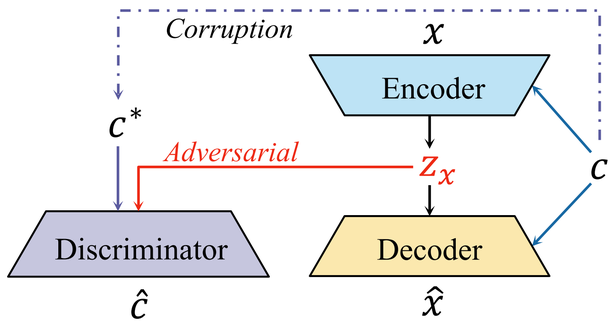
Domain Adversarial Training on Conditional VAE for Controllable Music Generation

Jingwei Zhao Gus Xia Ye Wang

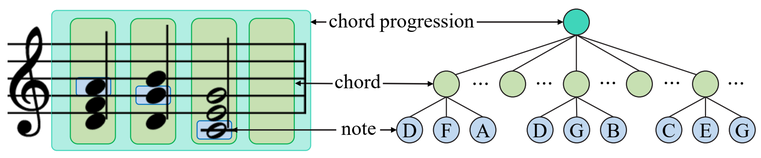
NUS National University of Singapore
Sound & Music Computing Lab
NYU Shanghai
MBZUAI
Music X Lab

Motivation

- The conditional VAE architecture can suffer from "condition collapse", because z_x is often too informative so the the decoder tends to ignore condition c.
- Domain adversarial training (DAT) can induce representation z_x to be disentangled from c, but it only applies to simple scenarios with categorical condition.
- We propose an adversarial condition denoising objective and generalize DAT to controllable music generation with complex sequential condition (e.g., melody).



Chord Representation



Objectives & Training

- VAE Objective

$$\mathcal{L}(\theta_{\text{enc}}, \theta_{\text{dec}}) = -\mathbb{E}_Q \left[\log P_{\theta_{\text{dec}}}(x \mid z_x, c)\right] + \alpha \mathbb{KL}(Q_{\theta_{\text{enc}}}(z_x \mid x, c) \parallel \mathcal{N}(\mathbf{0}, \mathbf{1})), \quad (1)$$

- Adversarial Objective

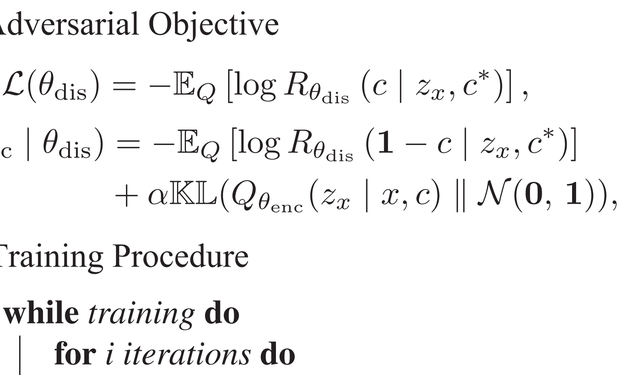
$$\mathcal{L}(\theta_{\text{dis}}) = -\mathbb{E}_Q \left[\log R_{\theta_{\text{dis}}}(c \mid z_x, c^*)\right], \quad (2)$$

$$\mathcal{L}(\theta_{\text{enc}} \mid \theta_{\text{dis}}) = -\mathbb{E}_Q \left[\log R_{\theta_{\text{dis}}}(\mathbf{1} - c \mid z_x, c^*)\right] + \alpha \mathbb{KL}(Q_{\theta_{\text{enc}}}(z_x \mid x, c) \parallel \mathcal{N}(\mathbf{0}, \mathbf{1})), \quad (3)$$

- Training Procedure

while *training* **do**
  for *i iterations* **do**
    Optimize VAE with $\mathcal{L}(\theta_{\text{enc}}, \theta_{\text{dec}})$,
  for *j iterations* **do**
    for *k iteration* **do**
      Optimize discriminator with $\mathcal{L}(\theta_{\text{dis}})$,
    for *l iterations* **do**
      Optimize encoder with $\mathcal{L}(\theta_{\text{enc}} \mid \theta_{\text{dis}})$.

