DITTO: Diffusion Inference-Time T-Optimization for Music Generation

Zachary Novack

UCSD, Adobe Research

znovack@ucsd.edu

Julian McAuley

UCSD

Taylor Berg-Kirkpatrick

UCSD

Nicholas J. Bryan

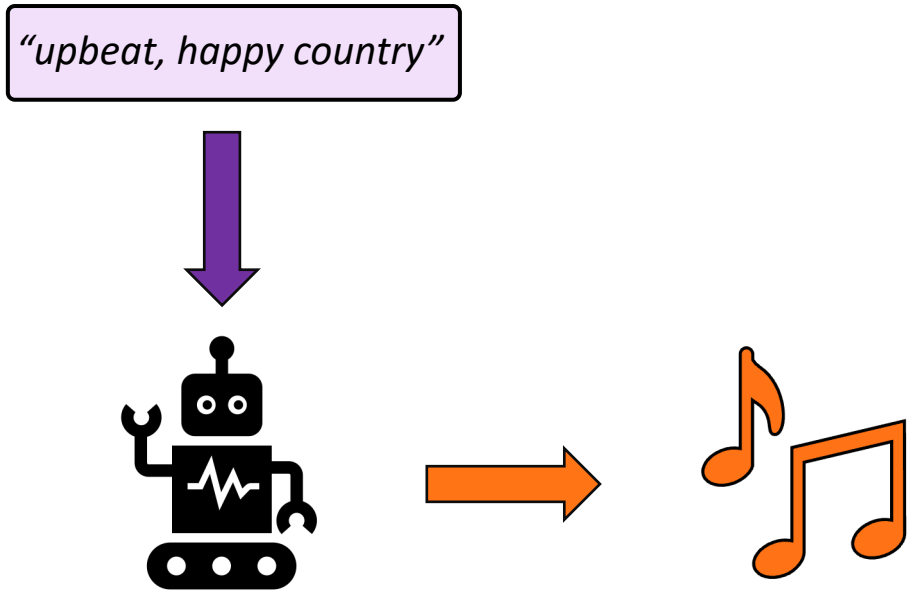
