (line) and standard deviation (shading) over attribution scores from 16 generated test samples. Minmax (left) and softmax (right) normalizations are shown (notice the logarithmic axes in the later).

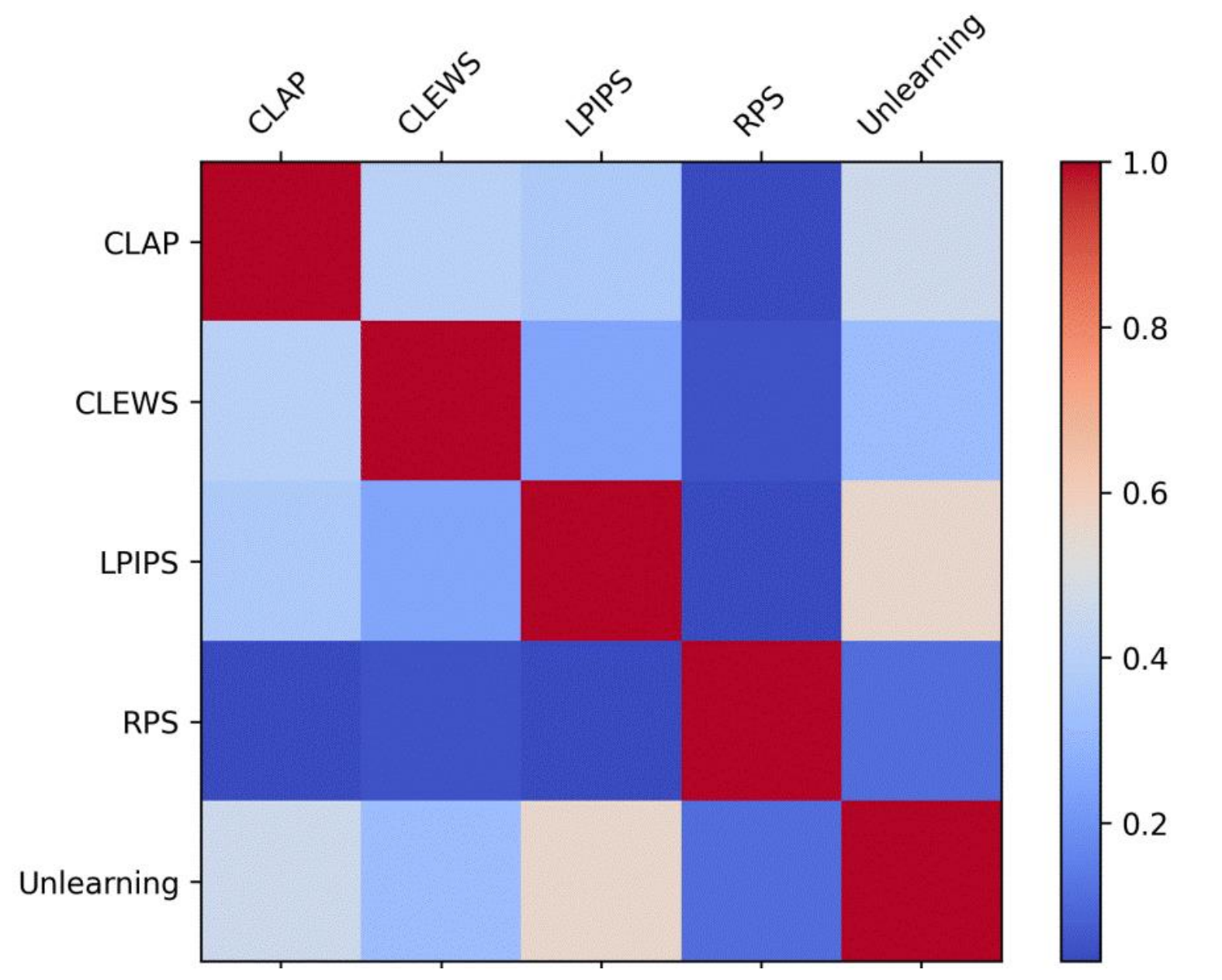

Figure 2: Correlation matrix between different attribution methods.