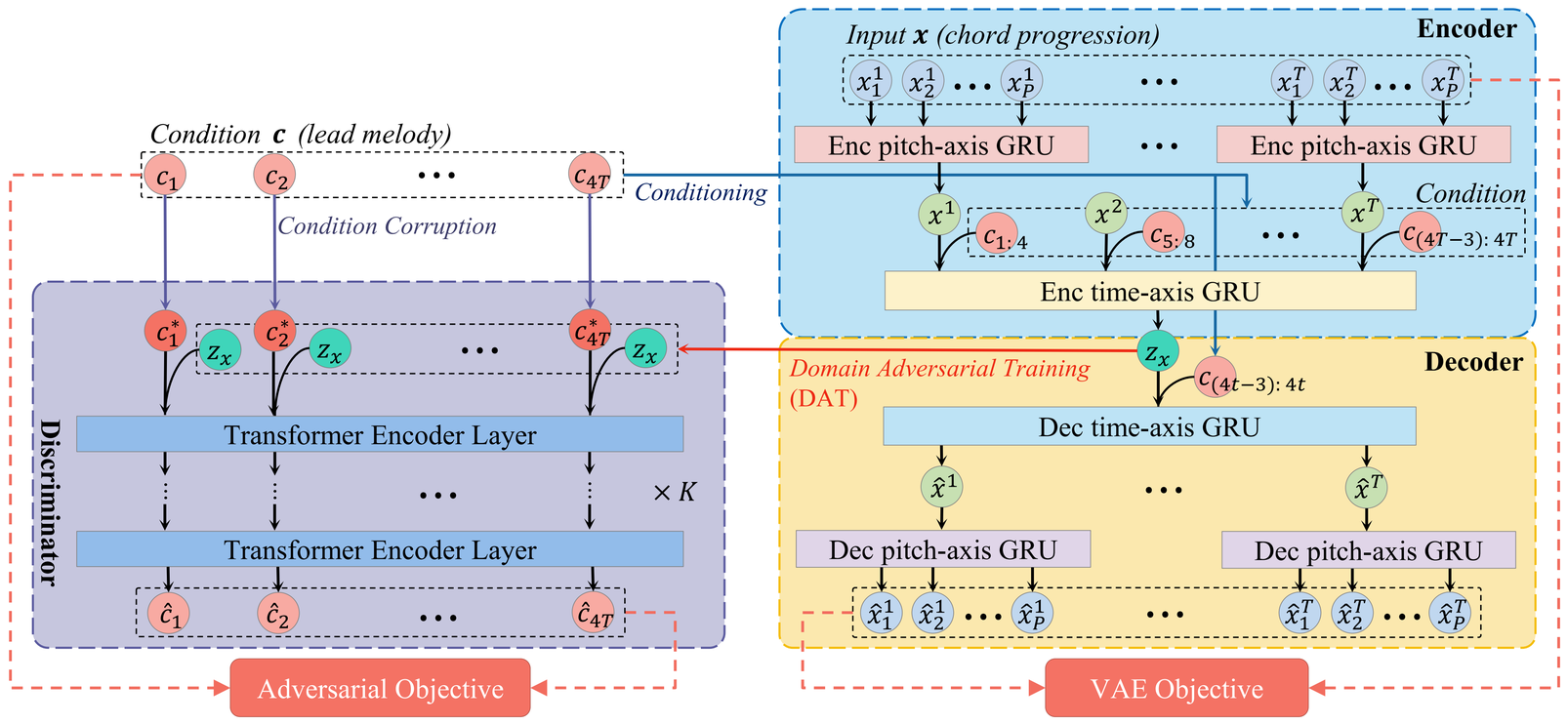
- Training Loss Curve



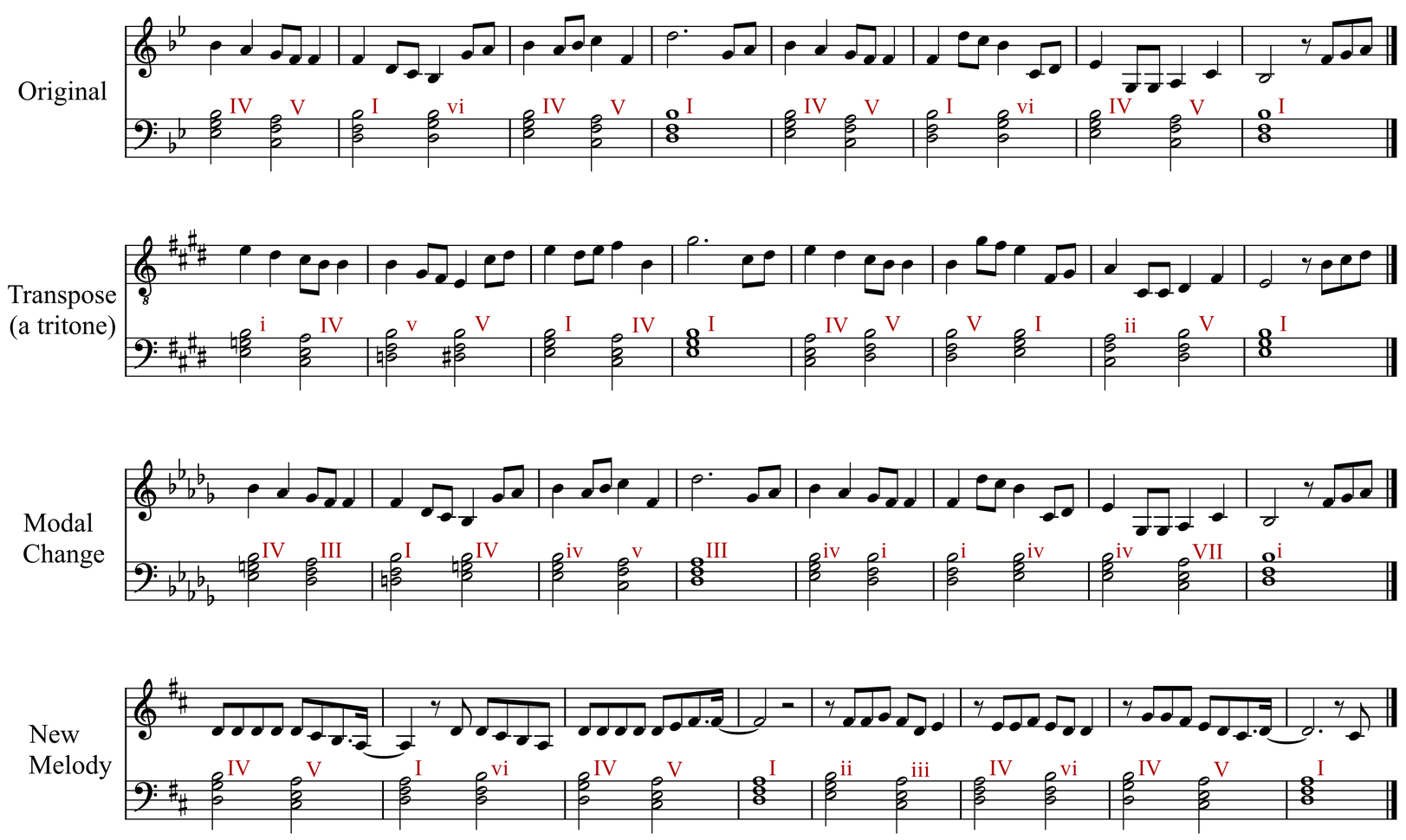
Resource

Paper Audio Demo GitHub

Model Architecture



Demo: Chord Generation with Varied Melody Conditions



Experiments

- **Ablation Models**

Non-DAT: Same VAE framework but without a discriminator. It does not explicitly try to disentangle z_x from c using domain adversarial training (DAT);

Mask-CR: Applying a general masking corruption instead of pitch transposition for condition corruption;