Adobe Research

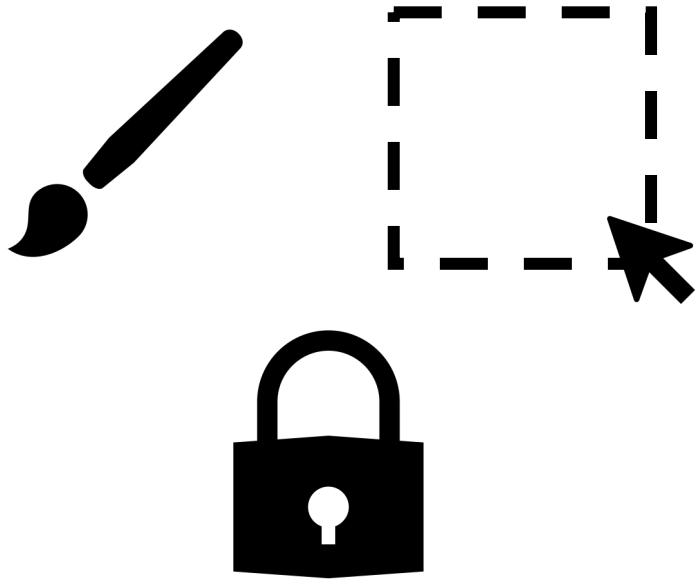
Project Website

Personal Website

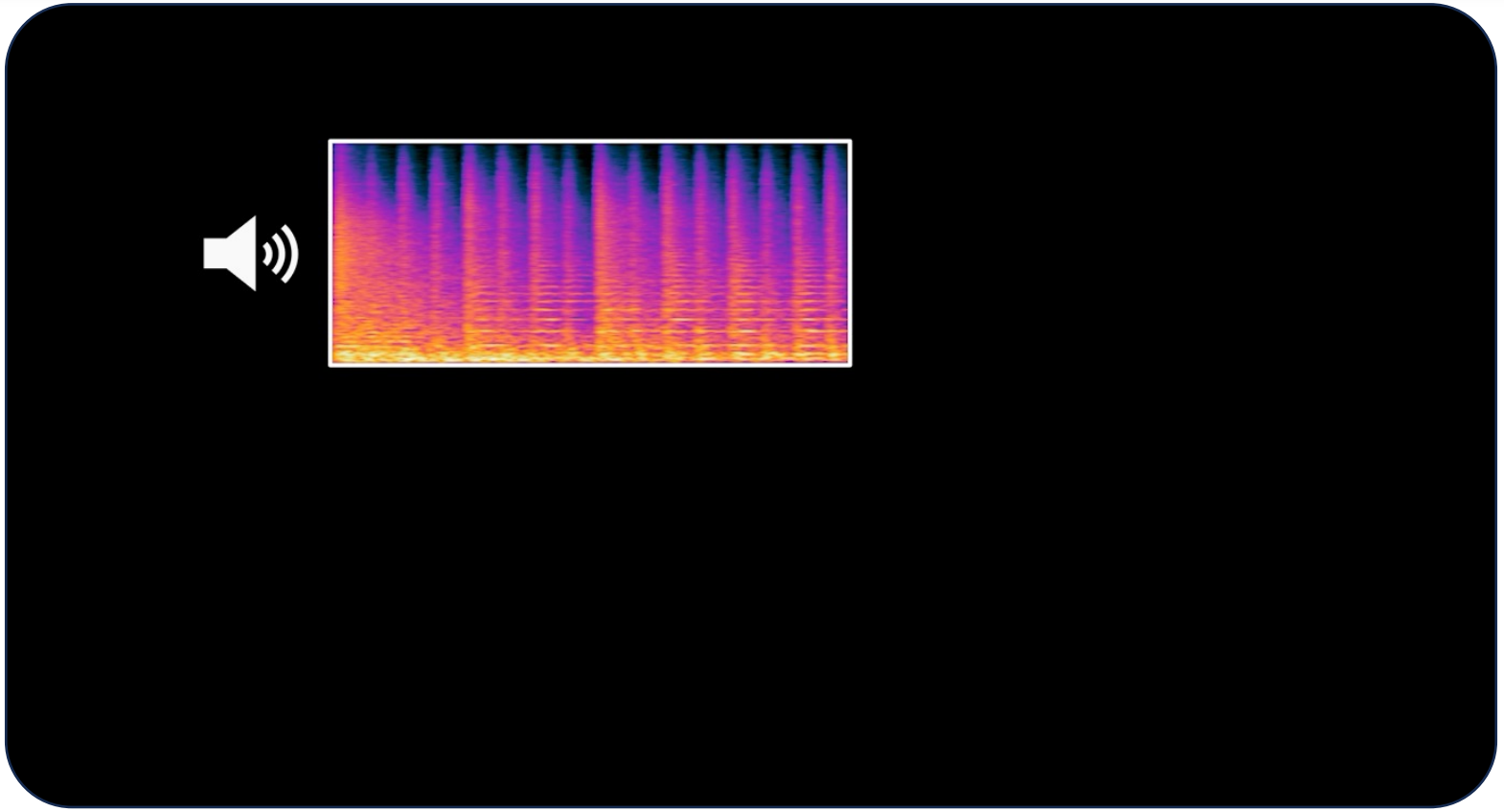
DITTO

Text-to-Music (TTM) Generation

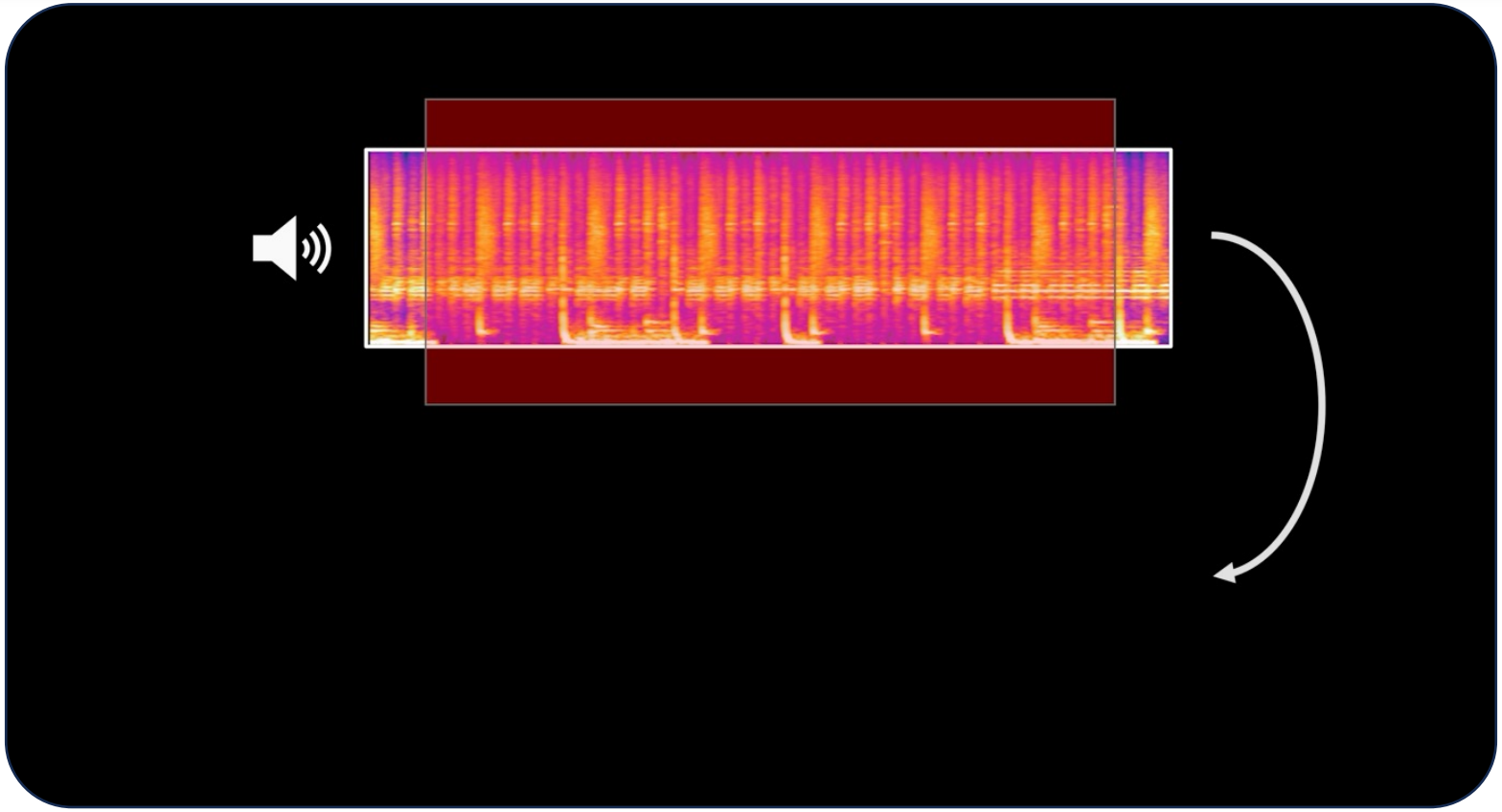




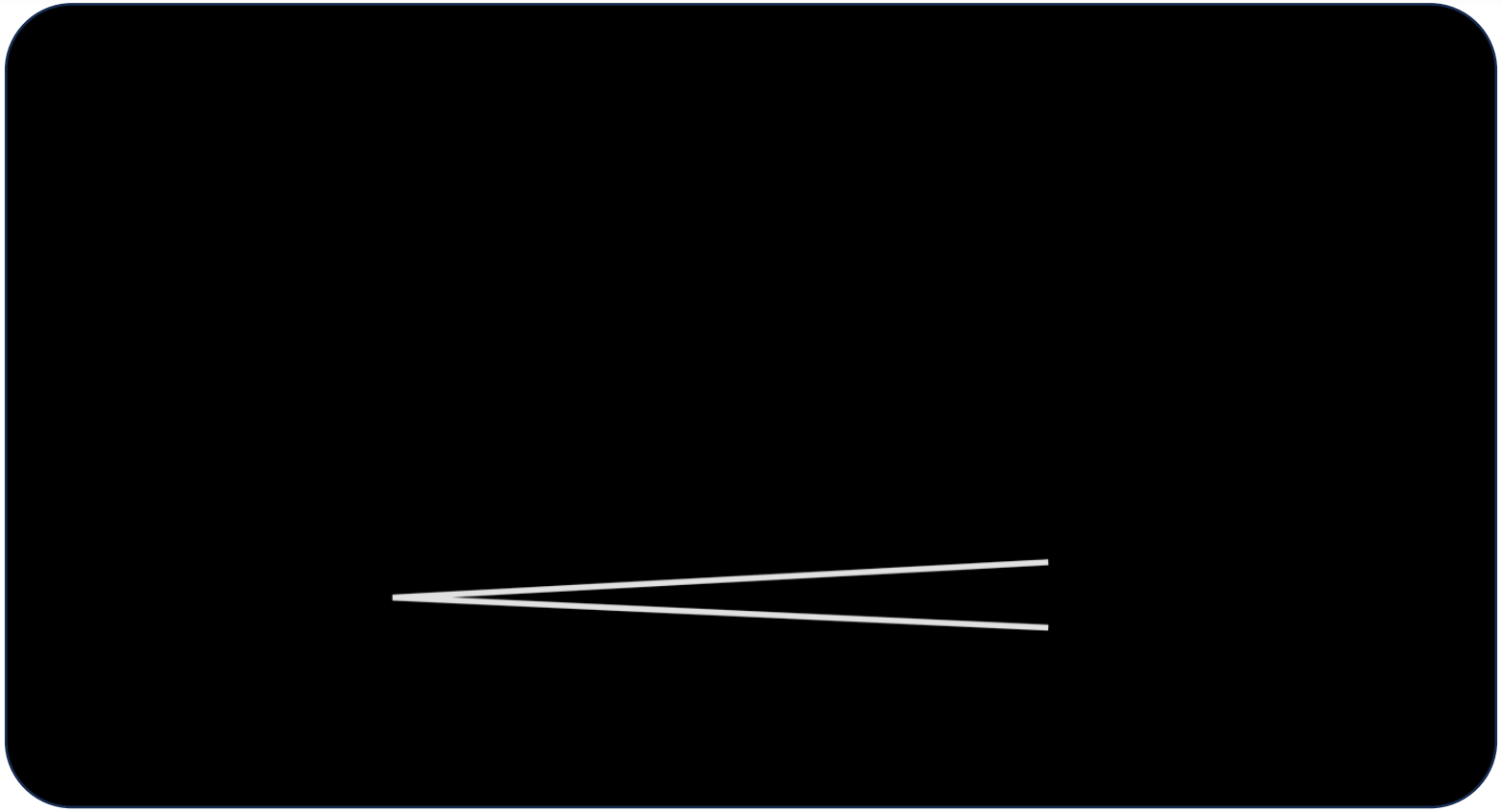
DITTO: Outpainting



DITTO: Inpainting

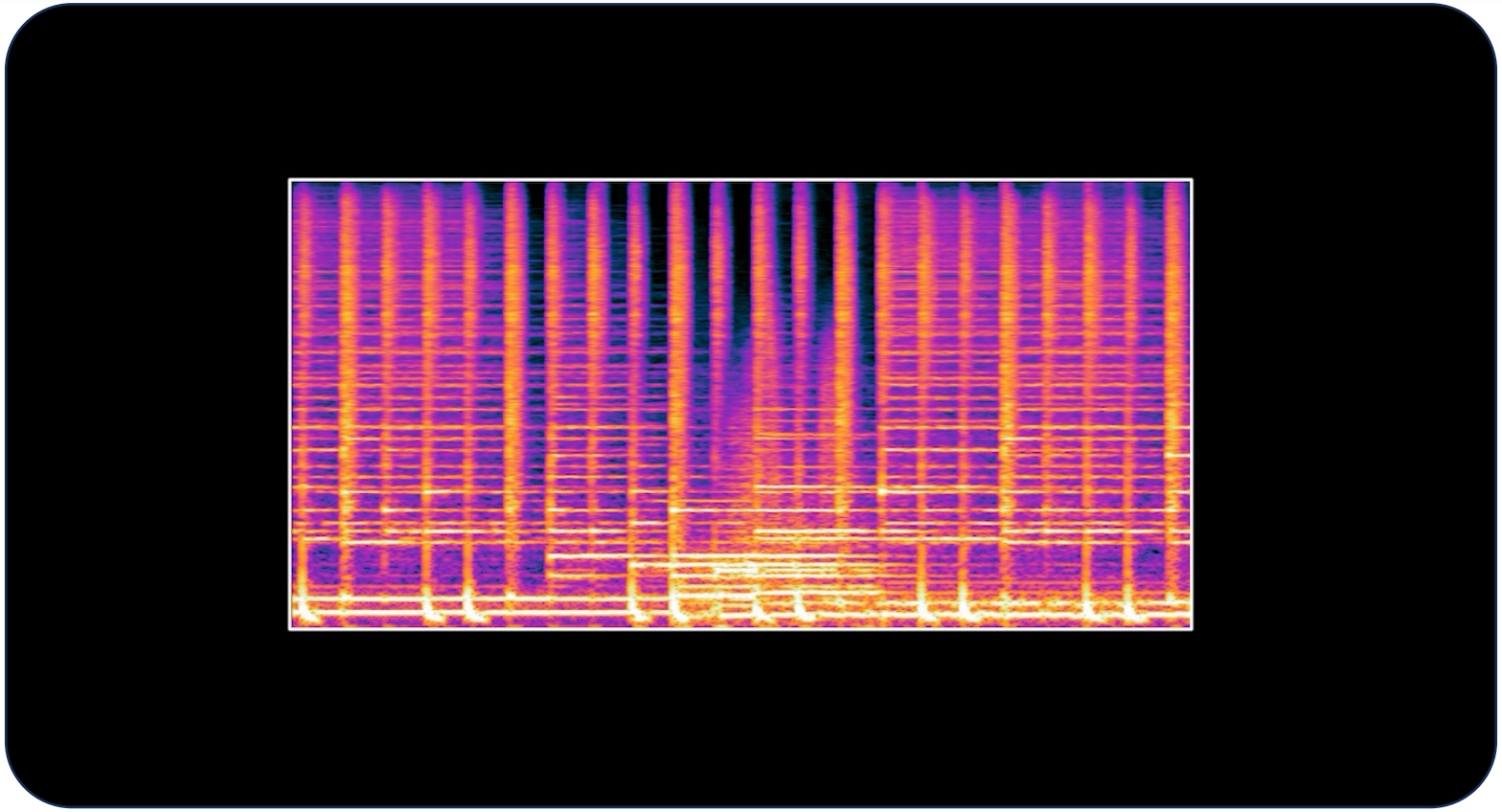


DITTO: Intensity Control



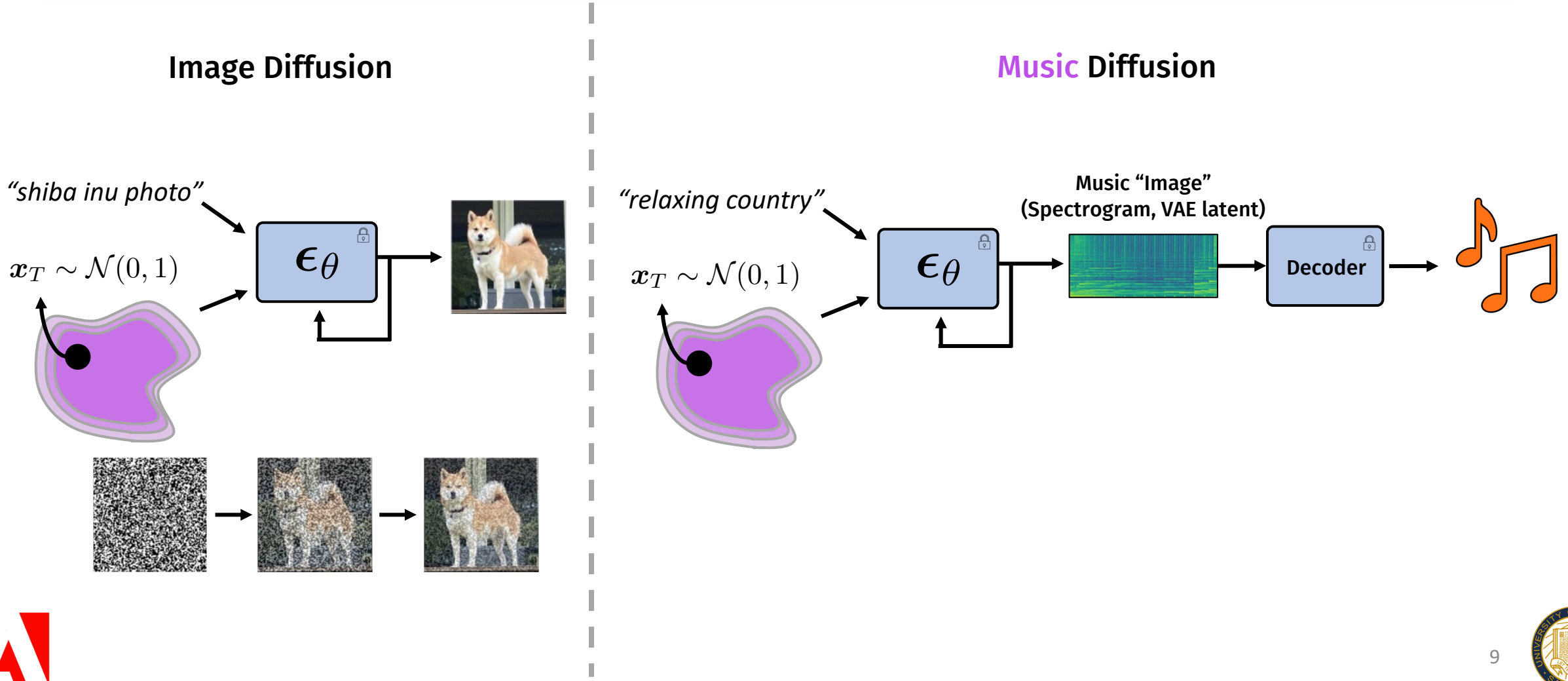
DITTO: Melody Control

DITTO: Structure Control



DITTO: Looping

Diffusion TTM

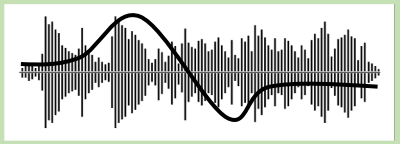


Beyond Text-Based Interactions

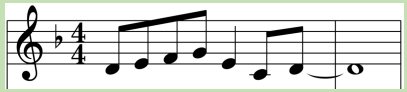
"writing about music is like dancing about architecture"