Non-CR: Using the conventional DAT objective without condition corruption. It predicts c directly from z_x using a GRU-based discriminator.

- **Subjective Evaluation**

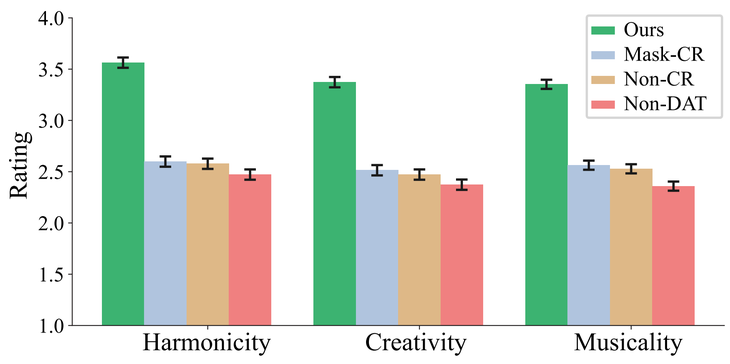


Figure 7: Subjective evaluation on the harmonization performance of our model and baseline models.

- **Objective Evaluation**

Figure 8: Object evaluation on representation similarity (invariance) against pitch transposition. A higher value denotes better disentanglement.

Figure 9: Objective evaluation on harmony histogram upon melody swapping. A higher ratio in root, 3rd, and 5th notes indicates a higher degree of controllability.