Local, Time-Varying Interactions

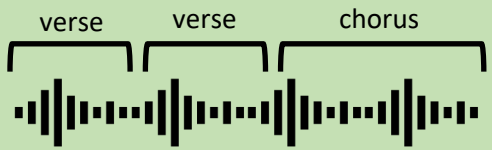
Feature Control



intensity

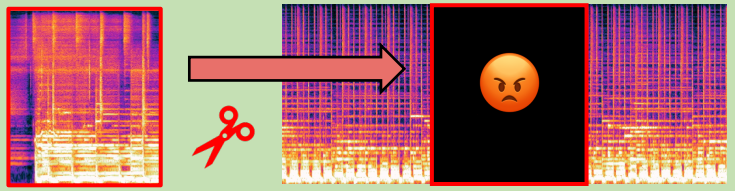


melody

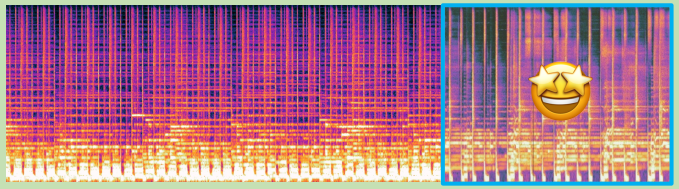


Musical structure

Editing

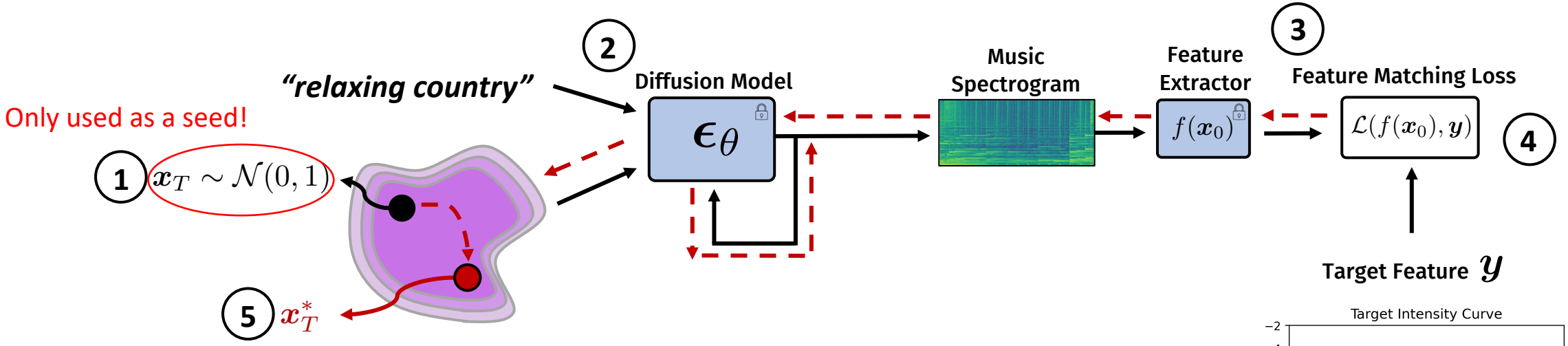


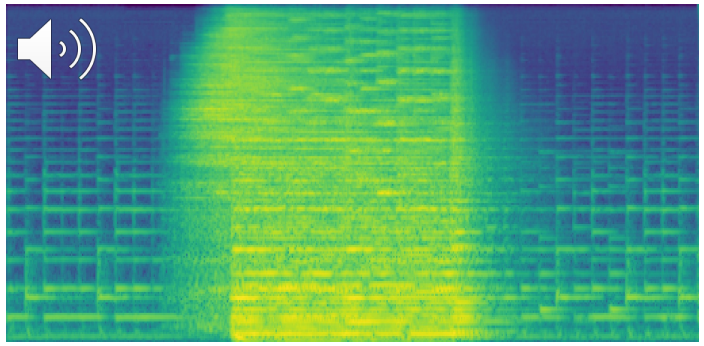
inpainting

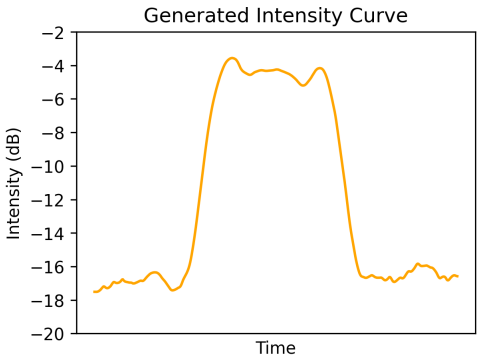


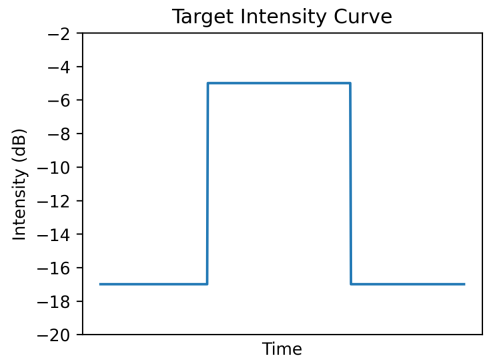
outpainting

DITTO: Diffusion Inference-Time T-Optimization

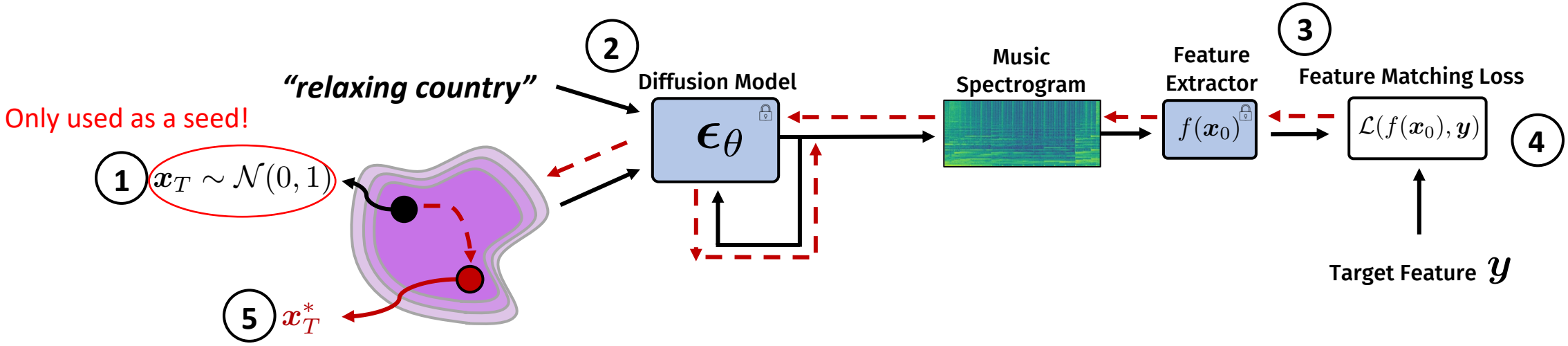








DITTO: Diffusion Inference-Time T-Optimization



- Any (differentiable) control
- Architecture/Sampler agnostic
- Zero training
- Exact control gradients

- Just latent optimization!
- Initialization->Structure

Training-Based

Music-ControlNet, JASCO

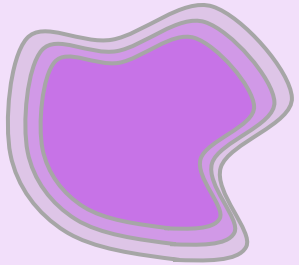
- ✓ Arbitrary Controls
- ✓ Control - Quality Balance
- ✓ Fast @ inference time

Large-scale training
Paired/labeled control data
Fixed controls @ training

DITTO

- ✓ Any (differentiable) control
- ✓ Architecture/Sampler agnostic
- ✓ Zero training
- ✓ Exact control gradients

Slow @ inference time



Training-Free (Guidance)

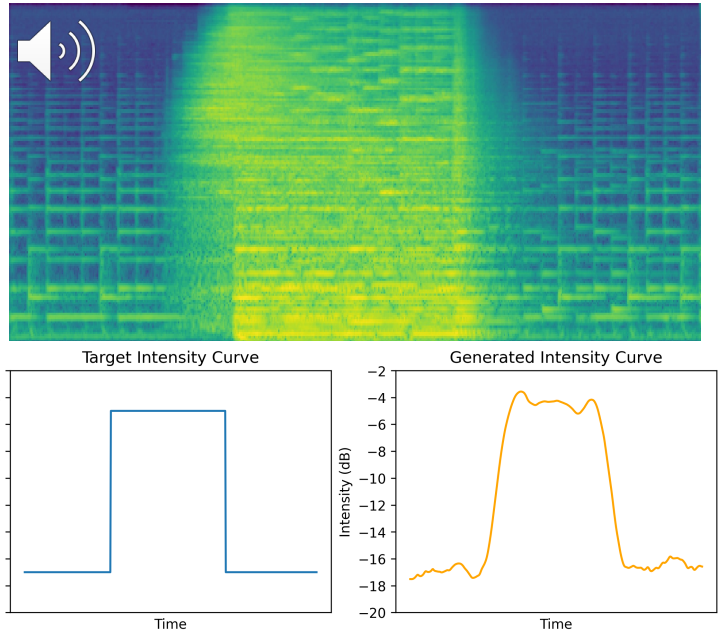
Classifier Guidance, DPS, FreeDoM

- ✓ Any (differentiable) control
- ✓ Zero training
- ✓ Moderate inference costs

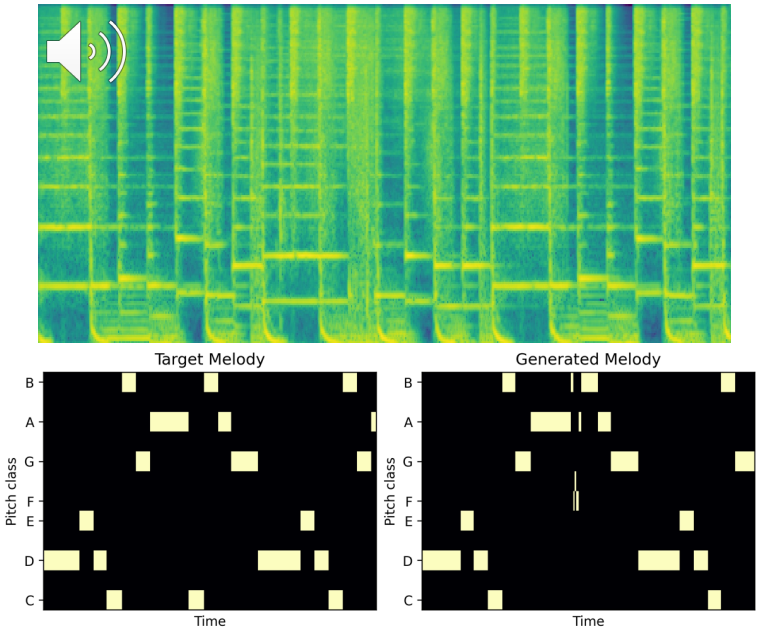
Approximate gradients
Limited control in low SNR
Bad at fine-grained controls

Qualitative Control Results

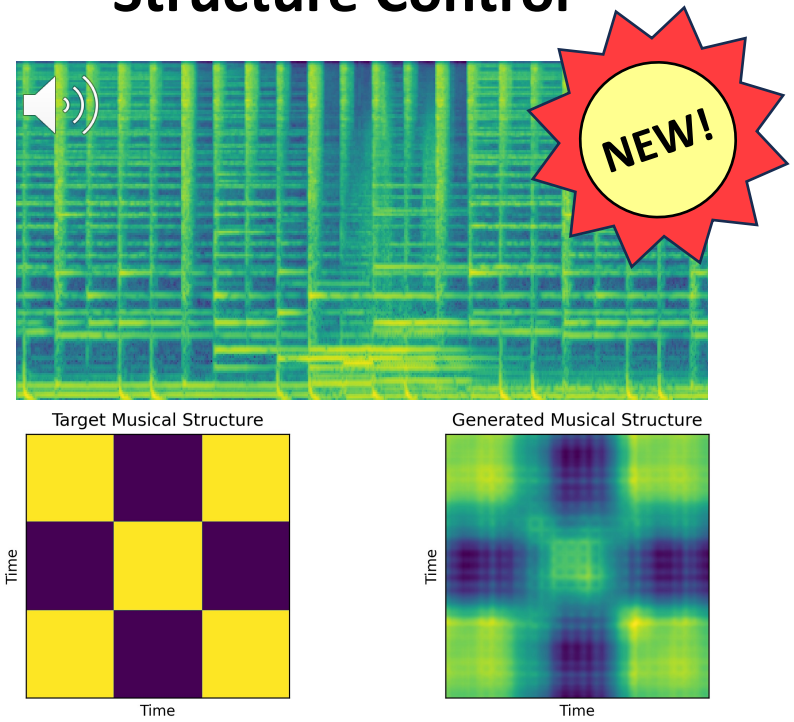
Intensity Control



Melody Control



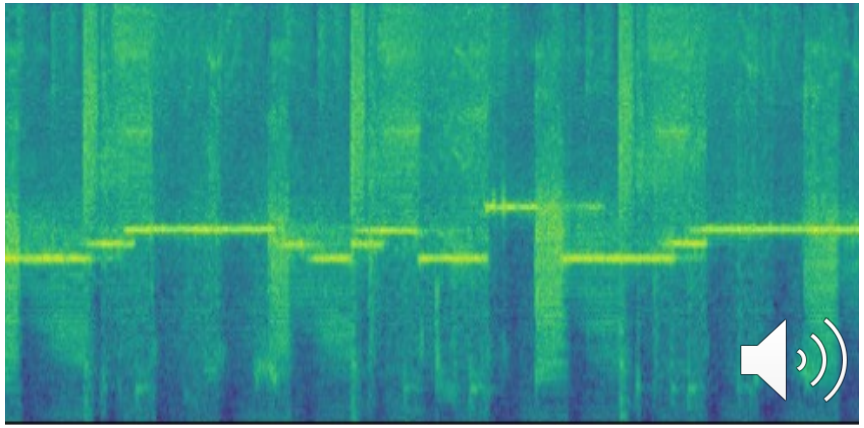
Structure Control



Quantitative Control Results

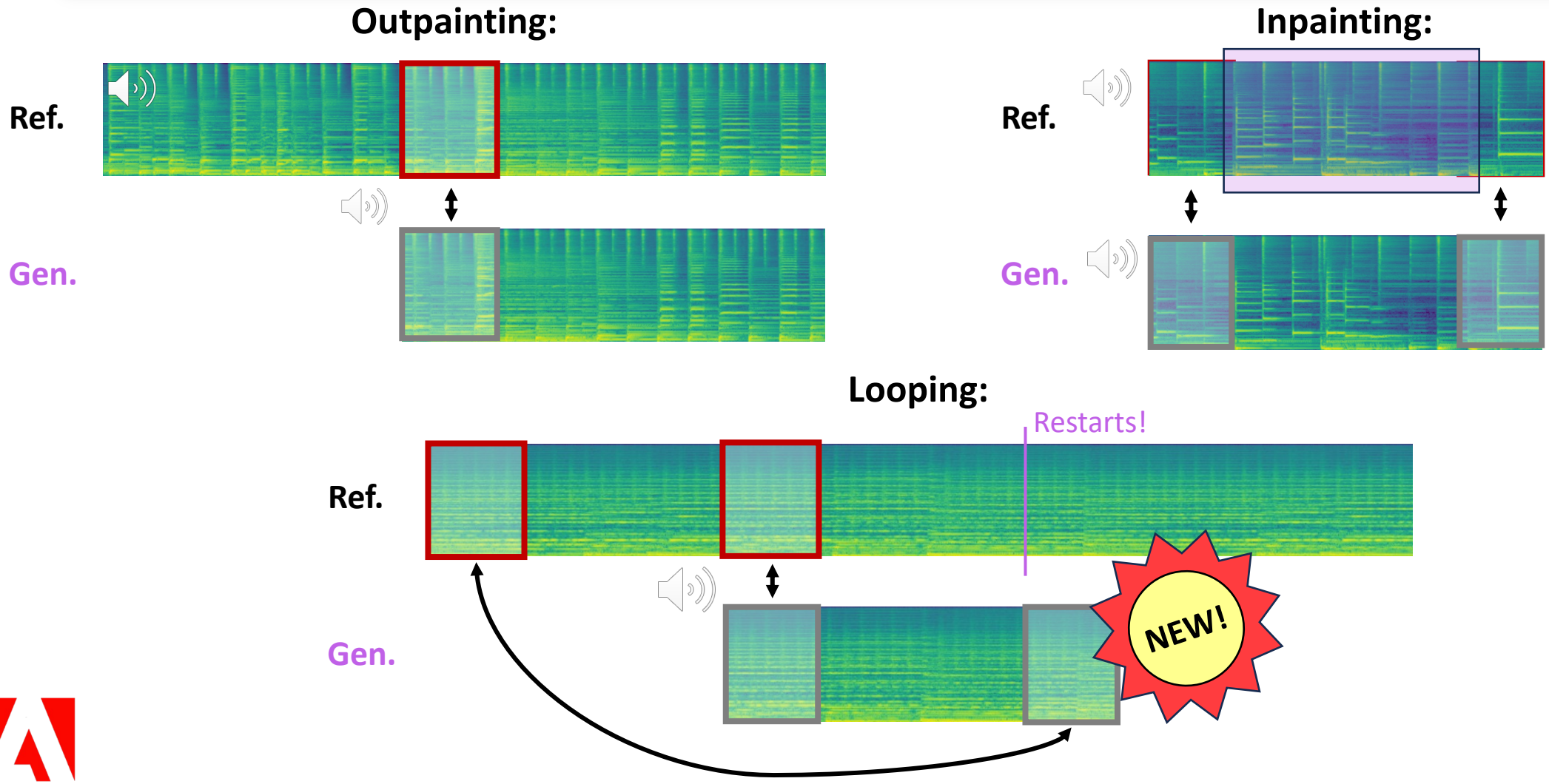
- Baselines:
  - **Music-ControlNet** (training-based)
  - **FreeDoM** (training-free guidance)
  - **DOODL** (training-free optimization)
- **DITTO** has SOTA Melody and Intensity Control
- **FreeDoM** struggles on complex controls
- **DITTO** avoids **DOODL** reward hacking

Reward Hacking



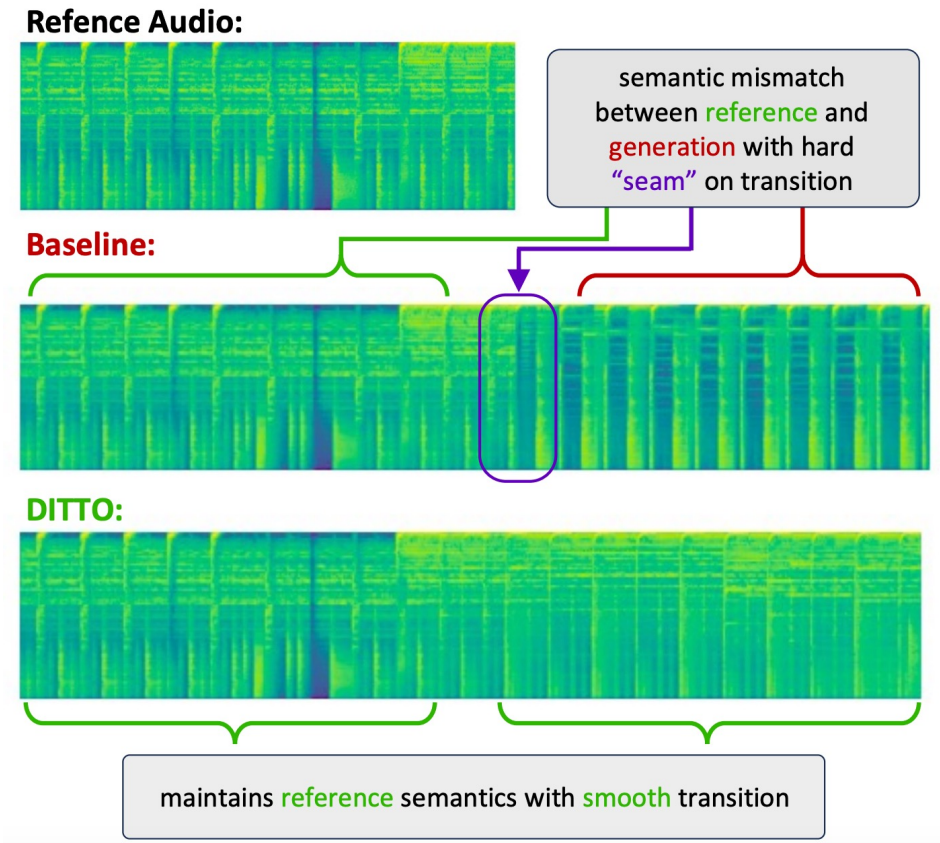
✓ High control accuracy
Low Quality
Low text relevance

Editing Tasks



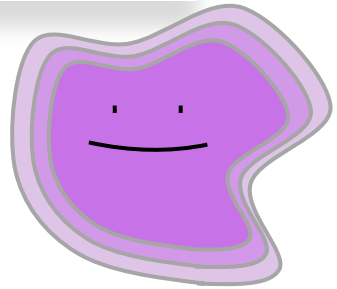
Quantitative Editing Results

- Baselines:
  - **Naïve/MultiDiffusion** (simple masking)
  - **FreeDoM/GG** (training-free guidance)
  - **DOODL** (training-free optimization)
- **DITTO** has SOTA FAD across mask widths



Conclusion

- **DITTO**: training-free editing/control for TTM models
- Simply $\boldsymbol{x}_T$ optimization + gradient checkpointing
- Array of tasks, new looping and structure control
- SOTA against training-based/free baselines
- Extra uses:
  - Reference-Free Looping
  - Structure Transfer
  - Multi-Feature Optimization
  - Optimized latent reuse
  - And more!



Any questions?

Email: **znovack@ucsd.edu**

𝕏: zacknovack

Project Website

Personal Website